mr manual de referencia

INFORMÁTICA

SISTEMAS DE AYUDA A LA DECISIÓN MÉDICA

Juan M. García-Gómez | Salvador Tortajada | Carlos Sáez

Editorial
Universitat Politècnica
de València

INFORMÁTICA

Sistemas de Ayuda a la Decisión Médica

Juan M. García-Gómez
Salvador Tortajada
Carlos Sáez

Editorial
Universitat Politècnica
de València

Colección Manual de Referencia

Los contenidos de esta publicación han sido evaluados mediante el sistema *doble ciego*, siguiendo el procedimiento que se recoge en http://bit.ly/Evaluacion_Obras

Para referenciar esta publicación utilice la siguiente cita: García-Gómez, Juan M.; Tortajada, Salvador; Sáez, Carlos, (2019). *Sistemas de Ayuda a la Decisión Médica*. Valencia: Universitat Politècnica de València

Autores:

Juan M. García-Gómez
Salvador Tortajada
Carlos Sáez

2019, Editorial Universitat Politècnica de València
Venta: www.lalibreria.upv.es / Ref.: 6520_01_01_01

ISBN: 978-84-9048-780-8 (versión impresa)

Si el lector detecta algún error en el libro o bien quiere contactar con los autores, puede enviar un correo a edicion@editorial.uvp.es

BIODATAS

JUAN M GARCÍA-GÓMEZ

Profesor de Ingeniería Biomédica y director del grupo de Ciencia de Datos Biomédicos en la Universitat Politècnica de València. Durante los últimos 20 años ha desarrollado y coordinado proyectos de Inteligencia Artificial aplicados a medicina, con especial interés en la ayuda a la decisión en oncología, farmacia y emergencias sanitarias.

SALVADOR TORTAJADA

Ingeniero informático y doctor en física aplicada por la Universidad Politécnica de Valencia. Ha centrado su trayectoria profesional e investigadora se ha centrado en la aplicación de la informática, la inteligencia artificial y los sistemas de ayuda a la decisión a la medicina y la imagen médica. Cuenta con varias contribuciones internacionales en libros y revistas de Tecnologías de la Información para la salud.

CARLOS SÁEZ

Investigador postdoctoral en el grupo de Ciencia de Datos Biomédicos en la Universitat Politècnica de València y docente de asignaturas de Ingeniería e Informática Biomédica. Durante los últimos 13 años ha desarrollado nuevas tecnologías y metodologías de Sistemas de Ayuda a la Decisión Médica y de extracción de conocimiento confiable y de calidad de Big Data biomédico, las cuales han sido aplicadas en múltiples problemas biomédicos reales.

RESUMEN

Los datos se han convertido en un torrente que fluye en todos los ámbitos de la sociedad. La aplicación de la Inteligencia Artificial al ámbito de la salud es posiblemente el de mayor impacto social, por las implicaciones que tiene para el bienestar de las personas. Muchos han depositado las esperanzas de la Medicina de Precisión en los Sistemas de Ayuda a la Decisión Médica, objeto de estudio de este libro.

El desarrollo de Sistemas de Ayuda a la Decisión Médica abarca múltiples aspectos procedentes de la medicina, la Inteligencia Artificial, la teoría de la decisión, la ingeniería biomédica, la biología, la farmacia, las organizaciones sanitarias, la ingeniería de procesos, la evaluación de sistemas y el sentido común. En este libro, elaboraremos los conceptos estratégicos, funcionales y técnicos necesarios para concebir, diseñar, adaptar, implantar y evaluar Sistemas de Ayuda a la Decisión Médica. Los textos están dirigidos a alumnos y profesionales de Tecnologías de Información para la Salud, Medicina, e Ingeniería Biomédica aprovechando casos reales de proyectos desarrollados por los propios autores.

Agradecimientos

Este texto es consecuencia indirecta del trabajo realizado en los proyectos de investigación del Laboratorio de Ciencia de Datos Biomédicos (`http://www.bdslab.eu/`) y la docencia en el Departamento de Física Aplicada de la Universitat Politècnica de València. Agradecemos por lo tanto a las personas y entidades involucradas en la financiación, definición y desarrollo de dichos proyectos durante cerca de veinte años.

Gracias a todos los alumnos de las asignaturas de Tecnologías de la Información para la Salud, Tratamiento Digital de Datos Biomédicos y Bioinformática que por su interés y entusiasmo desde el año 2001 han motivado la escritura de este libro.

Un sincero agradecimiento a Ramón Esteban i Romero por su ayuda en algunas formulaciones del capítulo 8.

Índice general

Prefacio

La figura 1 sintetiza la organización de los contenidos del libro describiendo las relaciones entre los capítulos a través de flechas y el carácter estratégico, funcional o técnico del contenido de los capítulos mediante el código de colores.

La visión del problema pretende ser holística, ya que la problemática que aborda el desarrollo de Sistemas de Ayuda a la Decisión Médica, o más comúnmente conocidos por su término en inglés Clinical Decision Support Systems (CDSS), abarca múltiples aspectos procedentes de la medicina, la ingeniería informática, la teoría de la decisión, la ingeniería biomédica, la biología, la farmacia, las organizaciones, la ingeniería de procesos, la evaluación y el sentido común. En los capítulos siguientes elaboraremos los conceptos estratégicos, funcionales y técnicos necesarios para concebir, diseñar, adaptar, implantar y evaluar CDSS. Los textos intentan dar una visión objetiva, pero siempre basada en la experiencia de los autores en el desarrollo de proyectos donde los Sistemas de Ayuda a la Decisión eran un instrumento y un fin.

El resto del libro se estructura de la siguiente forma. El capítulo 2 aborda los elementos involucrados en los procesos de decisión médica a partir de la identificación de casos reales hasta llegar a una traducción abstracta que nos permita modelar el problema mediante la teoría de la decisión.

El capítulo 3 introduce los conceptos clave acerca de la teoría de probabilidad y decision. Éste capítulo abre la rama de lectura del libro relacionada con CDSS basados en modelos de aprendizaje automático. Siguiendo esta rama, el capítulo 4 recoge técnicas de selección y extracción de características aplicables a los tipos de variables más comunes. En paralelo, los capítulos 5, 6 y 7 recogen técnicas de procesamiento de cadenas, señales e imágenes médicas que suelen ser necesarias para el tratamiento de datos biomédicos de diferentes niveles biológicos: desde secuencias genéticas hasta señales fisiológicas o incluso el análisis de características de la voz.

El capítulo 8 desarrolla los fundamentos del aprendizaje automático como metodología para el modelado predictivo. Éstos se desarrollan específicamente en los diferentes tipos de modelos de aprendizaje recogidos en los capítulos 9, 10, 11, 12 y 13, y se completan mediante las diferentes aproximaciones para la evaluación y selección de modelos en el capítulo 14. Esta exposición se desarrolla a partir de la teoría de la decisión, tomando el testigo del capítulo 2 para dar la solución que para nosotros resulta natural y óptima.

Los capítulos 15 y 16 recogen técnicas específicas de Inteligencia Artificial y Minería de Datos para el desarrollo de CDSS. El capítulo 15 analiza la tecnología *Datawarehouse* aplicada a medicina, tecnología ya introducida en los sistemas de gestión sanitaria junto a políticas sanitarias en base a objetivos, pero que también ofrece posibilidades interesantes para el uso clínico a nivel primario y secundario. El capítulo 16 expone la teoría y aplicación en medicina de razonadores semánticos, como base de los CDSS de Nivel III.

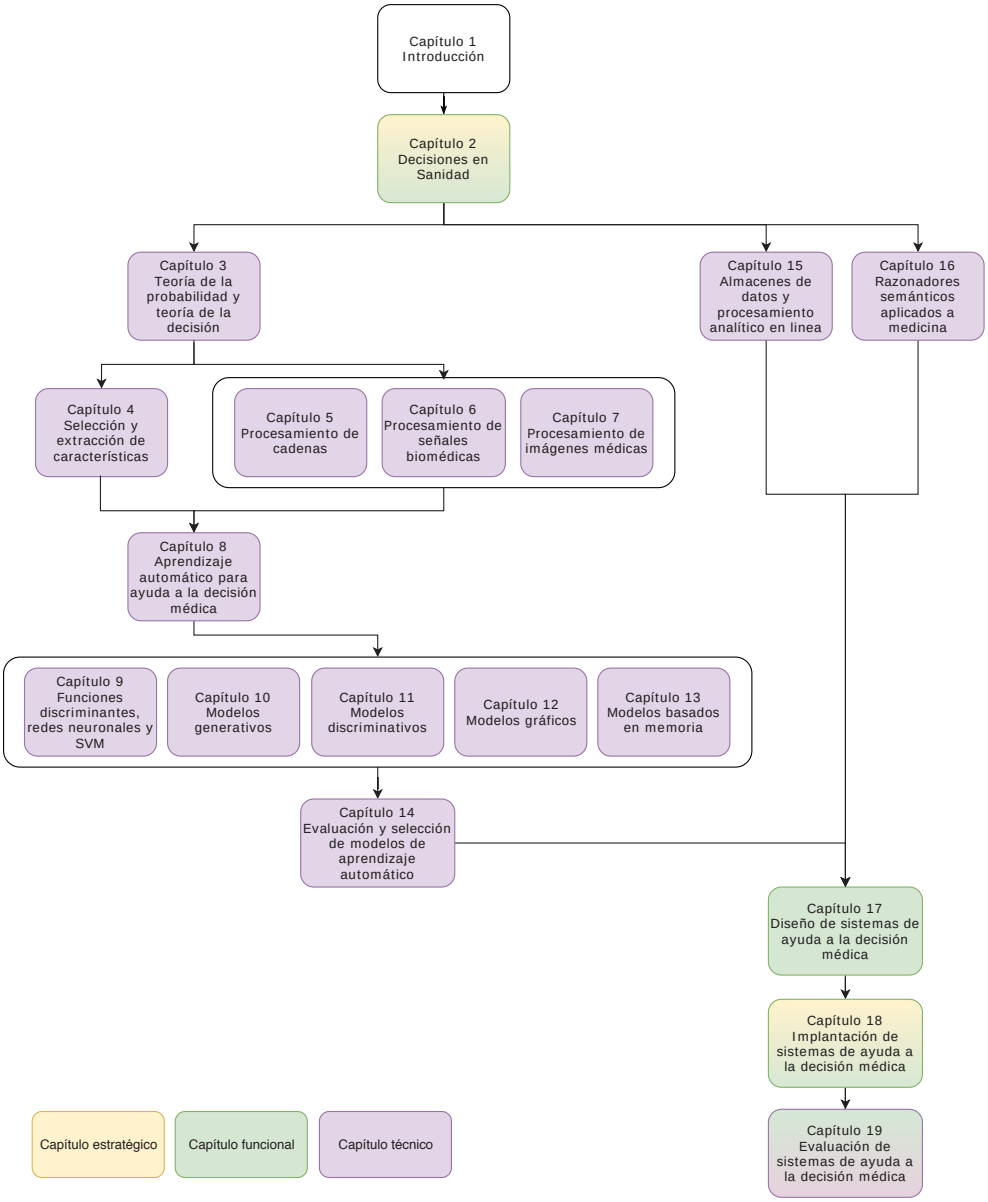


Figura 1: Organigrama de los capítulos del libro.

El capítulo 17 aborda los elementos clave de un software de CDSS, repasa el uso de estándares de datos y compartición de conocimiento y, finalmente, propone un diseño dual como arquitectura de CDSS. A continuación el capítulo 18 analiza los aspectos clave para el éxito en la implantación de un CDSS en la práctica médica. Seguidamente, se analizan las tendencias actuales de gestión de la innovación como estrategia de contratación de sistemas de ayuda a la decisión y se analiza la propuesta de HIMSS[a] para la implantación de programas de ayuda a la decisión en entornos clínicos. El final del capítulo resume la legislación actual sobre dispositivos médicos donde se engloban los CDSS.

El capítulo 19 completa el ciclo de vida de un CDSS mediante la evaluación de su valor añadido al servicio sanitario que pretende apoyar. Para ello se analizan diferentes métricas a utilizar y se estudian las técnicas necesarias para una evaluación del sistema.

[a]http://www.himss.org

Capítulo 1

Introducción

Los Sistemas de Ayuda a la Decisión Médica, comúnmente conocidos por su acrónimo en inglés CDSS[a], son sistemas computacionales que aportan conocimiento específico para las decisiones médicas que deben tomarse en los planes de tratamiento de pacientes, como por ejemplo durante el diagnóstico, pronóstico, tratamiento y administración.

La medicina está experimentando una evolución que tiende a transformar la asistencia sanitaria desde un modelo reactivo y general a un modelo preventivo y de precisión. Cada vez es más factible una medicina que incorpore diagnósticos multidisciplinares, tratamientos personalizados y planes preventivos individualizados para cada paciente, en los que además se sienta involucrado. En concreto, las iniciativas por una medicina de precisión, derivada de los conceptos de medicina P4 (preventiva, predictiva, personalizada y participativa) [1], asume un rol activo de todos los actores sanitarios para el cambio previsto durante las próximas dos décadas.

La medicina de precisión hace uso de la condición específica de las personas en el momento de decidir sus planes de tratamiento [2]. Debido a que la observación de la condición del paciente es siempre parcial, la decisión óptima de estos tratamientos debe basarse en los riesgos previstos condicionados a la observación de los datos biomédicos multinivel del paciente. Por otra parte, el plan de tratamiento debe ser visto como una secuencia de decisiones interconectadas, no como una secuencia de decisiones aisladas. Este punto de vista plantea varios retos a los diseñadores de los CDSS:

1. Modelar y optimizar los planes de tratamiento como secuencias de decisiones en las que se dispone de observaciones parciales del estado del paciente.

2. Incorporar herramientas de predicción individualizada en las etapas clave de los planes de tratamiento.

3. Actualizar continuamente el modelo de plan de tratamiento con el nuevo conocimiento adquirido a partir de los registros de datos y de los resultados de los procedimientos médicos.

Además, los CDSS están altamente conectados con el concepto de Medicina Basada en la Evidencia (MBE) [3, 4], ya que:

[a]CDSS, del inglés *Clinical Decision Support System.*

- Recogen el conocimiento médico del problema a través de los protocolos de adquisición, de los biobancos utilizados en el desarrollo de los sistemas, del control de calidad, de la anotación de los casos, de las especificaciones del proceso sanitario y/o de las guías de práctica clínica.

- Dan soporte computacional para la implantación en la práctica médica del conocimiento obtenido a partir de la evidencia científica fruto de proyectos empíricos (generalmente multicéntricos), o de consensos médicos (realimentación e investigación traslacional).

- Evaluan el rendimiento de la solución para cada problema específico, así como estudian la aceptación en el ámbito sanitario y la cuantificación del valor añadido para la solución del problema médico.

Los CDSS aportan funcionalidades para la práctica asistencial (uso primario en medicina) y para la explotación de la información médica (uso secundario). Las funcionalidades más destacadas de los CDSS son la alerta y/o recordatorio de situaciones de salud de pacientes concretos, la ayuda al diagnóstico y al pronóstico, la gestión de patologías crónicas, el soporte a reuniones multidisciplinares, el triaje, la calidad asistencial, la gestión de costes, tiempos y recursos, la interpretación de señales biomédicas, la definición de biomarcadores discriminantes, la búsqueda de casos similares, la búsqueda de información bibliográfica relacionada con el paciente, la búsqueda de patrones anómalos, las simulaciones de escenarios de salud y la educación de personal sanitario basado en casos.

El objetivo principal de la implantación de los CDSS en los servicios de salud es la mejora de la atención sanitaria individualizada. Para ello, los CDSS aportan conocimiento específico para la decisión en cada caso médico.

El segundo objetivo de la implantación de los CDSS es el aumento de la eficiencia de los servicios sanitarios. La recomendación de los procedimientos diagnósticos, la asistencia en reuniones multidisciplinares, la presentación intuitiva y fiable del conocimiento aplicable al caso médico, o la prescripción asistida de tratamientos son ejemplos de funcionalidades que pueden mejorar la eficiencia temporal y el aprovechamiento de los recursos sanitarios y reducir los errores médicos.

En definitiva, los CDSSs potencian las habilidades de los profesionales de sanidad para manejar múltiples variables biomédicas, a través de herramientas computacionales de valor añadido a sus tareas y al sistema sanitario en el que desarrollan los procesos asistenciales.

En la literatura especializada podemos encontrar diversas terminologías para referirse a los CDSS según el ámbito o momento del que proceden. En este libro hemos considerado *prognostication tools*, *clinical decision support system*, *computerized clinical decision support*, y *electronic clinical decision support system* como sinónimos de CDSS. Además, consideramos incluidos en CDSS a los *nomogramas*, las guías clínicas electrónicas, los *computerized physician order entry* (CPOE) y los *patient safety systems*. Además, algunas tecnologías que aportan funcionalidades a los CDSS son *bussiness intelligence*, sistemas expertos y cada vez con mayor relevancia aquellas basadas en ciencia de datos [5]. En el ámbito del *share decision making* y los *personal health systems* se puede también identificar el concepto de *meaning* y de *empowerment* como las funcionalidades de valor añadido que los CDSS dan a los sistemas de información sanitaria con el fin de involucrar activamente al paciente en su salud.

El anexo A contiene la recopilación realizada en el momento de la edición del libro sobre los foros internacionales, revistas y congresos donde se fomenta la investigación, el desarrollo y la integración de CDSS.

1.1. Características de los CDSS

Los CDSS deben obtener una solución óptima en base al conocimiento del problema médico y utilizando la información específica del caso a resolver. Por este motivo, es deseable que un CDSS cumpla con las siguientes características:

- Debe manejar adecuadamente situaciones con incertidumbre.

  Las decisiones siempre vienen acompañadas de la falta de información ya que no se suele saber a priori el diagnóstico de un enfermo, sino que se observan sus síntomas. La teoría de la probabilidad y la teoría de la decisión estudian la resolución objetiva de decisiones en situaciones de incertidumbre, que minimicen el riesgo dado el conocimiento disponible. Los CDSS diseñados mediante *aprendizaje automático* implementan soluciones basadas en estas teorías.

- Ser específico con el problema médico a resolver.

  Cuanto más específica sea la solución aportada por el CDSS al problema médico más útil y fiable será para su propósito. Esto supone estudiar detenidamente el objetivo del problema a resolver, así como obtener conocimiento en forma de guías de práctica clínica, reglas de decisión y/o casos retrospectivos del problema.

- Estar basado en las evidencias científicas.

  Las hipótesis de partida que permiten definir el problema a resolver por el CDSS deben estar basadas en evidencias científicas. Por otra parte, es una buena costumbre que el diseño de los CDSS venga avalado por la publicación de los resultados en revistas científicas del ámbito médico del problema, de informática biomédica, o de ayuda a la decisión. Además, si se han utilizado datos biomédicos durante el desarrollo del sistema, estos suelen venir acompañados de referencias bibliográficas a los protocolos consensuados para su adquisición, y de los estudios realizados mediante los biobancos.

- Ser interoperable a nivel semántico con las fuentes de información biomédica y la historias de salud.

  Para alcanzar el máximo despliegue posible de un CDSS, este debe interoperar con la historia clínica electrónica del paciente y otros sistemas de conocimiento del sistema sanitario. Vocabularios de terminología médica en forma de ontologías (como SNOMED CT, ICD9/10 y LOINC) han permitido la conceptualización de los diagnósticos, los procesos, el instrumental y demás términos involucrados en el proceso asistencial. Los estándares europeos ISO/EN 13606, definidos por el Comité Europeo de Normalización (CEN), y EN 13940 están específicamente diseñados para la representación de la historia clínica electrónica y para la representación de la continuidad del cuidado, respectivamente, asegurando la compatibilidad futura con otros sistemas existentes. Otros estándares de amplia difusión para el intercambio de información clínica electrónica son HL7-CDA y openEHR.

- Integrarse con los procesos asistenciales.

  Un factor fundamental de éxito en la implantación de los CDSS es su adecuado ajuste al proceso asistencial del entorno donde actúa. Así pues, el conocimiento suministrado por los CDSS debe transmitirse a los profesionales adecuados en el tiempo y forma necesarios. Una herramienta interesante para adaptarse al proceso asistencial son las Guías de Práctica Clínica (o Guías Clínica, GC). Una Guía Clínica es un documento que recoge directrices elaboradas sistemáticamente para asistir a clínicos y a pacientes en la toma de decisiones sobre la atención sanitaria adecuada para problemas clínicos específicos. En los últimos 25 años podemos encontrar ejemplos de sistemas de información que implementan GC, como EON[b], Asbru[c], PROforma[d], Prestige[e], PRODIGY[f], GLIF[g] y SAGE[h].

- Ofrecer fiabilidad en los resultados.

  Es deseable una respuesta fiable en el uso continuado de un CDSS, lo que conlleva la evaluación dinámica de la calidad de las decisiones y la actualización de los flujos de decisiones. Para llevar a cabo una evaluación dinámica de forma diferencial se ha propuesto recientemente un modelo de auditoría que tiene en cuenta el rendimiento de las decisiones en tiempo de diseño y en tiempo de ejecución[6]. Además, otros modelos miden lo adecuado de un sistema para el uso que se está haciendo mediante información contextual de los casos utilizados durante el diseño y los casos donde se está aplicando[6].

- Mantenerse actualizado.

  Por último, la aplicabilidad de los sistemas de ayuda en los entornos médicos puede ser mejorada mediante la adaptabilidad continua al medio, asumiendo los cambios observados en casos de nueva adquisición o la redefinición de nuevas decisiones a solucionar.

1.2. Taxonomías de sistemas de ayuda a la decisión médica

Los CDSS han sido catalogados por la National Electronic Health Records Taskforce Report en cuatro niveles según la complejidad de la generación del conocimiento específico del paciente [7]:

- Nivel I, engloba los CDSS que ofrecen acceso directo a información categorizada relativa a los pacientes, pero que requiere un procesado posterior[i].

[b] `http://bmir.stanford.edu/projects/view.php/eon`
[c] `http://www.asgaard.tuwien.ac.at/plan_representation/asbru_doc.html`
[d] `http://www.cossac.org/technologies/proforma`
[e] `http://www.openclinical.org/gmm_prestige.html`
[f] `http://prodigy.clarity.co.uk/home`
[g] `http://www.openclinical.org/gmm_glif.html`
[h] `http://sage.wherever.org/`
[i] En este libro no consideramos el estudio de este nivel por no realizarse una elaboración de conocimiento a partir de información específica del paciente.

Tabla 1.1: Soporte que dan los CDSS a las características enumeradas de la sección 1.1 categorizado por niveles de la National Electronic Health Records Taskforce Report.

Característica	- Soporte +
Manejar situaciones con incertidumbre	N2 N3 N4
Adaptado al problema	N1 N2 N3 N4
Basado en la evidencia/Referenciado	N1 N2 N3 N4
Interoperable	Según la interacción: estándares
Ajustado al proceso asistencial	Según la interacción: guías clínicas
Fidedigno/Preciso	N2 N3 N4
Actualizado/Dinámico	N4
Control del riesgo	N4
Escalable a múltiples centros	N1 N2 N3 N4
Interpretable en términos médicos	N3 N4 N2

- Nivel II, incluye los CDSS que presentan alertas basadas en el cambio de estado de algunas variables de los pacientes, pero que deben ser revisadas por los expertos.
- Nivel III, estos CDSS son sistemas deductivos que permiten inferir resultados según una base de conocimiento y la situación del paciente.
- Nivel IV, son CDSS basados en reglas de decisión que utilizan modelos predictivos inferidos a partir de casos del problema a resolver, generalmente basados en funciones discriminantes, modelos probabilísticos y/o lógica difusa.

La tabla 1.1 ilustra de forma comparativa el soporte que da cada nivel de CDSS a las características enumeradas en la sección 1.1.

Si atendemos a la interacción con el sistema sanitario podemos distinguir entre:

- CDSS autónomos

  Son aplicaciones de uso específico, que requieren la introducción manual de los datos de entrada y dan el resultado de forma visual. Este tipo de interacción debe ser tenida en cuenta cuando la funcionalidad del CDSS aporte gran valor añadido al profesional (por ejemplo, el cálculo del riesgo en decisiones quirúrgicas) y la cantidad de información biomédica propia del paciente es pequeña y de fácil acceso. En estos casos, la implementación de las herramientas mediante interfaces *responsive* pueden

aumentar la aceptación por los profesionales y pacientes involucrados en el cuidado de su salud al facilitar la asccesibilidad al sistema. Podemos por lo tanto considerar a las calculadores de salud y riesgo encontrados en la literatura de *mobile health* como CDSS autónomos implementadas para terminales móviles. Una modalidad no totalmente introducia en la actualidad son los servicios a terceros por internet, generalmente a través de navegadores o plataformas B2B, por lo que no requieren un despliegue específico por parte del usuario.

- Interconexión con dispositivos específicos.

  La conexión se realiza con el dispositivo de adquisición de datos biomédicos, como puede ser un hemograma o los resultados obtenidos por laboratorios computerizados. El CDSS aporta automáticamente un resumen de riesgos asociados a los resultados de los análisis.

- CDSS interoperables con la Historia Clínica Electrónica.

  La interacción natural de los expertos con un CDSS se realiza mediante los Sistemas de Información Hospitalaria y Poblacional. En esta modalidad, los datos clínicos y las señales biomédicas son incorporados en el CDSS mediante estándares de integración (p.e. HL7). Asimismo, los interfaces de usuario de los sistemas de información pueden incorporar los resultados de los CDSS. También algunas aplicaciones *mobile health* (por ejemplo, para el cuidado continuo de crónicos o la recuperación de depresiones) interactúan con la Historia Clínica Electrónica y las guías clínicas para implementar *personal health systems*.

Por último, si atendemos a la iniciativa para interactuar con el usuario, podemos clasificar los sistemas en:

- CDSS proactivo

  Un sistema proactivo es capaz de comenzar la interacción con el usuario por iniciativa propia, mostrando en el momento y lugar adecuados la información necesaria para la decisión en curso.

- CDSS reactivo

  Un sistema reactivo es aquel que debe ser interrogado para ofrecer una respuesta, y por lo tanto la iniciativa debe partir del usuario para comenzar la interacción.

Berlin et al. en [8] observaron dos grandes grupos de CDSS dependiendo del usuario al que iban dirigidos. En un conjunto de artículos publicados entre 1998 y 2003 encontraron un grupo formado por el 38 % de los casos que describían CDSS para la asistencia directa o indirecta (e.g. telefónica y correo) del cuidado preventivo de los pacientes. El segundo grupo encontrado en su estudio estaba formado por el 18 % de los artículos que describían sistemas de hospitalización dirigidas a los médicos.

En la sección 18.1 se analiza cómo la elección del tipo de CDSS supone un factor para el éxito o fracaso de la implantación del mismo.

1.3. Funcionalidades de los CDSS

Las funcionalidades de los CDSS encontrados en la literatura especializada y en especificaciones de productos disponibles en el mercado pueden agruparse en las siguientes categorías:

- Sistemas de alerta, vigilancia y/o recordatorios de situaciones de salud de pacientes.

  Estos sistemas proporcionan alertas para prevenir a los profesionales de condiciones que estén contraindicadas a las intervenciones sobre los pacientes, por ejemplo los sistemas de prescripción electrónica, Computerized Physician Order Entry (CPOE). Esta funcionalidad también incluiría la monitorización de datos biomédicos adquiridos de forma continua, como por ejemplo como resultado de los análisis de laboratorio.

- Ayuda al diagnóstico y al pronóstico.

  La ayuda al diagnóstico y al pronóstico médico han sido las funcionalidades más estudiadas por los proyectos de investigación sobre CDSS de Nivel IV. Los sistemas de ayuda al diagnóstico clasifican nuevos casos a partir de la información específica, asociándoles una probablidad de pertenencia a la clase y/o el riesgo de la clasificación. Por su parte, los sistemas de ayuda al pronóstico ofrecen un indicador de pronóstico del paciente ante la influencia de una serie de factores y/o tratamientos posibles.

- Gestión de patologías crónicas.

  El valor añadido que ofrecen los sistemas integrales de gestión de patologías crónicas viene dado por la gestión de alertas, monitorización continua y recomendaciones a los actores involucrados en dichas enfermedades: pacientes, cuidadores, médicos de primaria y médicos de especialidades. Por ejemplo, un sistema de gestión de personas con diabetes muestra a los pacientes las causas y actuaciones recomendadas a partir de reglas generales, riesgos y cumplimiento de objetivos, así como su evolución histórica. Esta información se puede completar con reglas de seguimiento y tratamiento para los servicios de primaria, además de las comorbilidades para los servicios médicos especializados.

- Planificación de procedimientos diagnósticos y tratamientos.

  Relacionado con la planificación eficiente de los procesos asistenciales, los CDSS han sido utilizados para ayudar a decidir el mejor procedimiento diagnóstico para un paciente dadas las circunstancias concretas en las que se encuentra. Como subproducto del diseño de los CDSS para la decisión de los procedimientos diagnósticos se pueden llegar a obtener pautas para mejorar la estructura de los procesos asistenciales.

  Las guías clínicas electrónicas son un buen soporte para la planificación y control de los procesos asistenciales, permitiendo introducir las conclusiones de la medicina basada en la evidencia en la práctica clínica diaria.

  La planificación de tratamientos puede abordarse de forma similar a la ayuda al diagnóstico y pronóstico, siendo común incluir análisis de riesgos y restricciones en el diseño del CDSS.

- Triaje.

  Una funcionalidad de los CDSS dirigida a la mejora de la eficiencia es su uso como sistema de triaje. En función de un conjunto mínimo de variables observacionales y poblacionales, estos sistemas procuran catalogar a los pacientes en base a criterios expertos para asignarles los tiempos y recursos óptimos en un servicio masificado, como puede ser el de urgencias. Para estos sistemas, es crítico el análisis de sensibilidad del decisor.

- Calidad asistencial.

  Las guías clínicas basadas en la evidencia científica permiten medir la variabilidad de la asistencia respecto a un estándar. Asimismo, el registro de las acciones asistenciales y la observación de la evolución del paciente permiten evaluar la calidad asistencial obtenida a partir de los procedimientos aplicados. De esta forma, un guía clínica electrónica implantada en un proceso asistencial e interoperable con los sistemas de información permite el estudio y optimización de la variabilidad asistencial, la detección de errores sistemáticos y la evaluación de decisiones críticas.

- Gestión de costes temporales y recursos.

  Los cuadros de mando basados en *business intelligence* son habituales como herramienta de ayuda a la decisión en la gestión de servicios e instituciones sanitarias. Estos sistemas permiten medir costes temporales y económicos y planificar requerimientos de los recursos para optimizar la eficiencia de la asistencia sanitaria.

- Interpretación de señales biomédicas.

  Típicamente, los CDSS han servido para la interpretación de señales biomédicas de cierta complejidad utilizados en procedimientos diagnósticos. Así pues, es extensa la literatura en aplicación de ayuda a la decisión mediante reconocimiento de formas para el diagnóstico mediante espectroscopía de resonancia magnética.

- Búsqueda de biomarcadores discriminantes.

  La investigación clínica es cada vez más común en los centros sanitarios. Esta investigación se apoya en la experimentación computacional, lo que permite contrastar aquellas hipótesis médicas que hagan uso de biomarcadores procedentes de datos biomédicos de múltiples niveles biológicos. Los equipos multidisciplinares de investigación suelen equiparse con librerías de análisis de datos de fácil uso en investigación clínica para el estudio estadístico de biomarcadores diagnósticos o pronósticos. El acceso comprensible a las técnicas de análisis de datos por los equipos multidisciplinares en salud facilitan la investigación clínica.

- Búsqueda de casos similares.

  Un buscador de casos similares puede dar mucha información a los médicos que realizan un diagnóstico diferencial para un nuevo paciente. La búsqueda puede realizarse mediante un conjunto de variables clínicas, señales biomédicas, imágenes o conceptos clínicos estructurados.

- Búsqueda de información bibliográfica relacionada con el paciente.

Los artículos científicos indexados en *pubmed* son la fuente de conocimiento no estructurada más rica que existe en medicina. Una búsqueda basada en minería de textos en las bases de datos científicas puede complementar el resultado de modelos predictivos y guías clínicas con documentos relevantes relacionados con el problema del paciente.

- Búsqueda de patrones anómalos.

  Complementario a la funcionalidad de alertas, la búsqueda de patrones anómalos permite la detección de casos cuyos datos biomédicos observados no siguen la distribución de casos del problema médico. Estos casos suelen ser de interés por su dificultad diagnóstica, su falta de calidad o por reflejar posibles subpatrones en grupos de pacientes. Esta funcionalidad permite a los CDSS la recomendación de revisiones de casos anómalos a los expertos.

- Simulaciones de escenarios.

  Los sistemas dirigidos a pacientes deben aportar una gran componente educacional que puedan servir para la prevención de la enfermedades o el control de hábitos en pacientes crónicos, por ejemplo personas con diabetes. Un sistema de ayuda a la decisión puede servir para simular escenarios donde los pacientes se sientan identificados y por lo tanto puedan visualizar lo que ocurriría si continuaran o cambiaran a cierto estilo de vida. De forma similar, la simulación puede ser de gran utilidad para la elección de terapias, mediante la visualización del cálculo de riesgos o supervivencias asociadas a las diferentes opciones durante la decisión.

- Sistema educativo basado en casos.

  Ciertos casos médicos o procedimientos diagnósticos son de difícil aprendizaje para los profesionales noveles, por ejemplo la interpretación de espectroscopía de resonancia magnética nuclear. El uso de sistemas de ayuda a la decisión que incorporen la predicción de diagnósticos y la interpretación de las señales es una herramienta de aprendizaje basado en casos que puede adaptarse al perfil de profesionales que están aprendiendo nuevas técnicas.

1.4. Antología de los sistemas de ayuda a la decisión médica

Los primeros CDSS usados en la práctica clínica fueron diseñados entre 1970 y 1980. Leaper et al. [9, 10] desarrollaron un CDSS para la ayuda al diagnóstico y la cirugía del dolor abdominal agudo basado en una aproximación naive Bayes. Al principio de la década de 1980, Shortliffe et al. [11, 12] diseñó un CDSS para la asistencia médica al tratamiento de pacientes de cáncer que recibían quimioterapia.

Los textos históricos en informática médica son una primera referencia para conocer la evolución de los CDSS. Shortliffe está considerado como uno de los pioneros de la Informática Biomédica. La decisión médica y los sistemas de ayuda a la decisión médica han sido temas recurrentes en las diferentes ediciones [13, 14] del libro `(Bio)Medical Informatics`. Berner recopila en [15] una serie de ensayos sobre Sistemas de Ayuda a la

Decisión Médica en dos secciones. La primera sección se centra en la teoría y tecnología necesaria para el desarrollo y evaluación de los sistemas, mientras que la segunda sección describe cuatro casos de uso en la práctica clínica.

Generalmente, los estudios encontrados en la literatura sobre CDSS son específicos de problemas médicos, como los siguientes ejemplos en detección, diagnóstico o pronóstico de cáncer: mama [16–25], gastrointestinal [26–29], hematológicos y linfáticos [30–36], piel [34, 37–42], oral [43, 44], pulmón [45–52], vejiga [53–57], o próstata [58, 59, 59, 60].

Ong recopila en [61] diversas experiencias de informática médica, incluyendo temas sobre prescripción electrónica, CPOEs, Retorno de la inversión (RoI), certificación de ICT en salud, gestión de la identidad, EHR, selección del software, dirección de proyectos, calidad y transición a organizaciones informatizadas.

Chen et al. profundizan en la gestión del conocimiento y la minería de datos en biomedicina en la edición de [62]. El recopilatorio incluye ensayos sobre técnicas de minería de datos, minería de textos, gestión y extracción de conocimiento, ontologías biomédicas, modelos de representación de la información, privacidad, ética, casos prácticos en centros de salud y otros retos en bioinformática. También incluyen varios ensayos con referencia a la inclusión y análisis de datos procedentes de biología molecular.

Escolar, en [63], identifica la ayuda a la decisión mediante la descripción, deducción e inferencia sobre los datos de los pacientes dentro de las consecuencias de la implantación de las historias clínicas electrónicas.

Zamorano et al. abordan en [64] temas relacionados con la telemática y los modelos de negocio relacionados con la telemedicina, que pueden inspirar posibles funcionalidades de los CDSS.

En 2016 Gulshan et al. desarrollaron y validaron un CDSS basado en *deep learning* para la detección de retinopatía diabética en fondo de ojo [65]. Un año más tarde la Food and Drug Administration (FDA) aprobó la comercialización del primer dispositivo médico usando inteligencia artificial para detectar retinopatía diabética.

La asistencia clínica y la gestión hospitalaria han incorporado la evidencia clínica y los paneles de mando como herramientas informativas en su práctica diaria. Además, desde hace ya varias décadas se ha estado recopilando información biomédica de los pacientes asociada a su asistencia clínica, certificación de fallecimiento, participación de ensayos clínicos, etc. En la actualidad, existen unos 16.000 hospitales en el mundo recopilando digitalmente datos biomédicos de pacientes [66]. El 80 % de esta información es no estructurada y está almacenada de forma distribuida en diversos formatos. La información genética ya supone la cuarta fuente de información masiva mundial [67]. Se estima que los servicios de radiología mundiales generan 69 PB/año y secuenciar a cada recién nacido supondría 100 PB/año [68]. Cada vez más pacientes están siendo monitorizados a través de unidades de seguimiento médico a domicilio, estimándose que 5 millones de pacientes usarán dichos servicios generando 1.000 lecturas por segundo y paciente [69]. En resumen, en 2020 se tardará sólo dos meses en doblar la cantidad de datos de salud almacenados por los servicios sanitarios, cuando hace 10 años se tardaba menos de tres años.

El futuro de la salud está en los datos. Efectivamente, el avance en el conocimiento de la etiología de las enfermedades y sus variantes, el resultado beneficioso o adverso de los tratamiento en poblaciones cada vez más específicas, la evaluación del efecto combinado de tratamientos y condiciones ambientales y en definitiva la personalización de la medicina a los individuos vendrán dados por el análisis preciso de los datos biomédicos.

En Europa se han comenzado a crear estructuras nacionales para coordinar las acciones industriales en *big data* en salud. Concretamente, en Alemania se ha creado la plataforma Smart Data Innovation Lab constituida en la región de Baden-Württemberg por 40 grandes empresas tecnológicas e industriales y centros y asociaciones de investigación con el fin de transferir los adelantos en *big data* directamente a los sectores estratégicos, como son la medicina personalizada. En Francia se ha creado la alianza Teralab entre Institut Mines-Télécom y Groupe des Écoles Nationales d'Économie et Statistique con el fin de acelerar la investigación e innovación en análisis *big data* para crear futuros profesionales en el nuevo sector. En Reino Unido, el NHS National Institute for Health Research (NIHR) y la Medicines and Healthcare products Regulatory Agency (MHRA) crearon el repositorio de registros de atención primaria anonimizados Clinical Practice Research Datalink (CPRD) con información desde 1987. El instituto Farr de la UCL ha desarrollado sobre el CPRD el repositorio de datos clínicos enlazados CALIBER para la investigación epidemiológica en enfermedades cardiovasculares. También es interesante destacar en Europa la creación de la *big data* Value Association (BDVA) que, fundada por 24 miembros, ya reúne a gran parte de la comunidad académica e industrial del sector con el propósito de promover el desarrollo y aplicación de las tecnologías *big data* en el programa Horizonte 2020.

En Estados Unidos, el National Institutes of Health (NIH) coordina 27 centros para el desarrollo de Ciencia de datos en salud a través del NIH Scientific Data Council y de la NIH Office of the Associate Director for Data Science (ADDS). Específicamente, el NIH ha creado el programa específico *big data* to Knowledge (BD2K) para el beneficio de la investigación sanitaria a través de la tecnología *big data*. BD2K ha creado 13 centros de excelencia en computación *big data* para el desarrollo de nuevas aproximaciones, métodos, software, infraestructura computacional y entrenamiento. El Informatics for Integrating Biology and the Bedside (i2b2) es un centro NIH para la biomedicina computacional dentro del Partners Healthcare System (Boston). I2b2 desarrolla un sistema escalable para la investigación mediante datos clínicos y genómicos. I2b2 se está consolidando como un estándar de facto entre la comunidad internacional. Las regiones con una alto potencial académico y grandes hospitales de referencia han visto la oportunidad de situarse a la vanguardia de los servicios *big data* en salud mediante la formalización de alianzas interdisciplinares. La Pittsburgh Health Data Alliance entre la Carnegie Mellon University, la University of Pittsburgh y el hospital UPMC. El modelo de colaboración, financiado con un presupuesto de $20M, espera la transferencia a través de UPMC Enterprises de los resultados producidos por dos nuevos centros tecnológicos en *big data* en salud (Center for Machine Learning and Health, CMLH y Center for Commercial Applications of Healthcare Data, CCA). A nivel privado, grandes farmacéuticas como AstraZeneca y Janssen han establecido convenios con HelthCore y EuroRec, respectivamente, para la explotación del Real World Data (RWD) con el objetivo de determinar los tratamientos más eficaces y eficientes en enfermedades crónicas y complejas. También el sector tecnológico está posicionando sus servicios *big data* en el sector salud, como son las iniciativas IBM Watson Health y diversos proyectos empresariales sobre la plataforma Azure de Microsoft.

Robert A. Greenes en [70] y en [71] recopiló y actualizó una visión de la ayuda a la decisión en salud cercana a los sistemas de información y a la gobernanza del conocimiento médico. El anexo B hemos clasificado según su nivel, interacción y funcionalidades principales algunos CDSS disponibles como productos, prototipos científicos o proyectos de investigación en el momento de la edición del libro.

1.5. Integración de los CDSS en entornos sanitarios

Los estudios sobre el valor añadido de las Historias Clínicas Electrónicas concluyen la necesidad de incorporación de funcionalidades de ayuda a la decisión clínica en los EHRs como estrategia para aportar mejoras en la calidad asistencial de forma continuada. Se espera que la siguiente generación de sistemas para la atención integral ambulatoria y sanitaria sean sistemas con funcionalidades de ayuda a la decisión médica.

Para una integración total de estos sistemas con los procesos asistenciales, estos interfaces deberían adaptarse a la etapa concreta donde se encuentra el caso de estudio obteniendo, a través de los motores de ayuda a la decisión, las consideraciones que optimicen la planificación del paciente. Por último, esta integración convertiría los Sistemas de Información Sanitaria en sistemas activos que generasen alertas ante condiciones detectadas automáticamente en las Historias Clínicas Electrónicas.

La industria de los sistemas de ayuda a la decisión tiene una de las mayores expectativas de crecimiento en el sector de las Tecnologías de la Información. Solo el mercado de Business Intelligence (BI) creció un 12 % en el año 2003, y un 22 % en 2008, con unos ingresos de 8.8 billones de dólares.

Concretamente, en medicina ya se tienen estudios de valor añadido sobre CDSS para la prescripción electrónica (CPOE), con resultados que demuestran una menor variabilidad en las actuaciones entre profesionales, un aumento de la supervivencia y un descenso de la tasa de error médico. Desde el punto de vista de la eficiencia, el uso de CPOE con funcionalidades de CDSS se estima superior a 44 billones de dólares anuales en el servicio ambulatorio de EEUU gracias a la reducción de medicación no requerida, de pruebas radiológicas y de tests de laboratorio, así como un descenso de Efectos adversos de fármacos (Adverse drug event) (ADEs). Johnston et al. en [72] estiman una reducción anual de 2 millones de ADEs mediante la implantación nacional de un CPOE ambulatorio con funcionalidades de CDSS.

1.6. Conclusión

En este capítulo hemos revisado los conceptos los conceptos básicos relacionados con los Clinical Decision Support Systems (CDSS). Tras definir los CDSS como sistemas computacionales que aportan conocimiento específico para las decisiones médicas que deben tomarse en los planes de tratamiento de pacientes, como por ejemplo durante el diagnóstico, pronóstico, tratamiento y administración, hemos comprobado su conexión con la medicina de precisión y su utilidad para convertir la práctica clínica en preventiva, predictiva, personalizada y participativa.

Como resultados, se estable que el objetivo principal de estos sistemas es la mejora de la atención sanitaria individualizada y el objetivo secundario es el aumento de la eficiencia de los servicios sanitarios.

Para proporcionar conocimiento objetivo para la solución de un problema médico ne base a la información específica de los pacientes, es deseable una serie de características de estos sistemas computacionales, siendo los más destacables la capacidad de manejar la incertidumbre, poder resolver problemas específicos, utilizar evidencia científica, ser interoperable semánticamente, integrarse en procesos asistenciales, ser fiable y mantenerse actualizado.

Estas características están sobretodo recogidas en los CDSS englobados en los Niveles III y IV de la clasificación elaborada por el National Electronic Health Records Taskforce Report.

Las funcionalidades más destacadas que los CDSS aportan a la práctica asistencial (uso primario en medicina) y a la explotación de la información médica (uso secundario) son la alerta y/o recordatorio de situaciones de salud de pacientes concretos, la ayuda al diagnóstico y al pronóstico, la gestión de patologías crónicas, el soporte a reuniones multidisciplinares, el triaje, la calidad asistencial, la gestión de costes, tiempos y recursos, la interpretación de señales biomédicas, la definición de biomarcadores discriminantes, la búsqueda de casos similares, la búsqueda de información bibliográfica relacionada con el paciente, la búsqueda de patrones anómalos, las simulaciones de escenarios de salud y la educación de personal sanitario basado en casos.

En definitiva, la industria de los sistemas de ayuda a la decisión tiene una de las mayores expectativas de crecimiento en el sector de las Tecnologías de la Información, siendo una de las tecnologías requeridas para la medicina de precisión y los sistemas de información para las reuniones multidisciplinares.

Capítulo 2

Decisiones en sanidad

Un plan de tratamiento del paciente está formado por la secuencia de decisiones que se van tomando con el objetivo de mejorar su salud en la medida de lo posible. Cada decisión médica implica a varios actores, incluido el paciente, y conlleva una serie de acciones generalmente complejas. Los conocimientos, técnicas y recursos necesarios para llevar a cabo las decisiones y acciones médicas requiere de grandes estructuras organizativas, denominadas servicios sanitarios, que suministran la asistencia sanitaria a la población a través de complejos y generalmente dispersos procesos asistenciales.

La dimensión de la asistencia sanitaria es enorme. En el plano social, la asistencia médica está recogida en el artículo 25 de Declaración Universal de los Derechos Humanos (1948) de la Organización de las Naciones Unidas. Desde el punto de vista económico, los servicios de salud rondan el 9 % del Producto Interior Bruto (PIB) en los países desarrollados. Desde el punto de vista del avance científico, la salud es uno de los potenciadores mundiales de la investigación, con programas propios como los de salud y tecnologías de la información para la salud en los programas marco de I+D de la UE.

La tendencia actual de los servicios sanitarios es la optimización de los recursos. Prácticamente todas las estrategias actuales comparten el estudio de los procesos asistenciales, la reutilización de la información clínica para la investigación y la incorporación de los avances científicos en la práctica clínica como marco de mejora de los servicios de salud.

En este capítulo repasaremos la estructura y los actores de la asistencia sanitaria, llegaremos a definir los procesos asistenciales y veremos las tendencias actuales para incorporar las mejores prácticas clínicas mediante guías clínicas. Finalmente, analizaremos los elementos de las decisiones médicas bajo la perspectiva de la teoría de la decisión, lo que nos llevará a justificar el uso de CDSS como herramienta objetiva de apoyo a los planes de tratamiento personalizados, predictivos, preventivos y participativos.

2.1. La asistencia sanitaria

La asistencia sanitaria se define como la prevención, tratamiento y manejo de las enfermedades, lesiones, y otros problemas físicos y mentales de las personas mediante servicios profesionales ofrecidos por la medicina, enfermería, farmacia, odontología, fisioterapia, biotecnología, tecnologías de la información para la salud y otras disciplinas afines. Según la Organización Mundial de la Salud (`www.who.int`), la asistencia sanitaria incluye los recursos y servicios asociados a la promoción de la salud de individuos y poblaciones, incluyendo intervenciones preventivas, curativas y paliativas.

2.1.1. Tipos de asistencia sanitaria

El acceso a la asistencia sanitaria varía entre regiones y grupos sociales, siendo un factor clave en la explicación de dicha variación la condición socio-económica y la política sanitaria de la región. En términos generales, se suele identificar una estructura de prestación de asistencia sanitaria formada por atención primaria, atención secundaria y salud pública. Los cambios sociales que aparecen conforme avanzan las sociedades han generado nuevas necesidades de asistencia sanitaria a la población, como por ejemplo la atención en el hogar, la atención residencial, la atención a personas en situación de dependencia y la promoción de su autonomía personal. En los servicios de salud, la unión de procesos asistenciales de diferentes niveles es uno de los retos para conseguir una asistencia sanitaria continuada, preventiva, personalizada y universal.

Atención primaria

La Atención Primaria se encarga de suministrar los servicios de salud fundamentales a la población local, por lo que resulta el nivel básico de asistencia sanitaria de cualquier sistema de salud. Desde el punto de vista del ciudadano, es el primer punto de consulta al servicio sanitario, generalmente a través de un médico de familia, médico de medicina general o pediatra. Según el modelo asistencial del sistema sanitario prestador de los servicios, este primer profesional puede ser también una enfermera, un farmacéutico o un trabajador sanitario.

La atención primaria incluye el rango más amplio de casos médicos a tratar, desde casos agudos a crónicos, así como problemas físicos, mentales o psicosociales. La atención primaria puede ser urgente, planificada o continuada como se hacen con los pacientes crónicos (esto es: hipertensión, diabetes, asma, EPOC, depresión, ansiedad, dolor de espalda, artritis y problemas de tiroides), infantiles y maternidad. Precisamente, la continuidad asistencial, cualidad fundamental del servicio y el primer nivel de la pirámide asistencial, tiene un aspecto intrínsecamente regulador de la carga del sistema sanitario, de promotor de la prevención de enfermedades y de la salud poblacional, de recuperación de la salud, paliación del dolor y de prescripción de medicación.

Atención secundaria

La atención secundaria es el conjunto de servicios sanitarios suministrados por especialistas médicos y profesionales de la salud que no suelen tener el primer contacto con los pacientes. Aunque se suele identificar la atención secundaria con la atención hospitalaria, esta última no incluye toda la atención secundaria, por ejemplo psiquiatría. Además, algunos servicios primarios se suministran en los hospitales. La atención secundaria planificada en los sistemas sanitarios públicos se prescribe desde atención primaria o desde otro servicio de atención secundaria. En los sistemas basados en seguros médicos privados también suele imponerse este modelo. La atención secundaria se encarga de la atención a pacientes agudos que requieren un tratamiento de urgencia durante un periodo corto de tiempo. También se encarga de los servicios de cirugía, maternidad, oncología, cuidados intensivos, diagnóstico por imagen o anatomía patológica, entre otros.

Algunos sistemas incluyen las especialidades médicas avanzadas como atención terciaria, ya que no suele ser prestada por todos los servicios de atención secundaria, sino por

algunos centros de referencia. Algunos de estos servicios son las unidades integrales de cáncer, neurocirugía, cirugía cardíaca, neonatología, foniatría y cuidados paliativos.

Atención domiciliaria y atención residencial

La inversión de la pirámide poblacional en los países desarrollados está creando nuevas necesidades en la atención sanitaria. Cada vez más se generalizan servicios sanitarios realizados fuera de las instalaciones sanitarias. La atención domiciliaria se centra en el seguimiento y tratamiento de pacientes crónicos, personas dependientes y personas mayores, pero también se crean modelos de hospitalización domiciliaria en procesos largos de recuperación posquirúrgica. Los objetivos europeos en TIC salud centrados en *Personal Health Systems* y *Ageing Well* se han centrado en la capacitación de las personas y sus entornos para una participación activa de los ciudadanos en la gestión de su salud. Por su parte, la Atención residencial suele identificarse con los servicios geriátricos llevados a cabo en residencias permanentes o residencias de día.

Salud pública

La salud pública no es un servicio sanitario directo al individuo, sino la encargada de proteger la salud a nivel poblacional. Por ello, sus funciones van dirigidas a políticas y actuaciones que engloban el total de la población. Estas funciones hacen que los procesos asociados a salud pública suelan ser a medio y largo plazo, incluyendo la generación de políticas para la protección y promoción de la salud, educación de la población y la planificación y capacitación para la actuación ante riesgos sanitarios. Si bien los sistemas sanitarios suelen separar la salud pública de la atención sanitaria, cada vez más se establecen enlaces entre ambos, atendiendo al carácter preventivo así como predictivo y personalizado, por ejemplo los programas de *screening* de mama.

2.2. Los CDSS como servicio auxiliar de la atención sanitaria

Los CDSS pueden considerarse tecnologías de la información en salud y, como tales, serían servicios auxiliares a la atención sanitaria. Estos sistemas pueden venir promovidos por las propias instituciones sanitarias dentro de sus estrategias de modernización de los procesos asistenciales.

En estos casos, los CDSS se plantean como servicios computacionales instalados en los sistemas centrales de los servicios de salud e integrados con los sistemas de información hospitalaria o historias de salud electrónicas. También las farmacéuticas están incorporando cada vez más CDSS para la detección de alertas asociadas a pacientes incluidos en ensayos clínicos, o inclusión de pacientes en los ensayos. Los laboratorios también son otro de los lugares donde los sistemas de alertas tienen buena acogida a nivel institucional. Muchos médicos especialistas suelen ser promotores de CDSS para resolver problemas generalmente difíciles en base a nuevas técnicas o protocolos relativos a su investigación particular que pueden ser publicados como prototipos científicos. Por último, empresas integradoras de soluciones de historia clínica electrónica o incluso de empresas de servicios en el hogar se plantean la inclusión de módulos de valor añadido basados en las tecnologías de CDSS, por ejemplo, módulos inteligentes de seguimiento de diabéticos en el hogar.

El anexo C presenta a los actores principales en salud y algunos ejemplos de servicios sanitarios que pueden ser de interés para conocer el contexto humano y organizacional donde se situan los CDSS.

2.3. Procesos asistenciales y guías de práctica clínica

La European Foundation for Quality Management (EFQM) define un proceso como la *organización lógica de personas, materiales, energía, equipos y procedimientos en actividades de trabajo diseñadas para generar un resultado específico*. En definitiva, un proceso es una sucesión de actividades realizadas por una organización en el tiempo con un fin determinado. Por ello, en el marco de la atención sanitaria, un proceso asistencial es el conjunto de decisiones y actividades que los actores principales realizan con el objetivo de incrementar el nivel de salud de la población.

En la actualidad ha habido un gran interés por representar de forma explícita los procesos asistenciales, principalmente con el fin de medir su eficacia y aplicar mejoras donde sea posible. La metodología Integration Definition for Function Modelling (IDEF) permite identificar cómo un proceso se relaciona e integra con el resto de procesos de la organización, así como definir las entradas de cada proceso, por ejemplo la solicitud o requerimiento de un servicio, los recursos consumidos, los factores condicionantes (en forma de guías o protocolos) y la salida de cada proceso, es decir el servicio facilitado al usuario. Los procesos asistenciales suelen representarse en varios diagramas de flujos a distinto nivel: nivel 0 (macroproceso), nivel 1 (clínico asistencial), nivel 2 (subprocesos), y nivel 3 (actividad de cada subproceso). En el nivel 3, representado como diagrama de flujos aparecen puntos de decisión, que son los puntos de inflexión en los caminos posibles a recorrer durante los procesos y es donde los CDSS inciden. La representación secuencial de dichos puntos de decisión son en definitiva el plan de tratamiento de un paciente.

Como resultado de la redefinición de los procesos asistenciales, se han puesto en marcha iniciativas para alcanzar una gestión por procesos asistenciales integrados, que aseguren la atención sanitaria continuada a los pacientes mediante equipos multidisciplinares y en base a guías de práctica clínica. Algunos de los procesos asistenciales integrados identificados por diversas organizaciones son: ataque cerebro vascular, atención al enfermo pluripatológico, cáncer de cérvix y útero, cáncer de mama, catarata, cefaleas, colecistitis, colelitiasis, cuidados paliativos, dolor torácico, dolor abdominal, amigdalectomía/adenoidectomía, anemias, arritmias, asma del adulto, asma infantil, cáncer colorrectal, cáncer de piel, disfonía, disfunción tiroidea, dispepsia, hemorragias uterinas anormales, brucelosis, tuberculosis, diabetes, demencias, embarazo, parto y puerperio, enfermedad pulmonar obstructiva crónica, fractura de cadera, hiperplasia prostática benigna/cáncer de próstata, insuficiencia cardíaca, patología ostearticular crónica, síndrome de ansiedad, depresión y somatizaciones, VIH/SIDA, hepatitis vírica, hernias abdominales, insuficiencia venosa crónica, otitis, politraumatizados, prótesis articulares, síndrome febril de larga duración, síndrome febril en el niño, trastornos de la conducta alimentaria y trasplante hepático.

La evaluación tras varios años de la implantación de los procesos asistenciales integrados destacan la utilidad para la continuidad asistencial, la mejora de la calidad asistencial y la reducción de la variabilidad en la práctica clínica, así como su utilidad para acercar los flujos de trabajo de la atención primaria y la atención secundaria.

2.3.1. Guías de práctica clínica

Las guías de práctica clínica son recomendaciones para clínicos sobre el cuidado de los pacientes con condiciones específicas, generalmente basadas en la mejor evidencia científica y la mejor práctica clínica posible. Las guías clínicas recogen la revisión sistemática de la literatura sobre la pregunta médica y las recomendaciones para cada condición específica, en base a un índice de evidencia. Las recomendaciones pueden ser A, B, C o E en función del nivel de evidencia (ver tabla 2.1). A lo largo del documento se describirá, siempre que sea posible, el nivel de evidencia asociado a cada regla. El sistema de gradación es el siguiente:

- Nivel de evidencia A: existe una evidencia clara de estudios bien definidos, generalizables y aleatoriamente controlados.
- Nivel de evidencia B: aporta una evidencia de estudios de cohorte bien definidos.
- Nivel de evidencia C: aporta una evidencia de estudios poco o nada controlados.
- Nivel de evidencia E: consenso de expertos o experiencia clínica.

El National Health Service (NHS) británico ha realizado una revisión intensiva de la forma en que lleva a cabo la atención sanitaria. Unos de los elementos más importantes han sido la redefinición de procesos asistenciales y ha generado una serie de recomendaciones para los planes de tratamiento de los pacientes en forma de guías de práctica clínica. Como resultado, NHS, a través de NICE[a] difunde aproximadamente mil guías de práctica clínica catalogadas en siete categorías: cáncer, tratamiento, diagnóstico, procedimientos intervencionistas, tecnologías médicas, salud pública y evaluación de tecnología.

2.4. Teoría de la decisión y Sistemas de Ayuda a la Decisión en la asistencia sanitaria

Como hemos visto en este capítulo, la asistencia sanitaria suele organizarse en procesos asistenciales que implican secuencias de decisiones realizadas por múltiples agentes, incluyendo profesionales sanitarios y pacientes, entre otros. Arcelay [73] pone de manifiesto la importancia de las decisiones en su definición de proceso: "concatenación de las decisiones, actividades y tareas llevadas a cabo por diferentes profesionales en un orden lógico y secuencial para producir un resultado previsible y satisfactorio". Estos flujos de decisiones se caracterizan por cuatro características muy marcadas: la incertidumbre, el coste de las consecuencias, el dinamismo del ambiente y la gran complejidad del sistema.

Todo proceso de decisión se realiza en un ambiente de incertidumbre ya que nunca se dispone del conocimiento completo de la situación del paciente. Por lo tanto, toda acción sobre el paciente producida tras la correspondiente decisión asume un riesgo. Por lo tanto, definimos el riesgo como el coste esperado al tomar una decisión teniendo en cuenta el conocimiento (siempre incompleto) del problema a resolver. El ámbito médico es experto en su gestión y atiende, como premisa, a la minimización del riesgo de la persona atendida, es decir, a llevar al máximo la salud de las personas atendidas. Así pues, es necesario

[a] `http://guidance.nice.org.uk/CG`

usar metodologías que estén basadas en la gestión del riesgo. Los CDSS deben incorporar metodologías que permitan gestionar el riesgo en circunstancias de incertidumbre y utilicen la información disponible de los pacientes junto con el conocimiento del problema para resolver el caso de forma óptima.

En segundo lugar, los ambientes como el de salud son inherentemente dinámicos. Esta característica no ha sido totalmente identificada en los sistemas de información sanitarios actuales y, sin embargo, es de gran relevancia para apoyar un correcto desarrollo de la práctica diaria. El acceso a las últimas tendencias de la evidencia científica, a los biobancos multicéntricos y a los datos biomédicos actualizados del caso a resolver, ofrece un conocimiento potencial para la decisión individualizada que los CDSS deben poder aprovechar para adaptarse a estas condiciones dinámicas que implica su continua actualización.

En tercer lugar, la relevancia y diversidad que implica la salud, su asistencia, promoción y prevención, hacen del sistema sanitario una de las mayores organizaciones desde el punto de vista de su complejidad. En la actualidad se han incorporado metodologías de gestión de procesos en la organización de las actuaciones y recursos necesarios para realizar los diferentes actos asistenciales. Además, la definición de los mejores procedimientos para la atención de cada patología son en sí mismo complejos. Con el objetivo de reducir la variabilidad asistencial y aplicar las mejores prácticas médicas basadas en la evidencia científica se adoptan, cada vez más, protocolos estandarizados en los centros sanitarios basados en las guías clínicas que han sido definidas por organizaciones multicéntricas de gran prestigio nacional e internacional. Sin embargo, los sistemas actuales todavía no tratan las secuencias de decisiones de los procesos asistenciales teniendo en cuenta la incertidumbre ni el coste, ni mucho menos el dinamismo de los entornos médicos.

Otras características inherentes al flujo de decisiones médicas son la heterogeneidad, la gran dimensión y la variabilidad de los datos biomédicos utilizados en todos los procesos de decisión en la asistencia médica como fuente de conocimiento específico del caso de estudio. Es una necesidad, y un reto en el desarrollo de los CDSS, la extracción de conocimiento desde la gran variedad de fuentes de datos biomédicos. Será necesario el estudio de las fuentes de datos, su procesamiento y análisis, para llegar a producir información de calidad que sirva para el fin último que es la ayuda a la decisión médica. En próximos capítulos veremos las técnicas que nos permitan abordar estas tareas.

2.5. Notas bibliográficas

El informe realizado por Telefónica en [74] analiza el sistema sanitario español y en las comunidades autónomas, recopilando estadísticas del CIS y de otros organismos.

Escolar et al. [75] identifican de forma no exhaustiva a los actores principales de los entornos clínicos. Escolar y Carnicero coordinan el monográfico [76] sobre la integración de los sistemas departamentales, clínico-administrativos, económico-financieros, de recursos humanos y servicios de soporte de los servicios de salud. Carnicero coordina en [77] un monográfico sobre la gestión del medicamento en los servicios de salud.

La dirección general de la agencia de calidad del sistema nacional de salud define en [78] los indicadores clave del sistema nacional de salud organizados por conceptos y atendiendo a las leyes y estructuras aplicables.

El Servicio Andaluz de Salud preparó en 2001 una interesante guía sobre sus procesos asistenciales integrados, que actualizó en el 2009 [79, 80]. La primera edición resulta de gran interés por su estudio sobre el diseño de procesos y la identificación de los principales

procesos asistenciales integrados. La segunda edición muestra el estado de implantación años después del diseño.

En 2018, E.H. Shortliffe, uno de los primeros creadores de CDSS usados en la práctica clínica ha recopilado las que en su opinión son las capacidades y características que a fecha de publicación deberían incorporarse en un CDSS para ser aceptado e integrado en la rutina clínica [81].

Nivel de evidencia	Descripción de los estudios
A	Estudios multicéntricos bien definidos y concluidos.
	Meta-análisis con índices de calidad incorporados.
	Evidencia convincente no experimental, como la regla "todo o nada" desarrollada por el *Center for Evidence Based Medicine de Oxford.*
	Estudios bien definidos y concluidos en una o más instituciones.
B	Estudios de cohorte prospectivos o registros bien definidos y concluidos.
	Meta-análisis de estudios de cohorte llevados a cabo correctamente.
	Estudios de control de casos bien definidos.
C	Estudios clínicos aleatorios con un fallo metodológico grave o tres o más fallos metodológicos menores que pudieran invalidar los resultados.
	Estudios observacionales con un sesgo potencial importante (como en series de casos comparados con casos históricos).
	Series de casos o informes de casos.
	Conflictos con el peso de la evidencia que apoya la recomendación.
E	Consenso entre expertos o experiencia clínica.

Tabla 2.1: Sistema de gradación de la evidencia para recomendaciones de práctica clínica.

Capítulo 3

Teoría de la probabilidad y teoría de la decisión

La teoría de la decisión Bayesiana desarrolla el procedimiento cuantitativo para la toma de decisiones óptima. Para ello utiliza la teoría de la probabilidad para cuantificar la verosimilitud de los sucesos inciertos y el coste o beneficio de las consecuencias de las acciones ante dichos sucesos. Además, el estudio de este capítulo asienta las bases del aprendizaje automático, en cuyas técnicas profundizaremos en capítulos posteriores.

3.1. Teoría de la probabilidad

La probabilidad es una forma de cuantificar el grado de certeza asociado a un suceso. Los procesos asociados a la decisión médica suelen venir acompañados de incertidumbre debido a la falta de información o al carácter inherentemente ruidoso de la que se dispone. La herramienta para cuantificar y manipular la incertidumbre es la teoría de la probabilidad. Su combinación con la teoría de la decisión (sección 3.3) permite resolver la predicción óptima de decisiones cuando se dispone de información incompleta y ruidosa.

3.1.1. Suceso, variable aleatoria y probabilidad

La probabilidad $p(A|H)$ de un suceso A bajo unas condiciones H, es el grado de certeza de A que sugieren las condiciones H a un observador. Por simplicidad de notación, solemos obviar las condiciones H cuando estas son evidentes, expresando lo anterior simplemente como $p(A)$.

Un suceso está formado por un subconjunto de resultados elementales de un *conjunto de referencia* Ω. Este conjunto Ω incluye todos los resultados elementales de los sucesos en cuya verosimilitud estamos interesados. Todo suceso elemental tiene una probabilidad en el rango $[0, 1]$ y la suma de las probabilidades de todos los resultados elementales es 1.

Con el fin de poder operar con el conjunto Ω, los resultados elementales de Ω suelen cuantificarse mediante una variable X, típicamente llamada variable aleatoria. Esta variable aleatoria tomará un valor $X = x$ que estará contenido en el conjunto de referencia, esto es $x \in \Omega$.

Ejemplo 3.1 (Test de O'Sullivan)
Los protocolos de seguimiento del embarazo de bajo riesgo incluyen el *Test de O'Sullivan* en

la semana 28 de gestación, dentro del programa de *screening* de diabetes gestacional. Este test consiste en la medición mediante analítica química de la glucemia en plasma a los 60 minutos de la ingesta de 50 g. de glucosa.

Un valor en la medición igual o inferior a $140mg/dl$ se considera normal, por lo que el test de O'Sullivan se considera negativo, mientras que si es superior a este valor, el test es positivo y debe realizarse a la gestante una prueba de tolerancia oral con glucosa de tres horas y 100 g de glucosa. De forma general, se ha observado que el 20 % de los test de O'Sullivan dan resultados positivos [82].

Así pues, podemos expresar el problema anterior mediante el conjunto de referencia $\Omega = \{\text{rango de valores de glucemia en plasma}\} = \mathbb{R}^+$, ya que la glucemia no puede ser negativa y no hemos establecido cota superior. Establecemos la variable aleatoria x como la medición de glucemia en plasma en mg/dl y definimos dos sucesos de interés, que en este caso son excluyentes y cubren totalmente Ω, $O+$ (test positivo) y $O-$ (test negativo):

$$O+ = \{x; x \leq 140\text{mg/dl}\},$$
$$O- = \{x; x > 140\text{mg/dl}\}.$$

La evidencia científica disponible, que con este enunciado podríamos establecer de nivel III y grado de recomendación B[a], nos permite establecer, obviando las condiciones H en la notación, que la probabilidad de los sucesos bajo las condiciones $H = \{\text{gestante, 28 semanas, tras 60' de la ingesta de 50 mg de glucosa}\}$ es:

$$p(O+) = 0{,}2$$
$$p(O-) = 1 - p(O+) = 0{,}8$$

3.1.2. Leyes aditiva y multiplicativa de la probabilidad y Teorema de la probabilidad total

La relación de sucesos en términos de probabilidad resulta de especial relevancia. Conocidas las probabilidades de un conjunto de sucesos, A y B, las leyes aditiva y multiplicativa permiten obtener las probabilidades de las relaciones entre ellos.

La *ley aditiva* dice que si, dado H, A y B son sucesos excluyentes, esto es, $A \cap B = \emptyset$, entonces

$$p(A \cup B|H) = p(A|H) + p(B|H). \tag{3.1}$$

Así pues, la probabilidad de que suceda A o B es la suma de las probabilidades de cada suceso dada la condición H.

La *ley multiplicativa* dice que

$$p(A \cap B|H) = p(A|H)p(B|A, H). \tag{3.2}$$

Es decir, la probabilidad de que sucedan conjuntamente A y B es igual a la probabilidad de que suceda A multiplicada por la probabilidad de que suceda B condicionado a que suceda A. $p(A \cap B|H)$ suele expresarse también como la probabilidad conjunta $p(A, B|H)$.

[a] En base a Agency for Healthcare research and Quality (AHRQ) de los Estados Unidos. Diseñado por la Universidad de Aberdeen-Health Services Research Unit

Supongamos ahora que conocemos las probabilidades conjuntas $p(A, B_i)$ del suceso A y cada uno de los sucesos mutuamente excluyentes $B_i, i = 1, \ldots, k$ que forman el conjunto de referencia Ω. Como consecuencia directa de la ley aditiva (3.1) y debido a que $A \cap B_i$ es excluyente con $A \cap B_j, \forall j \neq i$, podemos marginalizar sobre un conjunto de variables para encontrar la probabilidad marginal del resto:

$$p(A) = \sum_{i=1}^{k} p(A, B_i), \tag{3.3}$$

y por la ley multiplicativa (3.2), el *teorema de probabilidad total* permite calcular $p(A)$ como:

$$p(A) = \sum_{i=1}^{k} p(A|B_i)p(B_i). \tag{3.4}$$

Ejemplo 3.2 (Valor predictivo positivo del test de O'Sullivan)
En el ejemplo 3.1 hemos visto que la probabilidad de obtener un test de O'Sullivan positivo es $p(O+) = 0{,}2$. Sin embargo, únicamente el 2,2 % de las gestantes desarrollan diabetes gestacional. Sabiendo que ninguna gestante con test negativo de O'Sullivan ha desarrollado diabetes gestacional, ¿cuál es el valor predictivo positivo del test de O'Sullivan?

El valor predictivo positivo del test de O'Sullivan da una idea de la tasa de acierto del test para diagnosticar la diabetes gestacional, esto es, la probabilidad p(D+|O+) de haber desarrollado diabetes gestacional habiendo obtenido un test positivo de O'Sullivan..

Podemos expresar el conocimiento enunciado de la siguiente forma:

$$\begin{aligned} p(0+) &= 0{,}2 \\ p(D+) &= 0{,}022 \\ p(D+|0-) &= 0{,}001 \end{aligned}$$

Como se puede observar se ha asignado una probabilidad mínima pero no nula a $p(D+|O-)$ ya que ningún suceso es imposible como premisa. Mediante las leyes aditiva y multiplicativa y el teorema de probabilidad total, podemos desarrollar lo siguiente:

$$\begin{aligned} p(O-) &= 1 - p(O+) = 0{,}8 \\ p(D+, O-) &= p(D+|O-)p(O-) = 0{,}001 * 0{,}8 = 8e-4 \\ p(D+) &= p(D+, O-) + p(D+, O+) \\ 0{,}022 &= 8e-4 + p(D+, O+) \\ p(D+, O+) &= 0{,}022 - 8e-4 = 0{,}0212 \\ p(D+, O+) &= p(D+|O+)p(O+) \\ p(D+|O+) &= \frac{p(D+, O+)}{p(O+)} \\ p(D+|O+) &= \frac{0{,}0212}{0{,}2} = 0{,}106 \end{aligned}$$

Por lo que el valor predictivo positivo del test de O'Sullivan se sitúa en el 0,106. Por ello, podemos esperar que únicamente el 11 % de las gestantes con test positivo de O'Sullivan tengan realmente diabetes gestacional.

3.1.3. Teorema de Bayes

A través de la ley multiplicativa, ecuación (3.2), podemos calcular la probabilidad condicional de A habiendo observado B, $p(A|B)$, como

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{3.5}$$

El teorema de Bayes permite calcular la probabilidad final $p(A|B)$ tras observar B a partir de la probabilidad inicial $p(A)$ y de la relación que existe entre B y A.

Ejemplo 3.3 (Prueba complementaria de la diabetes gestacional tras O'Sullivan) Tras recibir el test de O'Sullivan, un médico puede tener sospechas de estar ante un caso de diabetes gestacional, por lo que para acercarse a su diagnóstico, realiza una prueba complementaria cuyo resultado positivo ($X+$) se asocia muy frecuentemente a la diabetes gestacional habiendo observado un test positivo de O'Sullivan, de tal forma que:

$$p(X+|D+, O+) = 0{,}95$$
$$p(X+|D-, O+) = 0{,}10$$

¿Como se modifica la probabilidad inicial $p(D+|O+)$ con el nuevo conocimiento que aporta la prueba complementaria?

Calcularemos la *probabilidad a posteriori* $p(D+|X+, O+)$ mediante el teorema de Bayes para incorporar el conocimiento nuevo que aporta $X+$ a la probabilidad inicial obtenida tras el test de O'Sullivan,

$$p(D+|X+, O+) = \frac{p(X+|D+, O+)p(D+|O+)}{p(X+|O+)},$$

lo que requiere calcular el denominador mediante la ley de probabilidad total

$$\begin{aligned} p(X+|O+) &= p(X+|D+, O+)p(D+|O+) + p(X+|D-, O+)p(D-|O+) \\ p(X+|O+) &= 0{,}95 \cdot 0{,}106 + 0{,}10 \cdot (1 - 0{,}106) = 0{,}1901. \end{aligned}$$

Por lo tanto, la probabilidad final de tener diabetes gestacional es

$$p(D+|X+, O+) = \frac{0{,}95 \cdot 0{,}106}{0{,}1901} = 0{,}53.$$

y la probabilidad de no tener diabetes gestacional es

$$p(D-|X+, O+) = 1 - 0{,}53 = 0{,}47.$$

Con los nuevos resultados unidos al test de O'Sullivan, la hipótesis de diabetes gestacional toma peso frente al diagnóstico negativo, lo que puede ayudar al médico a decidir poner en tratamiento a la gestante.

3.1.4. Densidad de probabilidad y distribución de probabilidad

Una variable aleatoria, X, se define a partir de los valores que puede tomar y la probabilidad asociada a dichos valores. Cuando la variable es continua se puede conocer la

probabilidad del suceso $\{X \leq x\}$, esto es $P(X \leq x)$, siendo $x \in \mathbb{R}$. Esta función recibe el nombre de **función de distribución de probabilidad acumulada** y se representa por $F(x) = P(X \leq x)$. Una función de distribución es monótona no decreciente y, además, cumple que

$$F(-\infty) = 0,$$

$$F(\infty) = 1,$$

$$P(a < X \leq b) = F(b) - F(a).$$

Se puede demostrar que, cuando $\Delta x \to 0$,

$$\lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = \frac{dF(x)}{dx} = p(x),$$

donde $p(x)$ recibe el nombre de **función de densidad de probabilidad**.

Por lo tanto, si un suceso A incluye un conjunto de resultados elementales en el rango $[a, b] \in \mathbb{R}$ de la variable aleatoria X, podemos calcular la probabilidad del suceso A a través de la suma de la probabilidad de los resultados elementales en un intervalo infinitesimal dx como

$$P(A) = P(a \leq x \leq b) = \int_a^b p(x)dx. \tag{3.6}$$

Además, la *función de distribución de probabilidad* $F(x)$ puede expresarse como

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} p(x)dx. \tag{3.7}$$

La figura 3.1 muestra una función de densidad de probabilidad $p(x)$ y su función de distribución de probabilidad $F(x)$ asociada.

Las leyes aditiva y multiplicativa y los teoremas de probabilidad total y de Bayes pueden extenderse para el uso de variables continuas o combinaciones de variables discretas y continuas. Así, por ejemplo, el teorema de la probabilidad total para dos variables continuas x e y queda como

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy. \tag{3.8}$$

Es importante observar que $p(x)$ no es una probabilidad, ya que puede tomar valores mayores que 1, al contrario que $F(x)$. Sin embargo, el elemento diferencial de probabilidad $P(x < X \leq x + dx) = p(x)dx$ sí es una probabilidad.

Cuando la variable aleatoria es discreta, esto es, los valores que pueden tomar son numerables, cada posible valor está asociado a una masa de probabilidad que representamos como $P(X = x)$. Esta masa de probabilidad se representará con una **función de masa de probabilidad**, $p(X)$. En adelante, con el fin de economizar la notación matemática, la probabilidad de un suceso, $P(X = x)$ se expresará como $p(x)$.

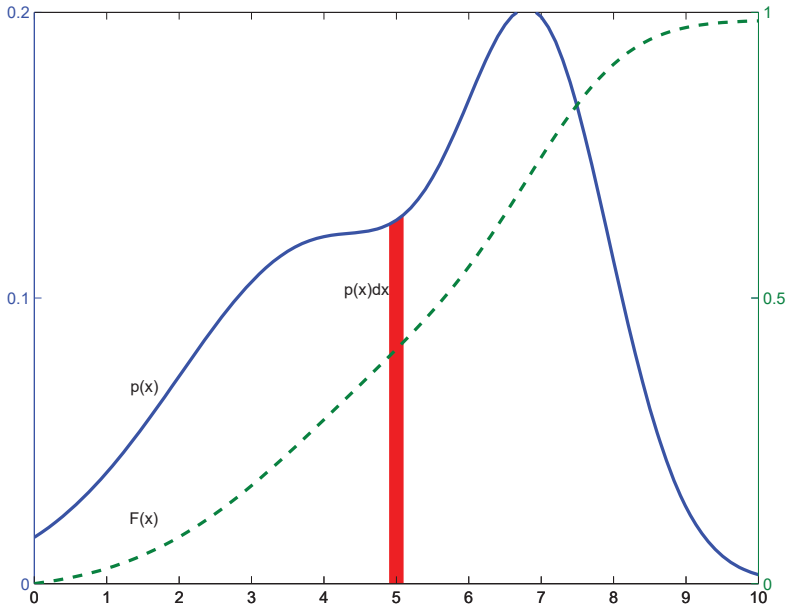


Figura 3.1: Función de densidad de probabilidad $p(x)$ en línea sólida y función de distribución de probabilidad $F(x)$ en línea discontinua. El área roja ilustra la probabilidad acumulada $p(x)dx$ en un intervalo infinitesimal dx.

3.1.5. Valor esperado de una función de variable aleatoria

Cualquier función $f(x)$ de una variable aleatoria x es a su vez una variable aleatoria. Una de las características más importantes de la distribución de $f(x)$ es el *valor esperado* que se define como

$$E_x[f(x)] = \sum_{x_i \in \Omega} p(x_i) f(x_i) \tag{3.9}$$

para variables discretas, y

$$E_x[f(x)] = \int p(x) f(x) dx \tag{3.10}$$

para variables continuas. Por lo tanto, el *valor esperado* es la media ponderada por la probabilidad de los valores que puede tomar x.

Cuando se dispone de una muestra limitada $\mathcal{S}$ de N casos de la distribución $F(x)$, el valor esperado puede aproximarse por el *valor esperado empírico* como

$$E_x[f(x)] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x) \tag{3.11}$$

Cuando dos variables, x, y, están relacionadas, puede ser de interés el cálculo del valor esperado de la función $f(x)$ sobre x condicionada a un valor de y, que se calculará como

$$E_{x|y}[f(x)] = \sum_{x_i \in \Omega} p(x_i|y_i) f(x_i) \tag{3.12}$$

para variables discretas, y

$$E_{x|y}[f(x)] = \int p(x|y) f(x) dx \tag{3.13}$$

para variables continuas.

3.1.6. Varianza y Covarianza

La *varianza* de $f(x)$ mide la variabilidad que hay en $f(x)$ en torno a su valor esperado $E[f(x)]$, por lo que se define como

$$E_x[(f(x) - E_x[f(x)])^2] = E_x[(f(x)^2] - E_x[f(x)]^2, \tag{3.14}$$

y, en particular,

$$\sigma^2 = E_x[(x - E_x[x])^2] = E_x[x^2] - E_x[x]^2. \tag{3.15}$$

La *covarianza* expresa la variación común de dos variables x e y y se puede calcular como

$$cov(x, y) = E_{x,y}[(x - E_x[x])(y - E_y[y])] = E_{x,y}[xy] - E_x[x]E_y[y]. \tag{3.16}$$

Un caso de particular interés es conocer la matriz de covarianzas $D \times D$ de un vector D-dimensional $\mathbf{x}$

$$\Sigma = cov(\mathbf{x}) = E_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T] - E_{\mathbf{x}}[\mathbf{x}]E_x[\mathbf{x}^T]. \tag{3.17}$$

3.2. Distribuciones estadísticas

En la sección 3.1.1, se ha definido el concepto de variable aleatoria y en la sección 3.1.4 se definen la *función de distribución de probabilidad*, $F(x)$, y la *función de densidad de probabilidad*, $p(x)$, que son dos formas alternativas para describir la distribución de los valores de la variable aleatoria. Generalmente, la descripción de la distribución de estos valores es más sencilla cuando se utilizan *funciones de densidad*. Según la variable aleatoria sea discreta o continua tendremos *funciones de densidad discretas* o *funciones de densidad continuas*, respectivamente. A continuación se detallarán algunas de las funciones de densidad discretas y continuas más empleadas.

3.2.1. Distribución de Bernoulli

Muchas de las variables con las que se va a trabajar tomarán únicamente 2 posibles valores complementarios. Se dice que la variable aleatoria X tiene una distribución de Bernoulli $X \sim Be(p)$ cuando se presentan estas dos alternativas, A con probabilidad p y $\neg A$, la negación de A, con probabilidad $q = 1 - p$. Estos sucesos se pueden codificar asignando, de manera arbitraria, el valor 1 cuando aparece A y 0 en caso contrario. La función de masa de probabilidad quedará:

$$p(X) = \begin{cases} 1-p & \text{cuando ocurre el suceso } \neg A \ (X=0) \\ p & \text{cuando ocurre el suceso } A \ (X=1) \end{cases} \tag{3.18}$$

En estos casos se suele decir que la variable aleatoria X sigue una distribución de Bernoulli, $X \sim Be(p)$. La función 3.18 se puede expresar como

$$p(X = x) = p^x(1-p)^{1-x}. \tag{3.19}$$

La función de distribución correspondiente es

$$F(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases} \tag{3.20}$$

La esperanza matemática y la varianza son $E[X] = p$ y $V[X] = p(1-p)$.
Ahora, supongamos un vector de variables independiente $\mathbf{X} = (X_1, \ldots, X_D)$ donde

$$X_1 \sim Be(p_1), \ldots, X_D \sim Be(p_D),$$

son independientes. Decimos entonces que $\mathbf{X}$ es Bernoulli D-dimensional de parámetros $p = (p_1, \ldots, p_D)^T$ y sun función de masa de probabilidad es:

$$p(\mathbf{x}) = \prod_d x_d = \prod_d p_d^{x_d}(1-p_d)^{(1-x_d)}.$$

3.2.2. Distribución binomial

Una variable discreta X sigue una distribución binomial $X \sim B(n; p)$ cuando mide el número de ocurrencias de un suceso en n ensayos independientes donde la probabilidad de

que dicho suceso ocurra es p. Esto es, la distribución $B(n;p)$ describe situaciones donde un mismo suceso dicotómico se repite n veces[b]. La variable X no es más que

$$X = X_1 + X_2 + \ldots + X_n, \tag{3.21}$$

donde cada variable independiente X_i toma el valor 1 cuando el suceso de interés A ocurre y 0 en caso contrario, $\neg A$. Por lo tanto, esta variable X podrá tomar cualquier valor natural entre 0 y n. Esto quiere decir que si se observa que $X = x$ entonces, en las n repeticiones que se han hecho del experimento, se ha observado x veces el suceso A. Como se puede deducir, pueden darse numerosas combinaciones en las que el suceso A apareciese exactamente x veces. Cada uno de los sucesos A se da con probabilidad p y cada suceso $\neg A$ se da con probabilidad $q = 1 - p$. Por lo tanto, la función de masa de probabilidad es:

$$p(x) = \binom{n}{x} p^x q^{n-x}. \tag{3.22}$$

La función de distribución correspondiente es:

$$F(x) = \sum_{i=0}^{x} \binom{n}{i} p^i q^{(n-i)}, \tag{3.23}$$

y la esperanza matemática y la varianza son, respectivamente, $E[X] = np$ y $V[X] = npq$.

Ejemplo 3.4

En un programa de prevención del cáncer de mama se diagnosticaron, durante el año 2007, 830 tumores malignos a partir de distintos cribados. De los tumores detectados un 86,75 % (720) fueron carcinomas invasivos y un 13,25 % (110) fueron carcinomas *in situ*. Conocidos estos datos y sin ningún tipo de información añadida, estimar la probabilidad de encontrar exactamente 4 pacientes con carcinoma in situ si se han cribado 20 pacientes con tumor maligno.

Para estimar la probabilidad habrá que proceder de la siguiente manera. En primer lugar, se detecta que el suceso de interés A es la aparición de carcinomas in situ que tienen una probabilidad $p = 0{,}1325$ de aparecer. Además, el proceso de cribado se ha repetido $n = 20$ veces, luego la variable sigue una distribución $X \sim B(20; 0{,}1325)$. Para estimar la probabilidad de encontrar exactamente 4 carcinomas in situ y, por tanto, 16 carcinomas invasivos, basta con aplicar la ecucación 3.22. Sin embargo, se hará una deducción previa para obtener el valor de la estimación.

Una posible combinación para obtener 4 carcinomas in situ, A, y el resto invasivos, $\neg A$ sería obteniendo la siguiente secuencia: $A, A, A, A, \neg A, \neg A, \ldots, \neg A$.

Este evento tendría como probabilidad el producto de las probabilidades de cada suceso independiente, es decir, $p(A, A, A, A, \neg A, \neg A, \ldots, \neg A) = \prod_{i=1}^{20} p(A_i) = p(A_1)p(A_2)\cdots p(A_{20}) = p \cdot p \cdot p \cdot p \cdot q \cdots q = p^4 \cdot q^{16} \approx 0{,}00003$.

Sin embargo, este evento no es el único que puede darse donde aparezcan 4 carcinomas in situ y 16 carcinomas invasivos. Otro ejemplo sería este: $A, A, \neg A, \neg A, A, A, \neg A, \ldots, \neg A$. De hecho, hay hasta $\binom{20}{4}$ combinaciones distintas en las que podemos encontrar los sucesos de interés. Esto significa que la probabilidad total que buscamos será:

$$p(X = 4) = \binom{20}{4} p^4 q^{16} = 0{,}1536.$$

[b] La distribución de Bernoulli es un caso particular de la distribución binomial para $n = 1$.

3.2.3. Distribución multinomial

La distribución multinomial expresa la probabilidad de aparición de que D sucesos aleatorios excluyentes $A_1, \ldots, A_D$, con probabilidades

$$p(A_i) = p_i \in [0,1], \text{ con } \sum_{d=1}^{D} p_d = 1, \tag{3.24}$$

ocurran $x_1, \ldots, x_D$ veces cada uno cuando se dan L ocurrencias. El conjunto de sucesos individuales observados pueden ser descritos mediante el vector de conteos $\mathbf{x} = (X_1 = x_1, \ldots, X_D = x_D)^T$, donde $x_1 + \ldots + x_D = L$. El número de secuencias posibles de L ocurrencias que pueden dar el vector de conteo $\mathbf{x} = (x_1, \ldots, x_D)^T$ es

$$\binom{L}{\mathbf{x}} = \frac{L!}{\prod_{d=1}^{D} x_d!},$$

y, al ser ocurrencias independientes, la probabilidad de cada secuencia de conteo $\mathbf{x}$ es $\prod_{d=1}^{D} p_d^{x_d}$, por lo que la probabilidad del vector de conteo $p(\mathbf{x})$ es

$$p(\mathbf{x}) = p(x_1, \ldots, x_D) = \frac{L!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D} p_d^{x_d}.$$

En estos casos se dice que el vector aleatorio sigue una distribución multinomial

$$(X_1, X_2, \ldots, X_D) \sim M(L; p_1, \ldots, p_D).$$

Se puede observar que la distribución multinomial es una generalización de la distribución binomial ya que cada suceso individual sigue una distribución binomial que se distribuye como $X_i \sim B(n; p_i)$, por lo tanto, la esperanza matemática y la varianza de cada suceso quedan $E[X_i] = np_i$ y $V[X_i] = np_i(1 - p_i)$, respectivamente.

3.2.4. Distribución uniforme

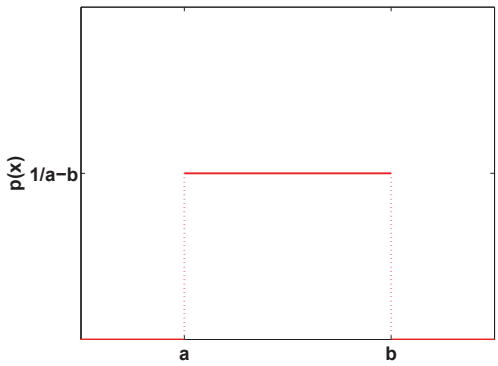


Figura 3.2: Función de densidad de probabilidad para una variable aleatoria $X \sim U(a; b)$.

Una variable aleatoria continua sigue una distribución uniforme (ver figura 3.2) en el intervalo $[a, b]$, $X \sim U(a; b)$ si su función de densidad de probabilidad es:

$$p(x) = \frac{1}{b-a} \qquad a \leq x \leq b, \tag{3.25}$$

siendo 0 en otro caso. La función de distribución de probabilidad es:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{3.26}$$

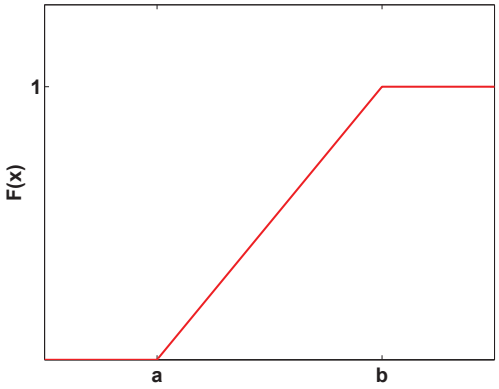


Figura 3.3: Función de densidad de probabilidad para una variable aleatoria $X \sim U(a; b)$.

La esperanza matemática y la varianza son, respectivamente,

$$E[X] = \frac{a+b}{2}$$

y

$$V[X] = \frac{(b-a)^2}{12}.$$

3.2.5. Distribución normal

Una variable aleatoria continua X sigue una distribución normal, $X \sim N(\mu; \sigma^2)$, con parámetros μ, la media, y σ^2, la varianza, cuando su función de densidad es:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{3.27}$$

La función de distribución de probabilidad es:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\text{ínf}}^{x} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{3.28}$$

El cálculo de probabilidades a partir de la función de distribución de probabilidad normal con una variable aleatoria $\mathcal{N}(0; 1)$ es muy complicado, ya que la función de densidad no tiene una primitiva conocida, por lo que la aplicación directa de la integral resulta complejo. Para obtener estas probabilidades es necesario utilizar procedimientos de aproximación muy tediosos. Hace unos años, para evitar estos cálculos, se disponía de tablas que proporcionaban valores aproximados de las probabilidades necesarias. A día de hoy, cualquier entorno matemático (R, Matlab, etc.) ofrece las probabilidades buscadas.

La figura 3.4 muestra el aspecto de las densidades de probabilidad en función de los parámetros. A la curva que se describe se la conoce como *campana de Gauss*. La curva es simétrica alrededor del punto medio, μ. Este punto medio coincide además con la mediana y la moda de la distribución. El parámetro μ de la distribución especifica la posición del pico de la campana y el parámetro σ^2 define la amplitud de la campana. Una característica interesante es que el 68,27 % de los posibles valores de la distribución normal se encuentran a una desviación estándar de la media, μ; un 95,45 % de los valores están a 2 desviaciones estándar y un 99,73 % a 3 desviaciones estándar. Esto se puede apreciar en la figura 3.6.

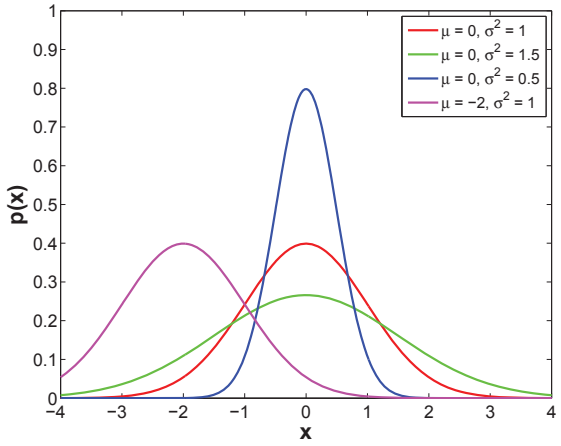


Figura 3.4: Función de densidad de probabilidad para distintas variables aleatorias normales.

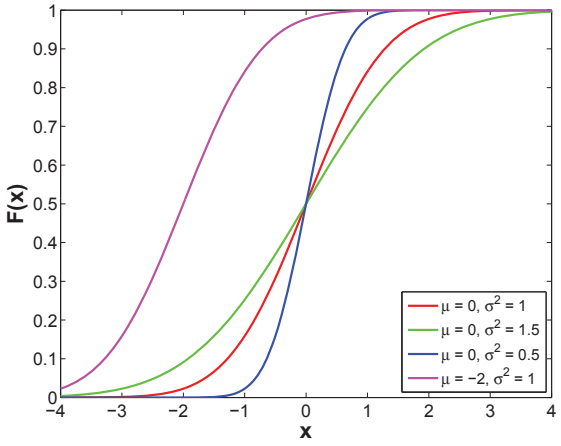


Figura 3.5: Función de distribución de probabilidad para distintas variables aleatorias normales.

Se entiende por distribución normal estándar cuando los parámetros son $\mu = 0$ y $\sigma^2 = 1$. La esperanza matemática y la varianza son, respectivamente, $E[X] = \mu$ y $V[X] = \sigma^2$.

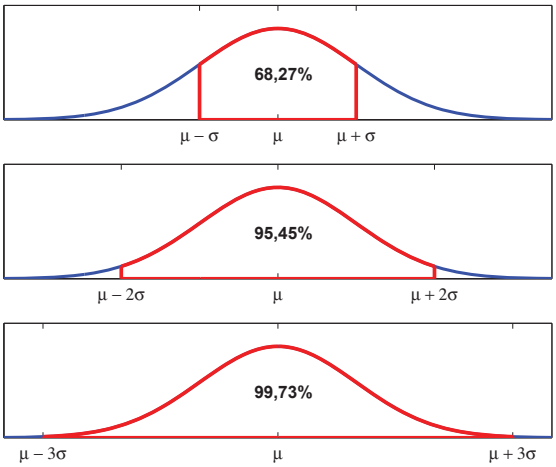


Figura 3.6: Arriba se muestra el área de la región $[\mu - \sigma, \mu + \sigma]$ que representa alrededor del 68 % de la masa de probabilidad. En el centro tenemos el área $[\mu - 2\sigma, \mu + 2\sigma]$ que encierra aproximadamente un 95 % de los valores. Por último, una masa de probabilidad de un 99,7 % está representada en la región definida por $[\mu - 3\sigma, \mu + 3\sigma]$. En algunos sitios conocen esta propiedad como la *regla 3 sigma*.

Hasta el momento se ha descrito la función de densidad normal para una única variable. Sin embargo, se puede generalizar la expresión de la función de densidad de probabilidad normal para múltiples variables. Dicho de otro modo, cuando la variable aleatoria que deseamos estudiar es un vector aleatorio D-dimensional, con $D > 1$. Entonces, un vector aleatorio sigue una distribución normal multivariante con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, $\mathbf{X} \sim N_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La función de densidad de probabilidad normal multivariante es:

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \tag{3.29}$$

Ejemplo 3.5

En un estudio sobre la diabetes de tipo II en una población homogénea, se tomó la presión sanguínea de 500 individuos. De las 500 observaciones se obtuvo un histograma por frecuencias, la media, $\hat{\mu} = 70{,}9$, y la desviación típica, $\hat{\sigma} = 11{,}9$ de la presión diastólica de la muestra. A partir de estos datos se puede calcular la probabilidad de que un individuo de dicha población tenga una presión diastólica mayor que 90 mmHg.

Para ello, puesto que deseamos hallar $P(\text{presión diastólica} > 90)$, deberemos buscar $1 - P(\text{presión diastólica} \leq 90) = 1 - F(90; \{70{,}9; 11{,}9\})$. En cualquier entorno matemático obtenemos que $F(90; \{70{,}9; 11{,}9\}) = 0{,}9458$, por lo tanto $P(\text{presión diastólica} > 90) = 1 - 0{,}9458 = 0{,}0542$.

3.2.6. Distribución gamma

Una variable aleatoria continua seguirá una distribución gamma en el intervalo $(0, \infty)$, con parámetros θ y k, si su función de densidad de probabilidad es:

$$p(x) = x^{k-1} \frac{e^{(-x/\theta)}}{\theta^k \Gamma(k)}, \tag{3.30}$$

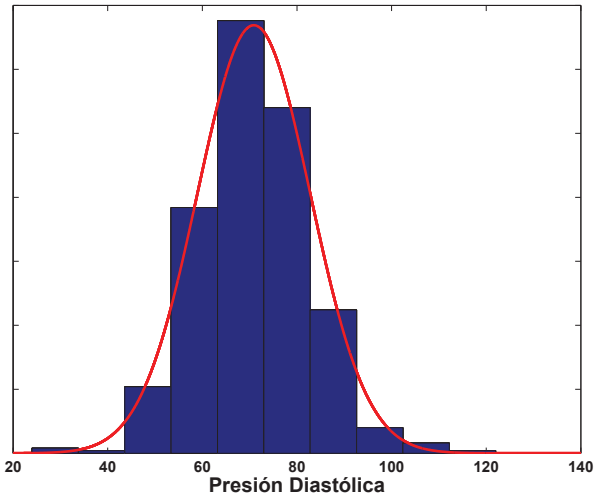


Figura 3.7: Histograma y función de densidad de probabilidad estimada a partir de las 500 observaciones para la presión diastólica medida en mmHg.

siendo $\Gamma(k)$ la función gamma expresada como:

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx, \tag{3.31}$$

donde $x > 0$ y $k > 0$. En la figura 3.8 se muestran distintas parametrizaciones de la función de densidad gamma.
La esperanza matemática es $E[X] = k\theta$ y la varianza $k\theta^2$. Conviene destacar que la distribución χ^2 es un caso particular de la distribución gamma cuando $\theta = 2$. En concreto, si $X \sim \Gamma(k, \theta = 2)$, entonces $X \sim \chi^2(2k)$. Además, en la inferencia bayesiana la distribución gamma es, precisamente, la distribución conjugada de la inversa de la varianza[c].

3.2.7. Distribución beta

Una variable aleatoria continua sigue una distribución beta en el intervalo $[0, 1]$, con parámetros α y β, si su función de densidad de probabilidad es:

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{3.32}$$

siendo $B(\alpha, \beta)$ una función conocida como función *beta* con expresión:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{3.33}$$

siendo $\Gamma(\cdot)$ la función *gamma* y $\alpha > 0$ y $\beta > 0$. De este modo, la función beta aparece como una constante de normalización para garantizar que la probabilidad total sume 1.

[c]En la teoría de la probabilidad bayesiana, cuando la densidad de probabilidad *a posteriori*, $p(\theta|x)$, es de la misma familia de funciones que la distribución de probabilidad *a priori*, $p(\theta)$, entonces se dice que son distribuciones conjugadas y a la *a priori* se le llama conjugada de la verosimilitud, $p(x|\theta)$.

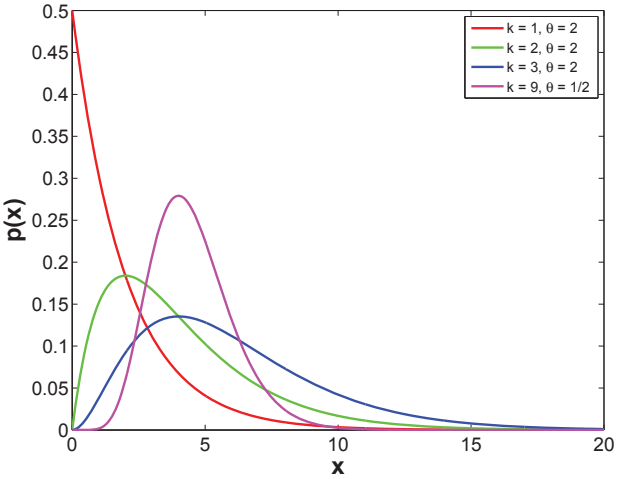


Figura 3.8: Función de densidad de probabilidad gamma con distintos parámetros k y θ.

La esperanza matemática y la varianza son, respectivamente, $E[X] = \alpha/(\alpha+\beta)$ y $V[X] = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. En la figura 3.9 se observa que la densidad es finita cuando $\alpha, \beta \geq 1$ y si se toma $\alpha = \beta = 1$ se obtiene la densidad de probabilidad uniforme estándar. Una propiedad importante, que se empleará en la teoría de la probabilidad bayesiana, es que la distribución beta es la distribución conjugada de la probabilidad binomial.

3.2.8. Distribución de Dirichlet

La distribución de Dirichlet es una generalización multivariante de la distribución beta. Es también la distribución conjugada de la distribución multinomial. Se dice que un vector aleatorio $\mathbf{X}$ sigue una distribución de Dirichlet con parámetros $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ y $K \geq 2$, $\mathbf{X} \sim Dir(\boldsymbol{\alpha})$ cuando la función de densidad de probabilidad es:

$$p(\mathbf{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_k x_k^{\alpha_k - 1}, \tag{3.34}$$

donde $x_k > 0$ y $\sum_k x_k = 1$. El factor de normalización se puede expresar en términos de la función gamma:

$$B(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}. \tag{3.35}$$

La esperanza matemática y la varianza de cada suceso individual es

$$E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$$

y

$$V[X_i] = \frac{E[X_i](1 - E[X_i])}{1 + \sum_k \alpha_k}.$$

Al igual que la distribución beta, la distribución de Dirichlet se empleará en la teoría de la probabilidad bayesiana, ya que la distribución de Dirichlet es la distribución conjugada de la probabilidad multinomial.

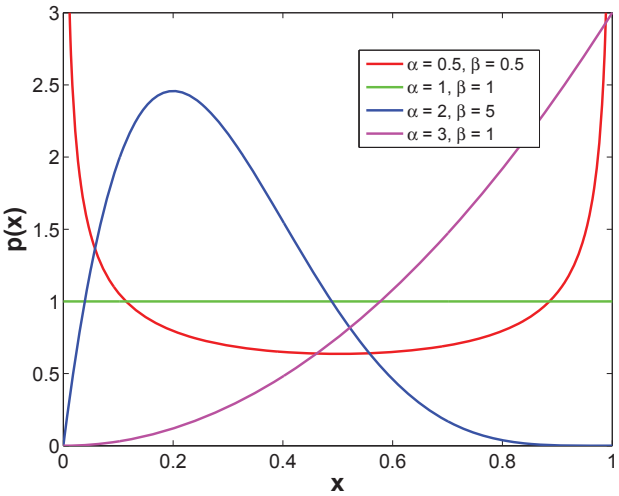


Figura 3.9: Función de densidad de probabilidad beta con distintos parámetros α y β.

3.3. Teoría de la decisión Bayesiana

La teoría de la decisión, junto con la teoría de la probabilidad, permite tomar decisiones óptimas en situaciones con incertidumbre, esto es, en ausencia de información completa y correcta. En este tipo de situaciones se puede incluir el proceso de decisión médica.

Un *problema de decisión* es aquel en "el que se debe elegir de forma razonada entre un determinado conjunto de alternativas, en presencia de incertidumbre sobre algunos de los factores que condicionan la consecuencias de la elección" [83].

El diseño de un problema de decisión implica las siguientes tareas:

1. Determinar el espacio de decisiones.
2. Definir los sucesos inciertos.
3. Definir los sucesos observados que aporten conocimiento a la resolución del problema.
4. Cuantificar la verosimilitud de la ocurrencia de los sucesos.
5. Determinar las consecuencias de tomar cada decisión bajo las circunstancias de producirse los sucesos inciertos.
6. Definir la función utilidad (o función pérdida) de las consecuencias.
7. Definir la utilidad (o pérdida) esperada de las decisiones.
8. Resolver el problema mediante la decisión que maximiza la utilidad esperada (o minimiza la pérdida esperada).

Estas tareas pueden estar relacionadas entre sí y requerir revisiones para llegar a definir el problema correctamente.

Ejemplo 3.6 (Diagnóstico de tumores cerebrales en la infancia)
El grupo interdisciplinar de diagnóstico precoz de tumores cerebrales en la infancia considera que un paciente menor de 14 años tiene un tumor cerebral y consideran que puede ser de tipo

Ependimoma (EP), Astrocitoma Pilocítico (AP) o Tumor Embrional (EM). Para este problema concreto, el grupo interdisciplinar considera que es extremadamente importante decidir uno y solo uno de los diagnósticos, y es importante no errar en la decisión.

Deciden que la probabilidad inicial de cada diagnóstico se realice mediante la prevalencia de estos tipos de tumores en niños entre 0 y 14 años:

Para obtener más información concreta del caso, recurren a un índice radiológico, obteniendo un valor de 8,5 en unidades arbitrarias ($u.a.$). La respuesta de dicho índice se ha caracterizado mediante estudios previos.

¿Qué decisión debe tomar el grupo interdisciplinar? Iremos solucionando el problema por etapas a lo largo de la sección.

Durante el diseño del problema de decisión, debemos elegir el conjunto de posibles decisiones y que denominaremos espacio de decisiones $\mathcal{D}$, del cual se elegirá un elemento d para dar solución al problema. Este espacio debe ser exhaustivo y los elementos que lo componen deberán ser excluyentes entre sí. Si bien el espacio de decisiones suele ser finito, este puede también contener infinitas alternativas.

Ejemplo 3.7 (Diagnóstico de tumores cerebrales en la infancia (cont. ej. 3.6))
El espacio de decisiones en el problema de diagnóstico que se plantea el grupo interdisciplinar es $\mathcal{D} = \{EP, AP, EM\}$ que representa la elección como diagnóstico de los tipos Ependimoma, Astrocitoma Pilocítico y Tumor Embrional, respectivamente. El espacio $\mathcal{D}$ es exhaustivo y sus elementos son excluyentes, ya que representan cada una de las posibilidades contempladas por el grupo de expertos.

Debemos ahora identificar el conjunto de sucesos inciertos para cada una de las posibles decisiones. Ante la situación de incertidumbre que plantea un problema de decisión, como el diagnóstico, pronóstico, tratamiento y gestión médica, los sucesos inciertos pueden ser muchos y muy variados, por lo que deben identificarse con sumo cuidado. Los sucesos inciertos en este problema pueden tener diferente naturaleza. Por ejemplo, sucesos como "remisión del tumor", "proliferación del tumor" o "muerte del paciente antes de 1 año" tendrían un carácter pronóstico. Sucesos como "supervivencia a la operación" o "secuelas", podrían ser sucesos inciertos asociados a tratamientos. Por último, sucesos como "anemia ferropénica" o "talasemia" podrían ser sucesos inciertos asociados a diagnósticos.

Ejemplo 3.8 (Diagnóstico de tumores cerebrales en la infancia, cont. ej. 3.6)
Los sucesos inciertos de interés para la resolución de este problema son los tres tipos de tumor: $\mathcal{Y} = \{EP, AP, EM\}$. En este problema los sucesos inciertos son los mismos con independencia de la decisión que se tome.

Si se tiene disponible mediante experimentación u observación algún conjunto de sucesos que puedan aportar información a la resolución del problema, este puede ser utilizado para reducir la incertidumbre de la decisión. Por ejemplo, disponer de una prueba diagnóstica puede reducir la incertidumbre sobre el diagnóstico gracias a la información específica que aporta del caso.

Ejemplo 3.9 (Diagnóstico de tumores cerebrales en la infancia)
Como observaciones del paciente, el equipo disciplinar dispone de un valor $x = 8{,}5$ del índice radiológico.

La cuantificación de la verosimilitud de los sucesos inciertos y de los sucesos observados condicionados a los sucesos inciertos se puede realizar mediante la teoría de la probabilidad (ver sección 3.1). Así pues, podemos expresar la verosimilitud de los sucesos inciertos mediante las probabilidades $p(y)$, y las probabilidades de las observaciones condicionadas a los sucesos inciertos a través de $p(x|y)$.

Ejemplo 3.10 (Diagnóstico de tumores cerebrales en la infancia)
Mediante la consulta de los datos históricos de tumores del sistema nervioso central estratificado por edad, el grupo interdisciplinar puede asignar las probabilidades a priori[d]:

$$\begin{aligned} p(EP) &= 0{,}15 \\ p(AP) &= 0{,}48 \\ p(EM) &= 0{,}37 \end{aligned}$$

Supongamos que para el índice radiológico, los estudios previos obtuvieron las siguientes densidades de probabilidad condicionales a cada diagnóstico:

$$\begin{aligned} p(x|EP) &= N(x|5,2) \\ p(x|AP) &= N(x|10,1) \\ p(x|EM) &= N(x|13,3) \end{aligned}$$

Determinadas las posibles decisiones y definidos los sucesos inciertos, se pueden determinar las consecuencias $c(d, y)$ de tomar cada decisión d ante la eventual realización de cada suceso incierto y.

Ejemplo 3.11 (Diagnóstico de tumores cerebrales en la infancia, cont. ej. 3.6)
Podemos definir las consecuencias del problema por enumeración de la siguiente forma:

$$\{c(EP, EP) = \text{“acierto”}, c(EP, AP) = \text{“error”}, \ldots, c(AP, EP) = \text{“error”}, \ldots\},$$

donde $c(d, y)$ representa la consecuencia de decidir el diagnóstico d cuando el paciente tiene un tumor de tipo y. Así pues, podríamos definir de forma concisa el conjunto de consecuencias de la siguiente forma:

$$\begin{aligned} c(d, y) &= \text{“acierto”}; \quad d = y \\ c(d, y) &= \text{“error”}; \quad d \neq y. \end{aligned}$$

Para poder operar con las consecuencias de las decisiones, a partir de los sucesos inciertos, será necesario establecer una función numérica de *utilidad*, o de *pérdida*, según el contexto del problema a resolver.

Una función de pérdida $L(d, y)$ asigna el coste que se produce como consecuencia de decidir d ante el suceso y, esto es, $c(d, y)$. Una de las funciones típicas de pérdida es la función 0-1.

$$L(d, y) = \begin{cases} 0 & \text{si } y = d \\ 1 & \text{si } y \neq d \end{cases} \tag{3.36}$$

Esta función suele utilizarse en problemas de *clasificación* donde y toma un valor de una lista de categorías y la decisión consiste en elegir un elemento de esa lista penalizando con 1 cuando la elección es errónea.

[d] Un ejemplo donde consultar esta información es el registro CBTRUS `http://www.cbtrus.org/`.

Ejemplo 3.12 (Diagnóstico de tumores cerebrales en la infancia, cont. ej. 3.6)
La función de pérdida 0-1 responde al enunciado del problema, ya que penalizará las decisiones cuyo diagnóstico no coincida con el diagnóstico real del paciente.

La utilidad (o pérdida) esperada al tomar la decisión d habiendo observado x es el valor esperado de la función utilidad (o función pérdida) sobre y condicionado al valor observado en x. Para el caso en el que y es una variable categórica con C posibles valores, podemos calcularlo mediante la ec. 3.12:

$$E_{y|x}[L(d)] = \sum_{i=1}^{C} p(y_i|x)L(d, y_i) \tag{3.37}$$

Para el caso en el que y es una variable continua, podemos calcularlo mediante la ec. 3.13:

$$E_{y|x}[L(d)] = \int_y p(y|x)L(d, y_i)dy \tag{3.38}$$

Ejemplo 3.13 (Diagnóstico de tumores cerebrales en la infancia, cont. ej. 3.6)
El valor esperado de pérdida vendrá dado por la expresión 3.37, por lo que debemos calcular por el teorema de Bayes, ecuación 3.5:

$$\begin{aligned} p(EP|x) \propto p(EP)p(x|EP) = 0{,}15 \cdot 0{,}0432 &= 0{,}0065, \\ p(AP|x) \propto p(AP)p(x|AP) = 0{,}48 \cdot 0{,}1295 &= 0{,}0622, \\ p(EM|x) \propto p(EM)p(x|EM) = 0{,}37 \cdot 0{,}0432 &= 0{,}0160, \\ p(x) = p(EP|x) + p(AP|x) + p(EM|x) &= 0{,}0846. \end{aligned}$$

Por lo que la estimación de la pérdida esperada para cada posible decisión será:

$$E_{y|x}[L(d_j)] = \sum_{i=1}^{C} p(y_i|x)L(d_j, y_i) = 1 - p(y_j|x),$$

por lo que obtenemos:

$$\begin{aligned} E[L(EP)] &= 1 - 0{,}0065/0{,}0846 = 1 - 0{,}08 = 0{,}92, \\ E[L(AP)] &= 1 - 0{,}0622/0{,}0846 = 1 - 0{,}73 = 0{,}27, \\ E[L(EM)] &= 1 - 0{,}0160/0{,}0846 = 1 - 0{,}19 = 0{,}81. \end{aligned}$$

La decisión óptima de un problema con incertidumbre es aquella que maximiza (o minimiza) la utilidad (o pérdida) esperada.
La regla de decisión basada en la utilidad esperada es:

$$d^* \leftarrow \arg\max_{\mathcal{D}} E_{y|x}[L(d)] \tag{3.39}$$

La regla de decisión basada en la pérdida esperada es:

$$d^* \leftarrow \arg\min_{\mathcal{D}} E_{y|x}[L(d)] \tag{3.40}$$

Ejemplo 3.14 (Diagnóstico de tumores cerebrales en la infancia, cont. ej. 3.6)
Aplicando la regla de decisión 3.40, el grupo interdisciplinar minimiza la pérdida esperada eligiendo como diagnóstico "Astrocitoma Pilocítico".

3.3.1. Teoría de la decisión Bayesiana para una secuencia de decisiones condicionales

Sea $\mathbf{d} = \{d_1, \ldots, d_i, \ldots, d_N\}$ el conjunto no ordenado de posibles decisiones que involucran un proceso decisivo complejo y que pueden aparecer en diferentes momentos t de una secuencia temporal 1 hasta T. En cada decisión d_i se puede elegir entre serie de alternativas, expresando como d_{ij} la alternativa j. Llamaremos decisión d_t a una decisión del conjunto $\mathbf{d}$ que se toma en el momento t. $\mathbf{d}_{t-1}$ es el vector ordenado de las alternativas seleccionadas desde el momento 1 hasta $t-1$. θ_{ijk} es cada uno de los sucesos inciertos que pueden darse tras tomar la alternativa d_{ij}. θ_t es el suceso acontecido en el momento t y $\mathbf{\Theta}_{t-1}$ es la secuencia ordenada de sucesos acontecidos desde el momento 1 hasta $t-1$.

Según la teoría de la decisión Bayesiana, la solución óptima a la decisión d_t condicionada a los sucesos $\mathbf{\Theta}_{t-1}$ y a las alternativas seleccionadas $\mathbf{d}_{t-1}$ en la decisiones anteriormente, será la alternativa que maximice la utilidad esperada en el momento t condicionada a los sucesos acontecidos y las decisiones tomadas anteriormente:

$$d_t^* | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1} = \arg\max_j u^*(d_{tj} | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1}), \tag{3.41}$$

con utilidad

$$u^*(d_t | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1}) = \max_j u^*(d_{tj} | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1}), \tag{3.42}$$

donde la utilidad de las alternativas de decisiones no finales se estiman a partir de las utilidades estimadas de sus sucesos inciertos utilizando las decisiones sucesivas

$$u^*(d_{tj} | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1}) = \sum_k p(\theta_{tjk} | \mathbf{\Theta}_{t-1}) u^*(d_{t+1} | \mathbf{\Theta}_{t-1}, \theta_{tjk}, \mathbf{d}_{t-1}, d_{tj}), \tag{3.43}$$

y para las alternativas de decisiones finales se utiliza la función de utilidad condicionada a los sucesos inciertos

$$u^*(d_{tj} | \mathbf{\Theta}_{t-1}, \mathbf{d}_{t-1}) = \sum_k p(\theta_{tjk} | \mathbf{\Theta}_{t-1}) u(d_{tj}, \mathbf{d}_{t-1} | \theta_{tjk}, \mathbf{\Theta}_{t-1}), \tag{3.44}$$

siendo la función $u(d_{tj}, \mathbf{d}_{t-1} | \theta_{tjk}, \mathbf{\Theta}_{t-1})$ la utilidad de las alternativas seleccionadas hasta el momento t. Si la utilidad de las alternativas seleccionadas es aditiva podemos escribir:

$$u(d_{tj}, \mathbf{d}_{t-1} | \theta_{tjk}, \mathbf{\Theta}_{t-1}) = u(d_{tj}, | \theta_{t,k}, \mathbf{\Theta}_{t-1}) + u(\mathbf{d}_{t-1} | \mathbf{\Theta}_{t-1}). \tag{3.45}$$

Cabe destacar que todas las alternativas de una decisión comparten el conjunto de sucesos inciertos, ya que todas las ramas deben comtemplar todas los posibles futuros tras la toma de la decisión, es decir todos los posibles sucesos inciertos.

3.3.2. Incorporación de observaciones realizadas durante las decisiones a la metodología de resolución

Las observaciones $\mathbf{o}_t$ realizadas al seleccionar una alternativa en una decisión en el momento t se incorporan como sucesos a los vectores $\mathbf{\Theta}'_t = [\mathbf{\Theta}_t, \mathbf{o}_t]$ del modelo, por lo que los sucesos inciertos posteriores quedan condicionados a dichas observaciones, e.g. $p(\theta_t | \mathbf{\Theta}'_{t-1}) = p(\theta_t | \mathbf{\Theta}_{t-1}, \mathbf{o}_{t-1})$, y las utilidades también pueden quedar condicionadas a dichas observaciones $u(d_{t,j}, | \theta_{tjk}, \mathbf{\Theta}_{t-1}, \mathbf{o}_{t-1})$.

Es importante recordar que para un modelo predictivo que vaya a funcionar en el momento t únicamente se podrán obtener observaciones asociadas a los momentos 1 hasta $t-1$.

Ejemplo 3.15 (Decisión de producción farmacológica)
Nos basaremos en un ejemplo del libro [83] como primer ejemplo de la metodología. En dicho ejemplo, una empresa farmacéutica se plantea producir un nuevo fármaco al mercado. Los beneficios de la producción del nuevo fármaco dependerán de su éxito alto, medio o bajo entre los médicos. Para estudiar la viabilidad de la nueva producción se plantea realizar una encuesta de campo con un coste asociado, siendo el resultado de la encuesta aconsejar o no aconsejar la producción del nuevo fármaco.

Claramente es un proceso de decisión que implica dos momentos de decisiones encadenadas. Identificamos los siguientes elementos del problema:

1. Decisión Encuesta (do): realizar la encuesta
2. Alternativas de do: sí (y), no (n).
3. Sucesos inciertos tras do=sí: aconsejar producir (x=1), aconsejar no producir (x=0).
4. : Decisión Producción (dp): comenzar producción
5. Alternativas de dp: sí (y), no (n).
6. Sucesos inciertos tras dp=sí: exito alto (e=a), éxito medio (e=m), éxito bajo (e=b)

La figura 3.10 identifica los posibles caminos de decisiones y sucesos inciertos, donde los cuadrados representan decisiones, las líneas tras las decisiones representan las alternativas, y las líneas tras los círculos representan los posibles sucesos inciertos tras cada alternativa. Al final de cada rama se identifica su utilidad, es decir de las decisiones condicionadas a los sucesos inciertos.

La resolución del problema viene dada por el recorrido en profundidad del árbol, obteniendo resultados parciales de la figura 3.11 según las ecuaciones de la sección 3.3.1.

Ejemplo 3.16 (Decisión de Biopsia de Ganglio Centinela)
Los nodos linfáticos constituyen el principal drenaje de las glándulas mamarias. Esto justifica las actuales guías clínicas que recomiendan una linfadenectomia completa (axillary lymph node dissection (ALND)) en aquellas pacientes en cuyos ganglios centinela (sentinel lymph node (SLN)) hay presencia de metástasis.

Deseamos estudiar lo adecuado que es realizar el protocolo quirúrgico basado en la biopsia intraquirúrgica del ganglio centinela (SNB) previa a una (ALND). Para ello especificamos el proceso de decisión quirúrgico que implicaría las dos decisiones secuenciales, según los posibles caminos indicados en la figura 3.12.

Identificamos los siguientes elementos en el árbol de decisión:

- Decisión: realizar biopsia de ganglio centinela (SNB)
- Alternativas de SNB: realizar SNB (y), no realizar SNB (n)
- Sucesos inciertos tras SNB=y y SNB=n: No metástasis más allá del SNB (met=M0), Metástasis más allá del SNB (met=M1)
- Decisión: realizar linfadenectomía (ALND)
- Alternativas de ALND: realizar ALND (y), no realizar ALND (n)
- Sucesos inciertos tras ALND=y y ALND=n: no linfedema (lim=n), linfedema (lim=y)

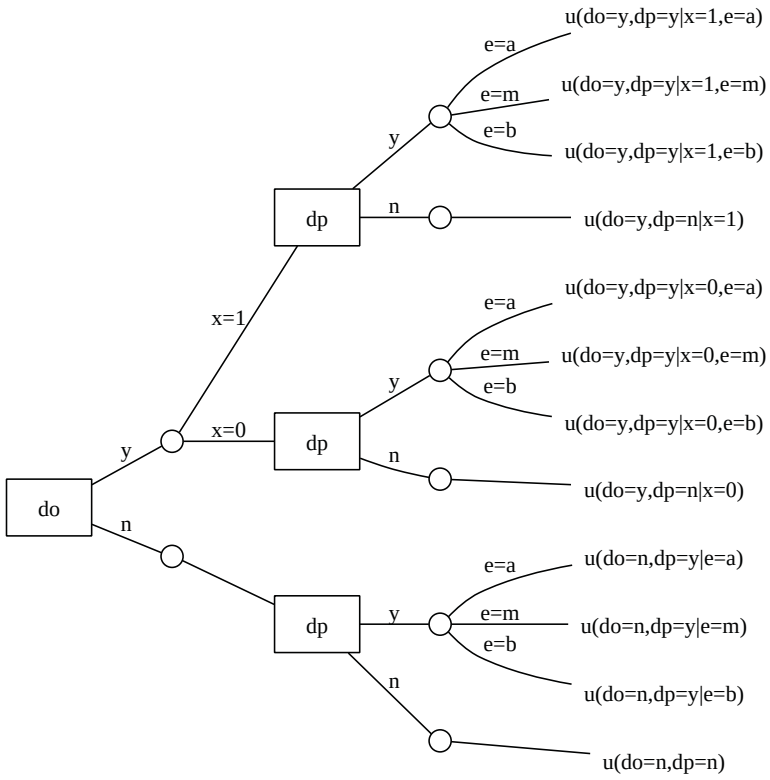


Figura 3.10: Árbol de decisiones y sucesos inciertos del proceso de decisión de realizar una encuesta para la producción de un nuevo fármaco.

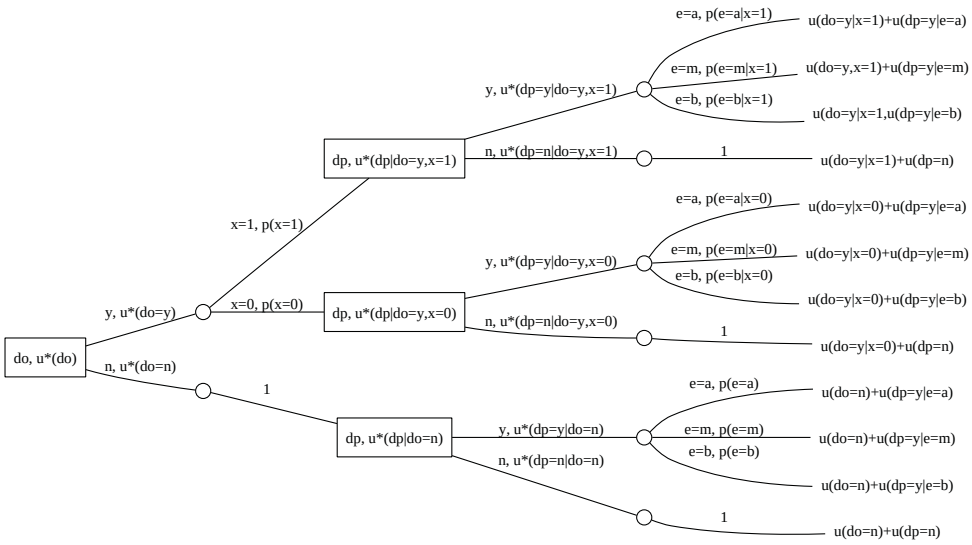


Figura 3.11: Resultados parciales obtenidos durante el recorrido en profundidad del árbol para resolver la decisión de realizar una encuesta para la producción de un nuevo fármaco.

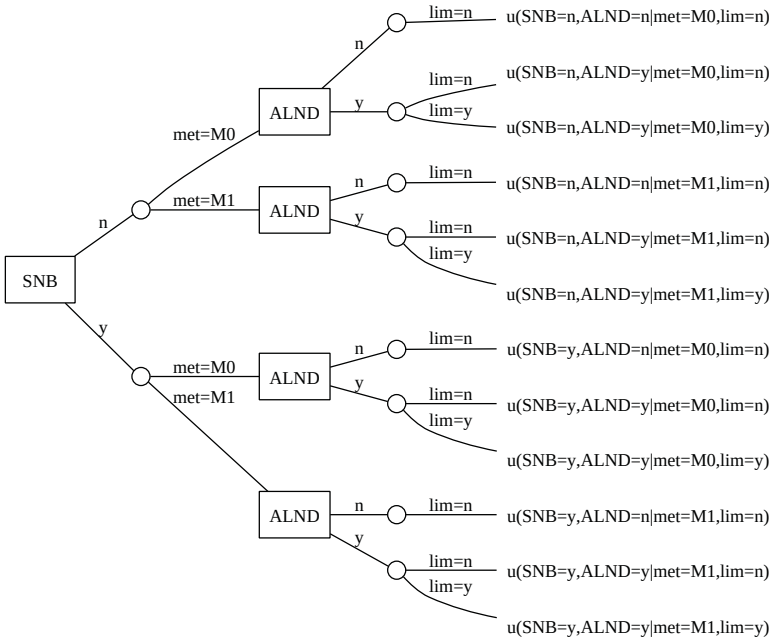


Figura 3.12: Árbol de decisiones y sucesos inciertos del proceso de decisión quirúrgico basado en la biopsia de ganglio centinela (SNB) previa a realizar una linfadenectomía (ALND).

Tabla 3.1: Costes económicos de los diferentes procedimientos del proceso de decisión SNB

Acrónimo	Procedimiento	Coste (kEUR)
costeSNB	Cirugía SNB	1,420
costeALND	Cirugía ALND	2,550
costePost	Postoperatorio	0,3
costeLim	Tratamiento Linfedema	4,0

Tabla 3.2: Utilidad de realizar una ALND según los sucesos inciertos de cada rama

ALND	met	lim	utilidad (QALY/kEUR)
y	M0	n	1/(costeALND+costePost)
y	M0	y	1/(costeALND+costePost+costeLim)
y	M1	n	10/(costeALND+costePost)
y	M1	y	(10*0.7)/(costeALND+costePost+costeLim)

Para especificar la utilidad de las alternativas utilizaremos el indicador formado por el ratio entre los años de vida aportados por la decisión ajustados por calidad de vida (QALY) entre el coste económico que implica la decisión. Como se puede observar, la métrica de utilidad $QALY/kEUR$ tiene en cuenta la diferencia de esperanza de vida al realizar la operación y la esperanza de vida al no realizarla y el coste que implica.

En primer lugar, la tabla 3.1 especifica los costes económicos de los diferentes procedimientos.

La utilidad directa para el paciente de decidir realizar la SNB es independiente del estado de la metástasis, por lo que $u(SNB = y|M0) = u(SNB = y|M1) = 1/costeSNB$. La utilidad de no realizar SNB será de 0 en cualquier caso, por lo que $u(SNB = n|M0) = u(SNB = n|M1) = 0$.

La tabla 3.2 especifica la utilidad de realizar una ALND dependiendo de los sucesos inciertos de cada rama.

En caso de no realizar ALND, no se conseguirá aumentar la esperanza de vida de la paciente, y tampoco habrá un coste por operación, por lo que definimos $u(ALND = n|met = M0) = u(ALND = n)|met = M1) = 0$.

Si no se realiza SNB, la probabilidad de tener metástasis o no tenerla será incondicional, y por lo tanto estimada por conocimiento de estudios poblacionales, será la circunstancia en la que menos conocimiento dispongamos. Asumiremos entonces que $p(met = M0) = p(met = M1) =$ 0,5. Al tomar la decisión de realizar SNB, dispondremos de una observación que nos aporta conocimiento de nuestros sucesos inciertos, ya que sabremos si hemos encontrado metástasis en los ganglios centinela ($X_{SNB} = M1$) o no ($X_{SNB} = M0$). Asumiremos la tabla 3.3 de probabilidades condicionales.

Si utilizaramos variables observaciones del estado del paciente y de tumor primario podríamos mejorar el conocimiento de nuestros sucesos inciertos. Para ello, sustituiríamos las funciones actuales por funciones que tuvieran en cuenta las observaciones previas disponibles. Por ejemplo, podríamos tener el grado G del tumor, el tipo tumoral T, y el perfil $HER2$, por lo que podríamos utilizar el modelo predictivo $p(met|X_{SNB}, G, T, HER2$ y $p(met|G, T, HER2)$ en lugar de $p(met|X_{SNB})$ y $p(met)$, respectivamente.

Tabla 3.3: Probabilidades condicionales de metástasis tras realizar SNB

met/X_{SNB}	M0	M1
M0	0.7	0.3
M1	0.3	0.7

Por último, la probabilidad de sufrir limfedema tras una ALND se estima poblacionamente en un 30 %, por lo que $p(lim = y) = 0{,}3; p(lim = n) = 0{,}7$.

La resolución del problema viene dada por el recorrido en profundidad del árbol, obteniendo resultados parciales de la figura 3.13 según las ecuaciones de la sección 3.3.1.

En caso de disponer de observaciones previas, podrían mejorarse los modelos predictivos de los sucesos inciertos condicionándolos también a dichas observaciones. En nuestra simulación, sin observaciones previas, asumiendo las utilidades y modelos de probabilidad especificado (sin uso de conocimiento previo del tumor o perfil genético), la decisión óptima es realizar la biopsia del ganglio centinela con una utilidad esperada de $2{,}23 QALY/kEUR$.

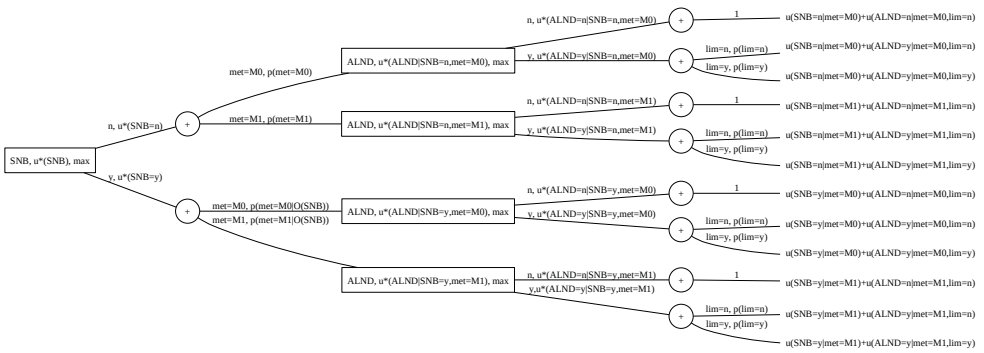


Figura 3.13: Resultados parciales obtenidos durante el recorrido en profundidad del árbol para resolver la decisión de realizar el procedimiento quirúrgico de biopsia de ganglio centinela previa a la linfadenectomia.

3.3.3. Extensión de la metodología para la resolución de decisiones de guías clínicas

Para aplicar a guías clínicas la metodología de resolución de secuencias de decisiones mediante la teoría de decisión Bayesiana debe realizarse la siguiente especificación:

1. Identificar las decisiones que componen la guía clínica
2. Identificar las alternativas posibles en las decisiones
3. Identificar observaciones del paciente previas a la guía que puedan ser relevantes para los estados del paciente
4. Identificar las observaciones obtenidas tras tomar cada alternativa posible para cada decisión

5. Identificar los sucesos inciertos que pueden acontecer en cada estado del paciente
6. Identificar los flujos de la guía clínica que definirán las posibles secuencias de decisión
7. Identificar las consecuencias que supone cada alternativa ante los sucesos inciertos
8. Definir la métrica adecuada para la guía clínica que deberá plasmar el beneficio conseguido por el paciente y el sistema sanitario por la aplicación de las decisiones
9. Definir las tablas de utilidad de las alternativas de las decisiones dependiendo de los sucesos inciertos que afectan a las consecuencias identificadas en el punto 4
10. Estimar los modelos predictivos de los sucesos inciertos condicionados a sucesos anteriores y observaciones previas
11. Aplicar la metodología expuesta en la sección 3.3.1 y 3.3.2.

Es necesario un trabajo bibliográfico profundo para realizar la especificación propuesta, ya que las guías clínicas actuales carecen de: 1) métricas, 2) utilidades, y 3) sucesos inciertos, siendo la última carencia la más relevante y sorprendente. Además, las decisiones y sus alternativas no suelen ser exhaustivas o son implícitas.

3.4. Notas bibliográficas

El libro de Bernardo [83] realiza una buena introducción a la teoría de la decisión mediante ejemplos de la medicina y la biología. Como alternativa, los libros de Hoff [84] y de Bailer-Jones [85] ofrecen una introducción práctica a la teoría de la probabilidad con especial hincapié en la inferencia bayesiana empleando ejemplos con código fuente.

Capítulo 4

Selección y extracción de características

4.1. Selección de características

Anteriormente se analizó el problema de la dimensionalidad y cómo un excesivo número de variables, lejos de aportar mayor información, pueden complicar sustancialmente el análisis de los datos. Se vio cómo un mayor número de variables requiere de más observaciones o, de lo contrario, los datos quedan dispersos por el espacio muestral. En este capítulo veremos cómo se puede reducir el número de variables para obtener un mejor análisis de los datos y, consecuentemente, mejores modelos predictivos. Existen dos paradigmas a la reducción de la dimensionalidad. La primera de ellas es la *selección de características* que vamos a ver a continuación. La segunda es la *extracción de características* que se verá en la siguiente sección.

La selección de características o de variables [86] consiste en seleccionar un subconjunto de todas las variables que, idealmente, será óptimo para comprender los datos y obtener los mejores modelos. Los beneficios potenciales son muchos: descarta variables irrelevantes o redundantes, facilita la comprensión de los datos y de los modelos obtenidos, permite la visualización de los datos, reduce los requerimientos de almacenamiento y los costes computacionales de los modelos estimados y desafía el problema de la dimensionalidad proporcionando modelos predictivos con mejor rendimiento.

Existen fundamentalmente dos aproximaciones distintas para la selección de características: la aproximación **indirecta** o *filters* y la aproximación **directa** o *wrappers*. La primera consiste en seleccionar aquellas variables que obtienen un mejor ajuste con respecto a una función objetivo distinta del acierto o error del modelo predictivo. La segunda consiste en seleccionar aquellas variables que obtienen *directamente* los mejores resultados para el modelo predictivo.

4.1.1. Filters

La aproximación indirecta se basa en la optimización de una medida de ajuste -una función objetivo- que es distinta de la medida de ajuste objeto de nuestra estimación de modelos. Por tanto, es un ajuste a una medida *indirecta* que permite establecer una clasificación entre las variables disponibles. De este modo, las variables se pueden ordenar de mayor a menor por orden de relevancia. Esto permite la selección de aquellas variables que se ajusten mejor a la función objetivo propuesta.

Formalmente, considérese un conjunto de N observaciones $\{\mathbf{x}_i, y_i\}$ con $i = 1, \ldots, N$, donde $\mathbf{x}_i$ se compone de D variables distintas e y_i es la variable a predecir. Se selecciona una función

objetivo $F(j)$, donde j es la variable j-ésima. Se asume que un valor alto de la función $F(\cdot)$ indica que la variable es relevante y así se puede establecer un orden entre las distintas variables. Obviamente, la selección de las variables más relevantes es independiente del modelo predictivo final y puede ser considerado un método de preproceso de los datos. Los métodos indirectos, aunque son subóptimos y se orientan hacia un objetivo distinto al objetivo que se busca con la estimación del modelo, son computacionalmente eficientes ya que solo es necesario realizar N cálculos y ordenarlos.

Cabe mencionar que una variable se considera **estadísticamente relevante** cuando su eliminación del conjunto de variables reduce el rendimiento predictivo del modelo final. Esto puede ocurrir por dos razones: o bien la variable está altamente correlacionada con alguna clase, o bien la variable es parte de un subconjunto de variables que está altamente correlacionada con alguna clase.

Existen multitud de funciones objetivo para llevar a cabo la selección de las variables más relevantes mediante métodos indirectos o *filters*. A continuación expondremos algunos de ellos.

- **Test de hipótesis clásicos**: los test de hipótesis clásicos como el test de χ^2 para comparación de proporciones o el test de comparación de medias, tanto paramétrico como el t-test como el no paramétrico como el test de Kruskal-Wallis, son muy utilizados para comparar si la distribución de las variables son iguales para cada clase. Si la probabilidad, el llamado p-valor, es menor que un nivel de significación α prefijado se rechaza la igualdad entre las distribuciones de las variables para cada clase. El p-valor es una probabilidad y, por tanto, $0 \leq p \leq 1$. Generalmente, un valor p cercano a 0 indica mayores diferencias y un valor p cercano a 1 indica diferencias menores.

- **Información mutua** [87]: cada variable x_j y cada clase y mide la dependencia entre la densidad de probabilidad de la variable j-ésima y la clase y. Se estima mediante

$$I(x_j; y) = \int p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} dx_j dy,$$

  donde $p(x_j)$ es la densidad de probabilidad de la variable j-ésima, $p(y)$ es la densidad de probabilidad de la clase y y $p(x_j, y)$ es la densidad conjunta. Cuando las variables son discretas o cuando se discretizan variables continuas se suele emplear la fórmula equivalente para funciones de masa de probabilidad

$$I(X_j; Y) = \sum_x \sum_y p(X_j = x, Y = y) \log \frac{p(X_j = x, Y = y)}{p(X_j = x)p(Y = y)},$$

  donde las probabilidades se estiman mediante conteo de frecuencias. Esta medida se relaciona con la *entropía* del siguiente modo:

$$I(X_j, Y) = H(X_j) - H(X_j|Y).$$

  Donde $H(X_j)$ es la entropía de la variable j-ésima y $H(X_j|Y)$ es la entropía de X_j condicionada a la observación de la variable Y.

- **Fisher Score** [88]: es una métrica que selecciona como más relevantes aquellas variables que maximizan la separación entre clases y minimizan la separación dentro de la misma clase mediante la fórmula

$$FS(x_j) = \frac{\sum_{c=1}^{C} N_c(\mu_{jc} - \mu_j)^2}{\sum_{c=1}^{C} N_c \sigma_{jc}^2},$$

donde μ_j es la media de la variable j, N_c es el número de instancias de la clase c, μ_{jc} es la media de la variable j para las instancias de la clase c y σ_{jc}^2 es la varianza de la variable j para la clase c. La métrica de Fisher otorga mayor valor a las variables que mejor separan las distintas clases y que agrupan de manera más homogénea las instancias que pertenecen a la misma clase. Es la misma idea que se aplicará en el análisis discriminante lineal en la sección 4.2.2.

- **Relief, ReliefF**: el algoritmo *Relief* [89] para problemas de dos clases y la variante multiclase *ReliefF* [90] son algoritmos que estiman la relevancia de las variables en función de su capacidad para distinguir instancias que son cercanas entre sí. En el algoritmo *Relief* se escogen aleatoriamente n instancias y se busca, para cada instancia escogida x_{ij}, su vecino más cercano que pertenezca a la misma clase, h_{ij}, así como el vecino más cercano que pertenezca a la clase contraria, m_{ij}. Así, la estimación de la relevancia es

$$R(x_j) = \frac{1}{2}\sum_{i=1}^{n} d(x_{ij} - m_{ij}) - d(x_{ij} - h_{ij}),$$

donde $d(\cdot)$ es una métrica o distancia entre los dos puntos. Este algoritmo da mayor peso a las variables que separan mejor las instancias que no pertenecen a la misma clase y a las variables que separan peor las instancias que pertenecen a la misma clase.

Para problemas multiclase se emplea el algoritmo *ReliefF*. En este caso, para cada instancia seleccionada aleatoriamente se escogen los K vecinos más próximos que pertenecen a la misma clase $\hat{c}$, h_{kj}. También, para cada una de las otras clases se escogen los K vecinos más próximos, m_{kjc}. Así, el criterio de relevancia queda

$$RF(x_j) = \frac{1}{n}\sum_{i=1}^{n}\Big(-\frac{1}{K}\sum_{h_{kj}} d(x_{ij} - h_{kj}) + \sum_{c\neq\hat{c}}\frac{1}{K}\frac{P(c)}{1-P(\hat{c})}\sum_{m_{kjc}} d(x_{ij} - m_{kjc})\Big).$$

La diferencia principal de la versión *ReliefF* es el empleo de una selección de K instancias vecinas lo que garantiza una robustez mayor del algoritmo. Además, se tienen en cuenta las diferencias con todas las clases existentes.

4.1.2. Wrappers

A diferencia de los métodos indirectos, los métodos directos [91] hacen uso de los algoritmos de aprendizaje, o modelos predictivos, creados mediante los subconjuntos de variables escogidos (ver capítulo 8). Es decir, los modelos ajustados se emplean como caja negra para evaluar los subconjuntos de variables con el que se ha creado el modelo final. Posteriormente, se escogerá el subconjunto de variables que hayan dado el mejor modelo predictivo en el proceso de validación.

Para llevar a cabo estos métodos directos es necesario definir cómo buscar en el espacio de posibles subconjuntos de variables. Otras cuestiones a definir son: cómo medir la capacidad predictiva de los modelos para que guíen la búsqueda de variables y qué tipo de modelos emplear para predecir.

La cuestión de la búsqueda de subconjuntos posibles es, quizás, la más problemática, ya que el coste computacional de una búsqueda exhaustiva de posibles subconjuntos de variables crece

exponencialmente con el número de variables. De hecho, la cardinalidad del conjunto potencia para D variables es 2^D. Para este problema existen diversas estrategias que contrarrestan, en la medida de lo posible, la búsqueda por fuerza bruta de todos los posibles subconjuntos de variables. Las estrategias que se han mostrado particularmente exitosas son las estrategias voraces[a], que aunque subóptimas, son computacionalmente más eficientes y robustas al problema del sobreaprendizaje.

Existen dos métodos básicos para las estrategias voraces: búsqueda hacia adelante o *forward selection* y búsqueda hacia atrás o *backward elimination*. En la primera, se escoge inicialmente una variable. Se entrena un modelo con esta variable y se comprueba su capacidad predictiva mediante una medida concreta M (puede ser un porcentaje de acierto, un error cuadrático medio, etcétera). Después, se incorporan iterativamente nuevas variables creando sendos modelos y evaluando los rendimientos correspondientes con la misma medida M. Se escoge un subconjunto de variables cuando se converge a un subconjunto de variables tal que añadirle una nueva variable no aporta ningún incremento respecto a la medida M.

El método *backward elimination* empieza creando un modelo partiendo del uso de todas las variables. Este modelo se evalúa empleando la medida M. Iterativamente se van retirando variables del subconjunto, creando nuevos modelos y evaluándolos con la medida M. Finalmente, se escogerá aquel subconjunto de variables cuyo modelo correspondiente haya obtenido el mejor resultado con la métrica M y no exista un modelo con mejores prestaciones que incluya menos variables. La figura 4.1 muestra esta metodología para un caso supuesto con tres variables disponibles.

A veces, la elección de uno de los dos métodos puede ser controvertida. Las características del método *forward selection* hacen que la construcción y evaluación de cada nuevo modelo sea computacionalmente más eficiente que el método *backward elimination* para generar subconjuntos de variables. Sin embargo, con el método *backward elimination* se suelen seleccionar subconjuntos de variables con más potencia predictiva debido a que las variables se escogen, desde el principio, teniendo en cuenta el contexto del resto de variables.

4.2. Extracción de características

La extracción de características está intrínsecamente relacionada con el análisis multivariante. Las múltiples variables observadas sobre un mismo objeto se representan como un vector D dimensional que incluye cada una de las observaciones singulares:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)^{\mathrm{T}}.$$

En general, el objetivo del análisis multivariante es explotar las relaciones entre las variables para encontrar transformaciones de los datos que faciliten su comprensión. Es muy habitual encontrar datos multivariantes donde la información relevante se encuentra oculta o latente entre las múltiples variables. Para extraer la información relevante se suelen aplicar transformaciones lineales sobre los datos para obtener un nuevo conjunto de variables que proporcione la información necesaria y suficiente para obtener conocimiento. Las transformaciones lineales son menos costosas y los resultados son más sencillos de interpretar. Estas transformaciones lineales se representan en forma algebraica como:

$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}\mathbf{x}.$$

[a]Una estrategia voraz es aquella que escoge en cada iteración el mejor resultado para una función objetivo de entre un conjunto de posibilidades sin tener en cuenta futuras combinaciones.

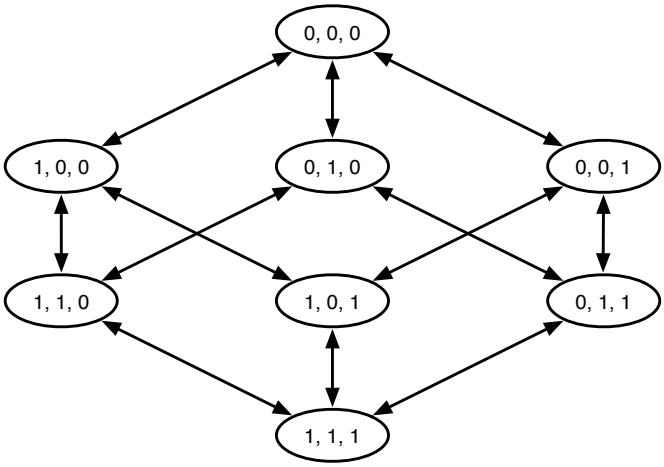


Figura 4.1: Espacio de búsqueda para la selección de un subconjunto de variables. Se dispone de tres posibles variables y cada nodo representa el número de variables que se introducen en el modelo: un 1 indica que la variable se introduce y un 0 indica lo contrario. Cada nodo se conecta con otros nodos que contienen una variable añadida o eliminada. Por tanto, una estrategia *forward selection* partirá del nodo superior y recorrerá el espacio hacia abajo por el camino que mejores prestaciones vaya obteniendo de manera voraz. La estrategia *backward elimination* partirá del nodo inferior y recorrerá el espacio hacia arriba del mismo modo.

La ventaja de algunas de las técnicas que se aplican para extraer características es que la proyección de los datos se puede hacer en unas dimensiones reducidas sin más que obtener una matriz de transformación $\mathbf{W}$ con dimensiones $D \times K$, con $K < D$. De este modo, si la transformación está bien escogida se puede llegar a reducir las dimensiones de los datos a cambio de perder un mínimo de información.

4.2.1. Análisis de Componentes Principales

El análisis de componentes principales (PCA, por sus siglas en inglés) proporciona una manera de facilitar la identificación de patrones y de expresar los datos resaltando las diferencias y similitudes de los mismos. Para ello, PCA busca la proyección de los datos en un espacio cuyos ejes sean ortogonales entre sí y representen la máxima varianza posible. Por lo tanto, las variables, originalmente correlacionadas, se transforman en un conjunto de variables linealmente no correlacionadas. Las nuevas variables son combinaciones lineales de las variables originales donde los coeficientes asociados están en orden decreciente, de tal modo que el primer coeficiente explica tanta varianza de los datos originales como sea posible y, así, sucesivamente.

Formalmente, se desea proyectar un conjunto de N datos d-dimensionales, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, en un espacio con base ortonormal, esto es, un espacio cuyos ejes son vectores unitarios y ortogonales entre sí. El algoritmo PCA exige que la media empírica de los datos sea cero. Esto se puede demostrar teóricamente, pero complica innecesariamente la explicación. Por lo tanto, en adelante asumiremos que la media de los datos es cero, $\mathbf{m} = \mathbf{0}$.

Para la primera componente $\mathbf{w}_1$ de la matriz $\mathbf{W}$ el objetivo de PCA es proyectar los datos sobre una recta que pase por la media, así,

$$\begin{aligned} \mathbf{x}_j &= \mathbf{m} + a_j \mathbf{w}_1 \\ &= a_j \mathbf{w}_1 \end{aligned} \tag{4.1}$$

donde a_j será el coeficiente asociado a la observación $\mathbf{x}_j$. Es posible encontrar un conjunto óptimo de coeficientes si se minimiza el error cuadrático:

$$J_1(a_1, \ldots, a_N, \mathbf{w}_1) = \sum_{j=1}^{N} ||a_j \mathbf{w}_1 - \mathbf{x}_j||^2. \tag{4.2}$$

Como los vectores deben ser unitarios, tenemos que $||\mathbf{w}_1|| = 1$, despejando, derivando parcialmente respecto a a_j e igualando a cero se obtiene que

$$a_j = \mathbf{w}_1^{\mathrm{T}} \mathbf{x}_j. \tag{4.3}$$

Si se sustituye la igualdad 4.3 en la ecuación 4.2, se puede encontrar la mejor dirección del vector $\mathbf{w}_1$ que pasa por la media muestral. De modo que la función criterio queda:

$$\begin{aligned} J_1(\mathbf{w}_1) &= -\sum_{j=1}^{N} \mathbf{w}_1^{\mathrm{T}} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} \mathbf{w}_1 + \sum_{j=1}^{N} ||\mathbf{x}_j||^2 \\ &= -\mathbf{w}_1^{\mathrm{T}} \mathbf{S} \mathbf{w}_1 + \sum_{j=1}^{N} ||\mathbf{x}_j||^2, \end{aligned} \tag{4.4}$$

donde $\mathbf{S} = \sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}$, es la *matriz de dispersión* cuando $\mathbf{m} = 0$. La matriz de dispersión no es más que la matriz de covarianzas muestral no sesgada multiplicada por $(N-1)$: $\mathbf{S} = (N-1)\mathbf{\Sigma}$.

En la ecuación 4.4, se observa que para minimizar la función criterio $J_1(\cdot)$ se debe maximizar $\mathbf{w}_1^{\mathrm{T}}\mathbf{S}\mathbf{w}_1$. La maximización de esta expresión se transforma en un problema de vectores propios donde $\mathbf{w}_1$ es el vector propio de $\mathbf{S}$:

$$\mathbf{S}\mathbf{w}_1 = \lambda\mathbf{w}_1. \tag{4.5}$$

Así pues, para maximizar la función criterio $J_1(\cdot)$ se debe seleccionar el vector propio asociado al mayor valor propio de la matriz de dispersión. De este modo, se podrían proyectar las observaciones en una única dimensión. Sin embargo, el análisis de componentes principales nos permite obtener el resto de componentes principales y, en consecuencia, aumentar la proyección a k dimensiones, donde $k \leq d$. Para ello, se generaliza la ecuación 4.1 para proyectar las observaciones en un espacio k-dimensional:

$$\mathbf{x}_j = \mathbf{m} + \sum_{i=1}^{k} a_{ji}\mathbf{w}_i. \tag{4.6}$$

Se demuestra que la función criterio

$$J_k(\mathbf{w}) = \sum_{j=1}^{N} ||\sum_{i=1}^{k} a_{ji}\mathbf{w}_i) - \mathbf{x}_j||^2 \tag{4.7}$$

se minimiza cuando los vectores $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ son los k vectores propios de la matriz de dispersión con los mayores valores propios asociados $\lambda_1, \lambda_2, \ldots, \lambda_k$. Dado que la matriz de dispersión es simétrica, los vectores propios son ortogonales. Los coeficientes a_i son los *componentes principales*. Nótese que el empleo de la matriz de covarianzas $\mathbf{\Sigma}$ en lugar de la matriz de dispersión $\mathbf{S}$ no introduce ningún cambio en la magnitud ni la dirección de los vectores $\mathbf{w}_i$, ya que estas dos matrices solo se diferencian en una constante. En cambio, la magnitud de los valores propios sí se verá afectada, aunque la relación entre ellos permanecerá constante. Esto quiere decir que el uso de la matriz de covarianzas es tan válido como el uso de la matriz de dispersión para el cálculo del PCA.

Se puede demostrar que la suma de las varianzas obtenidas mediante análisis de componentes principales es igual a la suma de varianzas de los datos originales. Es decir, $\sum_{i=1}^{d}\lambda_i = \sum_{i=1}^{d}\sigma_i^2$. Esto resultado es importante, ya que la variación total explicada por las componentes principales es igual a la cantidad total de variación medida por las variables originales. Esto permite ordenar los valores propios de las componente principales y poder escoger un número k de variables componentes principales de modo que representen un porcentaje de variación del total de varianza original. Es así como se puede escoger una matriz de transformación $\mathbf{W} = [\mathbf{w}_1\mathbf{w}_2 \ldots \mathbf{w}_k]$ con dimensiones $D \times K$ que permita transformar los datos en un espacio de menores dimensiones perdiendo el mínimo de información posible.

Ejemplo 4.1 (Análisis de Componentes Principales para tumores cerebrales) Disponemos de una base de datos con la concentración de diez metabolitos característicos de la actividad cerebral que se han obtenido mediante espectros de resonancia magnética. Algunos de estos metabolitos resuenan en distintas frecuencias, por lo que se dispone de un total de 15 variables. Estos datos se utilizan para discriminar distintos tipos de tumores cerebrales, en concreto, glioblastomas, astrocitomas y meningiomas. Sobre estos datos, aplicaremos un análisis de componentes principales para reducir la dimensionalidad y proyectar los datos en dos dimensiones cuyos ejes sean ortogonales.

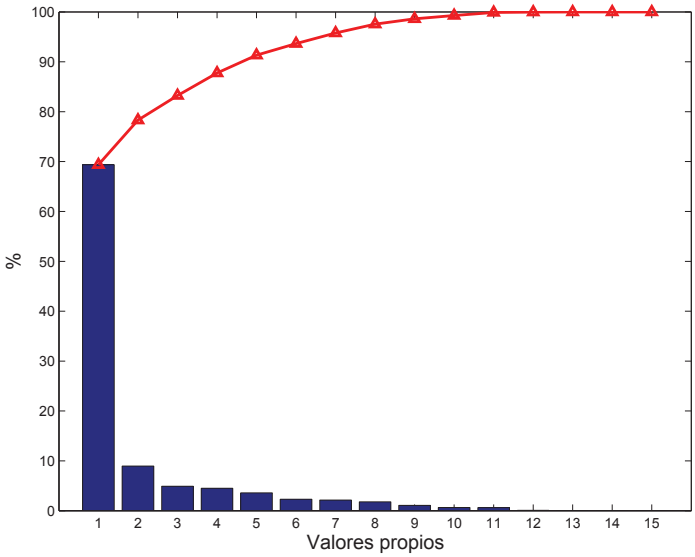


Figura 4.2: Es posible observar gráficamente la cantidad de variación que representa cada componente principal para los datos de tumores cerebrales. Aquí se observa cómo la primera componente principal representa casi el 70 % de la variabilidad de los datos. Mientras, la segunda componente principal y siguientes representan una variabilidad menor al 10 % del total. La línea roja muestra la cantidad de variabilidad acumulada que se representan con las componente principales. Solo las dos primeras representan casi el 80 % de la variación total de los datos. Si asumimos que el 20 % restante es ruido, entonces podremos representar los datos en dos dimensiones.

El primer paso es tipificar las variables. Como se explicó anteriormente, esto se necesita para que las variables se midan en unidades comparables. De lo contrario, si una variable tiene una varianza mucho más grande que las demás, dominará la primera componente principal, sin importar la estructura de las covarianzas de las variables.

Una vez estandarizadas las variables, si se aplica el análisis de componentes principales se obtienen los valores propios λ_i asociados a cada vector propio $\mathbf{w}_i$.
A partir de los valores de la varianza de cada componente principal se puede obtener una gráfica en la que se establece qué porcentaje de varianza explica cada componente principal (ver figura 4.2).
Si tomamos las dos primeras componentes principales, que representan cerca del 80 % de la variabilidad de los datos, podemos proyectar cada instancia en un espacio muestral bidimensional mediante una matriz de transformación $\mathbf{W}$ de dimensiones 2×15. De este modo, se han podido extraer las variables latentes detrás de las variables originales, reduciendo el número de variables de 15 a 2.

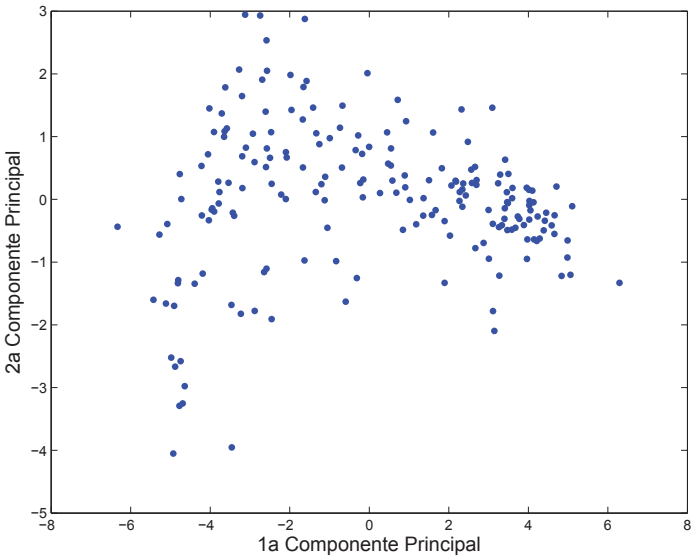


Figura 4.3: Una vez proyectados los datos mediante la matriz $\mathbf{W}$ utilizando las dos primeras componentes principales podemos observar la disposición de los casos disponibles en el espacio bidimensional latente en los datos.

4.2.2. Discriminante lineal de Fisher

El objetivo del análisis discriminante lineal de Fisher es encontrar una función lineal que proyecte las observaciones a un espacio donde se maximice la distancia entre las medias de las clases y se minimice la dispersión de las clases. Esta proyección busca, por tanto, aquellas direcciones que sean eficientes para la discriminación de clases.

Se considerará inicialmente un problema dicotómico, donde el objetivo es discriminar objetos d-dimensionales entre dos clases posibles, $\mathcal{C}_1$ y $\mathcal{C}_2$, con n_1 y n_2 observaciones respectivamente. En el análisis discriminante cada observación $\mathbf{x} \in \mathbb{R}^d$ será proyectada a un nuevo espacio a partir de una combinación lineal de $\mathbf{x}$:

$$\mathbf{z} = \mathbf{w}^{\mathrm{T}}\mathbf{x}, \tag{4.8}$$

donde el vector de proyección $\mathbf{w}$ determina la dirección del plano donde serán proyectadas las observaciones. Generalmente, se establece que $||\mathbf{W}|| = 1$ ya que la magnitud únicamente implica un escalado diferente de las proyecciones $\mathbf{z}$. El problema reside en encontrar la dirección adecuada de $\mathbf{W}$, de modo que se maximice la separación entre las clases y se minimice la dispersión dentro de cada clase. Una medida de la separación entre las clases es la *matriz de dispersión interclase* que, para un problema de dos clases, se define como

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}. \tag{4.9}$$

La dispersión dentro de cada clase se puede medir a partir de la *matriz de dispersión intra-clase*, que es la suma de las matrices de dispersión de cada clase:

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{|\mathcal{C}|} \mathbf{S}_i \\ &= \sum_{i=1}^{|\mathcal{C}|} \sum_{\mathbf{x}\in\mathcal{C}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^{\mathrm{T}}.\end{aligned} \tag{4.10}$$

La matriz de dispersión intraclase $\mathbf{S}_W$ es proporcional a la matriz de covarianzas muestral, es simétrica y semidefinida positiva y, si el número de muestras es mayor que el número de dimensiones $n > d$, normalmente es no singular.

Estas dos medidas permiten definir una función criterio a maximizar para encontrar los valores óptimos del vector de proyección $\mathbf{x}$:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}}, \tag{4.11}$$

lo que implica que la mejor solución será aquella que separe lo máximo posible las medias de las clases en relación a la suma de dispersiones de los datos de cada clase. La expresión 4.11 es conocida como cociente de Rayleigh generalizado. Es fácil demostrar que el vector $\mathbf{w}$ que maximiza la función criterio $J(\cdot)$ satisface:

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}, \tag{4.12}$$

que es un problema de valores propios generalizado. Si $\mathbf{S}_W$ es una matriz no singular, se puede obtener un problema de valores propios convencional a partir de 4.12,

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}. \tag{4.13}$$

Para el problema de las dos clases, no es necesario resolver el problema de los valores propios de $\mathbf{S}_W^{-1}\mathbf{S}_B$ ya que $\mathbf{S}_B\mathbf{w}$ estará en la dirección de $\mathbf{m}_1 - \mathbf{m}_2$ y, dado que el factor de escalado es despreciable, la solución que optimiza $J(\cdot)$ es:

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{4.14}$$

De este modo, se obtiene el vector de proyección $\mathbf{w}$ para el discriminante lineal de Fisher que maximiza la proporción entre la dispersión interclase y la dispersión intraclase. La clasificación de las observaciones proyectadas se puede llevar a cabo estableciendo un umbral ϕ y escogiendo la clase $\mathcal{C}_1$ si se excede dicho umbral, o la clase $\mathcal{C}_2$ en caso contrario.

4.2.3. Análisis discriminante para múltiples clases

En la generalización del análisis discriminante para múltiples clases podemos asumir que el número de dimensiones de las observaciones d es mayor que el número de clases $|\mathcal{C}|$. Además, la proyección se hará en un espacio de $K = |\mathcal{C}| - 1$ dimensiones por razones que se explicarán más adelante. Para ello se necesitan K vectores de proyección $\mathbf{w}_k$, que serán las columnas de la matriz de transformación $\mathbf{W}$, así

$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}\mathbf{x}. \tag{4.15}$$

En este caso, la generalización para la matriz de dispersión intraclase es trivial y sigue la misma expresión que en la ecuación 4.10. La matriz de dispersión interclase generalizada se obtiene a partir de la matriz de dispersión total $\mathbf{S}_T$ según proponen Duda y Hart [92], ya que

$$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}. \tag{4.16}$$

La matriz de dispersión total se puede descomponer en la suma de la matriz de dispersión intraclase $\mathbf{S}_W$ y una expresión que se identifica con la matriz de dispersión interclase generalizada $\mathbf{S}_B$,

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B, \tag{4.17}$$

donde

$$\mathbf{S}_B = \sum_{i=1}^{|\mathcal{C}|} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathrm{T}}. \tag{4.18}$$

De nuevo, el objetivo es establecer una función criterio $J(\mathbf{W})$ que nos permita maximizar la dispersión interclase y minimizar la dispersión intraclase. Este criterio puede depender explícitamente de la matriz de proyección $\mathbf{W}$ y está relacionada con la ecuación 4.11:

$$J(\mathbf{W}) = \frac{|\mathbf{W}^{\mathrm{T}} \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^{\mathrm{T}} \mathbf{S}_W \mathbf{W}|}. \tag{4.19}$$

El problema de encontrar la matriz de proyección que maximice $J(\mathbf{W})$ puede resolverse teniendo en cuenta que cada vector columna $\mathbf{w}_k$ de una matriz $\mathbf{W}$ óptima corresponde a los vectores propios asociados a los mayores valores propios:

$$\mathbf{S}_B \mathbf{w}_k = \lambda_k \mathbf{S}_W \mathbf{w}_k. \tag{4.20}$$

Como $\mathbf{S}_B$ es la suma de $|\mathcal{C}|$ matrices de rango menor o igual a uno, y dado que solo $|\mathcal{C}| - 1$ de estas matrices son independientes, se deduce que

$$\text{rango}(\mathbf{S}_B) \leq |\mathcal{C}| - 1.$$

Esto explica que el número de valores propios distintos de cero no sea mayor que $|\mathcal{C} - 1|$ y es la razón por la que el espacio donde se proyectan las observaciones tenga estas dimensiones.

Ejemplo 4.2 (Análisis Discriminante Lineal para tumores cerebrales (cont.))
En el ejemplo anterior se aplicó un análisis de componentes principales al problema de discriminación de tumores cerebrales. Con ello se consiguió reducir la dimensionalidad de un espacio de 15 dimensiones a otro bidimensional. Sin embargo, el PCA busca proyectar los datos en un espacio cuyos ejes sean ortogonales entre sí, reduciendo así la correlación entre variables, pero sin tener en cuenta la capacidad discriminante de las variables latentes encontradas. El análisis discriminante lineal, en cambio, extrae variables latentes que sean capaces de discriminar mejor los datos de cada clase (ver figura 4.4).

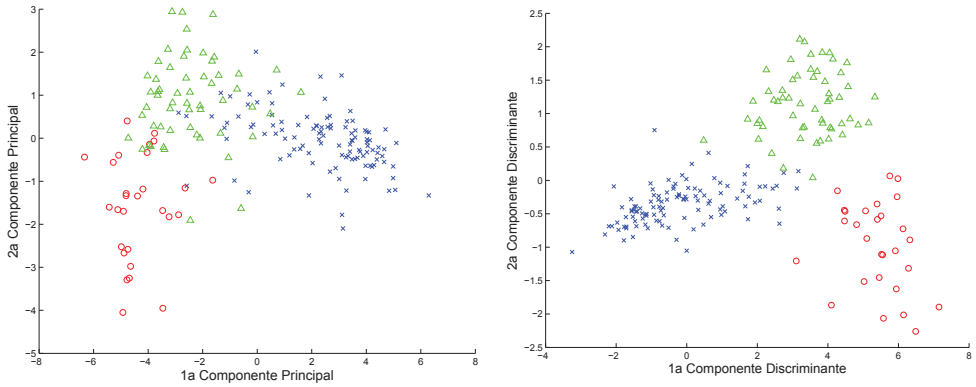


Figura 4.4: Comparación entre la proyección empleando las 2 primeras componente principales de la proyección PCA (izquierda) y empleando las 2 primeras componentes discriminantes de la proyección LDA (derecha). Se observa cómo la separación entre las distintas clases es mayor cuando se emplea LDA. Esto se debe a que el objetivo de LDA es encontrar las variables latentes que mejor separen los datos de las distintas clases. En cambio, el objetivo de PCA es encontrar una proyección donde los datos no estén correlacionados y cuyos ejes representen tanta variabilidad de los datos originales como sea posible.

4.3. Notas bibliográficas

Dos referencias básicas donde se estudian las técnicas de extracción de características son los libros de Duda, Hart y Stork de 2001 [92] y el libro de Bishop de 2006 [93]. El uso de la información mutua es bastante común para selección de características en el análisis automático de textos y lenguaje natural, pero también en otros campos. Sus fundamentos se basan en la Teoría de la Información fundada por Claude Shannon. Una buena introducción a la Teoría de la Información es el libro de Cover y Thomas [94]. Si el lector desea profundizar más en el análisis discriminante lineal, introducido por Ronald Fisher, puede acceder al artículo original [88]. Para profundizar en el análisis de componentes principales es recomendable el trabajo de Jollife [95]. Las técnicas de selección de características pueden estudiarse en [86] y [91]. Un análisis teórico y empírico de los métodos Relief y ReliefF y referencias a los métodos de selección de características se puede encontrar en Kononenko [96].

Finalmente, cabe mencionar la existencia de métodos de reducción de dimensionalidad no lineales los cuales tienen como fin capturar relaciones o posiciones relativas entre los puntos presentes en dimensiones superiores y proyectarlas en un espacio de dimensión inferior [97]. Entre ellos cabe destacar el método tSNE ampliamente empleado para la visualización de datos de grandes dimensiones [98].

Capítulo 5

Procesamiento de cadenas

La primera parte del capítulo se centra en el uso de expresiones regulares según el estándar IEEE POSIX 1003.2 (Sección 2.8) y en las aplicaciones de análisis sintáctico para la manipulación de ficheros de texto. La segunda parte del capítulo profundiza en algoritmos de programación dinámica para el análisis de similitud de cadenas, centrándose en su aplicación para el apoyo a la genómica.

5.1. Expresiones regulares para el tratamiento de datos

En repetidas ocasiones, es necesario tratar los conjuntos de datos clínicos mediante herramientas de búsqueda y sustitución de cadenas, con el fin de preparar los datos para su posterior tratamiento con algoritmos de reconocimiento de patrones o minería de datos. Generalmente, para conjuntos de datos pequeños puede ser suficiente con emplear un editor de textos y hacer búsquedas y sustituciones manualmente, sin embargo, esta tarea se convierte en titánica cuando el conjunto de datos es mayor. Para estos casos existen herramientas para el tratamiento de estos datos que habitualmente hacen uso de las expresiones regulares. Además, en el capítulo 15 veremos herramientas ETL para la transformación de registros procedentes de bases de datos u otras estructuras de datos.

Una expresión regular es un patrón que describe un conjunto de cadenas para la búsqueda codificada en textos planos. La sintaxis queda definida por las expresiones regulares extendidas (ERE) del estándar IEEE POSIX 1003.2 (Sección 2.8). Las ERE incluyen como subconjunto a las expresiones regulares básicas. De aquí en adelante nos centraremos en las ERE.

Empezaremos definiendo algunos términos importantes para componer expresiones regulares:

- literal: un literal es cualquier carácter que se use en una búsqueda. Es, literalmente, la cadena que queremos encontrar. Por ejemplo, para encontrar la subcadena “ent” en la palabra “paciente”, el literal será “ent”.

- metacarácter: un metacarácter es un carácter especial que tiene un significado único y que no se emplea como literal en la expresión de búsqueda. Por ejemplo, el carácter “^” es un metacarácter.

- secuencia de escape: una secuencia de escape sirve para indicar que se quiere utilizar un metacarácter como literal. En una expresión regular la secuencia de escape suele indicarse mediante una contrabarra “\” antes del metacarácter que se desea emplear como literal.

Las expresiones regulares se construyen combinando literales, metacaracteres y expresiones más pequeñas. La expresión regular más sencilla es la que se corresponde con un carácter único. La mayor parte de caracteres, incluyendo dígitos y letras, son expresiones regulares que se corresponden consigo mismas. Los metacaracteres con significado especial deben estar precedidos con una contrabarra "\".

Cuando una lista de caracteres se enmarca entre corchetes se realiza una búsqueda de cualquier carácter en dicha lista. Si el primer carácter es un circunflejo "^", entonces se buscará cualquier carácter que no esté en la lista. Por ejemplo, la expresión regular [0123456789] encontrará cualquier dígito entre 0 y 9. Dentro de una expresión entre corchetes, se puede indicar rangos si se emplean dos caracteres separados por un guión. Por ejemplo, [0-9] es equivalente a [0123456789]. También, el conjunto [a-d] equivale a [abcd].

Existen también algunas clases de caracteres predefinidas[a]:

- [:alnum:] cualquier carácter alfanumérico de 0 a 9, de "a" a "z" o de "A" a "Z". Se puede abreviar utilizando "\w". Emplear "\W" es equivalente a [^[:alnum:]].
- [:alpha:] cualquier carácter alfabético.
- [:cntrl:] caracteres de control (retorno de carro, tabulador, etc).
- [:digit:] dígitos del 0 al 9. Se puede abreviar utilizando "\d". Emplear "\D" es equivalente a [^[:digit:]].
- [:lower:] caracteres alfabéticos en minúsculas.
- [:upper:] caracteres alfabéticos en mayúsculas.
- [:print:] cualquier carácter imprimible.
- [:punct:] símbolos de puntuación.
- [:space:] caracteres con espacios en blanco.
- [:blank:] espacio y tabuladores.
- [:xdigit:] caracteres en notación hexadecimal, del 0 al 9, de "a" a "f" o de "A" a "F".

Para casar cualquier carácter se emplea el punto "." que sirve como comodín. Además, una expresión regular puede estar seguida por los siguientes operadores de repetición:

- ? el elemento anterior es opcional y a lo sumo se empareja una vez.
- * el elemento anterior aparece cero o más veces.
- + el elemento anterior aparece una o más veces.
- {n} el elemento anterior aparece exactamente n veces.
- {n,} el elemento anterior aparece n o más veces.
- {n,m} el elemento anterior aparece al menos n veces, pero no más de m veces.

[a]Es importante destacar que, según los comandos que se utilicen, no siempre son válidas estas clases predefinidas. Por lo tanto, se debe tener especial cuidado al usar estas simplificaciones.

Las expresiones regulares se pueden concatenar. También se puede emplear un operador disyuntivo | para que se empareje cualquier cadena con una de las dos expresiones. Además, se pueden emplear paréntesis para indicar la precedencia de las expresiones.

En algunos comandos que veremos a continuación, las expresiones regulares pueden sufrir variaciones en su sintaxis. Por esta razón se debe tener especial cautela al trabajar con estas expresiones. Por ejemplo, cuando se emplean expresiones regulares básicas, los metacaracteres "?", "+", "{", "|", "(" y ")", pueden necesitar una contrabarra que les preceda, esto es, "\?", "\+", "\{", "\|", "\(" y "\)".

5.2. Comandos para manipular ficheros de texto plano

A continuación se verán algunos comandos de Unix/Linux[b] muy útiles para trabajar con ficheros de texto plano, algunos de los cuales nos permitirán emplear expresiones regulares para manipular las cadenas de texto de los ficheros de datos si disponemos de ellos en formato ASCII. En todos los casos se recomienda hacer uso del manual de Unix/Linux para obtener una información completa de los diferentes comandos. Esto se consigue mediante la orden *man*.

5.2.1. cut

El comando **cut** permite seleccionar columnas de cada línea del fichero. Generalmente, los ficheros de datos contienen un paciente por fila y sus variables se muestran separadas por algún tipo de separador (un espacio, una coma, etc). Con **cut** podemos seleccionar aquellas variables que nos interesen por alguna razón concreta.

El comando tiene dos opciones muy interesantes. La primera (opción -d) nos permite especificar el tipo de delimitador de campos. Por defecto, se asume que el delimitador es el tabulado. La segunda opción es necesaria utilizarla (opción -f), ya que nos permite especificar qué campos queremos seleccionar. Su sintaxis sería:

```
cut [-d delim] -f <lista_campos> <fichero>.
```

La lista de campos se puede indicar de varias maneras, por ejemplo:

- `-f 1,3,5,7`

  seleccionará los campos 1, 3, 5 y 7.

- `-f 2-5`

  seleccionará los campos del 2 al 5.

- `-f 5-`

  seleccionará los campos del 5 hasta el final.

- `-f -5`

  seleccionará los campos del principio hasta el campo número 5.

También se pueden combinar las posibilidades anteriores mediante una coma, por ejemplo:

- `-f -3,5,7,9-`

permitirá seleccionar los campos del 1 al 3, el 5, 7 y del 9 al final.

Un comando complementario para el **cut** es el comando **paste**, que permite fusionar las columnas de dos ficheros distintos.

[b] También puede utilizarse la consola Cygwin para windows, `http://www.cygwin.com`.

Ejemplo 5.1 (Uso de cut)
Se dispone de una base de datos de una escuela de enfermería, de la que deben obtenerse modelos de clasificación para determinar a dónde deben derivarse los pacientes que están en recuperación postoperatoria: a cuidados intensivos, a una planta de recuperación general o prepararlos para ser enviados a casa. El conjunto de datos está en texto plano, lo cual debería verse como una ventaja, puesto que es un formato fácilmente manipulable y se puede importar en numerosas aplicaciones de tratamiento estadístico de datos.

Si quisiéramos seleccionar algunas variables, como las relacionadas con la temperatura del paciente, las de estabilidad del paciente y la decisión que se toma para derivar al paciente, podríamos usar el comando **cut**. Un vistazo rápido de los datos nos mostrará que cada fila del fichero es un paciente y sus variables están separadas por ";". Así, se procedería del siguiente modo:

```
$> cut -d ";" -f 1,2,6-8,10 post-operative.txt
```

5.2.2. uniq

El comando **uniq** nos permitirá filtrar las líneas que se repiten de manera consecutiva de un archivo de texto plano. Una opción de entrada interesante nos permitirá además contar el número de repeticiones de las líneas (opción -c). Esta opción es muy útil cuando se combina con el comando **sort** que nos permite ordenar alfabéticamente las líneas del fichero. Estos comandos se pueden combinar empleando tuberías "|". En entornos Unix/Linux, las tuberías permiten conectar la salida estándar de un comando con la entrada estándar del siguiente comando.

Por defecto, los sistemas tipo Unix/Linux dirigen la salida estándar de datos hacia la pantalla y la entrada estándar se lee del teclado. Sin embargo, existe la posibilidad de redireccionar la salida estándar y la entrada estándar. Esto se puede conseguir empleando ">" para redireccionar la salida estándar, de este modo, los resultados de un comando se pueden guardar en un fichero. Si se emplea ">>" en su lugar, se puede añadir el resultado del comando al final del fichero que se indique. Para cambiar la entrada estándar se emplea "<".

Ejemplo 5.2 (Uso de cut y uniq)
Una de las primeras acciones debería ser observar la prevalencia de cada una de las clases, donde cada clase es una de las posibles derivaciones del paciente y dependerá de las variables recogidas en el conjunto de datos. Para obtener el número de observaciones por clase podremos aplicar una combinación de los comandos vistos hasta ahora:

```
$> cut -d ";" -f 10 post-operative.txt | sort | uniq -c
```

5.2.3. sed

El comando **sed** proporciona la capacidad de modificar cadenas de un fichero de texto plano. Este comando puede hacer uso de expresiones regulares básicas, aunque existe la opción de emplear expresiones regulares extendidas (opción -E). Con el comando **sed** podemos eliminar, añadir o sustituir cadenas de ficheros de texto, entre otras posibilidades.

La sintaxis para sustituir una cadena del fichero por otra es:

```
sed 's/expresión_buscada/expresión_sustitución/g' <fichero>
```

la "s" indica que busque la cadena que se indica a continuación. Después se indica la expresión que debe sustituir a la buscada. El modificador "g" indica que debe aplicarse a todas las cadenas

que se encuentren en la misma línea. Si no utilizamos el modificador "g", la acción de eliminación de la cadena se aplicará únicamente a la primera cadena que encuentre en la línea.

Para eliminar una cadena la expresión es similar, simplemente le indicamos que debe sustituirse por una cadena vacía:

```
sed 's/expresión_buscada//g' <fichero>
```

El añadido de una cadena tras otra requiere del modificador "&" que indica que la expresión a sustituir debe aparecer en la expresión de sustitución. Por lo tanto, si deseamos añadir una expresión a la cadena buscada se utiliza:

```
sed 's/expresión_buscada/&expresión_añadida/g' <fichero>
```

Un comando que puede ser empleado como alternativa es el comando **tr**.

Ejemplo 5.3 (Uso de sed)
En el conjunto de datos de recuperación postoperatoria aparecen variables continuas con los decimales indicados mediante coma decimal. Sin embargo, muchos de los procesadores estadísticos emplean punto decimal para indicar los decimales. Una forma de procesar esta variable es emplear el comando **sed** del siguiente modo:

```
$> sed 's/,/\./g' post-operative.txt
```

Es importante observar que se ha empleado una secuencia de escape para el punto. Esto se debe a que el punto tal cual es un metacarácter en las expresiones regulares como se explica en la sección 5.1.

Además, puede ser de utilidad sustituir el símbolo de separación de campos actual ";" por un espacio. Esto se haría de un modo parecido:

```
$> sed 's/;/ /g' post-operative.txt
```

Los dos procesos anteriores pueden encadenarse mediante tuberías del siguiente modo:

```
$> sed 's/,/\./g' post-operative.txt | sed 's/;/ /g'
```

Las capacidades del comando **sed** son muchas más de las aquí expuestas. Sin embargo, estos ejemplos son suficientes para ilustrar algunos de los usos más habituales a la hora de procesar ficheros de datos multivariantes.

5.2.4. grep

El comando **grep** muestra las líneas del fichero que encajan con un patrón o expresión introducida por el usuario. Este patrón puede venir indicado como una expresión regular extendida. La sintaxis del comando es:

```
grep [opciones] <expresión_buscada> <fichero>
```

Algunas de las opciones interesantes son:

- -c: en lugar de mostrar las líneas que encajan con el patrón muestra el número de ellas.
- -n: muestra el número de la línea junto a cada línea filtrada con éxito.

- -v: invierte el filtro, de modo que muestra aquellas líneas que no encajan con el patrón. Esto es similar a usar "^" en las expresiones regulares.

- -w: selecciona únicamente aquellas líneas donde el patrón encaja para cadenas completas. Es decir, cadenas que están precedidas y sucedidas por espacios en blanco, tabulados o fin de línea.

Este comando es verdaderamente útil para filtrar conjuntos de datos por variables. Por ejemplo, si queremos seleccionar solo aquellos pacientes que pertenecen a una clase en concreto porque se desea hacer un estudio específico de ese tipo de pacientes. También nos permite filtrar pacientes con datos perdidos.

Ejemplo 5.4 (Uso de grep)
En los datos de recuperación post-operatoria existen pacientes con datos perdidos que, a falta de conocer técnicas de imputación de datos perdidos, podría ser necesario eliminarlos para no perjudicar el estudio. En el fichero de datos los datos perdidos se muestran con un interrogante "?". Una forma de seleccionar los pacientes sin datos perdidos es:

```
$> grep -v ? post-operative.txt
```

De este modo, podremos disponer solo de pacientes con los datos completos, lo que puede facilitar el análisis de los mismos si no tenemos herramientas tolerantes a datos perdidos o no se emplean técnicas de imputación de datos perdidos.

5.2.5. Awk

Awk es un lenguaje de programación interpretado como pueden ser Python, Perl o Lua. La ventaja de Awk sobre el resto de lenguajes, respecto al tratamiento de ficheros de texto plano, es su sencillez y velocidad. Sin embargo, Awk es menos versátil que otros lenguajes de programación interpretados. Awk está orientado para el procesado de líneas de un fichero de texto plano, ya que lleva a cabo un barrido de cada línea del fichero que encaje con el patrón especificado, si lo hubiere. A cada patrón se le puede asociar una serie de acciones a realizar como parte del proceso. Si no se indica un patrón concreto, las acciones se aplican a todas las líneas.

Una línea consiste en una serie de campos separados por un espacio en blanco o por algún tipo de expresión regular. En caso de estar separados por alguna expresión regular se debe emplear la opción "-F" para definir el separador de campos. También se pueden asignar valores a variables antes de ejecutar el programa mediante la opción "-v". La sintaxis general es:

```
awk [-F separador] [-v var=valor] '[/patrón/]{acciones}' <fichero>
```

Los campos de cada línea se indican con la notación $1, $2, etcétera. Si se utiliza $0 se referirá a toda la línea completa. La variable NF contiene el número de campos de la línea que se está procesando en ese momento. Igualmente, con $NF se puede referenciar al último campo de la línea. La variable NR contiene el número de línea que se está procesando.

Las acciones pueden contener secuencias de instrucciones, incluyendo instrucciones condicionales y de control de flujo, como *if*, *while*, *for*, etc. En la página de manual de Awk (*man awk*) se puede encontrar la sintaxis de estas instrucciones. Otra instrucción muy importante es *printf* que imprime en la salida estándar los argumentos que se le indiquen. El formato de la función *printf* se puede consultar en la página de manual (*man printf*).

Ejemplo 5.5 (Uso de awk para transformación de variables)
La primera variable de los datos sobre recuperación post-operatoria muestra la temperatura interna del paciente en grados centígrados. Como en ocasiones es interesante discretizar los datos continuos, se pasará la variable continua de temperatura a una variable categórica ordinal donde los valores serán {baja, media, alta} y estarán basados en las siguientes reglas:

- baja: si la temperatura es menor que 36ºC,
- media: si la temperatura está entre 36ºC y 37ºC,
- alta: si la temperatura es mayor que 37ºC.

Una posible forma de obtener esta discretización mediante Awk sería así:

```
$> awk -F ";" '{
    if($1<36) {
        printf("low;")
    } else {
        if($1<37) {
            printf("mid;")
        } else {
            printf("high;")
        }
    };
    for(i=2;i<NF;i++) {
        printf $i ";"
    };
    printf $NF "\n"
}' post-operative.txt
```

Ejemplo 5.6 (Uso de awk para selección de variables)
Se quiere hacer un estudio de los pacientes que son enviados a casa y la presión sanguínea de los mismos. Para ello, se desea aislar los datos de estos pacientes y escoger únicamente los datos relacionados con la presión sanguínea: presión sistólica, presión diastólica y estabilidad de la presión sanguínea.

Estas variables son los campos 4,5 y 8. Una forma de obtener los datos mediante Awk sería:

```
$> awk -F ";" '{
    if($NF=="S") {
        printf $4 " " $5 " " $8 "\n"
    }
}' post-operative.txt
```

Con Awk es posible introducir unas instrucciones de control antes y después de procesar la primera línea del fichero de datos. Esto se consigue mediante los entornos BEGIN y END. Estos entornos dotan de mayor flexibilidad al programa.

Ejemplo 5.7 (Uso de awk con entornos BEGIN y END)
Si, en lugar de obtener los datos de presión sanguínea de los pacientes que son enviados a casa, se desease comprobar la presión sanguínea sistólica media de dichos pacientes, los entornos BEGIN y END resultan necesarios. La presión media podría conseguirse del siguiente modo:

```
$> awk -F ";"
    'BEGIN {
        sum=0;
        n=0;
    }

    {
        if($NF=="S"){
            sum=sum+$4;
            n=n+1;
        }
    }

    END {
        printf("%3.2f \n",sum/n)
}' post-operative.txt
```

Como se puede observar, el entorno BEGIN inicia las variables *sum* y *n* a 0. La primera variable contiene la suma de los valores de presión sistólica de los pacientes; la segunda variable contiene el número de pacientes. En el cuerpo del programa, se recorren las líneas del fichero de datos y se suman los valores de presión sistólica de aquellos pacientes que cumplen la condición de ser enviados a casa. Además, se van contabilizando en la variable *n*. Finalmente, el entorno END nos permite calcular la media aritmética al dividir la suma total de presiones sistólicas entre el número de pacientes encontrados.

Se debe tener en cuenta que también es posible guardar el código fuente de las acciones a realizar, en un fichero que se puede invocar mediante el comando **awk** directamente. En este caso, la llamada al programa sería del siguiente modo:

```
awk -f <fuente> <fichero>
```

Este tipo de llamada es conveniente si las acciones a realizar van a llevarse a cabo con frecuencia.

5.3. Los algoritmos de similitud entre cadenas

Los algoritmos de similitud realizan la comparación entre pares o conjuntos de cadenas de texto. El interés por comparar cadenas de texto en biomedicina es múltiple. Actualmente, las terminologías médicas (p.e. SNOMED) se están consolidando como nomenclatura estándar de la información médica, sin embargo, las descripciones que las personas entendemos son texto libre, siendo de interés una búsqueda eficiente de términos por una cadena de búsqueda del documentalista del hospital o servicio de salud. La comparación entre cadenas supone además la tecnología básica para inferir relaciones funcionales o evolutivas entre genes y/o proteínas, dando lugar a los sistemas de alineamiento de cadenas en bioinformática.

La programación dinámica es el marco ideal para entender los algoritmos de alineamiento de cadenas. El algoritmo de Needleman-Wunsch obtiene el alineamiento óptimo global entre dos cadenas, mientras que el algoritmo de Smith-Waterman obtiene los alineamientos óptimos entre subcadenas de las dos cadenas de entrada (alineamiento local). Si bien estos algoritmos son óptimos en la resolución de las tareas de alineamiento entre dos cadenas, su aplicación para comparar una cadena con una base de datos de millones de cadenas es computacionalmente

inviable. Esto ha llevado al diseño de algoritmos como BLAST o q-Grams, que no aseguran el alineamiento óptimo, pero obtienen resultados de búsqueda aceptables a coste computacional asumible a partir de la aplicación de heurísticos. La tarea de alineamiento de cadenas puede abordarse entre pares (1 a 1), en la búsqueda en una base de datos (1 a n) o entre múltiples cadenas (n a n). En este texto nos centraremos en las dos primeras, remitiendo a textos del ámbito de la genómica y la filogenia computacional para una explicación de alineamiento de cadenas múltiples.

Utilizaremos para nuestras explicaciones ejemplos sencillos de biología molecular, si bien, los algoritmos y la explicación de los mismos es directamente aplicable a cualquier cadena de símbolos de cualquier alfabeto.

Ejemplo 5.8 (Alineamiento de cadenas de nucleótidos y aminoácidos)
Una de las principales tareas en bioinformática es la comparación de dos cadenas genéticas [99], compuestas por un símbolos de un alfabeto de representación de los nucleótidos en cadenas de ADN (ver ejemplo en el tabla 5.1) o aminoácidos en caso de cadenas de proteínas (ver ejemplo en el tabla 5.2).

Tabla 5.1: Cadena de nucleótidos correspondiente las posiciones 98041..98160 del Cromosoma II de la bacteria Brucella

```
ggcgtgtcgt tccacgtcgg ctcgcagcag acggatctca cggcttggga tcgtgcgctg  98100
gctgacgcgg cagcggtctt ccgcacgctt gccgatgagg gcatcatctt gcgcatggtc  98160
```

Tabla 5.2: Cadena de aminoácidos de la proteína Hemoglobin alpha chain (Human, chimpanzee, and pygmy chimpanzee)

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV
AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

Los diferentes proyectos de secuenciación han proporcionado cadenas de diferentes organismos, cuya función y posible evolución ha sido estudiada por la comunidad científica ámpliamente. La comparación de una cadena recién obtenida con cadenas cuyas funciones son conocidas es uno de los mecanismos habituales de inferencia de conocimiento en bioinformática.

5.3.1. El algoritmo del turista en Manhattan

Antes de abordar los algoritmos de alineamiento de cadenas, el problema del turista en Manhattan nos ayudará a entender la resolución de problemas mediante programación dinámica. Un turista se encuentra en la esquina noroeste de la calle 59 con la 8ª avenida y quiere caminar por Manhattan (ver figura 5.1) hacia el Sur y el Este hasta el edificio Chrysler en la esquina de la calle 42 con la avenida Lexington de tal forma que visite el mayor número de atracciones posibles. Desde una vista cenital, el turista está en la esquina noroeste de Manhattan (vértice origen) y desea llegar a la esquina sureste (vértice destino) andando únicamente en las direcciones Este y Sur. Nuestro turista podrá tomar múltiples caminos, eligiendo andar únicamente hacia el sur o el este en cada manzana, pero estas decisiones le llevarán a visitar más o menos número de atracciones. Aún podemos complicar algo más el problema, asignando cierto valor de interés a cada atracción, como por ejemplo, visitar el Moma puede tener gran interés (10) y visitar Times Square algo menos (7).

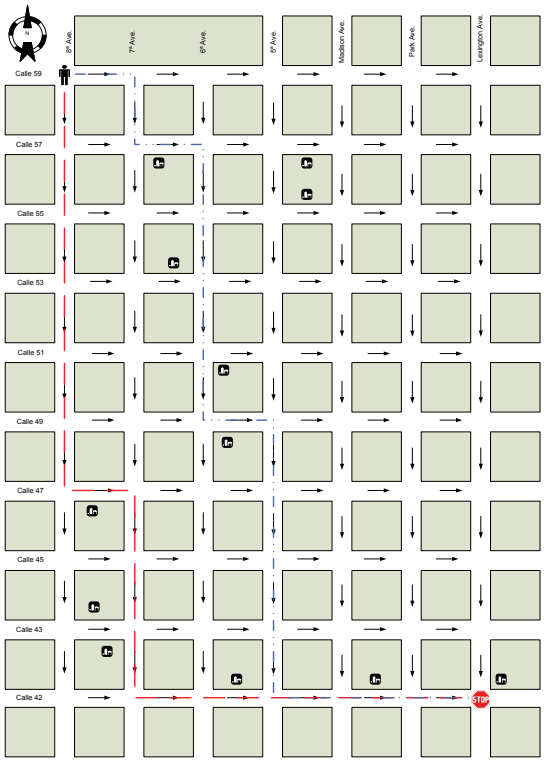


Figura 5.1: Mapa simplificado de Manhattan que el turista quiere recorrer desde la esquina Noroeste hasta la Sureste andando únicamente en dirección y sentido Sur o Este de tal forma que consiga visitar el mayor número de atracciones.

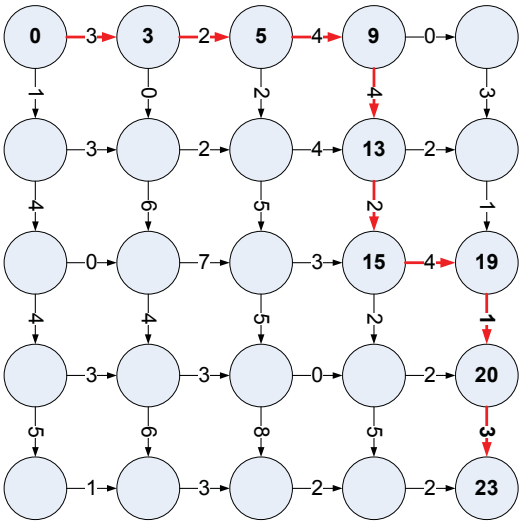


Figura 5.2: Grafo del problema del turista en Manhattan, donde cada vértice es una esquina y cada arista contiene el valor de interés de cada calle. Se ha incluido en cada nodo el interés acumulado al seguir el camino marcado en rojo.

La formulación del objetivo del problema es: encontrar el camino más valioso para ir desde el vértice origen $(0,0)$ al vértice destino (n,m);

La entrada es: la rejilla G con pesos en las aristas ($\vec{w}_{i,j}$) para las aristas que unen el vértice $(i,j-1)$ con (i,j) y $\downarrow w_{i,j}$ para las aristas que unen el vértice $(i-1,j)$ con (i,j)), un vértice origen $(0,0)$ y un vértice destino (n,m);

Y la salida es: el camino más valioso en G desde el vértice origen $(0,0)$ hasta el vértice destino (n,m) con una puntuación de $s_{n,m}$.

La representación del problema como grafo (ver figura 5.2) nos ayudará a plantear el problema de forma esquemática. Cada esquina se representa por un vértice y cada calle en dirección este o sur con una arista con un peso correspondiente al valor de interés de las atracciones de dicha calle.

La solución por fuerza bruta del problema con $n \times m$ esquinas recorrería todos los posibles caminos desde el vértice origen hasta el vértice destino uno detrás de otro, que viene dado por la expresión recursiva:

$$c(n,m) \leftarrow \begin{cases} 1, \text{ si n=1 o m=1} \\ c(n,m-1) + c(n-1,m), \text{ otro caso} \end{cases}$$

Esta solución no parece viable ni siquiera con rejillas pequeñas; por ejemplo, 10 calles y 10 avenidas nos supodría recorrer 48620 posibles caminos. Pensando en el problema como la composición en subproblemas podemos plantear una solución basada en programación dinámica.

En lugar de resolver directamente el problema del turista en Manhattan, es decir, encontrar el camino más valioso desde el vértice $(0,0)$ hasta el vértice (n,m), resolveremos un problema más general: encontrar el camino más valioso (con una puntuación $s_{i,j}$) para ir desde el vértice origen $(0,0)$ hasta cualquier vértice (i,j), donde $0 <= i <= n$ y $0 <= j <= m$. Con este planteamiento, resolveremos $i \times j$ problemas, que mediante la ayuda de la programación dinámica serán de bajo coste. El truco consistirá en ingeniárselas para obtener la solución del problema (i,j) utilizando las soluciones de los problemas anteriores, de tal forma que no repitamos trabajo ya realizado.

En primer lugar asignaremos la puntuación al vértice origen $(0,0)$, en nuestro caso el interés en este es claramente 0, por lo que $s_{0,0} = 0$.

Comenzaremos por los problemas más sencillos, por lo que resolveremos las puntuaciones de los vértices $(0,j)$ (para $0 < j <= m$). Para llegar a estos vértices, el turista no tendrá ninguna flexibilidad en el camino a elegir ya que siempre irá hacia el Este. Por lo tanto, la puntuación del camino $s(0,j)$ será la suma de los primeros j pesos de la fila 0 de nuestra rejilla (ver figura 5.3). Incluso podremos expresar $s(0,j)$ de forma recursiva como la suma de la puntuación del vértice anterior $s_{0,j-1}$ y el valor de la arista que los une $\overrightarrow{w}_{0,j}$.

$$s(0,j) \leftarrow s(0,j-1) + \overrightarrow{w}_{0,j}$$

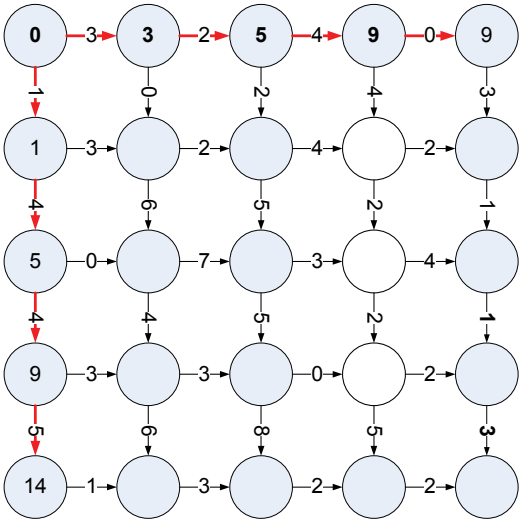


Figura 5.3: Solución de los problemas $(0,j)$ y $(i,0)$.

De forma similar, resolveremos los problemas $(i,0)$, (para $0 < i <= n$). Para llegar a estos vértices, el turista no tendrá ninguna flexibilidad en el camino a elegir ya que siempre irá hacia el Sur. Por lo tanto, la puntuación del camino $s(0,j)$ será la suma de los primeros i pesos de la columna 0 de nuestra rejilla (ver figura 5.3). Incluso podremos expresar $s(i,0)$ de forma recursiva como la suma de la puntuación del vértice anterior $s_{i-1,0}$ y el valor de la arista que los une $\overset{\downarrow}{w}_{i,0}$.

$$s(i,0) \leftarrow s(i-1,0) + \overset{\downarrow}{w}_{i,0}$$

Ejemplo 5.9 (Camino desde $(0,0)$ hasta $(1,1)$ por programación dinámica.) En la expresiones anteriores ya hemos visto un adelanto de la forma de resolver cada problema utilizando la resolución de los problemas anteriores, ¿podríamos ahora resolver el mejor camino desde $(0,0)$ hasta el vértice $(1,1)$ sabiendo $s_{0,1}$ y $s_{1,0}$?

Como en la resolución de los problemas anteriores ya disponíamos de los valores $s_{0,1}$ y $s_{1,0}$ no necesitaremos calcularlos de nuevo. El mejor camino hasta $s_{1,1}$ que pasa por $(0,1)$ tendrá una puntuación de $s_{0,1} + \overset{\downarrow}{w}_{1,1}$ Alternativamente, el mejor camino hasta $s_{1,1}$ que pasa por $(1,0)$ tendrá una puntuación de $s_{1,0} + \overset{\downarrow}{w}_{1,1}$ El mejor camino será aquel que tenga la mayor puntuación posible de los dos que llegan al vértice $(1,1)$, por lo que mediante dos sumas y la función máximo podemos resolver el problema $(1,1)$.

Ejemplo 5.10 (Camino desde $(0,0)$ hasta $(3,4)$ por programación dinámica.) ¿Podríamos ahora resolver el mejor camino desde $(0,0)$ hasta el vértice $(3,4)$ sabiendo $s_{0,1}$, $s_{1,0}$ y $s_{1,1}$?

En este caso deberemos esperar un poco hasta poder resolver el problema $(3,4)$,ya que todavía no sabemos cual es la puntuación de los vértices desde lo que podemos llegar a él. Estos vértices son $(3,3)$ y $(2,4)$.

Ahora llega la hora de resolver los problemas (i,j), (para $0 < i <= n$ y $0 < j <= m$). Para llegar al vértice (i,j), el turista tiene dos posibles caminos. Desde el norte hacia el sur llegamos a (i,j) desde $(i-1,j)$, con una puntuación compuesta por el mejor camino ya calculado $s_{i-1,j} + \overset{\downarrow}{w}_{i,j}$ Desde el oeste hacia el este llegamos a (i,j) desde $(i,j-1)$, con una puntuación compuesta por el mejor camino ya calculado $s_{i,j-1} + \vec{w}_{i,j}$. Como el objetivo es encontrar el camino más valioso, elegiremos aquel cuya suma sea mayor, obteniendo como resultado el camino más valioso entre $(0,0)$ y (i,j). El resultado del problema puede expresarse como una operación con los resultados de los problemas anteriores, por lo que podemos utilizar los resultados anteriores y evitar resolver operaciones ya realizadas anteriormente.

$$s(i,j) \leftarrow max \begin{cases} s_{i,j-1} + \vec{w}_{i,j} \\ s_{i-1,j} + \overset{\downarrow}{w}_{i,j} \end{cases}$$

Vemos ahora, que la resolución del problema del turista en Manhattan no es más que resolver el problema (n,m) del caso general.

Estamos ahora en disposición de escribir el algoritmo 5.1 del turista en Manhattan.

Como vemos en el algoritmo 5.1, los vértices se recorrerán fila a fila (ver figura 5.4(izq)). Esta topología resuelve el problema $(i-1,j)$ y $(i,j-1)$ antes de resolver (i,j), que es la única restricción que debe cumplirse para su correcta resolución por programación dinámica.

Ejemplo 5.11 (Recorridos de vértices mediante programación dinámica.) ¿Existen otras topologías posibles en el recorrido de los vértices (i,j)? Adapta el algoritmo Turista en Manhattan para recorrer los vértices de forma diferente. Otra topología es el recorrido de los vértices columna a columna (ver figura 5.4(centro)), que de forma equivalente a la topología fila a fila cumple el orden necesario.

Otra topología, todavía más interesante, es el recorrido por las antidiagonales (ver figura 5.4(derecha)) sucesivas partiendo de $s_{0,0}$. Un posible código para la resolución mediante esta topología es:

Algoritmo 5.1 TuristaEnManhattan

function = TuristaEnManhattan($\overset{\downarrow}{w}$, $\vec{w}$, n, m)
$s_{0,0} \leftarrow 0$
for i in 1:n **do**
  $s_{i,0} \leftarrow s_{i-1,0} + \overset{\downarrow}{w}_{i,0}$
end for
for j in 1:m **do**
  $s_{0,j} \leftarrow s_{0,j-1} + \vec{w}_{0,j}$
end for
for for i in 1:n **do**
  for for j in 1:m **do**

$$s(i,j) \leftarrow max \begin{cases} s_{i,j-1} + \vec{w}_{i,j} \\ s_{i-1,j} + \overset{\downarrow}{w}_{i,j} \end{cases}$$

    return $s_{n,m}$
  end for
end for

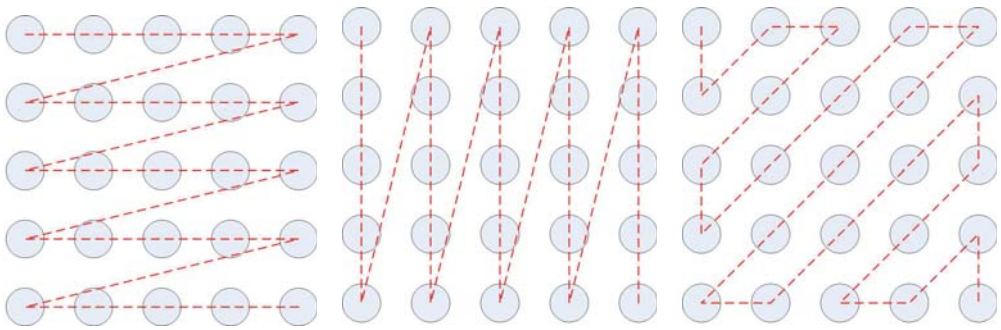


Figura 5.4: Posibles recorridos para resolver el Problema del turista en Manhattan. (izquierda) topología fila a fila, (centro) topología columna a columna, (derecha) topología antidiagonales.

Algoritmo 5.2 TuristaEnManhattanDiag

function TuristaEnManhattanDiag($\overset{\downarrow}{w}$, $\vec{w}$, n, m)
$s_{0,0} \leftarrow 0$
for k in 1:n+m **do**
  for j in 0:k **do**
    $i \leftarrow k - j$

$$s_{i,j} \leftarrow max \begin{cases} s_{i,j-1} + \vec{w}_{i,j} \\ s_{i-1,j} + \overset{\downarrow}{w}_{i,j} \end{cases}$$

    return $s_{n,m}$
  end for
end for

Ejemplo 5.12 (Recuperación del camino óptimo)
Con los algoritmos vistos hasta el momento, hemos conseguido averiguar el valor del mejor camino desde el vértice origen hasta el vértice destino. Sin embargo, nos costaría un buen esfuerzo recuperar cual es ese camino directamente del grafo. ¿Cómo modificarías el algoritmo TuristaEnManhattan para poder recuperar la secuencia de movimientos del turista para recorrer el camino más valioso desde el vértice origen $(0,0)$ al vértice destino (n,m)?

Para poder recuperar el camino más valioso debemos guardar en cada (i,j) un puntero al vértice predecesor desde el que hemos alcanzado el mejor camino.

Algoritmo 5.3 TuristaEnManhattanBackTracking

function = TuristaEnManhattanBackTracking($\overset{\downarrow}{w}$, $\overrightarrow{w}$, n, m)
$s_{0,0} \leftarrow 0$
for i in 1:n **do**
  $s_{i,0} \leftarrow s_{i-1,0} + \overset{\downarrow}{w}_{i,0}$
  $b_{i,0} \leftarrow (i-1,0)$
end for
for j in 1:m **do**
  $s_{0,j} \leftarrow s_{0,j-1} + \overrightarrow{w}_{0,j}$
  $b_{0,j} \leftarrow (0,j-1)$
end for
for i in 1:n **do**
  for j in 1:m **do**

$$s_{i,j} \leftarrow max \begin{cases} s_{i,j-1} + \overrightarrow{w}_{i,j} \\ s_{i-1,j} + \overset{\downarrow}{w}_{i,j} \end{cases}$$

$$b_{i,j} \leftarrow \begin{cases} (i,j-1), if(s_{i,j} = s_{i,j-1} + \overrightarrow{w}_{i,j}) \\ (i-1,j), if(s_{i,j} = s_{i-1,j} + \overset{\downarrow}{w}_{i,j}) \end{cases}$$

    return $(s_{n,m}, b)$
  end for
end for

El algoritmo 5.4 escribe el algoritmo que imprime la secuencia de vértices del camino más valioso.

Algoritmo 5.4 PrintTuristaEnManhattan

function = PrintTuristaEnManhattan(b,(i,j))
if i=0 and j=0 **then**
  exit
end if
PrintTuristaEnManhattan(b,$b_{i,j}$)
$print(i,j)$
PrintTuristaEnManhattan(b,(n,m))

Ejemplo 5.13 (Generalización del algoritmo del turista a 3 aristas.)
Intenta generalizar la resolución del problema del turista en Manhattan a un grafo donde a cada vértice llegan 3 aristas (ver figura 5.5). Por ejemplo, puedes pensar que existen calles en diagonal

que cruzan las manzanas de Manhattan: $(i-1, j)$, $(i-1, j-1)$ y $(i, j-1)$). En este grafo existe una topología que permite un recorrido ordenado, de tal forma que cuando se visita (i, j) todas las aristas que inciden en él, tienen su vértice origen resuelto.

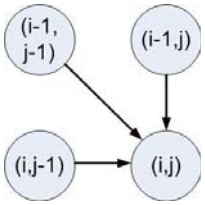


Figura 5.5: Ahora en cada vértice inciden 3 aristas, la resolución del problema es similar al problema con 2 aristas.

¿Cual será la función que resuelva el valor del vértice (i, j)?

Extendiendo el caso general del turista en Manhattan encontramos la solución por programación dinámica:

$$s_{i,j} \leftarrow max \begin{cases} s_{i,j-1} + \overrightarrow{w_{i,j}} \\ s_{i-1,j} + \overset{\downarrow}{w_{i,j}} \\ s_{i-1,j-1} + \overset{\searrow}{w_{i,j}} \end{cases}$$

A través del ejemplo del turista en Manhattan hemos analizado el procedimiento a seguir para la resolución de problemas mediante programación dinámica. En el siguiente apartado plantearemos el problema del alineamiento de cadenas y la distancia de edición, que nos permitirá justificar completamente el papel de la programación dinámica en los problemas biomédicos.

5.3.2. Distancias de edición y alineamiento de cadenas

Dadas dos cadenas u y v, nos preguntamos cual es la similitud o distancia que tienen dichas cadenas entre si. La distancia de Hamming D_g se define como el número de posiciones que difiere la cadena u de la cadena v. D_g de $u = ATATATATATA$ y $v = ATATATATCTA$ es $D_g(u, v) = 1$, que es muy razonable. Sin embargo, si tenemos las cadenas $u' = ATATATATATA$ y $v' = TATATATATAT$ la distancia $D_g(u', v')$ será 11, que parece absolutamente excesiva, ya que un desplazamiento de 1 posición de toda la cadena u', haría corresponder 10 símbolos de los 11 de cada cadena. Así pues, vemos que la distancia de Hamming resulta insuficiente para conseguir una comparación adecuada de cadenas de texto libre o procedentes de secuenciación genética.

En 1966, Vladimir Levenshtein definió la *distancia de edición* entre dos cadenas de un mismo alfabeto, como el mínimo número de operaciones de edición necesarias para transformar la cadena u en v. Siendo las operaciones de edición: la inserción de un símbolo, el borrado de un símbolo y la sustitución de un símbolo por otro.

El alineamiento de la cadena u de longitud $\mid u \mid$ y v de longitud $\mid v \mid$ es una matriz (ver tabla 5.3) de 2 filas y un máximo de $\mid u \mid * \mid v \mid$ columnas, donde cada fila tiene una cadena con los caracteres en el mismo orden, pero no necesariamente adyacentes. Esta matriz representa la disposición óptima de los símbolos según su distancia de edición. En el alineamiento aparecen 4 tipos diferentes de columnas: match, que indica un emparejamiento entre los símbolos; mismatch, que indica una la necesidad de realizar una sustitución ya que no existe emparejamiento de los

símbolos; deleción: indica el borrado del símbolo en la cadena u; e inserción, que indica la inserción de un símbolo en u. Las operaciones de inserción y deleción se denominan conjuntamente indel[c].

Tabla 5.3: alineamiento de la cadena $u = ATGTTATC$ y la cadena $v = ATCGCAC$, m: match; mm: mismatch; d: deleción; i: inserción

u	A	T	-	G	T	T	A	T	C
v	A	T	C	G	C	-	A	-	C
seq	m	m	i	m	mm	d	m	d	m

Ejemplo 5.14 (alineamiento de cadenas de nucleótidos o aminoácidos.)
Las cadenas genéticas es el nombre genérico con el que denominamos tanto a las cadenas de ADN (ver figura 5.1) como a las cadenas de proteínas (ver figura 5.2). Uno de los factores más importantes para decidir qué tipo de cadenas comparar radica en que el parecido entre cadenas de nucleótidos con un origen común se pierde más rápidamente que el parecido en las cadenas de aminoácidos correspondientes.

Esto es debido, por una parte a que el alfabeto es más reducido (cuatro letras frente a veinte) y por otra porque cada proteína puede ser codificada por varios tripletes de nucleótidos. Así pues, la comparación de cadenas de nucleótidos se considera apropiada cuando se comparan cadenas muy parecidas (con uno o dos nucleótidos de diferencia), como en estudios filogenéticos de poblaciones o SNPs. También se utiliza para identificar genes, comparando zonas equivalentes entre genoma de diferentes especies (p.e: ratón vs. humanos) y vemos que las regiones exónicas están más conservadas que las intrónicas. Por supuesto, en el caso de querer comparar cadenas no codificantes será necesario utilizar comparación de cadenas de nucleótidos.

Por otro lado, la comparación de cadenas de aminoácidos se utiliza cuando se buscan homólogos más o menos cercanos, o cuando se desea identificar regiones importantes de las proteínas. En las proteínas, el parecido en la cadena aminoacídica se pierde más lentamente y algunos aminoácidos tienen propiedades más parecidas que otros, por lo que podemos darle más sentido a los cambios que observamos.

El grafo de edición. Una representación útil para esquematizar el proceso de alineamiento de cadenas es el *grafo de edición* (ver figura 5.6). De forma similar al grafo construido en el problema del turista en Manhattan, pero con aristas diagonales, se construye una rejilla de alineamiento, donde cada vértice (i, j) corresponde al alineamiento de los prefijos $u(1, i)$ con $v(1, j)$. Las aristas diagonales representan emparejamientos (matches) los símbolos u_i con v_j o sustituciones del símbolo u_i por el símbolo v_j, las aristas horizontales representan inserciones del símbolo v_j delante del u_i, y las aristas verticales representan deleciones del símbolo u_j. Cada arista tendrá asociada una puntuación que da cuenta del valor de dicha operación de edición. Cada camino entre el vértice $(0, 0)$ y el vértice $(\mid u \mid, \mid v \mid)$ es un posible alineamiento entre las cadenas u y v, y tendrá asociada una puntuación, siendo tarea de los algoritmos de alineamiento descubrir cual de los caminos es el óptimo.

El grafo de edición ofrece una visión clara sobre los posibles alineamientos que pueden establecerse entre las dos cadenas u y v. Desde los alineamientos más drásticos como borrar toda la cadena u e insertar toda la cadena v o viceversa, tenemos todas las combinaciones que dan las posibles sucesiones de operaciones de edición que pueden ser más o menos acertadas.

[c]directamente adaptado del término inglés

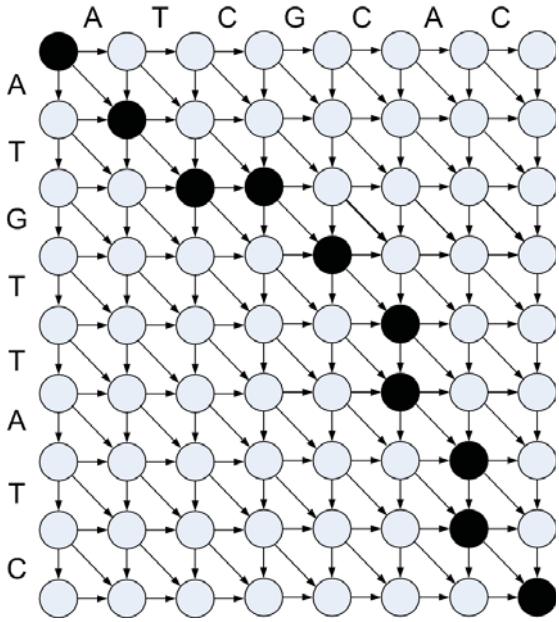


Figura 5.6: En cada vértice del grafo de edición inciden 3 aristas, la arista diagonal es un match o una sustitución, la arista horizontal es una deleción y la arista vertical una inserción.

5.3.3. El valor de las operaciones de edición y las matrices de puntuación

En la elección del camino óptimo que recorra el grafo de edición influye en gran medida el valor de cada operación de edición. Por ejemplo, en el alineamiento de cadenas genéticas el valor asociado a las operaciones de edición está determinado por evidencias biológicas de la operación. En la edición de texto natural, la distancia de edición puede venir determinada por la frecuencia de errores para cada letra que un usuario medio comete al mecanofriar.

Ejemplo 5.15 (Operaciones de edición en nucleótidos y aminoácidos.)
Algunas evidencias utilizadas para generar las matrices de puntuación responden a propiedades químico-físicas de los nucleótidos en cadenas de ADN o de los aminoácidos en proteínas. Se han estudiado puntuaciones basadas en la carga y propiedades estructurales, propiedades estructurales y genéticas y patrones hidrofóbicos.

Otras evidencias responden a tasas de sustitución observadas a través de grandes distancias evolutivas. En la comparación de cadenas de ADN generalmente el valor de la operación de edición depende únicamente de la operación y no de los símbolos implicados. Así pues, la operación tendrá un valor positivo δ en caso de emparejamiento (match), un valor negativo μ en caso de sustitución y un valor negativo σ en caso de inserción o borrado (indel).

En el alineamiento de cadenas protéicas, la mutación de un aminoácido puede o no afectar a la estructura de la proteína. Esto implica que algunas mutaciones son fáciles de encontrar a lo largo de la evolución y otras son realmente difíciles. Asn, Asp, Glu y Ser son los aminoácidos más mutables, mientras que Cys y Trp son los que menores tasas de mutuación tienen. Para hacernos una idea, la probabilidad de mutación de Ser en Phe es 3 veces la probabilidad de mutación de Trp en Phe. Las *matrices de puntuación* expresan las evidencias de conservación y sustitución de un aminoácido por otro a lo largo de periodos evolutivos, así como el valor de las operaciones de borrado e inserción para cada aminoácido. Para diseñar las matrices de puntuación PAM, Dayhoff realizó el alineamiento de cadenas con similitud superior al 90 %. Dayhoff definió PAM1 (Point Accepted Mutation) como la probabilidad de mutación de un residuo durante la cantidad de tiempo aceptada para mutar el 1 % de los aminoácidos de una cadena. Así pues, PAMn (ver tabla 5.4) será la medida de mutación de un residuo durante la cantidad de tiempo aceptada para mutar el n % de los aminoácidos de una cadena. Matrices PAMn con n bajo puede ser utilizada con cadenas con alta similitud, pero para la comparación de cadenas homólogas lejanas, deben utilizarse matrices de mayor tiempo de mutación. Cuanto más lejanas son las cadenas a comparar, mayor n deberá utilizarse (la tabla 5.5 contiene la matriz de puntuación PAM250).

Henikoff y Henikoff en 1992 observaron las limitaciones de las matrices PAM, al ser derivadas exclusivamente de cadenas con alta similitud entre ellas. Así pues, calcularon diferentes matrices de puntuación llamadas BLOSUM (BLOcks SUbstitution Matrix) a partir del alineamiento entre 71 bloques (base de datos BLOCKS) compuestos cada uno de ellos de cadenas altamente relacionadas.

Estas matrices de puntuación son utilizadas por los algoritmos de alineamiento de cadenas para valorar cada camino posible en el grafo de edición, pudiendo obtener diferencias significativas en los alineamientos óptimos obtenidos según las matrices de puntuación[d]. Las diferencias se acentuan en alineamientos entre homólogos lejanos.

[d]El servidor FTP del NCBI dispone de un repositorio de matrices de puntuación accesible de forma anónima en: .

Tabla 5.4: Matriz de puntuación PAM10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	7	-10	-7	-6	-10	-7	-5	-4	-11	-8	-9	-10	-8	-12	-4	-3	-3	-20	-11	-5	-6	-6	-6	-23
R	-10	9	-9	-17	-11	-4	-15	-13	-4	-8	-12	-2	-7	-12	-7	-6	-10	-5	-14	-11	-11	-7	-9	-23
N	-7	-9	9	-1	-17	-7	-5	-6	-2	-8	-10	-4	-15	-12	-9	-2	-5	-11	-7	-12	7	-6	-6	-23
D	-6	-17	-1	8	-21	-6	0	-6	-7	-11	-19	-8	-17	-21	-12	-7	-8	-21	-17	-11	7	-1	-9	-23
C	-10	-11	-17	-21	10	-20	-20	-13	-10	-9	-21	-20	-20	-19	-11	-6	-11	-22	-7	-9	-18	-20	-13	-23
Q	-7	-4	-7	-6	-20	9	-1	-10	-2	-11	-8	-6	-7	-19	-6	-8	-9	-19	-18	-10	-6	7	-8	-23
E	-5	-15	-5	0	-20	-1	8	-7	-9	-8	-13	-7	-10	-20	-9	-7	-9	-23	-11	-10	-1	7	-8	-23
G	-4	-13	-6	-6	-13	-10	-7	7	-13	-17	-14	-10	-12	-12	-10	-4	-10	-21	-20	-9	-6	-8	-8	-23
H	-11	-4	-2	-7	-10	-2	-9	-13	10	-13	-9	-10	-17	-9	-7	-9	-11	-10	-6	-9	-4	-4	-8	-23
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9	-4	-9	-3	-5	-12	-10	-5	-20	-9	-1	-9	-9	-8	-23
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7	-11	-2	-5	-10	-12	-10	-9	-10	-5	-12	-10	-9	-23
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7	-4	-20	-10	-7	-6	-18	-12	-13	-5	-6	-8	-23
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12	-7	-11	-8	-7	-19	-17	-4	-16	-8	-9	-23
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9	-13	-9	-12	-7	-1	-12	-14	-20	-12	-23
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8	-4	-7	-20	-20	-9	-10	-7	-8	-23
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7	-2	-8	-10	-10	-4	-8	-6	-23
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8	-19	-9	-6	-6	-9	-7	-23
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13	-8	-22	-13	-21	-16	-23
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	-10	-9	-13	-11	-23
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8	-11	-10	-8	-23
B	-6	-11	7	7	-18	-6	-1	-6	-4	-9	-12	-5	-16	-14	-10	-4	-6	-13	-9	-11	7	-3	-8	-23
Z	-6	-7	-6	-1	-20	7	7	-8	-4	-9	-10	-6	-8	-20	-7	-8	-9	-21	-13	-10	-3	7	-8	-23
X	-6	-9	-6	-9	-13	-8	-8	-8	-8	-8	-9	-8	-9	-12	-8	-6	-7	-16	-11	-8	-8	-8	-8	-23
*	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	-23	1

Tabla 5.5: Matriz de puntuación PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

5.3.4. Alineamiento global: algoritmo de Needleman-Wunsch

El objetivo del alineamiento global es: **encontrar el mejor alineamiento entre dos cadenas u y v bajo una matriz de puntuación** δ; por lo tanto, la entrada del problema es **la cadena u de longitud** $|u|$ **y v de longitud** $|v|$ compuestas por símbolos del alfabeto Σ, una matriz $(length(\Sigma)+2)\times(length(\Sigma)+2)$ de puntuación δ; y la salida del problema es **un alineamiento de u y v** cuya puntuación $s_{|u|,|v|}$ (según δ) sea máximo entre todos los posibles alineamientos entre u y v.

Resolveremos el problema mediante programación dinámica, de forma similar a como resolvimos el problema del turista en Manhattan. Así pues, iremos resolviendo los alineamientos entre los prefijos $u(1,i)$ y $v(1,j)$, para $1 <= i <= |u|$ y para $1 <= j <= |v|$. El uso de los problemas anteriormente resueltos evitará el cálculo de operaciones efectuadas con anterioridad, lo que resultará una resolución del problema eficiente además de óptima.

El cálculo de la puntuación $s_{i,j}$ del alineamiento óptimo entre los prefijos $u(1,i)$ y $v(1,j)$ es el siguiente:

$$s_{i,j} \leftarrow max \begin{cases} s_{i,j-1} + \delta(-, v_j) (\text{inserción de } v_j) \\ s_{i-1,j} + \delta(u_i, -) (\text{borrado de } u_i) \\ s_{i-1,j-1} + \delta(u_i, v_j) (match, si u_i = v_j, \text{sino sustitución } (u_i, v_j)) \end{cases}$$

donde asumimos que la matriz de puntuación indica similitudes entre pares de símbolos, ya que al utilizar la función max realizará una maximización de la puntuación del alineamiento.

A partir del caso general $s_{i,j}$ y del algoritmo TuristaEnManhattan 5.1, podemos escribir el algoritmo de Needleman-Wunsch para alineamiento global de cadenas.

Algoritmo 5.5 Needleman-Wunsch

```
function = Needleman-Wunsch(u,w,δ)
s_{0,0} ← 0
for i in 1:| u | do
    s_{i,0} ← s_{i-1,0} + δ(u_i, −) (borrado de u_i)
end for
for j in 1:| v | do
    s_{0,j} ← s_{0,j-1} + δ−, v_j (inserción de v_j)
end for
for i in 1:| u | do
    for j in 1:| v | do
        s_{i,j} ← max { s_{i,j-1} + δ(−, v_j), insercion de v_j
                      { s_{i-1,j} + δ(u_i, −)borrado de u_i
                      { s_{i-1,j-1} + δ(u_i, v_j), match, si u_i = v_j, sino sustitucion(u_i, v_j)
        return s_{|u|,|v|}
    end for
end for
```

Ejemplo 5.16 (Alineamientos con matrices de puntuación diferentes)
Aplicaremos el algoritmo de Needleman-Wunsch a las cadenas $u = PAWHEAE$ y $v = HEAGAWGHEE$ con la matriz BLOSUM50 (ver tabla 5.6) y con la matriz PAM250 (ver tabla 5.5)

Tabla 5.6: Matriz de puntuación BLOSUM50

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

La tabla 5.7 presenta el trellis del alíneammiento, además del resultado final obtenido mediante BLOSUM50 es:

Tabla 5.7: Trellis y el alineamiento obtenido mediante BLOSUM50.

	-	H	E	A	G	A	W	G	H	E	E
-	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50
P	-5	-2	-6	-11	-16	-21	-26	-31	-36	-41	-46
A	-10	-7	-3	-1	-6	-11	-16	-21	-26	-31	-36
W	-15	-12	-8	-6	-4	-9	4	-1	-6	-11	-16
H	-20	-5	-10	-10	-8	-6	-1	2	9	4	-1
E	-25	-10	1	-4	-9	-9	-6	-3	4	15	10
A	-30	-15	-4	6	1	-4	-9	-6	-1	10	14
E	-35	-20	-9	1	3	0	-5	-10	-6	5	16

```
7,10 <- 6,9:    16 MoS  E E
6,9 <-  5,9:    10 B    A -
5,9 <-  4,8:    15 MoS  E E
4,8 <-  3,7:     9 MoS  H H
3,7 <-  3,6:    -1 I    - G
3,6 <-  2,5:     4 MoS  W W
2,5 <-  1,4:   -11 MoS  A A
1,4 <-  1,3:   -16 I    - G
1,3 <-  0,2:   -11 MoS  P A
0,2 <-  0,1:   -10 I    - E
0,1 <-  0,0:    -5 I    - H
```

La tabla 5.7 presenta el trellis del alíneammiento, además del resultado final obtenido mediante PAM250 es:

Tabla 5.8: Trellis y el alineamiento obtenido mediante BLOSUM50.

```
      -   H   E   A   G   A   W   G   H   E   E
-     0  -8 -16 -24 -32 -40 -48 -56 -64 -72 -80
P    -8   0  -8 -15 -23 -31 -39 -47 -55 -63 -71
A   -16  -8   0  -6 -14 -21 -29 -37 -45 -53 -61
W   -24 -16  -8  -6 -13 -20  -4 -12 -20 -28 -36
H   -32 -18 -15  -9  -8 -14 -12  -6  -6 -14 -22
E   -40 -26 -14 -15  -9  -8 -16 -12  -5  -2 -10
A   -48 -34 -22 -12 -14  -7 -14 -15 -13  -5  -2
E   -56 -42 -30 -20 -12 -14 -14 -14 -14  -9  -1

7,10 <- 6,9:    -1  MoS   E E
6,9 <-  5,8:    -5  MoS   A E
5,8 <-  4,7:    -5  MoS   E H
4,7 <-  3,6:    -6  MoS   H G
3,6 <-  2,5:    -4  MoS   W W
2,5 <-  1,4:    -21 MoS   A A
1,4 <-  1,3:    -23 I     - G
1,3 <-  0,2:    -15 MoS   P A
0,2 <-  0,1:    -16 I     - E
0,1 <-  0,0:    -8  I     - H
```

5.3.5. Alineamiento local: algoritmo de Smith-Waterman

El objetivo del alineamiento global es encontrar el mejor alineamiento entre cadenas enteras. El hecho que dos cadenas completas sean similares de forma global es muy raro. Por ejemplo, en bioinformática, el alineamiento global solo se da en alineamientos entre miembros de familias de proteínas, estudios filogenéticos de poblaciones o SNPs.

En muchas aplicaciones la puntuación de alineamientos locales entre regiones de las cadenas es mayor que el alineamiento global, esto es debido a que solo esas pequeñas regiones son las conservadas entre las cadenas de la comparación. El gen homeobox (que desempeña un papel crucial en los primeros pasos de los embriones) es muy variable entre especies. Sin embargo, una región del mismo llamada homeodominio está muy conservada entre especies. El alineamiento global de dos cadenas sería totalmente ineficaz en la búsqueda de esta región conservada en dos cadenas que difieren en el resto de símbolos. Esto es debido a que el objetivo del alineamiento global es encontrar el mejor camino entre los vértices (0,0) y (|u|,|v|) del grafo de edición, y no se detiene en pequeñas regiones con gran similitud.

El objetivo del alineamiento local es: **encontrar el mejor camino entre dos vértices arbitrarios** (i, j) **y** (i', j') **del grafo de edición. entre dos cadenas** u **y** v **bajo una matriz de puntuación** δ; por lo tanto, la entrada del problema es **la cadena** u **de longitud** $|u|$ **y** v **de longitud** $|u|$ compuestas por símbolos del alfabeto Σ, una matriz $(card(\Sigma)+1) \times (card(\Sigma)+1)$ de puntuación δ; y la salida del problema es **la puntuación de dos subcadenas de** u **y** v cuyo alineamiento global según δ sea máximo entre todos los alineamientos globales de las subcadenas de u y v un alineamiento de u y v cuya puntuación $s_{|u|,|v|}$ (según δ) sea máximo entre todos los posibles alineamientos entre u y v.

Ejemplo 5.17 (alineamiento local vs. alineamiento global.)
Observa las diferencias entre el alineamiento global y local entre las cadenas u y v.
$u = TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC$.
$v = AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC$.

- alineamiento global

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

- alineamiento local

```
                    tccCAGTTATGTCAGgggacacgagcatgcagagac
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Ejemplo 5.18
Número de alineamientos de pares de subcadenas. ¿Cuántos posibles pares de subcadenas de más de 1 símbolo existen en dos cadenas de 4 y 5 símbolos respectivamente?

El número de subcadenas de longitud mayor que 2 para una cadena de 4 símbolos 7 (3+2+1). y para una cadena de 5 símbolos será 10 (4+3+2+1). Por lo tanto el número de posibles pares de subcadenas es 70 ($7 * 10$).

Por lo tanto, abordar el problema como un alineamiento global de cada par de subcadenas resulta inviable, ya que tendríamos que alínear desde el principio 70 cadenas de longitud variable desde 2 hasta 5 (el máximo de las longitudes de las cadenas).

Una solución adecuada para resolver el problema será incluir aristas directas desde el vértice $(0, 0)$ hasta cada uno de los vértices (i, j) con una puntuación 0 (ver figura 5.7). Esto hará que desde el comienzo de la cadena exista un camino directo a cada vértice, por lo que el prefijo no influirá en el alineamiento de la subcadena.

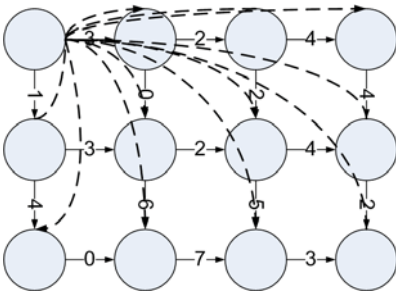


Figura 5.7: Las flechas punteadas representan las aristas directas desde $(0, 0)$ hasta cada vértice.

$$s_{i,j} \leftarrow max \begin{cases} 0 \text{ nuevo alineamiento local} \\ s_{i,j-1} + \delta(-, v_j) (\text{insercion de } v_j) \\ s_{i-1,j} + \delta(u_i, -) (\text{borrado de } u_i) \\ s_{i-1,j-1} + \delta(u_i, v_j) (\text{match, si } u_i = v_j, \text{ sino sustitucion}(u_i, v_j)) \end{cases}$$

La puntuación máxima $s_{i,j}$ sobre el grafo de edición total es el mejor alineamiento local. Podemos observar que el único cambio entre la función del alineamiento global y la nueva para alineamientos locales es la inclusión de una nueva alternativa de valor 0 correspondiente a la arista directa desde $(0, 0)$.

A partir del caso general $s_{i,j}$ podemos escribir el algoritmo 5.6 de Smith-Waterman para el alineamiento local de cadenas.

Algoritmo 5.6 Smith-Waterman

function = Smith-Waterman(u,w,δ)
$s_{0,0} \leftarrow 0$
for i in 1:| u | **do**
  $s_{i,0} \leftarrow 0$
end for
for j in 1:| v | **do**
  $s_{0,j} \leftarrow 0$
end for
for i in 1:| u | **do**
  for j in 1:| v | **do**

$$s_{i,j} \leftarrow max \begin{cases} 0 & \text{nuevo alineamiento local} \\ s_{i,j-1} + \delta(-, v_j), & \text{insercion de } v_j \\ s_{i-1,j} + \delta(u_i, -), & \text{borrado de} u_i \\ s_{i-1,j-1} + \delta(u_i, v_j) & \text{match, si } u_i = v_j, \text{sino sustitucion } (u_i, v_j) \end{cases}$$

    return $s_{|u|,|v|}$
  end for
end for

Ejemplo 5.19
Aplica el algoritmo de Smith-Waterman a las cadenas
$u = TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC$ y
$v = AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC$

utilizando la matriz de edición $\delta = \begin{pmatrix} & A & T & C & G & - \\ A & 1 & -1 & -1 & -1 & -1 \\ T & -1 & 1 & -1 & -1 & -1 \\ C & -1 & -1 & 1 & -1 & -1 \\ G & -1 & -1 & -1 & 1 & -1 \\ - & -1 & -1 & -1 & -1 & 0 \end{pmatrix}$

El mejor alineamiento local obtenido es:

```
15,33 <-        14,32:  12 MoS   G G
14,32 <-        13,31:  11 MoS   A A
13,31 <-        12,30:  10 MoS   C C
12,30 <-        11,29:  9 MoS    T T
11,29 <-        10,28:  8 MoS    G G
10,28 <-        9,27:   7 MoS    T T
9,27 <-         8,26:   6 MoS    A A
8,26 <-         7,25:   5 MoS    T T
7,25 <-         6,24:   4 MoS    T T
6,24 <-         5,23:   3 MoS    G G
5,23 <-         4,22:   2 MoS    A A
4,22 <-         3,21:   1 MoS    C C
3,21 <-         2,20:   0 MoS    C G
2,20 <-         2,19:   1 I      - A
2,19 <-         1,18:   2 MoS    C C
1,18 <-         0,17:   1 MoS    T T
```

5.3.6. Alineamiento con huecos

En algunas tareas de alineamiento de cadenas nos podemos encontrar con que la penalización de borrados (o inserciones sucesivas) no responde a la realidad del problema. Por ejemplo, el borrado de varios nucleótidos en una cadena de ADN es normalmente causado por errores en la replicación del ADN. La naturaleza generalmente borra o inserta subcadenas enteras como unidad, en contraposicón a la deleción o inserción de nucleótidos individualmente conforme realizan los algoritmos anteriormente vistos.

Ejemplo 5.20
En el alineamiento de cadenas de nucleótidos, las deleciones o inserciones de múltiples símbolos se da como fenómeno corriente, por lo que al alínear $u = ATAGC$ con $v = ATATTGC$ es preferible el alineamiento:

```
ATA__GC
ATATTGC
```

al alineamiento

```
ATAG_GC
AT_GTGC
```

aunque el algoritmo visto hasta el momento calcula idénticas puntuaciones en ambos.

Un *gap* en un alineamiento se define como una cadena contigua de espacios en una de las filas. Desde el punto de vista evolutivo, en el que se basa el alineamiento de cadenas genéticas, una penalización de un *gap* de x espacios por $-\sigma * x$ es excesiva y no representa el proceso evolutivo que produjo dicha mutación.

La penalización de huecos contiguos (*affine gap penalties*) se define como la puntuación línealmente ponderada de huecos largos. De esta forma, la puntuación de un *gap* de longitud x será $-(\rho + \sigma x)$, donde ρ es la penalización inicial del hueco y $\sigma > 0$ es la penalización por cada símbolo en el hueco. Generalmente ρ será un valor grande respecto a σ.

Esta nueva propuesta de puntuación en las deleciones e inserciones modifica el cálculo de la puntuación $s_{i,j}$ del alineamiento entre el prefijo $u(0,i)$ y $v(0,j)$, no pudiendo únicamente depender de los valores $s_{i-1,j}$, $s_{i,j-1}$ y $s_{i-1,j-1}$ de las funciones vistas hasta el momento. Incorporando aristas horizontales de larga duración desde los vértices $(i, j-x)$ hasta (i,j) con puntuación $-(\rho + \sigma x)$ solucionamos la penalización de huecos contiguos producidos por deleciones. Incorporando aristas verticales de larga duración desde los vértices $(i-x, j)$ hasta (i,j) con puntuación $-(\rho + \sigma x)$ solucionamos la penalización de huecos contiguos producidos por inserciones.

Una implementación directa de las funciones de puntuación con estas nuevas fórmulas incrementarían el coste del algoritmo, ya que cada nodo depende de todos los antecesores de su columna, todos los antecesores de su fila y del elemento anterior en su diagonal. Sin embargo, la función de penalización de huecos contiguos $-(\rho + \sigma x)$ pueden descomponerse para realizar su cálculo mediante la siguiente fórmula recursiva:

$$\overset{\downarrow}{s}_{i,j} \leftarrow max \begin{cases} \overset{\downarrow}{s}_{i-1,j} - \sigma \text{ (continuar gap)} \\ s_{i-1,j} - (\rho + \sigma) \text{(iniciar gap)} \end{cases}$$

$$\vec{s}_{i,j} \leftarrow max \begin{cases} \vec{s}_{i,j-1} - \sigma \text{ (continuar gap)} \\ s_{i,j-1} - (\rho + \sigma) \text{(iniciar gap)} \end{cases}$$

$$s_{i,j} \leftarrow max \begin{cases} s_{i-1,j-1} + \delta(u_i, v_j)(\text{match, si} u_i = v_j(\text{, sino sustitución}(u_i, v_j) \\ \text{(match o sustitución)} \\ \overset{\downarrow}{s}_{i,j} \text{ (deleción: puede ser hueco)} \\ \vec{s}_{i,j} \text{ (inserción: puede ser hueco)} \end{cases}$$

donde $\overset{\downarrow}{s}_{i,j}$ contiene los alineamientos entre el prefijo $u(0,i)$ y el prefijo $v(0,j)$ que finaliza con una deleción (posible *gap* en u). De forma similar, $\vec{s}_{i,j}$ contiene los alineamientos entre el prefijo $u(0,i)$ y el prefijo $v(0,j)$ que finaliza con una inserción (posible *gap* en v). En un grafo de edición, $\overset{\downarrow}{s}_{i,j}$ y $\vec{s}_{i,j}$ son las puntuaciones de los caminos óptimos que llegan al vértice (i,j) mediante aristas verticales u horizontales.

Por medio de la nueva fórmula, es posible resolver el problema con el mismo coste temporal que el algoritmo original, a costa de multiplicar por 3 el coste espacial.

Ejemplo 5.21 (Comparación de alineamiento con y sin huecos.)
El alineamiento de las cadenas $u =$ "$AAACGCTCGGGAA$" y $v =$ "$AAAGGAA$" mediante el algoritmo Needleman-Wunsch utilizando la matriz de puntuación $\delta =$

$$\begin{pmatrix} & A & T & C & G & - \\ A & 1 & -1 & -1 & -1 & -1 \\ T & -1 & 1 & -1 & -1 & -1 \\ C & -1 & -1 & 1 & -1 & -1 \\ G & -1 & -1 & -1 & 1 & -1 \\ - & -1 & -1 & -1 & -1 & 0 \end{pmatrix}$$

produce el alineamiento:

```
    -   A  A  A  G  G  A  A
-   0  -1 -2 -3 -4 -5 -6 -7
A  -1   1  0 -1 -2 -3 -4 -5
A  -2   0  2  1  0 -1 -2 -3
A  -3  -1  1  3  2  1  0 -1
C  -4  -2  0  2  2  1  0 -1
G  -5  -3 -1  1  3  3  2  1
C  -6  -4 -2  0  2  2  2  1
T  -7  -5 -3 -1  1  1  1  1
C  -8  -6 -4 -2  0  0  0  0
G  -9  -7 -5 -3 -1  1  0 -1
G -10  -8 -6 -4 -2  0  0 -1
G -11  -9 -7 -5 -3 -1 -1 -1
A -12 -10 -8 -6 -4 -2  0  0
A -13 -11 -9 -7 -5 -3 -1  1


13,7 <-         12,6:   1 MoS    A A
12,6 <-         11,5:   0 MoS    A A
11,5 <-         10,5:  -1 B      G -
10,5 <-         9,5:    0 B      G -
9,5  <-         8,4:    1 MoS    G G
8,4  <-         7,4:    0 B      C -
7,4  <-         6,4:    1 B      T -
6,4  <-         5,4:    2 B      C -
5,4  <-         4,3:    3 MoS    G G
4,3  <-         3,3:    2 B      C -
3,3  <-         2,2:    3 MoS    A A
2,2  <-         1,1:    2 MoS    A A
1,1  <-         0,0:    1 MoS    A A
```

donde vemos que el alineamiento producido realiza borrados ((11,5), (10,5)) , ((8,4), (7,4), (6,4)) y ((4,3)); alternados con matches ((9,5)) y ((4,3)).

El alineamiento del algoritmo alineamiento con huecos con los parámetros $\delta(\text{"}i\text{"},\text{"}i\text{"}) = 5$ (match), $\delta(\text{"}i\text{"},\text{"}j\text{"}) = -5$ (sustitución), $\rho = -15$ (comienzo de gap), $\sigma = -1$ (factor sumatorio de continuación de gap) resuelve el alineamiento de la siguiente forma:

```
"Trellis Borrado"
    -   A   A   A   G   G   A   A
- -15 -31 -32 -33 -34 -35 -36 -37
A -16 -32 -33 -34 -35 -36 -37 -38
A -17 -11 -27 -28 -29 -30 -31 -32
A -18 -12  -6 -22 -23 -24 -25 -26
C -19 -13  -7  -1 -17 -18 -19 -20
G -20 -14  -8  -2  -6 -19 -20 -21
C -21 -15  -9  -3  -7  -1 -17 -18
T -22 -16 -10  -4  -8  -2  -6 -19
C -23 -17 -11  -5  -9  -3  -7 -11
G -24 -18 -12  -6 -10  -4  -8 -12
G -25 -19 -13  -7 -11  -5  -9 -13
G -26 -20 -14  -8 -12  -6 -10 -14
A -27 -21 -15  -9 -13  -7 -11 -15
A -28 -22 -16 -10 -14  -8  -7 -11

"Trellis Insercion"
    -   A   A   A   G   G   A   A
- -15 -16 -17 -18 -19 -20 -21 -22
A -31 -32 -11 -12 -13 -14 -15 -16
A -32 -33 -27  -6  -7  -8  -9 -10
A -33 -34 -28 -22  -1  -2  -3  -4
C -34 -35 -29 -23 -17  -6  -7  -8
G -35 -36 -30 -24 -18 -12  -1  -2
C -36 -37 -31 -25 -19 -20 -17  -6
T -37 -38 -32 -26 -20 -21 -18 -19
C -38 -39 -33 -27 -21 -22 -19 -20
G -39 -40 -34 -28 -22 -16 -17 -18
G -40 -41 -35 -29 -23 -17 -11 -12
G -41 -42 -36 -30 -24 -18 -12 -13
A -42 -43 -37 -31 -25 -26 -23  -7
A -43 -44 -38 -32 -26 -27 -24 -18

"Trellis Sustitucion"
     -    A    A    A    G    G    A    A
-    0 -Inf -Inf -Inf -Inf -Inf -Inf -Inf
A -Inf    5  -11  -12  -13  -14  -15  -16
A -Inf  -11   10   -6   -7   -8   -9  -10
A -Inf  -12   -6   15   -1   -2   -3   -4
C -Inf  -13   -7   -1   10   -6   -7   -8
G -Inf  -14   -8   -2    4   15   -1   -2
C -Inf  -15   -9   -3   -7   -1   10   -6
T -Inf  -16  -10   -4   -8   -2   -6    5
C -Inf  -17  -11   -5   -9   -3   -7  -11
G -Inf  -18  -12   -6    0   -4   -8  -12
G -Inf  -19  -13   -7   -1    5   -9  -12
G -Inf  -20  -14   -8   -2    4    0  -13
A -Inf  -21  -15   -9  -13   -7    9    5
A -Inf  -22  -16  -10  -14   -8   -2   14

13,7 <-         12,6:   14 MoS   A A
12,6 <-         11,5:    9 MoS   A A
11,5 <-         10,4:    4 MoS   G G
10,4 <-         9,3:    -1 MoS   G G
9,3 <-          8,3:    -6 B     G -
8,3 <-          7,3:    -5 B     C -
7,3 <-          6,3:    -4 B     T -
6,3 <-          5,3:    -3 B     C -
5,3 <-          4,3:    -2 B     G -
4,3 <-          3,3:    -1 B     C -
3,3 <-          2,2:    15 MoS   A A
2,2 <-          1,1:    10 MoS   A A
1,1 <-          0,0:     5 MoS   A A
```


En estos resultados observamos un único gap en la parte central de alineamiento, ya que favorece los huecos largos a los cortos gracias a la menor penalización en caso de la continuidad del hueco respecto a la penalización de la iniciación de un nuevo hueco tras un match o sustitución.

5.3.7. El algoritmo BLAST

Una vez hemos revisado los métodos del programación dinámica,estudiaremos BLAST, algoritmo utilizado por la herramienta homónima ampliamente utilizada en biología molecular.

Los algoritmos de programación dinámica obtienen la solución óptima de los problemas de alineamiento de cadenas con un coste temporal $O(n \times m)$, donde n y m son las longitudes de las cadenas a alínear. Este coste resulta prohibitivo cuando se trata de comparar una cadenas con toda una base de datos. Por ejemplo, una cadena de 1000 símbolos, comparada con una base de datos que contiene 100 millones de símbolos (10^8) requiere un total de 10^{11} comparaciones, que, suponiendo una capacidad de cálculo de 10^7 comparaciones/segundo, tardaría 17 minutos en finalizar. Es por ello que los algoritmos de alineamiento de cadenas dedicados a la búsqueda de cadenas similares en bases de datos realizan alineamientos heurísticos que encuentran buenas soluciones pero sin poder asegurar el alineamiento óptimo.

BLAST es la herramienta de búsqueda en bases de datos dominante en la biología molecular. Su primera versión [100] fue desarrollada por Alschul, Gish, Miller, Myers y Lipman en 1990, motivados por el convencimiento de poder acelerar la búsqueda de cadenas eligiendo menos, pero mejores, puntos calientes de las cadenas durante una primera fase de ventanado. En 1997, Altschul, Madden, Schaffer, Zhang, Zhang, Miller y Lipman, publicaron [101] la nueva versión gapped-BLAST. que incorporaba el alineamiento con *gaps* y aceleraba la búsqueda en un factor 3 respecto al BLAST original.

Desde el mismo servidor del NCBI definen BLAST como "La herramienta de alineamiento local básico que busca regiones de similitud local entre cadenas. El programa compara cadenas de nucleótidos o proteínas con bases de datos de cadenas y calcula la significancia estadística de los emparejamientos. BLAST puede ser usado para inferir relaciones funcionales o evolutivas entre cadenas, así como ayudar a identificar miembros de familias genéticas". Sin embargo, cabe destacar que el algoritmo BLAST puede configurarse de tal modo que su aplicación a otras aplicaciones es perfectamente viable.

Necesitaremos algunas definiciones para entender el funcionamiento del algoritmo BLAST:

- Hit: subcadenas de longitud w de la base de datos que se alínean con subcadenas de la cadena problema con una puntuación mayor que supera el umbral t
- Par de segmentos (Segment Pair): par de subcadenas de la misma longitud que forman un alineamiento sin GAPs
- Par de segmentos localmente máximo: par de segmentos cuya puntuación de alineamiento no puede ser mejorado por extensión o acortamiento
- Par de segmentos máximo: par de segmentos con la máxima puntuación sobre todos los segmentos de las cadenas
- HSP (High-Scoring Segment Pairs): pares de segmentos con puntuación de alineamiento mayor que un umbral s

El procedimiento básico de la herramienta BLAST consta de los siguientes pasos:

1. Alínear la cadena problema con todas las cadenas de la base de datos

2. Establecer un rango de similitud basado en las puntuaciones de alineamiento
3. Mostrar los HSP (High-Scoring Segment Pairs)

A continuación veremos el algoritmo BLAST, tanto en su versión original (1990) como versión gapped BLAST(1997).

1. Elegir los parámetros de longitud w y umbral t.
   - El umbral t se utilizara para la selección de hits, un aumento de t hará que el algoritmo funcione de forma más rápida; una t más baja aumentará la habilidad de detectar relaciones distantes (con un efecto de aumento de ruido).
   - w establece la longitud de los hits. Para la comparación de proteínas w generalmente está entre 3 y 5. Para la comparación de ADN w tendrá un valor entre 11 y 12 nucleótidos.
2. Hacer una lista de todas las subcadenas de la cadena problema de longitud w. Antes de cualquier búsqueda en la base de datos.
3. Para cada subcadena de la cadena problema,
   a) evaluar la puntuación del alineamiento con el emparejamiento exacto utilizando una matriz de puntuación (generalmente BLOSUM62), incluir las subcadenas en la lista SW.
   Para w=3:
   P Q G
   P Q G
   $7 + 5 + 6 = 18$
   b) evaluar la puntuación del alineamiento con emparejamiento muy cercanos que superen el umbral t; incluir las subcadenas cercanas en la lista SW.
   Por ejemplo, si $w = 3$, habrá $20^w = 8000$ posibles subcadenas de longitud 3; sim embargo, solo unas 50 tienen una puntuación superior al umbral t, las otras directamente se no las consideramos.
4. Para toda subcadena sw de la lista SW, buscar los emparejamientos (HIT) exactos que existan en la base de datos mediante un árbol de búsqueda[e] Es importante la utilización de un método rápido para este paso, ya que utiliza todas las subcadenas de longitud w de la base de datos.
   a) Versión BLAST original:
   Extender el alineamiento local de cada HIT hacia ambos lados sin deleciones ni inserciones, hasta alcanzar el Par de segmentos localmente máximo. La extensión en cada dirección termina cuando la puntuación cae por debajo de cierta distancia de la mejor extensión obtenida por una extensión más corta. Una vez finalizado, se debe comprobar si su puntuación es mayor que s.
   En el artículo original [100] los autores comentan, también, una implementación de la extensión mediante programación dinámica que permitía inserciones y deleciones, que mejora la sensibilidad[f], siempre a costa de perder selectividad y sobretodo el aumento del coste temporal.

[e] Un árbol de búsqueda interesante puede ser el Árbol de Palabras Clave propuesto por Aho-Corasick ([102])

[f] habilidad para encontrar cadenas lejanas

b) Versión Gapped BLAST:

1) Método "2-hits": realizar una extensión sin deleciones ni inserciones entre 2 HITS que se encuentre en la misma diagonal dentro de una distancia A el uno del otro. El objetivo es reducir el número de extensiones a realizar, esto disminuye la sensibilidad de la búsqueda, por lo que se utiliza una t de 11 en lugar de 13. Como concadena, se tendrá un mayor número de HITS, pero solo se extenderán aquellos que se encuentra cercanos, por lo que habrá menor número de regiones extendidas.
2) Extender las regiones de alta similitud por programación dinámica con huecos. Si la puntuación de la región supera el umbral s_g, entonces alínear mediante programación dinámica comenzando por el centro de la región. s_g se elige de forma que 1/50 de las cadenas sean comparadas.

Para valorar si un alineamiento constituye una evidencia de una homología, es bueno saber como de fuerte puede ser esperado un alineamiento por casualidad (por azar) únicamente. En el contexto de la comparación de cadenas homólogas, casualidad puede interpretarse mediante tres significados: (i) la comparación de cadenas reales pero no homólogas; (ii) la comparación de cadenas reales que han sido barajadas para preservar las propiedades de su composición; (iii) comparación de cadenas que son generadas aleatoriamente basadas en un modelos de cadenas de ADN o proteína. Los resultados estadísticos obtenidos analíticamente usan la definicón (iii) mientras que los empíricos o simulaciones pueden utilizar cualquier definición.

Ejemplo 5.22 (La herramienta BLAST)
BLAST puede ser utilizado online desde múltiple servidores, p.e desde el NCBI[g]. También puede descargarse libremente[h] para su instalación en un servidor local en el cual podemos incorporar bases de datos estándar réplica de las disponibles públicamente o crear nuestras propias bases de datos de propósito particular.

La tabla 5.9 muestra un ejemplo de utilización de BLAST desde línea de comandos para buscar homólogos de una cadena de Hemoglobina en la base de datos Swissprot.

Existen varias versiones de la herramienta BLAST, que realizan la búsqueda de cadenas de nucleótidos y aminoácidos en bases de datos. Según el tipo de búsqueda que se desee realizar se debe utilizar una versión de la herramienta :

- blastn: Compara una cadena de nucleótidos frente a una base de cadenas de nucleotidos.
- blastp: Compara una cadena de aminoácidos frente a una base de cadenas de proteínas.
- blastx: Compara una cadena de nucleótidos traducida en todas las fases de lectura abierta frente a una base de cadenas de proteínas. Esta opción se utiliza para encontrar productos potenciales de la traducción de una cadena de nucleotidos desconocida.
- tblastn: Compara una cadena de aminoácidos frente a una base de cadenas de nucleótidos traducida dinamicamente en todas las fases de lectura abiertas.
- tblastx: Compara una cadena de nucleótidos traducida en las 6 fases de lectura abierta frente a la traducción de las 6 fases de lectura abierta de una base de cadenas de nucleótidos. No se puede usar on line con la base de datos nr por ser computacionalmente intensivo.

[g] `http://www.ncbi.nlm.nih.gov/BLAST/`
[h] `ftp://ftp.ncbi.nih.gov/blast`

Tabla 5.9: Ejemplo de utilización de BLAST desde línea de comandos para buscar homólogos de una cadena de Hemoglobina en la base de datos Swissprot

```
cat > hahu.aa
>HAHU | 1114 |  Hemoglobin alpha chain - Human, chimpanzee, and pygmy chimpanzee
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV
AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

../bin/blastall -p blastp -i hahu.aa -d swissprot > hahu.blast

more hahu.blast

BLASTP 2.2.13 [Nov-27-2005]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= HAHU | 1114 |  Hemoglobin alpha chain - Human, chimpanzee, and
pygmy chimpanzee
         (141 letters)

Database: swissprot
           186,234 sequences; 69,023,086 total letters
Searching.....................................................done
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value
gi|57013850|sp|P69905|HBA_HUMAN Hemoglobin alpha subunit (Hemogl...   285   3e-77
gi|122407|sp|P01923|HBA_GORGO Hemoglobin alpha subunit (Hemoglob...   283   7e-77
gi|14194806|sp|Q9TS35|HBA1_HYLLA Hemoglobin alpha-1 subunit (Hem...   281   5e-76
gi|122467|sp|P01924|HBA_PREEN Hemoglobin alpha subunit (Hemoglob...   280   1e-75
gi|122466|sp|P06635|HBA_PONPY Hemoglobin alpha subunit (Hemoglob...   279   1e-75

...

>gi|57013850|sp|P69905|HBA_HUMAN Hemoglobin alpha subunit
           (Hemoglobin alpha chain) (Alpha-globin)
 Hemoglobin alpha subunit (Hemoglobin alpha chain) (Alpha-globin)
 Hemoglobin alpha subunit (Hemoglobin alpha chain) (Alpha-globin)
          Length = 142
 Score =  285 bits (728), Expect = 3e-77
 Identities = 141/141 (100%), Positives = 141/141 (100%)
Query: 1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK 60
           VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
Sbjct: 2   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK 61

Query: 61  KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA 120
           KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
Sbjct: 62  KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA 121

Query: 121 VHASLDKFLASVSTVLTSKYR 141
           VHASLDKFLASVSTVLTSKYR
Sbjct: 122 VHASLDKFLASVSTVLTSKYR 142

...
```


Tabla 5.10: Versiones de la herramienta BLAST. Según el alfabeto de la cadena problema, de la base de datos y del alfabeto que se desee utilizar para realizar la comparación, utilizaremos diferentes versiones. aa: aminoácidos, nt: nucleótidos

Programa	Seq problema	Base de Datos	Comparación
blastp	aa	aa	aa
blastn	nt	nt	nt
blastx	nt	aa	aa
tblastn	aa	nt	aa
tblastx	nt	nt	aa

Capítulo 6

Procesamiento de señales biomédicas

En este capítulo estudiaremos técnicas básicas de procesamiento de señales biomédicas utilizadas típicamente para el uso médico. Seguiremos una aproximación clásica, basada en el texto de referencia sobre análisis de señales biomédicas escrito por Rangayyan [103]. Así pues, veremos las propiedades básicas de las señales, seguida de una introducción a filtros digitales para la eliminación de artefactos de la señal y métodos para la detección de eventos. A continuación, veremos una introducción al análisis espectral y para finalizar analizaremos señales no-estacionarias. Una señal determinista es aquella cuyo valor en cada instante de tiempo puede ser calculada mediante una ecuación con solución en forma cerrada en función del tiempo. Por lo contrario, las señales que no cumplen esta condición se denominan señales aleatorias.

En la última parte del capítulo, estudiaremos la voz humana, con el fin de utilizarlo en problemas de foniatría o psicología.

6.1. Propiedades de las señales

Una señal estocástica es un proceso aleatorio (o estocástico) que se caracteriza por la observación de valores de una (o varias) variable aleatoria en función de otra variable (generalmente el tiempo). Por ser variables aleatorias responden a la teoría de la probabilidad, estudiada en la sección 3.1, por lo que vienen caracterizadas por sus distribuciones de probabilidad 3.7.

Las medidas estadística de las variables aleatorias de los procesos estocásticos tienen sentido físico. Así pues, el valor medio (i.e. 3.10) es la componente continua de la señal (CC), y la media cuadrática $E[\eta^2] = \int_{-\infty}^{\infty} p(\eta)\eta^2 d\eta$ es la potencia media de la señal. También es de utilidad conocer la varianza σ^2 (ecuación 3.15) del proceso estocástico η.

Como casi siempre pasa con las variables aleatorias, será raro conocer la distribución de probabilidad de los procesos estocásticos, por lo que será común estimar las medidas estadísticas a través de observaciones del proceso que originan las señales. Supongamos que disponemos de M observaciones de un proceso aleatorio x con función de densidad de probabilidad $p(x)$ como M funciones del tiempo $x_1(t), x_2(t), \ldots, x_M(t)$.

La media en cada instante específico t_i puede ser calculada mediante promediado muestral de forma similar al cálculo del valor esperado empírico 3.11:

$$\bar{x}(t_i) = \lim_{M\to\infty} \frac{1}{M} \sum_{n=1}^{M} x(t_i),$$

siendo el prototipo del proceso x una función media que está compuesta por la media muestral en cada instante:

$$\bar{x}(t) = \lim_{M\to\infty} \frac{1}{M} \sum_{n=1}^{M} x(t),$$

De forma general, si elevamos x al exponente k obtendremos de forma similar el estadístico de orden k.

La función de autocorrelación (ACF, autocorrelation function) se define como el valor estimado de la multiplicación de dos valores de $x(t)$ separados un retardo de tiempo τ:

$$\Phi_{xx}(t_i, t_i + \tau) = \int_{-\infty}^{\infty} x(t_i)x(t_i + \tau p(x)dx,$$

que podemos estimar de forma empírica como

$$\hat{\Phi}_{xx}(t_i, t_i + \tau) = \sum_{i=1}^{M} x_i(t_i)x_i(t_i + \tau).$$

La autocorrelación indica la relación estadística entre los valores de x en dos instantes de tiempo.

6.2. Transformaciones y propiedades temporales de señales

Decimos que una transformación sobre una señal es una operación que intenta enfatizar una información sobre el proceso estocástico que la genera que no es fácilmente observable con la señal original. Muchas transformaciones se obtienen mediante la comparación de la señal de interés con una función patrón sumada (o integrada) a lo largo del tiempo:

$$\hat{x}(m) = \int_{-\infty}^{\infty} x(t)f_m(t)dt,$$

donde $f_m(t)$ suele ser una familia de funciones similares (llamada base). Un ejemplo de transformación es la transformada de Fourier, donde

$$\hat{x}(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi tf}dt, \forall f,$$

siendo f el conjunto de frecuencias.

Generalmente, la señal viene expresada en tiempo discreto como $x(n) = x(nT) = [x(1), \ldots, x(N)]$, donde n es el índice del vector de valores que toma la señal muestreada a intervalos, y Δt representa el intervalo entre dos muestreos consecutivos (por lo que $f_s = 1/\Delta t$ es la frecuencia de muestreo). Si la señal puede tomar valores cuantizados de un rango, entonces se denomina señal digital.

Para operar con señales discretas, la integral se convierte en el sumatorio sobre un rango finito[a]:

$$\hat{x}(m) = \sum_{n=1}^{N} x(n)f_m(n),$$

sobre el que se puede aplicar además una función de ventanado para limitar la operación a un espacio determinado, lo que se expresa como

$$\hat{x}(m) = \sum_{n=1}^{N} x(n)f_m(n)W(n),$$

[a] o un producto escalar $x(n) \cdot f_m(n)$

pudiendo ser $W(n)$ una ventana rectangular entorno a n, o una triangular, de Hamming, de Hanning, etc.

La operación convolución sobre dos señales $x(t)$ y $h(t)$ se define como la superposición de una de ellas y una versión trasladada e invertida de la otra,

$$x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau,$$

y de forma discreta

$$x(n) * h(n) = \sum_{k=1}^{N} x(k)h(n-k).$$

Ejemplo 6.1 (Convolución de dos funciones simples)
Comprueba que en la implementación en matlab el resultado de la convolución de x y h se obtiene como $x(n) * h(n) = \sum_{k=1}^{N} x(k)h(n+1-k)$.

```
x=[0,1,3,4,5,6,7];
h=[1,3,2];
conv(x,h)
ans =
     0     1     6    15    23    29    35    33    14
```

La correlación cruzada identifica la similitud entre el valor de dos señales en diferentes instantes, para ello desplaza una señal respecto a la otra. Al igual que la correlación, el valor 1 indica una relación lineal positiva, el -1 indica una relación lineal negativa, y 0 indica que no hay correlación.

$$\theta_{xh}(t) = \int_{0}^{T} x(t)h(t+\tau)d\tau,$$

o de forma discreta

$$\theta_{xh}(n) = \sum_{k=0}^{N} x(n)h(n+k),$$

que se convierte en el estadístico temporal autocorrelación Φ_{xx} si ambas señales son la misma.

La función de correlación asume que las funciones están normalizadas, si no es así, hablaremos de covarianza.

Ejemplo 6.2 (Correlación cruzada de dos funciones senoidales)
Buscamos mediante correlación cruzada del retraso entre las señales senoidales x y $x2$ de periodo $10s$ digitalizadas con frecuencia de muestreo $f_s = 250Hz$.

```
sampling_frequency=250; %Hz
sampling_interval=1/sampling_frequency;
time_length=5*60; %in secs. 5 min
time=0:sampling_interval:(time_length-sampling_interval);
x=sin(2*pi/10*time);
x2=sin(2*pi/10*time-pi/2);
%el desfase es de desfase de 45º (1/4 de periodo)
xxx2=xcorr(x,x2,'coeff');
figure
subplot(2,1,1);
plot(time,x);
hold on;
```

```
plot(time, x2,'r');
subplot(2,1,2);
timexxx2 = -time(size(time,2)):sampling_interval:time(size(time,2));
plot(timexxx2,xxx2);
[maxxxx2,imaxxxx2]=max(xxx2);
timexxx2(imaxxxx2)
ans =
   -2.4920
```

Por lo tanto, la señal $x2$ está retrasada $2{,}5s$ respecto a $x1$.

De forma similar, podríamos calcular la media de un rango de tiempo de la señal, mediante la ponderación temporal de la señal a lo largo del rango.

6.3. Procesos estacionarios y no estacionarios

1. Proceso estacionario en el sentido estricto (stacionary in the strict sence/strongly staciony): un proceso es estacionario (de orden k, p.e. $k = 1, 2$) si el promediado muestral no depende del tiempo (por lo tanto es constante).

2. Proceso estacionario en el sentido amplio (stacionary in the wide sense/weakly stacionary): un proceso es estacionario en sentido amplio si la media y su autocorrelación no varían con el tiempo.

3. Proceso estacionario y ergódico: un proceso estacionario es ergódico si sus estadísticos temporales son independientes de la muestra i que se elija.

Generalmente, se suele asumir que las señales biomédicas son observaciones de un proceso estacionario y ergódico. Resumiendo, un proceso estacionario y ergódico cumple que: 1) la media temporal calculada con cualquier muestra i es igual al valor medio de la señal, 2) el valor de autocorrelación solo depende del intervalo τ (o k en discreto) y no de la posición donde se calcule, y además es independiente de la muestra i elegida.

6.4. Ruido blanco, ruido estructurado e interferencia fisiológica

Generalmente, la observación de la señal biomédica $x(t)$, fruto de la realización de un proceso estocástico x con densidad de probabilidad $p(x)$, se realiza en un entorno con ruido aleatorio $\epsilon(t)$, por lo que la señal adquirida $y(t)$ puede considerarse la realización de otro proceso estocástico y, caracterizable por una distribución de probabilidad $p(y)$. Lo más común es que el ruido se sume a la señal biomédica (ruido aditivo), por lo que $y(t) = x(t) + \epsilon(t)$.

La relación señal/ruido (SNR, signal to noise ratio) se calcula como la división entre la potencia de la señal y la potencia del ruido. De forma alternativa el SNR también se define como el ratio entre el rango de amplitud pico a pico y la potencia del ruido.

El ruido blanco es una interferencia procedente de un proceso estocástico como el ruido térmico de un dispositivo electrónico. Viene caracterizado por una densidad de probabilidad normal de media cero y desviación estándar σ, por lo que la ponderación muestral con N muestras del mismo proceso disminuye el ruido blanco un factor de $\sqrt{N}$.

El ruido estructurado es aquel que posee un patrón preestablecido, y por lo tanto puede predecirse. Por ejemplo, el ruido estructurado típico superpuesto a señales biomédicas es la

interferencia de 50 o 60 Hz producida por la red eléctrica, que constituye un ruido estructurado en los harmónicos fundamentales de dichas frecuencias.

Además, el cuerpo humano es una mixtura de diversos procesos, de los cuales podemos estar interesados en aislar la señal de uno de ellos. Sin embargo, varios procesos pueden estar presenten en las capturas que realicemos, por lo que estarán superpuestas en la señal obtenida. Algunos ejemplos de estas interferencias fisiológicas son el registro de EGG (electrogastrograma) en ECG precordial, o el registro del ECG materno al tomar el ECG fetal, o la interferencia del ECG en el EEG (electroencefalograma).

Ejemplo 6.3 (Promediado muestral de ECG para la eliminación de ruido blanco) Supondremos que tenemos disponible la adquisición de múltiples muestras del ECG de un paciente realizadas en diferentes momentos de tiempo. Buscamos desarrollar un procedimiento para la reducción del ruido blanco de nuestro ECG basado en promediado muestral, que sabemos que reduce el nivel de ruido en un factor $\sqrt(N)$, siendo N el número de muestras utilizadas en el promediado.

El promediado muestral aplicado a nuestro problema tiene dos inconvenientes principales a tener en cuenta: i) al realizar un promediado muestral, debemos suponer que el proceso que origina el ECG es ergódico, pero no tiene por qué serlo, por ejemplo, simplemente por el cambio de ritmo cardíaco del paciente; ii) las muestras de ECG utilizadas no tienen por qué estar en fase, es decir, los ciclos de cada muestra no tienen que comenzar necesariamente en el mismo instante.

La correlación cruzada entre señales (ecuación 6.2) nos dará información valiosa para ambos problemas:

- Para solucionar el segundo problema, podemos desarrollar un algoritmo de alineamiento de señales que adelante o retrase una señal el invervalo τ que obtiene el máximo de correlación entre señales. Así pues, alinearemos todas las señales respecto a una señal de referencia (p.e. la primera) y cuantificaremos la correlación entre señales mediante el valor máximo de correlación calculado.

- Podemos establecer un umbral sobre el índice de correlación por debajo del cual descartar las muestras a la hora de realizar el promediado muestral. Este método descarta las señales que difieren de forma excesiva de la señal de referencia y por lo tanto acerca el muestreo a las generadas por un proceso ergódico.

Por lo tanto, el promediado muestral se realizará mediante señales alineadas cuyos valores máximos de correlación cruzada superan un umbral preestablecido. Podemos realizar el procedimiento descrito mediante el siguiente código en Matlab, donde suponemos que la matriz ecg11 tiene n muestras de longitud *length*, muestreadas con $T = sampling\_interval$:

```
time1=time(1:l);
ecg11=zeros(n,l);
ecg11aligned=zeros(n,l);
xcmax=zeros(n,1);
ecg11aligned(1,:)=ecg11(1,:);
xcmax(1)=1;
subplot(2,1,1)
plot(time1,ecg11)
for i=2:n
    xc=xcorr(ecg11(1,:),ecg11(i,:),'coeff');
    time1xc = -time1(size(time1,2)):sampling_interval:time1(size(time1,2));
```

```
    subplot(2,1,1)
    plot(time1,ecg11(1,:))
    hold on
    plot(time1,ecg11(i,:),'r')
    hold off
    subplot(2,1,2)
    plot(time1xc,xc);
    %buscamos el maximo en un entorno de +-1 ciclo cardiaco aprox. 1s.
    range=[l-1/sampling_interval:l+1/sampling_interval];
    [maxxc,imaxxc]=max(xc(range));
    xcmax(i)=maxxc;
    shift=time1xc(l-1/sampling_interval+imaxxc);
    shift_positions=round(shift/sampling_interval)
    if shift_positions <0
        ecg11aligned(i,1:(l+shift_positions))=ecg11(i,(1-shift_positions):l);
    end
    if shift_positions > 0
        ecg11aligned(i,(1+shift_positions):l)=ecg11(i,1:(l-shift_positions));
    end
    if shift_positions ==0
        ecg11aligned(i,:)=ecg11(i,:);
    end
    subplot(2,1,2)
    plot(time1,ecg11aligned(1,:))
    hold on
    plot(time1,ecg11aligned(i,:),'r')
    hold off
end
ecg11mean=mean(ecg11aligned);
subplot(2,1,1)
plot(time1,ecg11mean)
ecg11meant=mean(ecg11aligned(xcmax>threshold,:));
subplot(2,1,2)
plot(time1,ecg11meant)
```

La figura 6.1 muestra 75 adquisiciones de $4s$ de una derivación de ECG, muestreadas con un intervalo de muestreo de $0{,}004s$ y a las que se les ha superpuesto un ruido blanco de desviación estándar 0,1.

El procedimiento descrito anteriormente obtiene como resultado una reducción del ruido blanco de la señal, tal como vemos en la figura 6.2. Este algoritmo es sensible a la longitud de las señales promediadas, ya que a mayor longitud, mayor dificultad para alinear las secuencias. También la elección del umbral será un factor clave en la obtención de un buen promedio. Vemos que en el cuadro superior no hemos aplicado el umbral sobre el índice de correlación cruzada, y esto provoca la distorsión de componentes del ECG (p.e. el complejo QRS), La aplicación de un umbral de 0.7 implica reducir el número de muestras ponderadas a 11, y por lo tanto menor atenuación del ruido blanco; sin embargo, se mantiene más estable la forma de las componentes del ECG. Teniendo en cuenta que el ruido se reduce por la raíz cuadrada del número de muestras utilizadas, es mejor utilizar muestras con una correlación cruzada alta que aumentar el número de muestras no correlacionadas.

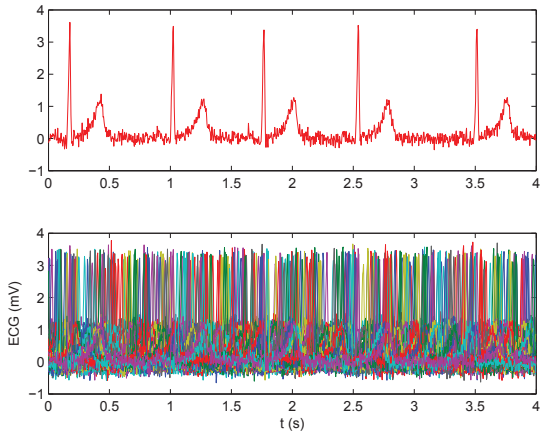


Figura 6.1: Setenta y cinco adquisiciones de 4s de una derivación de ECG, muestreadas con un intervalo de muestreo de 0,004s y a las que se les ha superpuesto un ruido blanco de desviación estándar 0,1. Como vemos, las señales no están en fase. La señal mostrada en el cuadro superior es la primera de las señales, que se utiliza como referencia del algoritmo de eliminación de ruido blanco por promediado de señales alineadas.

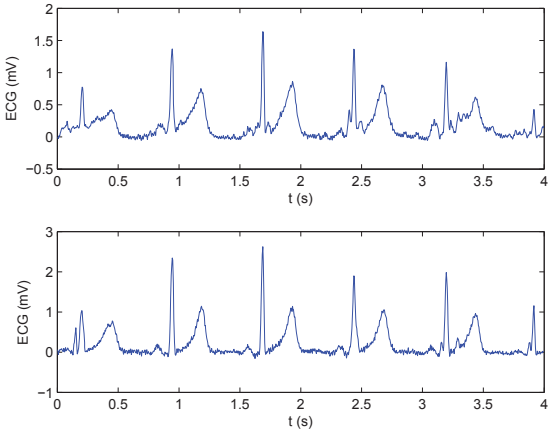


Figura 6.2: Resultado de la reducción del ruido blanco de la señal de ECG mediante promediado muestral.

6.5. Análisis en frecuencia

Como hemos visto, en la sección 6.2, la transformada de Fourier (ecuación 6.2) representa la señal $x(t)$ mediante un vector de las amplitudes de funciones senoidales de un rango de frecuencias. En definitiva, el análisis de Fourier se basa en que toda función puede expresarse como la suma de un número infinito de funciones senoidales de diferentes frecuencias, cada una con su amplitud y su fase.

Una función continua $x(t)$ al ser muestreada en N muestras a intervalos de T segundos, queda representada por una secuencia $x(n) = x(nT), n = 1, \ldots, N$. Para estimar la transformada de Fourier de esta señal a partir de sus valores muestreados, la integral de la trasnformada continua se sustituye por un sumatorio:

$$\hat{x}(f_k) = \frac{1}{N} \sum_{n=1}^{N} x(n) e^{-i2\pi nk/N}, \forall f_k.$$

Para optimizar el proceso de cálculo se han desarrollado algoritmos específicos que han dado lugar a la llamada Transfomada Rápida de Fourier (Fast Fourier Transform, FFT), que es la que utilizaremos usualmente en el análisis de las señales. La FFT reduce el coste temporal de la transformada desde N^2 hasta $N \log_2(N)$.

6.5.1. Resolución frecuencial o espectral

La resolución espectral $\Delta f = 1/NT = f_s/N$ nos indica la capacidad de distinguir dos componentes espectrales muy próximas, y depende del tiempo muestreado NT y de la frecuencia de muestreo f_s. Así, si se quiere distinguir entre dos componentes que difieran en $0{,}001Hz$, Δf tendrá que ser menor que $0{,}001Hz$. El "zero-padding" es una técnica que aumenta la definición del espectro pero no su resolución, es equivalente a realizar una interpolación en el dominio de la frecuencia. No da mayor resolución frecuencial, para ello deberíamos aumentar el tiempo de adquisición.

6.6. Eliminación de artefactos de la señal

Como ya hemos visto, las señales adquiridas distan de ser observaciones puras de los procesos fisiológicos que deseamos estudiar, ya que suelen estar superpuestas a interferencias de diferente naturaleza. Antes de comenzar con un proceso de detección o clasificación para la ayuda a la decisión médica, deberemos eliminar los artefactos de la señal. El filtro digital es la herramienta principal para esta tarea cuando procesamos señales digitales.

Un filtro digital $h(k)$ es un sistema discreto que transforma una secuencia entrada $x(n)$, en una secuencia salida $y(n)$.

$$y(n) = \sum_{k} h(k) x(n-k).$$

Como su nombre indica, la tarea de un filtro es filtrar o separar partes de una señal de acuerdo con un criterio. Por ejemplo, separar una señal del ruido con el que está mezclada, como es el caso de los ECGs, donde el ruido es la interferencia de 50 Hz de la red eléctrica y del armónico de 100 Hz.

El diseño de un filtro digital consiste en identificar los coeficientes de la función que relaciona la entrada con la salida, para ello, la transformada Z es de gran utilidad.

6.6.1. Transformada Z

Dada la señal $x(t)$, muestreada como $x(n) = x(nT)$, podemos expresar la secuencia en forma de sumatorio e introducir un operador en cada sumando que indique el orden de la muestra. Esta operación se denomina Transformada Z y facilita el trabajo con las señales digitales.

Así pues, la transformada Z de una secuencia $x(n)$ se define como:

$$X(z) = \sum_{n=0}^{\infty} x(n) z^{-n},$$

donde z^{-n} indica que el valor $x(n)$ está retardado n intervalos T desde el comienzo del proceso de muestreo.

Ejemplo 6.4 (Transformada Z de una secuencia)
La transformada Z de la secuencia $x(n) = \{1, 2, 5, 3, 0, 0, 0, \ldots\}$ es

$$X(z) = 1 + 2z^{-1} + 5z^{-2} + 3z^{-3}.$$

Ejemplo 6.5 (Transformada Z de una función impulso)
Sea la función impulso $x(t) = \{1, 0, 0, 0, 0, \ldots\}$, su transformada X es $X(Z) = 1$.

Ejemplo 6.6 (Transformada Z de una función escalón unitario)
Sea la función impulso $x(t) = \{1, 1, 1, 1, 1, \ldots\}$, su transformada X es

$$X(z) = 1 + z^{-1} + z^{-2} + z^{-3} + \ldots,$$

que si multiplicamos ambos lados por $(z - 1)$ se convierte en

$$(z-1)X(z) = (z + 1 + z^{-1} + z^{-2} + \ldots) - (1 + z^{-1} + z^{-2} + \ldots) = z,$$

por lo que

$$X(z) = \frac{z}{z-1} = \frac{1}{1 - z^{-1}}.$$

6.6.2. Operadores básicos de un filtro digital

Encontraremos los siguientes filtros básicos. Estos filtros corresponden a operaciones básicas sobre secuencias: retraso de una secuencia, multiplicación de una secuencia por un escalar y suma de secuencias. Estos operadores simples podrán combinarse para diseñar filtros más complejos

- Retardo unidad: z^{-1}

$$Y(z) = X(z)z^{-1}$$

$$y(n) = x(n-1)$$

- Amplificación: k

$$Y(z) = kX(z)$$

$$y(n) = ky(n)$$

- Suma de secuencias

$$Y(z) = X_1(z) + X_2(z)$$

$$y(n) = x_1(n) + x_2(n)$$

6.6.3. Función de transferencia de un filtro digital

La función de transferencia (o fdt) de un filtro digital es el cociente entre la transformada Z de la salida y la transformada Z de la entrada:

$$H(z) = \frac{Y(z)}{X(z)}$$

Ejemplo 6.7 (Función de transferencia de un filtro dado $X(z), Y(z)$)
Si al aplicar la entrada $X(z) = 1 + z^{-1} + z^{-2}$ obtenemos la salida $Y(z) = 2z^{-1} + 2z^{-2} + 2z^{-3}$, ¿cuál es la función de transferencia $H(z)$ del filtro?

$$H(z) = \frac{2(z^{-1} + z^{-2} + z^{-3})}{1 + z^{-1} + z^{-2}} = \frac{2z^{-1}(1 + z^{-1} + z^{-2})}{1 + z^{-1} + z^{-2}} = 2z^{-1}$$

Por lo que el filtro es una combinación de un retardo y una amplificación $k = 2$.

6.6.4. Tipos de filtros digitales

Existen dos tipos de filtros digitales, los no recursivos y los recursivos. En los no recursivos la fdt contiene un número finito de elementos y están en forma polinomial. También se llaman filtros FIR (finite impulse response), all-zero o moving-average (MA).

$$H(z) = \sum_{i=0}^{N} h_i z^{-i}.$$

En los recursivos, la fdt se expresa como cociente de dos polinomios. También se llaman filtros IIR (infinite impulse response), all-pole, recursivos o autoregresivos (AR). En estos filtros el valor de la salida depende de los valores de la entrada y de los valores previos de salida.

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{j=0}^{m} b_j z^{-j}}{\sum_{i=0}^{N} a_i z^{-i}}$$

que puede expresarse como polinomios en z^{-1} como

$$Y(z) = B(z)X(z) + A(z)Y(z),$$

con lo que vemos que la salida depende de la entrada y de salidas anteriores.

6.6.5. Polos y ceros de un filtro digital

Los valores de z para los cuales $H(z)$ se hace cero son llamados ceros de la fdt, y los valores para los cuales $H(z)$ tiende a infinito son llamados polos.

Si $H(z) = \frac{A(z)}{B(z)}$, para encontrar los ceros resolveremos la ecuación $A(z) = 0$, los que anulan el denominador. Para encontrar los polos resolveremos la ecuación, $B(z) = 0$, los que anulan el denominador. Los filtros no recursivos como no tienen $B(z)$ solo tienen un polo en $z = 0$.

Para estudiar la respuesta de los filtros desde un punto de vista práctico, se sustituye

$$z = e^{iwT} = cos(\omega T) + isen(\omega T)$$

y se halla la fase y el módulo de la función de transferencia. Si $|z| = 1$ la expresión anterior es la ecuación de un círculo de radio unidad en el plano z (ver figura 6.3), por lo que podemos centrarnos en la fase wT del punto.

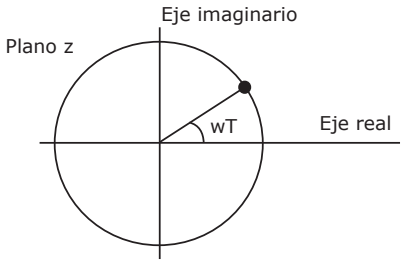


Figura 6.3: Circulo unidad en el plano z.

En el plano z representaremos el círculo unidad, así como los ceros y los polos que, en general, serán números complejos. El ángulo correspondiente al cero o al polo en el plano complejo nos indicará de una forma gráfica y sencilla, qué frecuencias atenuarán (cero) o amplificarán el filtro (polo).

Como $\omega = 2\pi f$ y $T = 1/f_s$, siendo f la frecuencia específica y f_s la frecuencia de muestreo, la fase del punto en el plano z es $\omega T = 2\pi \frac{f}{f_s}$, siendo $\frac{f}{f_s}$ la frecuencia normalizada que toma valores entre 0 y 0.5[b]. Por ejemplo, $\omega T = 0, f = 0Hz$, $\omega T = \pi/2 = 90, f = f_s/4$, y $\omega T = \pi = 180, f = f_s/2$. Por lo tanto, la frecuencia específica del filtro dependerá de la frecuencia de muestreo. Por ejemplo, un filtro con un cero en $\pi/2$ tendría una frecuencia específica de 50 Hz para una señal muestreada a 200 Hz, pero para una señal muestreada a 100 Hz, la frecuencia específica sería de 25 Hz.

Ejemplo 6.8 (Análisis de polos y ceros de un filtro)
Sea el filtro $H(z) = \frac{1}{3}(1+z^{-1}+z^{-2})$, para encontrar sus polos y ceros multiplicamos y dividimos por z^2.

$$H(z) = \frac{1}{3}(1 + z^{-1} + z^{-2})\frac{z^2}{z^2} = \frac{z^2 + z + 1}{z^2}$$

Igualando el numerador a cero, obtenemos los ceros

$$z^2 + z + 1 = 0 \Rightarrow z = -0{,}5 \pm i0{,}866$$

Igualando el denominador a cero, obtenemos los polos

$$z^2 = 0 \Rightarrow z = 0$$

Situándolos en el circulo unidad del plano z de la figura 6.4 observamos la situación de polos y ceros.

Sabemos que los ceros de la H(z) anulan la salida, por tanto para la frecuencia específica de los ceros la salida del filtro tenderá a cero eliminando por lo tanto cualquier señal a esa frecuencia. El ángulo específico será

$$\omega T = 180 - arctan\frac{0{,}866}{0{,}5} = 120 = \frac{2\pi}{3},$$

y como $\omega T = 2\pi \frac{f}{f_s}$, entonces $f = \frac{f_s}{3}$. Si la señal de entrada está muestreada a 180 Hz, el cero en 120^0 eliminará las señales de 60 Hz.

[b]Esto es debido a que si suponemos que $f = f_s$, entonces $\omega T = 2\pi = 360$, siendo la frecuencia específica igual a la de muestreo, pero por el teorema de Nysquit $f_s \geq 2f, \forall f$, esta situación no puede darse.

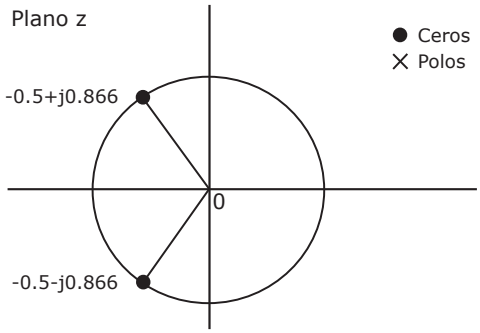


Figura 6.4: Polos y ceros de H(z) en el circulo unidad en el plano z.

Ejemplo 6.9 (Eliminación de artefactos de baja frecuencia de ECG)
Los artefactos de baja frecuencia y el nivel de continua de las derivaciones precordiales de un ECG son causados por la tos o respiraciones profundas, mientras que en las derivaciones periféricas pueden deberse al movimiento de un brazo o una pierna. La conexión incorrecta de los electrodos y los contactos también pueden causar ruido de baja frecuencia. Por su parte, el aumento del nivel de continua puede proceder de la variación de la temperatura o desajustes del instrumental o de los amplificadores.

En el ECG de la figura 6.5 observamos que la componente de continua es muy elevada. Además, gracias a la representación en frecuencia, vemos que las frecuencias muy cercanas a 0 tienen una energía considerable. Por último, se observa también el rizado de la red eléctrica en los 60 Hz.

El nivel de continua puede elmininarse cuando calculamos la derivada de la señal, por lo que podemos plantear un filtro de artefactos de nivel de continua mediante un operador derivada:

$$y(n) = \frac{1}{T}(x(n) - x(n-1)) \tag{6.1}$$

$$Y(z) = \frac{1}{T}(X(z) - X(z)z^{-1}) \tag{6.2}$$

$$\frac{Y(z)}{X(z)} = \frac{1}{T}(1 - z^{-1}) \tag{6.3}$$

$$H(z) = \frac{1}{T}1 - z^{-1} \tag{6.4}$$

En el plano z vemos que, efectivamente, el filtro tiene un cero en la frecuencia $0Hz$, por lo que habrá una atenuación en esta zona.

Una forma inmediata de identificar la respuesta de un filtro será la representación gráfica de la amplitud y fase de su respuesta en función de la frecuencia. En la figura 6.7 vemos la respuesta del filtro para cada frecuencia.

Podemos realizar en Matlab el filtrado de nuestro espectro de la siguiente forma:

```
b=[1 -1]
a=[sampling_interval 0]
zplane(b,a);
ecg1_fd1=filter(b,a, ecg1);
```

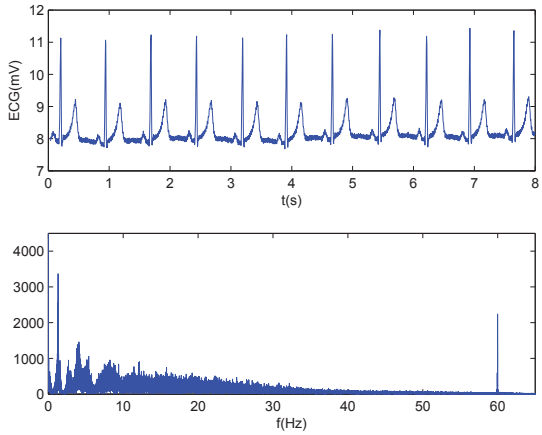



Figura 6.5: Primeros 8s y espectro en frecuencia de una derivación de ECG adquirida con una frecuencia de muestreo $f_s = 250Hz$ durante $300s$.

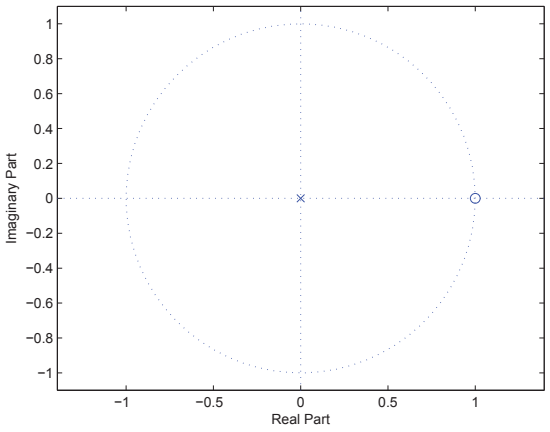


Figura 6.6: Cero en la frecuencia $0Hz$ del filtro derivada.

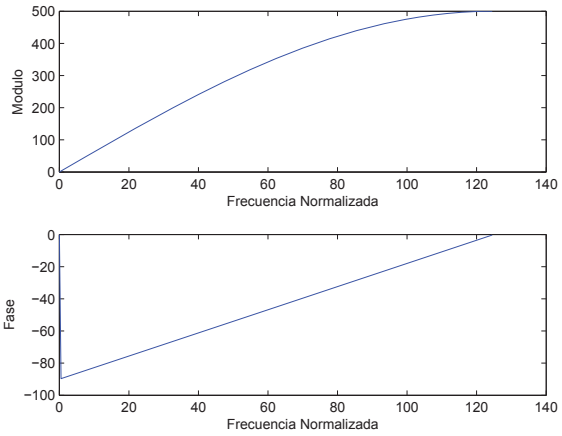


Figura 6.7: Respuesta en frecuencia del filtro derivada.

La figura 6.8 muestra el resultado del filtro sobre la señal de ECG. Observamos que efectivamente el nivel de continua se ha corregido gracias al filtro. Sin embargo, muchas de las componentes, como la onda P y la onda T, se han perdido en la salida, y el complejo QRS (de alta frecuencia) ha sido amplificado.

Ejemplo 6.10 (Mejoras del filtro derivada (cont. ej. 6.9))
El filtro derivada del ejemplo anterior provoca un aumento considerable de ruido de alta frecuencia. Esto es debido a que la respuesta crece continuamente con la frecuencia. Una mejora del filtro derivada consiste en realizar la media de dos salidas sucesivas de un filtro derivada.

$$\begin{aligned} y_2(n) &= \frac{1}{2}(y(n) + y(n-1)) && (6.5)\\ &= \frac{1}{2T}(x(n) - x(n-1) + x(n-1) - x(n-2)) && (6.6)\\ &= \frac{1}{2T}(x(n) - x(n-2)) && (6.7)\\ H(z) &= \frac{1}{2T}(1 - z^{-2}) && (6.8)\\ &= \frac{1}{T}(1 - z^{-1})\frac{1}{2}(1 + z^{-1}) && (6.9)\\ & && (6.10) \end{aligned}$$

En la figura 6.7 vemos la respuesta del filtro para cada frecuencia y el plano z. Ahora existen ceros tanto en las frecuencias bajas como en las frencuencias altas de la señal.

El resultado que obtenemos no es nada óptimo, y seguimos teniendo una atenuación grande en la mayoría de frecuencias con información de la señal ECG.

Si queremos mantener los niveles de las componentes de la señal entre las frecuencias 0.5 y 1.0 Hz debemos mantener el nivel de ganancia de esas regiones. Podemos mantener la ganancia de las regiones de baja frecuencia incluyendo un polo en $z = 0{,}995$, obteniendo la fdt

$$H(z) = \frac{1}{T}(\frac{1 - z^{-1}}{1 - 0{,}995z^{-1}}).$$

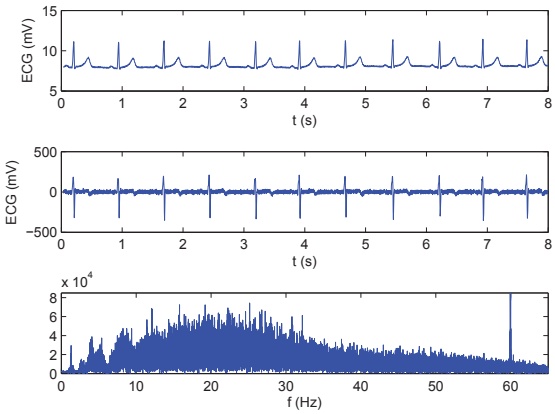


Figura 6.8: Resultado del filtro del operador derivada sobre el ECG.

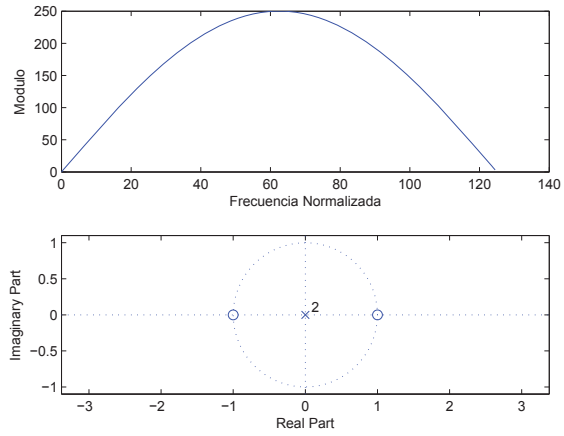


Figura 6.9: Respuesta en frecuencia del filtro derivada promediado. Ahora existen ceros tanto en las frecuencias bajas como en las frencuencias altas de la señal.

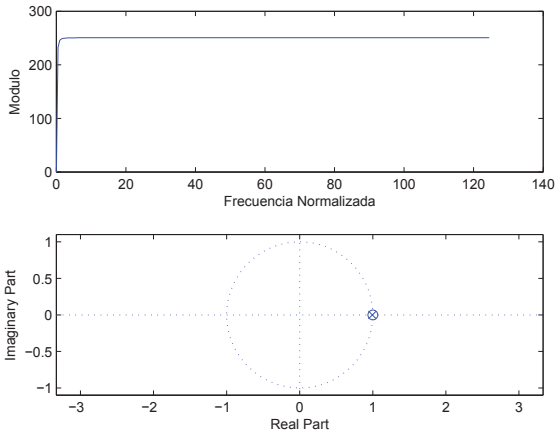


Figura 6.10: Respuesta del filtro de continua con un polo en 0.995 para la recuperación de las ganancia de frencuencias bajas.

La respuesta del filtro (ver figura 6.10) será un rechazo de la línea de continua, con la recuperación rápida del nivel de ganancia para las frecuencias inmediatamente superiores.

Finalmente, aplicado a la señal de ECG, vemos en la figura 6.11 que efectivamente hemos eliminado el nivel de continua, y conservamos bastante la forma original de las componentes de la señal cardiaca. Sin embargo, vemos que el filtro requiere un transitorio hasta conseguir la eliminación del nivel del continua.

Ejemplo 6.11 (Filtrado del rizado de red eléctrica de 60 Hz)
Con el resultado obtenido en el ejemplo 6.10 hemos mejorado la señal respecto al ruido de baja frecuencia y nivel de continua. Sin embargo, nos sigue apareciendo un rizado de alta frecuencia provocado por la señal de 60 Hz de la red eléctrica.

Podremos eliminar este ruido estructura de alta frecuencia mediante un filtro rechazo banda centrado en la frecuencia de 60 Hz. El filtro $H(z) = 1 - 0{,}1256z^{-1} + z^{-2}$ tiene la respuesta mostrada en la figura 6.12, por lo que es adecuado para eliminar el ruido de $60Hz$ de nuestra señal muestreada a $256Hz$.

Si aplicamos el filtro rechazo-banda a la señal obtenida en la figura 6.11, obtendremos la señal de la figura 6.13. En el espectro de frecuencias vemos cómo se ha eliminado la componente de $60Hz$ y sus armónicos.

6.7. Detección de eventos

Una de las tareas más demandadas en el procesamiento de señales fisiológicas es la detección de eventos. Esta detección puede ser de interés para caracterizar una señal (p.e. calcular la frecuencia cardíaca), o para seleccionar características relevantes sobre las que realizar una clasificación (p.e. detectar el complejo QRS para la clasificación de arritmias cardiacas), etc.

La detección de eventos dependerá de la señal tratada y el evento a detectar y, generalmente, combinará la aplicación de métodos de mejora de la señal por ponderación, extracción de estadísticos temporales y filtros para enfatizar las características deseadas.

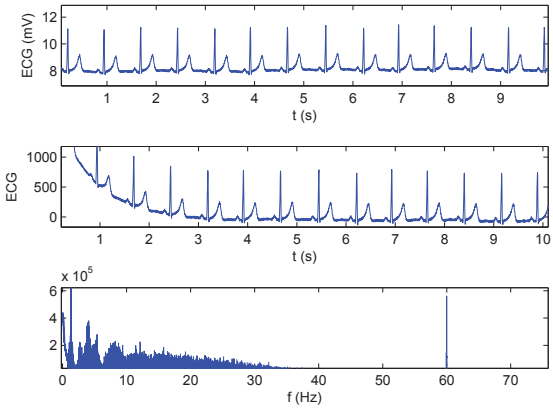


Figura 6.11: Resultado del filtro del operador derivada mejorado sobre el ECG.

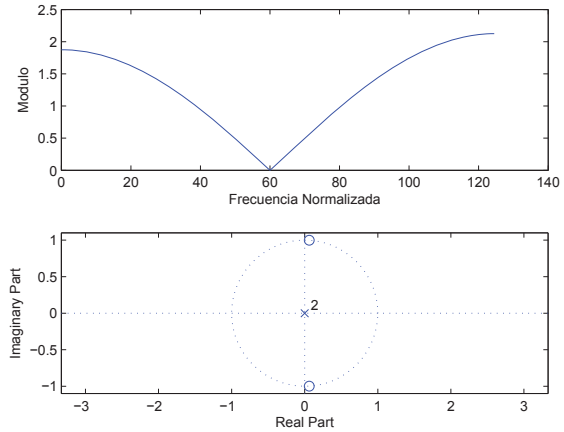


Figura 6.12: Respuesta en frecuencia del filtro rechazo-banda de 60Hz.

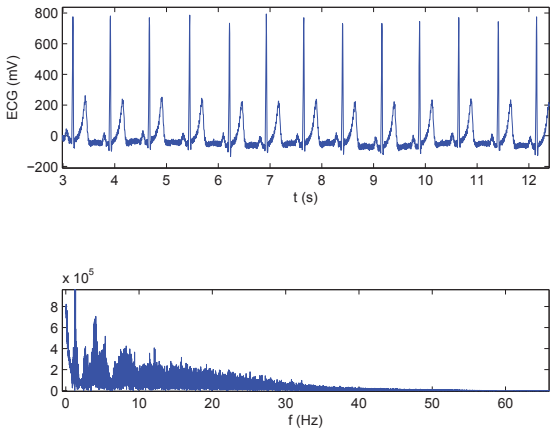


Figura 6.13: Resultado del filtro rechazo-banda de 60Hz sobre el ECG.

Ejemplo 6.12 (Detección del complejo QRS)
Balda et al. propusieron un algoritmo para la detección del QRS basado en operadores derivada. El filtro, preparado para una señal muestreada a $200Hz$, limpia de artefactos de baja frecuencia y ruido de red eléctrica, se basa en la ponderación de un filtro derivada

$$y_0(n) = |x(n) - x(n-2)|,$$

y un filtro de segunda derivada

$$y_i(n))|x(n) - 2x(n-2) + x(n-4)|,$$

para dar lugar al filtro

$$y_2(n) = 1{,}3y_0(n) + 1{,}1y_1(n).$$

El resultado de $y_2(n)$ se escanea de forma continua, y si seis muestras de ocho seguidas superan un umbral de 1.0, entonces, el segmento de ocho muestras se considera parte de un complejo QRS. El algoritmo puede ser sensible a ruido de alto nivel, ya que está basado en los operadores derivada, por lo que suele requerir un suavizada antes de la aplicacción del heurístico de detección.

6.8. Análisis de señales no-estacionarias

Las señales no estacionarias son las más comunes en el cuerpo humano. Para el análisis de señales no estacionarias, no es útil estudiar el contenido frecuencial de toda la señal. Su análisis requiere utilizar técnicas tiempo/frecuencia que analizan el contenido frecuencial de una señal en función del tiempo.

La idea básica de una distribución tiempo-frecuencia (o espectrograma) es la de tomar ventanas de tiempo contiguas y calcular el espectro de frecuencias asociado a cada una de ellas. Si representamos los espectros en función del tiempo nos queda una representación bidimensional que puede ser visualizada con mapas de color o con representaciones 3D.

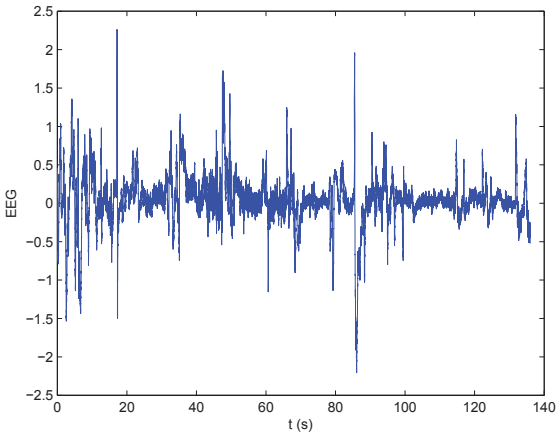


Figura 6.14: EEG muestreado a $256Hz$, claramente muestra que es un proceso no-estacionario.

Ejemplo 6.13 (Distribución tiempo-frecuencia de un EEG)
Representaremos la distribución tiempo-frecuencia del EEG muestreado a $256Hz$ de la figura 6.14. Para ello, utilizaremos la función CTF de Matlab, que toma la señal de entrada, el tamaño de la ventana y el paso para ir desplazando la ventana.

```
function [T] = CTF(signal, win, inc)
i=1;
n=1;
while(n+win-1<length(signal)),
     ven=signal(n:n+win-1);
     val=abs(fft(ven))
     for j=1:128, T(i,j)=val(j); end;
     n=n+inc;
     i=i+1;
end;
mesh(T), xlabel('Frecuencia'), ylabel('Tiempo'), zlabel('Amplitud');
```

Como resultado obtenemos la distribución tiempo-frecuencia de la figura 6.15, que claramente muestra que las componentes frecuenciales de la señal varían según el instante de muestreo.

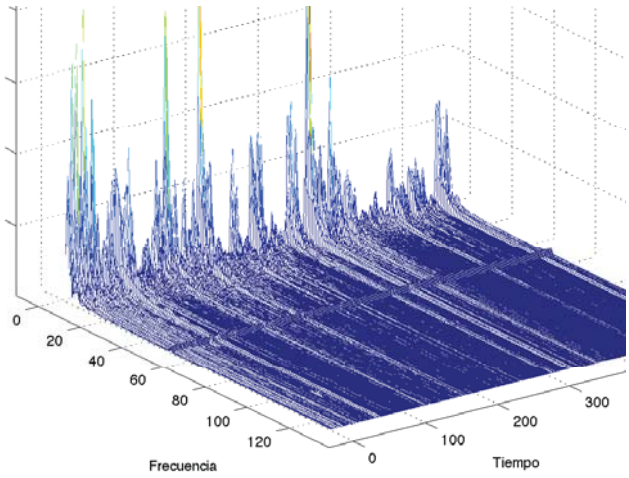


Figura 6.15: Distribución tiempo-frecuencia del EEG muestreado a $256Hz$, claramente muestra diferentes componentes frecuenciales según la ventana de muestreo.

6.9. Análisis de la voz

6.9.1. El proceso de comunicación, acústica, la voz y el oído humano

En la comunicación lingüistica podemos identificar los siguientes elementos de un sistema de comunicación:

- Emisor: el cerebro del hablante.
- Transmisor: órganos articulatorios del hablante que generan ondas sonoras.
- Canal: aire, y otras etapas intermedias.
- Receptor: aparato auditivo del receptor.
- Destino: cerebro oyente.

Las ondas sonoras se originan por el movimiento vibratorio de un cuerpo. Dicha vibración provoca el choque de partículas cercanas, que comprimen y descomprimen partículas cercanas, provocando la transmisión de la vibración. La vibración puede caracterizarse por un péndulo que repite un movimiento de amplitud A con un ciclo de periodo T (y frecuencia ($f = 1/T$), por lo que su propagación se observará como una onda senoidal.

Como curiosidad, las ondas sonoras emitidas por una persona pueden transmitir del orden de $100kb/s$, sin embargo la comunicación hablada basada en el lenguaje humano transmite aproximádamente $50b/s$.

La voz humana forma una señal compuesta de ondas de diversas frecuencias y por lo tanto puede tratarse mediante el procesamiento de señales visto en este capítulo. En efecto, una de las herramientas básicas para el análisis de la voz es el espectrograma, ya visto en la sección 6.8. Sin

embargo, los patrones observables en la voz humana hace posible un análisis más especifico de sus características, que estudiaremos en esta sección.

El oído recoge las ondas sonoras mediante el pabellón auditivo y las canalizan a través del conducto acústico externo que actúa como resonador de las frecuencias entre 3 y $4kHz$. El tímpano transforma la vibración aérea en vibración sólida que se amplifica de 25 a 30 veces por los huesecillos. Por último, el caracol óseo trasnforma las vibraciones mecánicas en impulsos nerviosos.

La intensidad de sonido percibida es logarítmica, es decir, un crecimiento geométrico de la intensidad de las ondas de presión se manifiesta en un crecimiento aritmético de la sensación. La presión del aire es de $10^5 Pa$, y el oído humano puede percibir modulaciones de presión de $10^{-6} Pa$. El sonido se mide en variaciones relativas de presión (decibelios, dB).

6.9.2. Aplicaciones del procesamiento de voz en medicina

El procesamiento de la voz tiene múltiples aplicaciones en el ámbito médico. Enumeraremos una lista no exhaustiva de ámbitos de uso:

1. Dictado de informes médicos. Existen sistemas, por ejemplo en radiología, que permiten la transcripción de los informes médicos a texto libre escrito a través del reconocimiento automático del habla.

2. Análisis al diagnóstico de foniatría y apoyo al servicio de logopedia. El análisis de las características de la señal acústica puede revelar problemas en el aparato fonador o auditivo. Las características que veremos en la sección 6.9.4 pueden revelar evidencias para el diagnóstico.

3. Monitorización de depresiones. Las características de la voz indican estados anímicos de la persona, que pueden dar pistas sobre la evolución de su estado depresivo, y por lo tanto de la efectividad de los tratamientos. En el ejemplo 6.14 veremos trabajos científicos sobre el tema, un posible protocolo de adquisición de voz y en análisis de algunas caracetirísticas para el estudio de la depresión.

4. Ayuda a discapacitados. Personas con dificultades para usar interfaces típicas como el teclado podrían comunicarse por habla, o personas con dificultades auditivas podrían obtener transcripciones de lo dicho por sus interlocutores, personas inmovilizadas podrían manejar dispositivos domóticos mediante voz.

6.9.3. Adquisición de la señal

La adquisición de la voz se realiza mediante un micrófono, que transforma las ondas de presión en una señal eléctrica. Básicamente un micrófono consta de un diafragma que responde a a las variaciones de presión del aire y un transductor que puede funcionar mediante la variación de la capacitancia de un conductor, o mediante el efecto piezoeléctrico o mediante inducción electromagnética.

Una vez adquirida, la señal se transforma mediante un conversor Analógico/Digital, que discretiza la señal con una frecuencia de muestreo (p.e. $48kHz$) y se cuantifica en unos niveles predeterminados (p.e. $32bits$), obteniendo un vector de valores como resultado.

La figura 6.16 muestra la señal digitalizada, llamada sonograma, correspondiente a la lectura algunas frases en castellano.

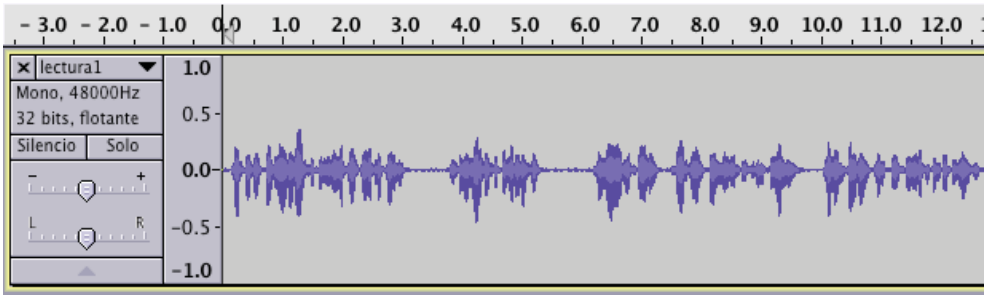


Figura 6.16: Sonograma de voz digitalizada.

Para realizar un análisis de la voz para fines dignósticos será necesario definir un protocolo de grabación que incluyan los parámetros de adquisición y digitalización de la señal. Además será común encontrarse un texto que especifique los ejercicios de voz (lecturas) que el paciente debe realizar para homogeneizar la muestra a analizar.

6.9.4. Características de la voz

Una vez digitalizada la señal y obtenido el sonograma **s** se pueden extraer características útiles para el análisis de la voz. Veremos algunas de ellas de utilidad en medicina. Algunas de estas características pueden obtenerse directamente de la representación temporal de la señal, mientras que las más interesantes suelen obtenerse del espectrograma de la misma.

Antes de comenzar, discribiremos el ejemplo que nos servirá para ilustrar algunas de las características estudiadas.

Ejemplo 6.14 (Procesamiento de la voz en depresiones)
La psicología ha visto en la voz una expresión del estado anímico de las personas. Esto se ha utilizado para caracterizar estados depresivos, esquizofrenias y otras afecciones.

Disfunciones en el flujo acústico de la voz, como lentitud, retardos, monotonía, imposibilidad de expresar el rango habitual de respuestas afectivas, y retardo psicomotor, pueden dar evidencias de problemas psicológicos[c].

Ellgring y Scherer describen algunos indicadores de la voz para detectar cambios de humor en depresión. Algunas de las características que observaron fueron el incremento de velocidad del habla y el decremento en la duración de las pausas con la mejora del humor y por lo tanto una remisión de la depresión. Además, específicamente en mujeres se observó que el descenso de frecuencia del pitch correlacionaba con la mejora del tratamiento[d].

EmoVoice es un software reconocedor de emociones basado en análisis de la voz[e]. EmoVoice, utiliza caracaterísticas como el pitch, la energía, la calidad de la voz e información espectral para realizar el reconocimiento de las emociones.

Durante los ejemplos sucesivos ilustraremos el uso de las características de la voz en depresión leve. Se ha utilizado el programa Audacity para grabar y editar el sonido y el programa Praat para el análisis de voz.

Ejemplo 6.15 (Protocolo de grabación de la voz en depresiones (cont 6.14).)
El grupo Stassen en el "Psychiatric University Hospital Zurich" diseñó un protocolo de grabación de la voz para el estudio clínico de la voz en pacientes con problemas psicológicos. Las

[c] http://www.bli.uzh.ch/vox04.html
[d] http://www.springerlink.com/content/y477726741200u5q/
[e] http://mm-werkstatt.informatik.uni-augsburg.de/EmoVoice.html

características principales de protocolo son las siguientes[f]:

- La grabación se realiza en un laboratorio insonorizado preparado para un rango dinámico de $60dB$. La persona está sola en la habitación. El micrófono está a $50cm$.
- La grabación se calibra con la lectura de un texto, de tal forma que la amplitud máxima quede entre $0dB$ y $-2dB$.
- La grabación digital se realiza con una resolución de $48kHz$ a 20 bits de resolución.
- Las grabaciones se realizan en las mismas horas del día para evitar variaciones debidas a fluctuaciones circadianas.
- Los pacientes siguen un guion que incluye un conteo de 1 a 40, pausa de 30 segundos, lectura de un texto, pausa de 30 segundos, y conteo de 1 a 40.

Para nuestros experimentos, hemos simplificado el protocolo, si bien, se ha intentado seguir al máximo el protocolo anteriormente descrito. La digitalización de la señal se realizó a $48kHz$ y 32 bits, si bien se guardó en ficheros wav de 16 bits. En nuestro guion se ha incluido el siguiente texto:

diez, nueve, ocho, siete, seis, cinco, cuatro, tres, dos, uno, cero.

(10 segundos de respiración)

Las cosas podían haber sucedido de cualquier otra manera y, sin embargo, sucedieron así. Daniel, el Mochuelo, desde el fondo de sus once años, lamentaba el curso de los acontecimientos, aunque lo acatara como una realidad inevitable y fatal. Después de todo, que su padre aspirara a hacer de él algo más que un quesero era un hecho que honraba a su padre. Pero por lo que a él afectaba...

(15 segundos de respiración y saliva)

a (silencio), aaa (silencio), aaaaaaaa (silencio)

e (silencio), eee (silencio), eeeeeeee (silencio)

i (silencio), iii (silencio), iiiiiiii (silencio)

o (silencio), ooo (silencio), oooooooo (silencio)

u (silencio), uuu (silencio), uuuuuuuu (silencio)

(10 segundos de respiración y saliva)

diez (silencio), nueve (silencio), ocho (silencio), siete (silencio), seis (silencio), cinco (silencio), cuatro (silencio), tres (silencio), dos (silencio), uno (silencio), cero.

(FIN)

Un actor leyó el guion con su propia voz (voz normal) y en una segunda grabación fingió un estado depresivo leve (voz depresión).

Energía. La energía puede calcularse en el dominio del tiempo, y se calcula como la integral del cuadrado de la señal $E = \sum_{n=0}^{N-1} s_n^2$. La energía puede utilizarse para detectar intervalos de voz e intervalos de silencio en la muestra.

Ejemplo 6.16 (Lentitud, retardos y monotonía en la voz en depresiones (cont 6.15).) La figura 6.17 representa los fonogramas de la voz normal y voz depresión Observamos que el mismo texto ha sido leído con mayor lentitud en el segundo caso, de $30s$ a $40s$. El análisis mediante niveles de energía nos permite detectar que en la voz normal se realizan 70 cortes y los silencios ocupan el 43 % del tiempo, mientras que en la voz depresión hay 74 cortes y un 47 % del tiempo se utiliza en silencios.

[f] `http://www.bli.uzh.ch/vox01.html`

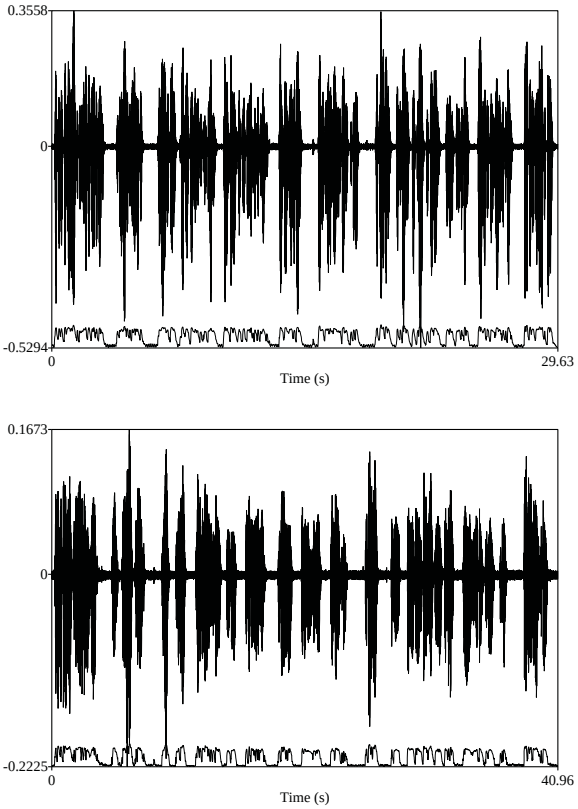


Figura 6.17: Voz normal (arriba) y voz simulando depresión (abajo) de un barón de 30 años. Sonograma y energía (representada bajo cada sonograma). Observamos que el mismo texto ha sido leído con mayor lentitud en el segundo caso, de 30s a 40s. La cantidad de retardos (70 cortes y 43 % del tiempo frente a 74 cortes y un 47 % del tiempo) es mayor en la voz depresión que en la voz normal.

Pitch. El Pitch es la frecuencia fundamental de vibración del sistema vibrante del aparato fonador (laringe y cuerdas vocales).

Ejemplo 6.17 (La variación del pitch (cont 6.16).)
Para este estudio y los siguientes vamos a utilizar la pronunciación sostenida de la vocal 'a'. Compararemos las características que encontramos en el tramo central de la pronunciación entre la voz normal y la voz depresión recogida mediante nuestro protocolo.

La figura 6.18 presenta el sonograma, el espetrograma y algunas de las características principales de la voz normal. El pitch está representado por la línea contínua azul, superpuesta en el gráfico de los formantes. La figura 6.19 presenta la misma información pero para la voz depresión.

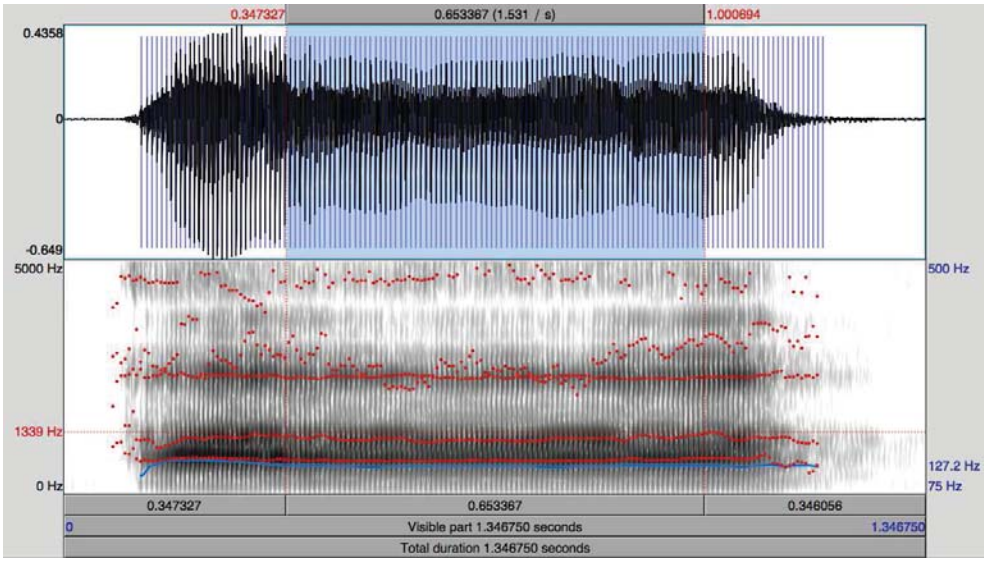


Figura 6.18: Pronunciación sostenida de la vocal 'a' con voz normal. Sonograma (cuadro superior), espectrograma (cuadro inferior), pulsos glóticos (líneas vertiles azules), pitch (linea azul) y formantes (cinco puntos rojos por cada instante).

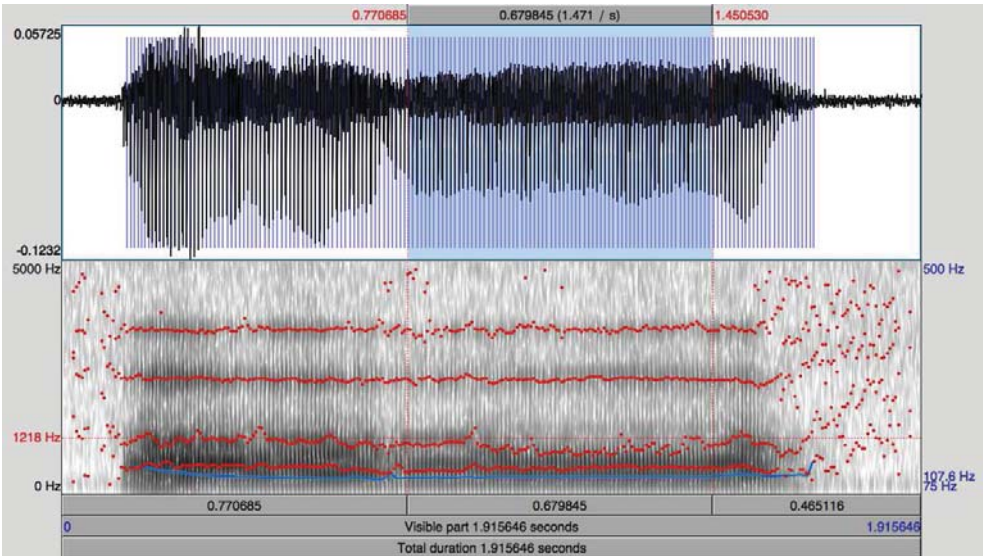


Figura 6.19: Pronunciación sostenida de la vocal 'a' con voz depresión. Ver etiqueta de 6.18 para leyenda.

En nuestra muestra observamos que el pitch medio del locutor con voz normal es de $127Hz$, sin embargo el mismo locutor simulando depresión bajó la frecuencia principal a $108Hz$. Sin embargo, el estudio de Ellgring y Scherer no observaron una variación signitiva del pitch en hombres. Con este resultado únicamente podemos afirmar que el pitch desciende para un mismo locutor simulando depresión respecto a un voz normal.

Formantes. Los formantes son los armónicos de la frecuencia principal con una alta intensidad. En el espectrograma se suelen observar como bandas oscuras (mayor intensidad). Suelen presentarse 5 formantes siendo los dos primeros fundamentales para la percepción de las vocales.

Ejemplo 6.18 (La observación de los formantes (cont 6.17).)
Seguimos con el estudio de la 'a' presentado para la voz normal en la figura 6.18 y para la voz depresión en la figura 6.19. Podemos observar que en la zona seleccionada de la voz normal, tanto el formante primero como el segundo están bien definidos, pudiéndose ver líneas continuas. Sin embargo, en la voz depresión, observamos que el segundo formante no está representado por una línea continua a una misma frecuencia, pudiendo caracterizar menor control de la voz en la pronunciación sostenida de la vocal.

Pulsos glóticos. Los pulsos glóticos son la expresión sonora de la apertura y cierre de la glotis. Las figuras 6.18 y 6.19 representan los pulsos glóticos como líneas verticales azules. Cuando un pulso glótico aparece más allá del 125 % del periodo medio entre pulsos (pitch), se considera que se ha producido un corte de voz. El grado de cortes de voz se utiliza en foniatría con fines diagnósticos.

Jitter. La variación media del periodo entre pulsos glóticos consecutivos respecto al periodo medio se denomina Jitter. Es una medida de la calidad de la voz, y el software MDVP considera patológico un Jitter mayor de 1.04 %.

Shimmer. El Shimmer es la variación absoluta media de la amplitud entre dos periodos consecutivos, dividido por la amplitud media.

Ejemplo 6.19 (Jitter y Shimmer en depresión (cont 6.18).)
En la ventana seleccionada en la figura 6.18, que intenta abarcar el momento de mayor estabilidad de la pronunciación sostenida de la vocal 'a', hemos medido el Jitter y el Shimmer, obteniendo un valor de 0.187 % y 2.390 %, respectivamente. Cuando medimos ambas características en el intervalo seleccionado de la figura 6.19, el Jitter aumenta hasta 0.537 %, y el Shimmer pasa a ser de 4.861 %. Sería interesante demostrar que las variaciones de Jitter y Shimmer sirven de marcadores de las variaciones de estados depresivos. Esto puede estar producido por un mayor o menor control de la voz según el humor del paciente.

Estas características, así como parametrizaciones de la voz para representar los fonemas pueden ser utilizados mediante algoritmos de programación dinámica (ver sección 5.3) para el reconocimiento de estados clínicos o reconocimiento de palabras aisladas.

6.10. Notas bibliográficas

Rangayyan en [103] desarrolla el procesamiento de señales biomédicas mediante un buen conjunto de aplicaciones bien documentadas sobre las señales biomédicas más conocidas.

Capítulo 7

Procesamiento de imágenes médicas

7.1. Introducción

Las imágenes médicas vienen siendo una importante herramienta de diagnóstico prácticamente desde el descubrimiento de los rayos X por Röntgen en 1895. Desde entonces han surgido numerosas técnicas para obtener imágenes de la anatomía interna de los pacientes. Desde los rayos X, pasando por la tomografía computerizada, la ecografía, la imagen nuclear hasta la resonancia magnética, todas han proporcionado evidencias que pueden ser empleadas para ayudar a los médicos a tomar las decisiones pertinentes cuando el proceso de las imágenes ha sido completado. Las imágenes médicas son, pues, un conjunto de técnicas y procesos que se emplean para crear imágenes anatómicas del cuerpo humano o partes de él, con el fin de revelar, diagnosticar o examinar enfermedades o para ayudar al estudio de la anatomía no patológica y su función.

Las distintas técnicas para obtener imágenes médicas pueden ser caracterizadas según su interacción con los tejidos, según su capacidad para separar objetos a diferentes profundidades o según la naturaleza del contraste. Así pues, en función de la interacción con los tejidos podemos tener técnicas con:

- Radiaciones ionizantes: son radiaciones con energía suficiente como para ionizar la materia. Por ejemplo, la radiografía o las imágenes de medicina nuclear como la tomografía por emisión de positrones (PET) o la tomografía computerizada por emisión de fotones individuales (SPECT).
- Radiaciones no ionizantes: no tienen suficiente energía como para ioinizar la materia. Únicamente pueden provocar calentamiento en los tejidos. Por ejemplo, la resonancia magnética nuclear (MRI, MRSI) o la ecografía.

En función de la capacidad para separar objetos se tienen (ver figura 7.1):

- Imágenes proyectivas: son técnicas que superponen los objetos en un mismo plano como las radiografías.
- Imágenes tomográficas: son técnicas que pueden proporcionar varios cortes de los objetos bajo estudio. Como la tomografía computerizada, las tomografías de imagen nuclear o las resonancias magnéticas.

Según la naturaleza del contraste se disponen de:

- Imágenes morfológicas: tratan de representar partes de la anatomía con una buena resolución.

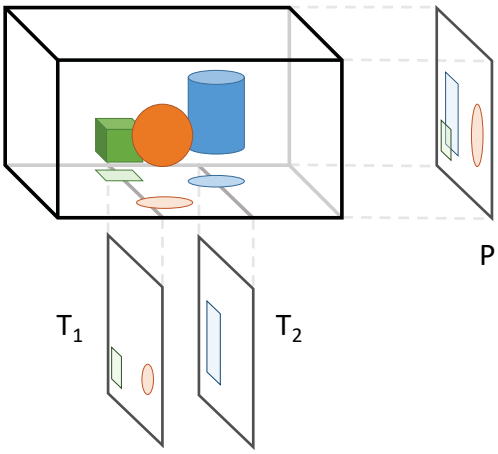


Figura 7.1: Las imágenes proyectivas, como su nombre indica, proyectan los objetos del espacio bajo estudio sobre un mismo plano P. En cambio, las imágenes tomográficas obtienen distintas secciones, o tomos, de los objetos del espacio bajo estudio. De este modo, se pueden obtener distintos cortes S_1, S_2, etcétera. Posteriormente, si se tienen suficientes cortes de los objetos bajo estudio se puede llevar a cabo una reconstrucción tridimensional de los objetos analizados.

- Imágenes funcionales: tratan de representar el funcionamiento del metabolismo. Por ejemplo, la resonancia magnética funcional o las tractografías.

En este tema no se va a profundizar en los fundamentos físico-químicos de las técnicas de adquisición de imágenes sino que partiremos del supuesto en el que las imágenes ya han sido adquiridas pero no procesadas para nuestros propósitos.

7.2. Representación digital de la imagen

Anteriormente se ha visto que las imágenes permiten tener una información morfológica o funcional acerca de las texturas, brillos, colores y formas de los objetos analizados. El procesamiento o tratamiento digital de las imágenes adquiridas pretende facilitar e incrementar la información potencial que puede ser extraída a partir de dichas imágenes.

Las imágenes suelen ser tratadas digitalmente como datos bidimensionales. Generalmente, muchas de las técnicas de procesamiento de señales son aplicables a imágenes, aunque algunas de estas técnicas deben adaptarse para tratar con datos en dos dimensiones, por ejemplo, convolución, filtros digitales o transformada de Fourier.

Es muy habitual representar las imágenes como una matriz de píxeles[a], aunque estas representaciones dependen del tipo de imagen que estemos tratando. Por ejemplo, las imágenes en color que sigan el modelo de codificación RGB estarán representadas por tres matrices, donde la primera indicará la intensidad del color rojo de cada pixel, la segunda la intensidad de color verde y la tercera la intensidad de color azul. Las imágenes en escala de grises podrán ser representadas en una única matriz, así como las imágenes en blanco y negro. Sobre estas matrices se podrán llevar a cabo multitud de operaciones para transformar las imágenes. Para construir

[a]Un *pixel* (del inglés, *picture element*) es un elemento de imagen mínimo.

imágenes tridimensionales bastará con combinar imágenes bidimensionales tomográficas tomadas de distintas secciones correlativas a partir de un mismo objeto.

Así pues, se dispone de una o varias imágenes que pueden entenderse como funciones bidimensionales de intensidad de luz $f(x, y)$, donde x e y son las coordenadas de un pixel y la función $f(x, y)$ determina la intensidad de luz para el par (x, y). La intensidad de luz tiene una escala $[0, L]$, donde L es un valor máximo tal que $0 \leq f(x, y) \leq L$. El número máximo de píxeles en los ejes X e Y determinará la resolución espacial de la imagen, mientras que la resolución en intensidad vendrá determinada por el valor máximo L. Los dos tipos de resolución provienen de los conceptos de muestreo y de cuantificación, respectivamente. Al digitalizar las imágenes analógicas continuas, las coordenadas espaciales deben transformarse a coordenadas discretas, este proceso es conocido como muestreo. Del mismo modo, la cuantificación es la discretización de la intensidad de los colores o de una escala de grises de la imagen analógica continua.

Existen diferentes esquemas de codificación para la representación de las imágenes. En la gran mayoría de programas con capacidades para procesar imágenes digitales podemos encontrar la codificación en color RGB, ya mencionada, la codificación por índices, codificación por intensidad y codificación binaria. Las dos primeras clases de codificación están orientadas a la representación de imágenes en color. En las imágenes indexadas los valores de los píxeles son índices a una tablas que asocian a dicho índice un color. Aunque es una buena forma de almacenar la información, este tipo de codificación no es apta para realizar operaciones aritméticas, por lo que el procesamiento de las imágenes codificadas de este modo no siempre producen como resultado imágenes con sentido. Además, las imágenes indexadas necesitan estar relacionadas con una matriz que contenga el mapa de colores (*colormap*).

La representación de imágenes en escala de grises se lleva cabo mediante una codificación por intensidades, donde el valor de cada pixel representa el brillo o el valor de gris de la imagen en el punto concreto. Si la imagen es en blanco y negro exclusivamente, entonces el pixel se codifica con un valor binario que toma el valor 0 si es negro y el valor 1 si es blanco.

Como se ha comentado, la resolución espacial de la imagen depende del muestreo. Una imagen de $M \times N$ píxeles se representará con una matriz de M filas y N columnas. Cuanto mayor sea la resolución en píxeles mayor será la calidad de la imagen y menor la pérdida de información (ver figura 7.2). Sin embargo, será necesaria más memoria y, puesto que la matriz que represente la imagen será mayor, también conllevará un mayor coste computacional a la hora de procesarla.

La resolución en intensidad depende de la cuantificación. El valor de estas intensidades puede ser más o menos preciso en función del formato de datos que se emplee. Es decir, una intensidad se puede representar empleando un número concreto de bits. Cuanto mayor sea el número de bits mayor será la calidad de la imagen y menor la pérdida de información, aunque también será necesaria más memoria y más esfuerzo computacional para procesar las imágenes. Normalmente, se suele emplear un *byte*[b], dos *bytes* o un tipo de variable que es el doble precisión (*double*) que equivale a 8 *bytes* para representar las intensidades. En nuestros ejemplos con imágenes médicas trabajaremos con un formato de 1 *byte*, esto es, 8 bits. De este modo, en cada pixel se empleará una palabra de 8 bits, con lo que se podrán codificar $2^8 = 256$ niveles distintos de intensidad, desde 0 hasta $L = 2^8 - 1$. En general, si se representa una imagen con n bits, se podrán representar 2^n niveles de intensidad: desde 0 hasta $L = 2^n - 1$.

Ejemplo 7.1

El espacio requerido en memoria variará en función del formato de los datos y de la resolución espacial, es decir, los píxeles de la imagen. La tabla 7.1 resume los requisitos en memoria de algunos formatos de imagen.

[b] Un *byte* son ocho bits contiguos.

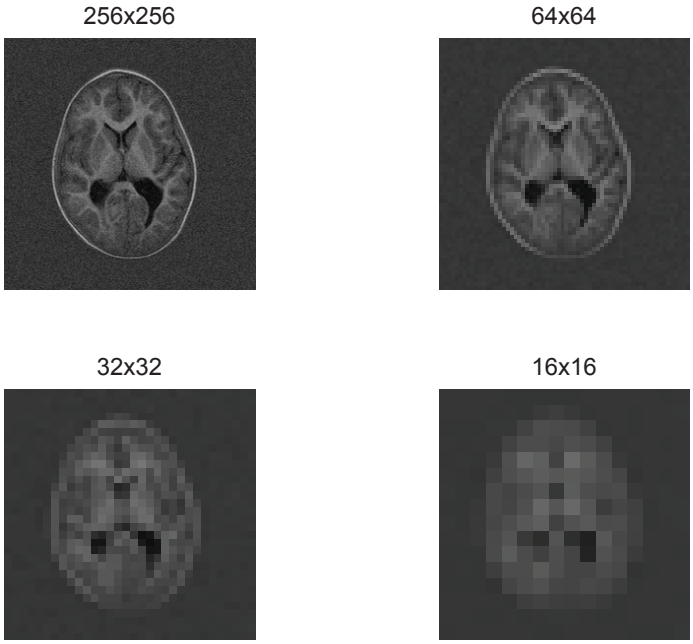


Figura 7.2: Imagen de Resonancia Magnética de un corte axial del cerebro. En las distintas imágenes se puede apreciar la pérdida de información y de calidad de imagen al reducir la resolución espacial.

$M \times N$ píxeles	n bits	Memoria requerida (*bytes*)
16×16	1	32
16×16	8	256
16×16	16	512
32×32	1	128
32×32	8	1024
32×32	16	2048
64×64	1	512
64×64	8	4096
64×64	16	8192
256×256	1	8192
256×256	8	65536
256×256	16	131072

Tabla 7.1: Los requisitos en memoria varían en función de la resolución espacial y la resolución en intensidad.

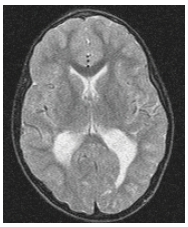
7.3. Tratamiento digital de imágenes médicas

Los objetivos principales del tratamiento digital de imágenes médicas son: mejorar la apariencia visual de las imágenes, preparar las imágenes para un posterior análisis e identificar, en la medida de lo posible, formas o patrones relevantes dentro de las imágenes. Por ejemplo, un tumor o anomalías fisiológicas, entre otras.

Muchas de las técnicas que se emplean para tratar las imágenes se basan en la información que proporcionan los histogramas de las imágenes. Un **histograma** es una función que asocia a cada nivel de gris su frecuencia absoluta de aparición en la imagen. Esto es,

$$H(x) = \frac{NP(x)}{NP},$$

donde $NP(x)$ es el número de píxeles que encontramos con intensidad x y NP es el número total de píxeles. En definitiva, el histograma es una gráfica que muestra las frecuencias de aparición de cada posible intensidad de gris. Cuando tenemos imágenes en formato RGB es posible, con algunas aplicaciones, obtener los histogramas de los niveles de intensidad de rojo, verde y azul.



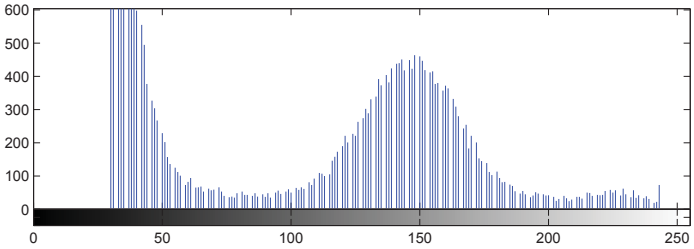


Figura 7.3: Arriba, imagen de resonancia magnética de un corte axial del cerebro. Debajo, el histograma de intensidades de grises asociado a la imagen de resonancia magnética anterior. Se puede observar que la imagen está representada con 1 byte puesto que las intensidades de grises van de 0 a 255.

7.3.1. Brillo y contraste

Los cambios de brillo y contraste en las imágenes son métodos que se basan en una transformación lineal de los valores de intensidad de los distintos píxeles que conforman una imagen. Esto es, sobre cada pixel $P(i, j)$, con $i \in [0, M-1]$ y $j \in [0, N-1]$, se da una transformación T de modo que los nuevos valores de los píxeles son $S(i, j) = T(P(i, j))$.

El brillo se puede entender como la cantidad de luminancia de una imagen. Esto es, mayor brillo implica niveles de gris más claros. Como habitualmente el valor del negro es 0 y el valor del blanco es 255 (con representación de 1 byte), para aumentar el brillo se debe sumar una cantidad constante c a cada pixel de la imagen. Es decir,

$$S(i, j) = P(i, j) + c.$$

El contraste se puede ver como la diferencia relativa en intensidad entre distintos objetos de la imagen. Un contraste alto implica una diferenciación clara entre los distintos objetos (fondo y figura, por ejemplo). Un contraste bajo hará más difícil la diferenciación entre objetos. Aumentar el contraste significa multiplicar el valor de cada pixel de la imagen por una constante c. Esto es,

$$S(i, j) = P(i, j) \cdot c.$$

Obviamente, si el valor $S(i, j) > L$, donde L es el máximo valor de intensidad, entonces $S(i, j) = L$. Así pues, un contraste alto implica un histograma ancho y un contraste bajo implica un histograma estrecho.

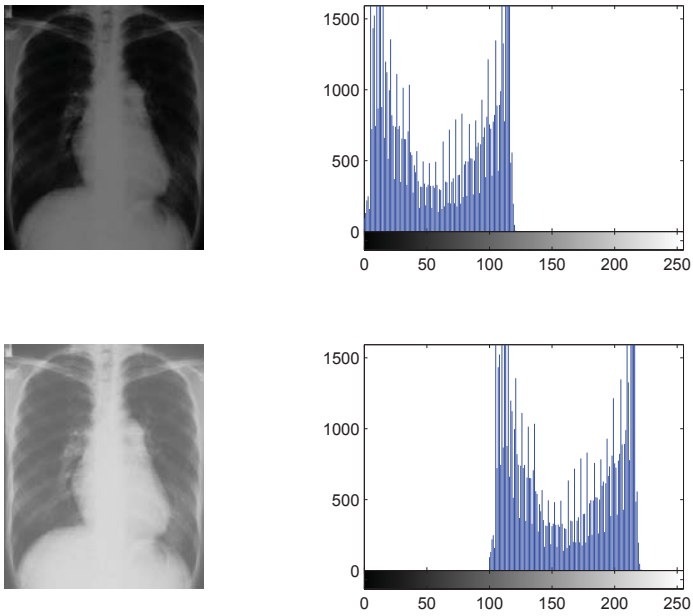


Figura 7.4: Arriba, una radiografía del pecho digitalizada con resolución espacial de 363×264 y 1 byte de resolución de intensidad. Abajo, la misma imagen con más brillo. El histograma se ve desplazado 100 posiciones a la derecha puesto que la intensidad de todos los píxeles ha sido aumentada en 100 puntos.

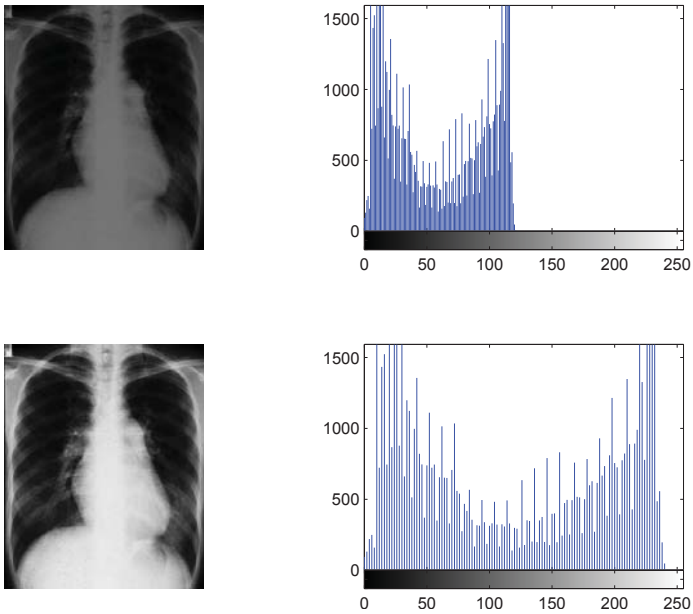


Figura 7.5: Arriba, la radiografía con un contraste menor. Su histograma es más estrecho que la imagen de abajo, donde se ha multiplicado la intensidad de todos los píxeles por 2. De este modo, la forma se distingue mejor del fondo.

7.3.2. Ecualización

La ecualización es una transformación que se aplica a todo el histograma. Su objetivo es obtener un histograma cuya distribución sea lo más uniforme posible conservando su entropía. Es decir, la ecualización maximiza el contraste de una imagen sin perder información de tipo estructural. Dicho de otro modo, se busca que todos los niveles de gris tengan un número de píxeles lo más equilibrado posible.

Aunque la ecualización puede producir imágenes poco realistas, es muy útil para tratar imágenes médicas. La ecualización se basa en una normalización de la función de distribución de probabilidad (cdf). La transformación para cada pixel es:

$$h(v) = \text{round}\left(\frac{cdf(v) - cdf_{min}}{(M \times N) - cdf_{min}} \cdot (L-1)\right),$$

donde $cdf(v)$ es el valor de la función de distribución para el valor de intensidad de gris v. Cualquier pixel con una intensidad v, $P(i,j) = v$ será transformado de manera que su nuevo valor de intensidad será $P(i,j) = h(v)$. Además, cdf_{min} es el valor mínimo de la función de distribución del histograma, M y N son las dimensiones de la imagen y L es el nivel máximo de intensidad de la imagen.

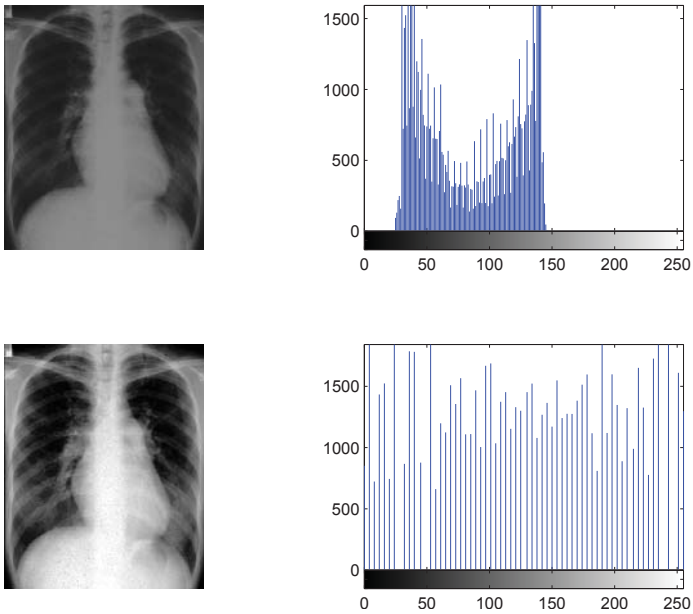


Figura 7.6: Arriba, la imagen original con un poco más de brillo. Abajo, se aprecia el resultado de ecualizar la imagen y el histograma correspondiente.

Ejemplo 7.2 (Ecualización de una imagen.)
Supongamos que se dispone de una imagen en escala de grises de 7×7 píxeles con 1 byte para la representación de la intensidad. La matriz para representar esta imagen es:

$$A = \begin{bmatrix} 48 & 50 & 59 & 87 & 77 & 52 & 73 \\ 68 & 55 & 56 & 90 & 108 & 84 & 61 \\ 60 & 63 & 68 & 109 & 146 & 98 & 65 \\ 61 & 49 & 71 & 118 & 150 & 104 & 71 \\ 70 & 70 & 70 & 112 & 128 & 90 & 75 \\ 81 & 66 & 58 & 78 & 76 & 60 & 62 \\ 82 & 75 & 64 & 62 & 60 & 57 & 54 \end{bmatrix}$$

Se puede obtener la frecuencia absoluta de cada nivel de gris con el histograma:

Así como los valores de la función de distribución acumulada:

Si quisiéramos obtener el nuevo valor de los píxeles con una intensidad de gris de 82 se debería aplicar la fórmula de la siguiente manera:

$$cdf(82) = 36.$$

$$h(78) = \text{round}\left(\frac{36-1}{49-1} \cdot 255\right) = 186.$$

Int	Frec	Int	Frec	Int	Frec	Int	Frec
48	1	49	1	50	1	52	1
54	1	55	1	56	1	57	1
58	1	59	1	60	3	61	2
62	2	63	1	64	1	65	1
66	1	68	2	70	3	71	2
73	1	75	2	76	1	77	1
78	1	81	1	82	1	84	1
87	1	90	2	98	1	104	1
108	1	109	1	112	1	118	1
128	1	146	1	150	1		

Tabla 7.2: En esta tabla se puede observar la frecuencia absoluta de los niveles de intensidad de gris de la imagen.

Esto quiere decir que las intensidades de gris con valor 82 deberán tomar el valor 186. Así, la matriz de la imagen ecualizada quedaría:

$$E = \begin{bmatrix} 0 & 12 & 49 & 194 & 170 & 16 & 150 \\ 113 & 28 & 32 & 202 & 223 & 190 & 73 \\ 61 & 89 & 113 & 227 & 251 & 210 & 101 \\ 73 & 4 & 142 & 239 & 255 & 219 & 142 \\ 125 & 125 & 125 & 235 & 243 & 202 & 158 \\ 182 & 105 & 45 & 174 & 166 & 61 & 81 \\ 186 & 158 & 97 & 81 & 61 & 36 & 20 \end{bmatrix}$$

7.3.3. Segmentación

La segmentación es un proceso que trata de dividir una imagen digital en múltiples segmentos con el objetivo de detectar estructuras o simplificar la representación de la imagen para poder mejorar el análisis de la misma. Existen multitud de métodos para llevar a cabo una buena segmentación y es, todavía, un campo de investigación de métodos nuevos para segmentar imágenes. Para profundizar en estos métodos de segmentación, aplicados a imagen en tumores cerebrales, pero extrapolables a otras aplicaciones se puede consultar la cita [104].

Las técnicas más sencillas para segmentar son las basadas en umbrales. Estas son técnicas que modifican los valores de aquellos píxeles cuyo valor de intensidad de gris se encuentra entre unos umbrales inferior y superior. A estos píxeles se les da un valor extremo (negro o blanco) y a los píxeles cuyos valores de intensidad caigan fuera de los umbrales se los sustituye por el otro extremo.

Int	cdf	Int	cdf	Int	cdf	Int	cdf
48	1	49	2	50	3	52	4
54	5	55	6	56	7	57	8
58	9	59	10	60	13	61	15
62	17	63	18	64	19	65	20
66	21	68	23	70	26	71	28
73	29	75	31	76	32	77	33
78	34	81	35	82	36	84	37
87	38	90	40	98	41	104	42
108	43	109	44	112	45	118	46
128	47	146	48	150	49		

Tabla 7.3: Las frecuencias absolutas acumuladas de la imagen conforman la distribución acumulada. En base a esta tabla se pueden obtener los valores de transformación de cada pixel en base a su nivel de intensidad.

Ejemplo 7.3 (Segmentación por umbrales)
Se tiene una imagen de resonancia magnética de un corte coronal del cerebro. El objetivo de una segmentación podría ser conseguir detectar o separar lo que es materia blanca del tálamo, el ventrículo lateral y la materia gris.

Para ello, basta aplicar un umbral inferior cercano a 80 y un umbral superior en 255. De este modo se obtiene el resultado que se muestra en la figura.

Existen algoritmos para establecer umbrales óptimos con los que minimizar la varianza intra-clase de los píxeles blancos y negros obteniendo una segmentación óptima. Algunos algoritmos para segmentar partes de la imagen son mucho más sofisticados, permitiendo seleccionar varios pares de umbrales.

El algoritmo de segmentación por crecimiento de regiones permite establecer un criterio para admitir píxeles a partir de una semilla, de modo que la región pueda aumentar si los píxeles de alrededor cumplen la condición definida por el criterio de admisión. Generalmente, se proporciona una semilla s y un umbral θ. De este modo, solo se incorpora el pixel $P(i, j)$ a la región segmentada cuando se cumple que $|s - P(i, j)| \leq \theta$.

Ejemplo 7.4 (Segmentación por crecimiento de regiones.)
A un paciente se le realiza un corte sagital del cerebro mediante una imagen de resonancia magnética. Se sospecha que pueda tener un tumor cerebral y para ello se emplea un algoritmo de segmentación por crecimiento de regiones. En una resonancia magnética potenciada en T1 los tumores suelen aparecer más oscuros que la materia gris del cerebro. Si se establece un valor de intensidad de gris de $s = 65$ y un umbral $\theta = 40$, se puede conseguir una segmentación de la zona tumoral. La zona segmentada se ha pintado con un color amarillo para resaltarla.

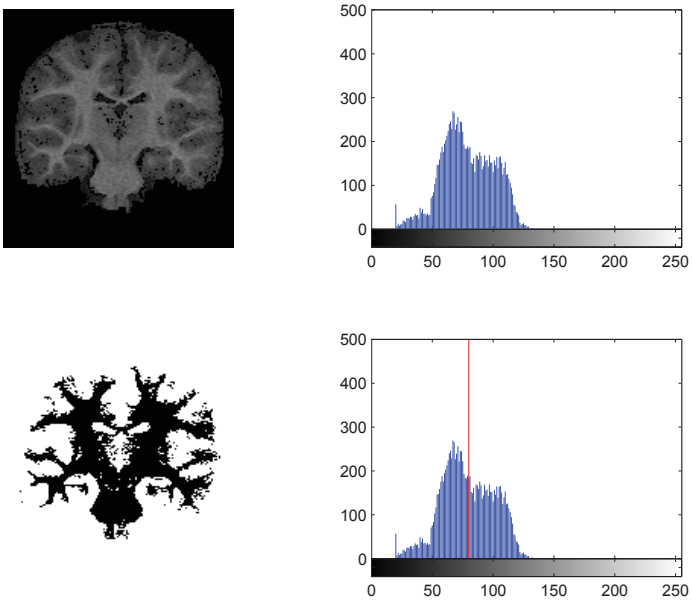


Figura 7.7: Segmentación de la imagen. Los píxeles cuyas intensidades caen fuera de los umbrales se tornan blancos y los píxeles cuyas intensidades están dentro del rango de los umbrales, se sustituyen por negro. De este modo, se ha diferenciado una estructura del cerebro del resto.

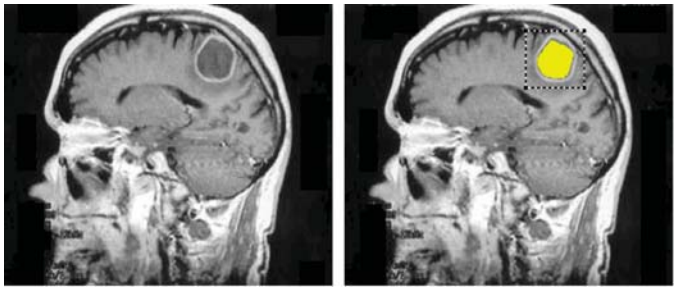


Figura 7.8: Resultado de emplear el algoritmo de segmentación por crecimiento de regiones para segmentar el tumor localizado en la zona parietal del cerebro a partir de una imagen de resonancia magnética potenciada en T1.

7.4. Técnicas de filtrado

El filtrado de imágenes es una operación local que se basa en la información que aportan los píxeles vecinos. Es decir, el valor del pixel filtrado se calcula como una combinación lineal de los píxeles de alrededor. Estos filtros utilizan, por lo tanto, una convolución en dos dimensiones de manera análoga al filtrado de señales en una dimensión.

En el filtrado de imágenes digitales se necesita una matriz $\mathbf{W}$, conocida como **máscara**, con los coeficientes del filtro. Con esta matriz se realiza la convolución con la matriz de la imagen siguiendo la expresión:

$$S(i,j) = \sum_{m=-d_m}^{d_m} \sum_{n=-d_n}^{d_n} w(m,n) \cdot P(i+m, j+n).$$

Ejemplo 7.5 (Aplicación de un filtro.)
Un filtro clásico para imagen digital es el filtro *Sobel*, que es un operador diferencial discreto -horizontal o vertical- que calcula una aproximación al gradiente de la función de intensidad de una imagen. Este filtro se presenta mediante la siguiente matriz:

$$\mathbf{W}_{Sobel} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Suponiendo que tuviésemos una matriz que representa un subconjunto de los píxeles de una imagen:

$$\mathbf{I} = \begin{bmatrix} 124 & 72 & 52 & 48 & 0 \\ 101 & 49 & 62 & 48 & 12 \\ 69 & 59 & 54 & 10 & 22 \\ 45 & 53 & 43 & 17 & 68 \\ 67 & 55 & 8 & 33 & 122 \end{bmatrix}$$

podríamos aplicar el filtro tal y como se describe en la figuras 7.9 y 7.10.

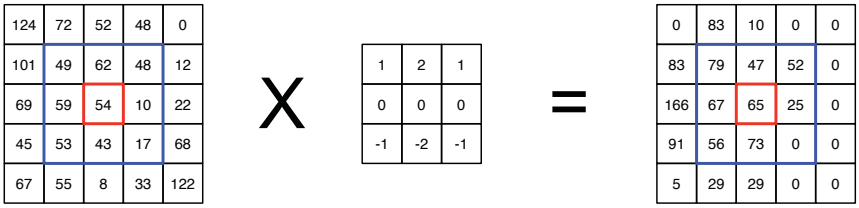


Figura 7.9: Al aplicar sobre el pixel de la tercera fila y la tercera columna el filtro *Sobel*, se obtiene el valor correspondiente a aplicar la siguiente operación: $49 \cdot 1 + 62 \cdot 2 + 48 \cdot 1 + 59 \cdot 0 + 54 \cdot 0 + 10 \cdot 0 + 53 \cdot (-1) + 43 \cdot (-2) + 17 \cdot (-1) = 65$.

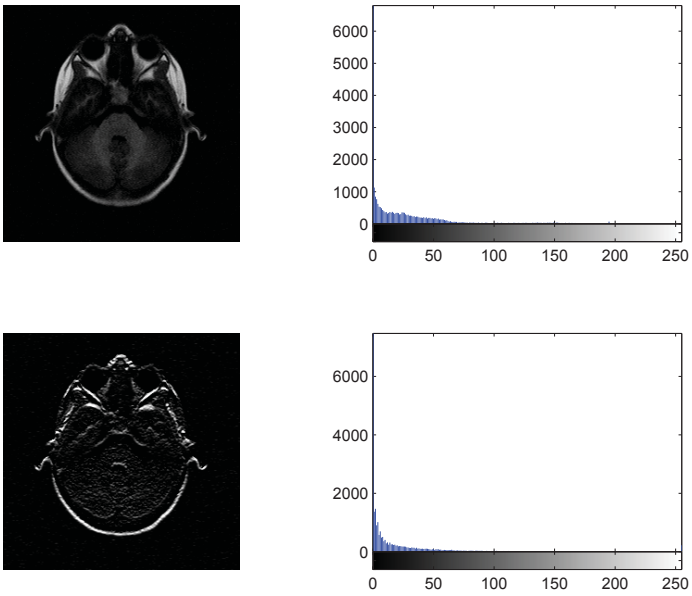


Figura 7.10: Al aplicar el filtro de contorno horizontal de *Sobel* sobre una imagen de resonancia magnética completa de un corte axial del cerebro se obtiene el resultado observado.

Como es natural, los filtros que se pueden aplicar a una imagen son prácticamente infinitos. Sin embargo, la aplicación para todos ellos sigue la misma estrategia. Entre los múltiples tipos de filtros podemos distinguir al menos tres clases: filtros de contorno, filtros paso bajo y filtros paso alto.

7.4.1. Filtros de contorno

Los **filtros de contorno** son aquellos que realzan los márgenes de las figuras de la imagen. Se ha visto como ejemplo el efecto del filtro *Sobel* horizontal. Este filtro puede realzar también los márgenes verticales sin más que aplicar la matriz traspuesta del filtro. Casi todos estos filtros pueden realzar los márgenes horizontales o verticales según se aplique la matriz traspuesta o no.

Algunos de estos filtros son:

- Filtro Sobel:

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

- Filtro de diferencia de píxeles:

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- Filtro Roberts:

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- Filtro Prewitt:

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

En la figura 7.11 se puede ver qué resultados se obtienen al aplicar algunos de estos filtros a una imagen médica.

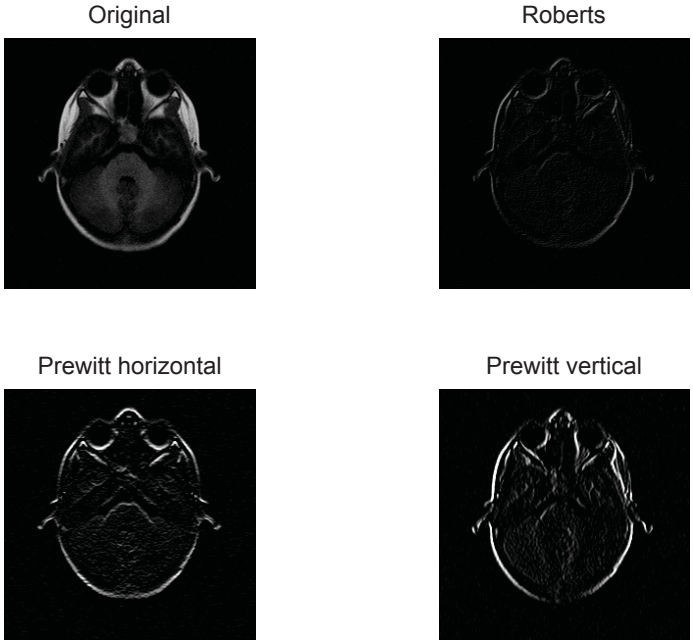


Figura 7.11: Resultados al aplicar distintos filtros de contorno sobre la imagen de resonancia magnética.

7.4.2. Filtros de paso bajo

El objetivo de los **filtros de paso bajo** es eliminar el ruido de las imágenes. Esto es, suavizar la imagen. El ruido son variaciones sobre los niveles de gris de la imagen que corresponden a señales de alta frecuencia. Por ello, las matrices que se utilizan como filtros para eliminar ruido están simulando la aplicación de filtros paso bajo. Existen varios tipos de filtros posibles:

- Filtro de media: se combinan linealmente todos los píxeles vecinos del pixel central otorgando el mismo peso a cada uno de ellos. Si la matriz del filtro es de 3×3, cada pixel se multiplicará por $1/9$.

$$\begin{bmatrix} {}^1/_9 & {}^1/_9 & {}^1/_9 \\ {}^1/_9 & {}^1/_9 & {}^1/_9 \\ {}^1/_9 & {}^1/_9 & {}^1/_9 \end{bmatrix}$$

- Filtro de media ponderada: es similar al filtro anterior, pero se le da más peso al pixel central para evitar que la imagen aparezca borrosa.

$$\begin{bmatrix} {}^1/_{10} & {}^1/_{10} & {}^1/_{10} \\ {}^1/_{10} & {}^2/_{10} & {}^1/_{10} \\ {}^1/_{10} & {}^1/_{10} & {}^1/_{10} \end{bmatrix}$$

- Filtro Gaussiano: es un filtro que proporciona un buen suavizado manteniendo la nitidez de la imagen. El filtro Gaussiano simula una distribución gaussiana bivariante empleando la expresión

$$w(x, y) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma^2}\right\},$$

  donde x es la distancia horizontal al punto central de la matriz u origen e y es la distancia vertical. Además, la matriz máscara debe normalizarse para que sume 1. Por ejemplo, un filtro Gaussiano de tamaño 3×3 y parámetro $\sigma = 0{,}5$ quedaría:

$$\begin{bmatrix} 0{,}0113 & 0{,}0838 & 0{,}0113 \\ 0{,}0838 & 0{,}6193 & 0{,}0838 \\ 0{,}0113 & 0{,}0838 & 0{,}0113 \end{bmatrix}$$

7.4.3. Filtros de paso alto

El objetivo de los **filtros de paso alto** es resaltar las zonas de mayor variabilidad. Por lo tanto, realizan el efecto contrario a los filtros paso bajo, eliminando las bajas frecuencias en lugar de las altas frecuencias. Si los elementos de los filtros paso alto suman menos de 1, entonces el resultado de aplicar el filtro debe sumarse a la matriz original para obtener un efecto de enfoque. También existen varios filtros posibles:

- Filtro media opuesta: este filtro se consigue restando a la matriz del filtro indentidad[c] la matriz del filtro de media. Los elementos de la máscara suman 0. Esto quiere decir que, si el objetivo es enfocar la imagen, se deben sumar la imagen original y el resultado de aplicar el filtro a dicha imagen.

[c] La matriz del filtro indentidad pondera con 1 el pixel central y los píxeles de alrededor son multiplicados por 0, de modo que una imagen filtrada con este filtro no varía.

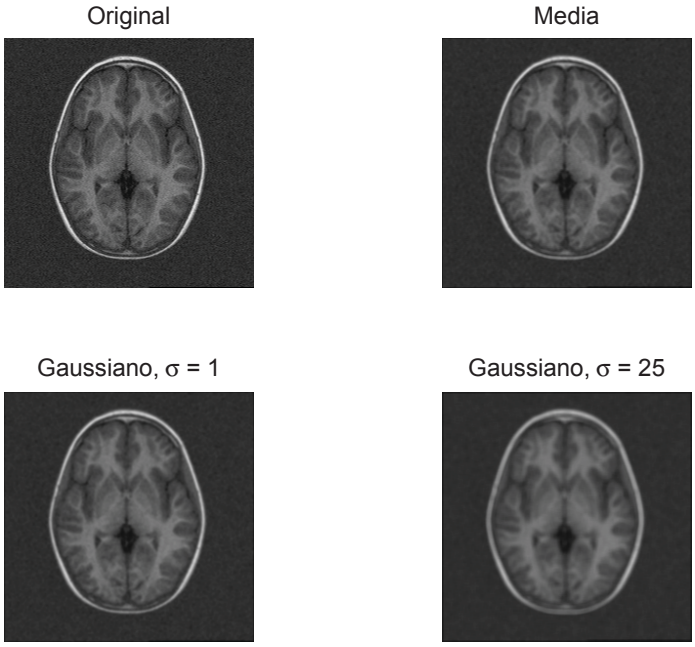


Figura 7.12: Resultados al aplicar distintos filtros paso bajo sobre la imagen de resonancia magnética.

$$\begin{bmatrix} {}^{-1}/_{9} & {}^{-1}/_{9} & {}^{-1}/_{9} \\ {}^{-1}/_{9} & {}^{8}/_{9} & {}^{-1}/_{9} \\ {}^{-1}/_{9} & {}^{-1}/_{9} & {}^{-1}/_{9} \end{bmatrix}$$

- Filtro Laplaciano: se obtiene aproximando la forma de un operador laplaciano bidimensional,

$$w(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}.$$

  su representación matricial, cuya suma de elementos es 0, es:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

- Filtro Laplaciano inverso: suele emplearse para mejorar el contraste. Esto se consigue eliminando las bajas frecuencias espaciales que son las que restan nitidez a la imagen. Se calcula a partir de un filtro negativo laplaciano en base a un parámetro $\alpha \in [0,1]$:

$$\frac{1}{\alpha+1}\begin{bmatrix} -\alpha & \alpha-1 & -\alpha \\ \alpha-1 & \alpha+5 & \alpha-1 \\ -\alpha & \alpha-1 & -\alpha \end{bmatrix}.$$

Si $\alpha = 0$ la máscara quedará así:

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

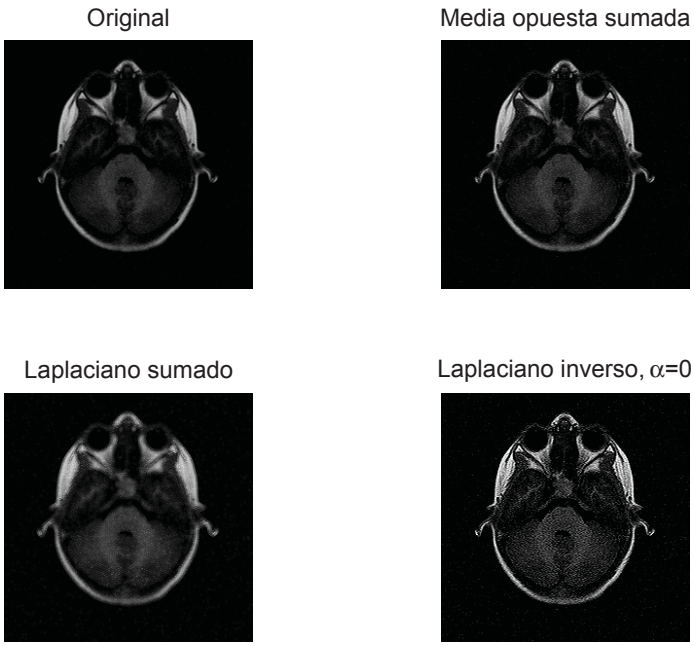


Figura 7.13: Resultados al aplicar distintos filtros paso alto sobre la imagen de resonancia magnética.

7.4.4. Operaciones con imágenes

En ocasiones, las información relevante puede estar contenida en varias imágenes en lugar de una única imagen. Para obtener la información deseada puede ser útil operar con las distintas imágenes empleando algunos operadores como los que se explican a continuación. Originalmente, los operadores trabajaban con formatos de doble precisión. Pero esto puede suponer cuatro veces más requerimientos computacionales. Actualmente, para reducir los requisitos computacionales, muchas herramientas proporcionan operadores para formatos de 1 y 2 bytes.

Sean dos imágenes I_1 e I_2, con el mismo formato de representación de la información, sobre las que se aplicarán las operaciones dando como resultado una tercera imagen S. Los píxeles en la posición (i, j) de las imágenes se representarán como $I_1(i, j)$, $I_2(i, j)$ y $S(i, j)$. Por último, sea L la máxima intensidad que se puede representar según el formato de las imágenes. Entonces, algunas de las operaciones que se pueden aplicar a imágenes serán:

- Suma de imágenes: $S(i, j) = I_1(i, j) + I_2(i, j)$.
- Diferencia de imágenes: $S(i, j) = I_1(i, j) - I_2(i, j)$.
- Diferencia absoluta de imágenes: $S(i, j) = |I_1(i, j) - I_2(i, j)|$.
- Complemento de una imagen: $S(i, j) = L - I_1(i, j)$.
- Producto de imágenes: $S(i, j) = I_1(i, j) \cdot I_2(i, j)$.
- División de imágenes: $S(i, j) = I_1(i, j)/I_2(i, j)$.
- Máximo de imágenes: $S(i, j) = \max(I_1(i, j), I_2(i, j))$.
- Mínimo de imágenes: $S(i, j) = \min(I_1(i, j), I_2(i, j))$.

7.5. Formato de datos DICOM

Aunque no se va a profundizar en este tema, es conveniente mencionar el estándar que se está imponiendo para la transmisión y almacenamiento de las imágenes médicas: el estándar DICOM[105][d]. DICOM responde a las siglas *D*igital *I*maging and *Co*mmunications in *M*edicine. El método estándar para la transmisión de imágenes médicas y la representación de la información asociada apareció en 1985 a partir de la colaboración del American College of Radiology (ACR) y el National Electrical Manufacturers Association (NEMA). Además, el estándar DICOM lleva desde 1999 trabajando en conjunto con el estándar de interoperabilidad en medicina HL7.

DICOM ha sido adoptado ampliamente por hospitales y otros sistemas sanitarios. Además, múltiples fabricantes y proveedores de maquinaria para la obtención de imágenes médicas tienen una declaración de conformidad con el formato DICOM que establece claramente las clases DICOM que soportan. DICOM permite integrar la información que se obtiene a partir de escáneres médicos, servidores en red, impresoras, etcétera en un sistema de almacenamiento llamado PACS (Picture Archiving and Communication System).

Un objeto de datos DICOM consiste en un conjunto de atributos más la información de la imagen. Un objeto DICOM sencillo puede contener solo la imagen. Sin embargo, también puede contener múltiples imágenes de un mismo objeto permitiendo almacenar imágenes en tres dimensiones e, incluso, imágenes en movimiento. Los píxeles de la imagen se pueden comprimir usando diversos formatos de compresión, incluyendo JPEG o RLE.

Al agrupar la información en conjuntos de datos DICOM difiere de otros formatos de imagen. Por ejemplo, un fichero que contenga un imagen de rayos-X contendrá, además, el identificador del paciente dentro del mismo fichero, así como los parámetros de adquisición de la imagen. De este modo, la imagen nunca irá separada de la información del paciente o de la forma en que se adquirió, evitando posibles errores en el futuro y facilitando, en algunos casos y si fuese necesario, el procesado posterior de la imagen.

[d] Se puede encontrar toda la información necesaria en la URL `http://www.dicomstandard.org/`.

7.6. Notas bibliográficas

Beutel, Sonka y Fitzpatrick recopilan las fuentes de imagen médica y las principales técnicas de su procesamiento en [106]. En Semmlow [107] encontramos técnicas de procesamiento de señales e imágenes biomédicas basadas en Matlab.

Capítulo 8

Aprendizaje automático para la ayuda a la decisión médica

Supongamos un problema típico de decisión médica: diagnosticar a un paciente tras observar un conjunto de datos biomédicos. El mismo médico podría tener que decidir el mejor tratamiento que debe aplicar a la persona que tiene esperando su decisión o plantearse el pronóstico del paciente (por ejemplo, dosis de un fármaco). El sistema de ayuda a la decisión más intuitivo de imaginar es aquel que predice el diagnóstico, pronóstico o tratamiento adecuado utilizando la información disponible relacionada con el paciente: conocimiento previo y datos biomédicos procedentes del paciente.

Para llegar a disponer de esta ayuda debemos implementar una función que asigne la respuesta que maximice el beneficio del paciente teniendo en cuenta la información observada. En este capítulo, trataremos el diseño de esta función mediante un conjunto limitado de casos previamente observados. Abordaremos por lo tanto el diseño de sistemas de ayuda a la decisión mediante el *aprendizaje automático*.

Una vez descrita la metodología de aprendizaje automático a nivel de bloque, profundizaremos en el diseños de los modelos predictivos. Partiremos de la teoría de la decisión y la teoría de la probabilidad, que nos proporcionarán una base racional para tomar las decisiones óptimas en los problemas.

8.1. Diseño de modelos de predicción

Un Clinical Decision Support Systems (CDSS) produce un resultado útil para la decisión de un profesional sanitario sobre un paciente a partir de la información disponible. El aprendizaje automático se basa en el uso de patrones o regularidades en la información de los casos para ajustar los modelos predictivos y, posteriormente, adoptar decisiones tales como su clasificación automática (por ejemplo, diagnosticar el tipo de arritmia de un paciente mediante hallazgos en su ECG).

Estas regularidades pueden ser especificadas en forma de reglas mediante conocimiento experto de mayor o menor nivel de evidencia. El capítulo 16 profundiza en la especificación de conocimiento experto para su uso en razonadores semánticos para la ayuda a la decisión médica. Esto permitiría aplicar las reglas a nuevos casos y realizar la decisión sobre el caso concreto basado en la evidencia acumulada en las fuentes de conocimiento.

Podríamos también descubrir estas regularidades en casos del dominio del problema. Para ello, se utiliza un conjunto representativo de casos del dominio del problema, llamado *corpus* o *dataset*, que permite estimar las regularidades de la muestra con la que especificar una regla

con la que resolver el problema de decisión en nuevas observaciones. Esta forma de proceder se estudia en la disciplina del aprendizaje automático, que suele constar de las etapas descritas en la figura 8.1. Entre estas etapas consta una fase de entrenamiento o ajuste de los modelos predictivos para el CDSS. Esta fase partirá de un corpus de entrenamiento con datos del dominio del problema. Posteriormente, los modelos deberán ser validados con un corpus de evaluación para poder estimar la capacidad predictiva de los modelos entrenados.

En primer lugar se realizará el preprocesado de los casos para adecuarlos al uso posterior. Esta etapa suele incluir diferentes pasos de procesado de datos multivariantes o de procesado de señales biomédicas como los estudiados en el capítulo 6.

Una vez preprocesados los casos, se realizará una etapa de selección y extracción de características que tiene como objetivo la obtención de la mejor representación de los casos para resolver el problema. Esta etapa implica generalmente la comparación de varios métodos alternativos de selección y extracción de características, como los estudiados en el capítulo 4. Alternativamente, también pueden extraerse características interesantes mediante métodos de detección de hallazgos en señales biomédicas (por ejemplo, parámetros del complejo QRS de un ECG) o cuantificación de componentes (por ejemplo, concentración de metabolitos en un Magnetic Resonance Spectroscopy (MRS)).

El paso siguiente a la selección de características consiste en la estimación del modelo predictivo, que establecerá la regla de decisión del CDSS. Este es el propósito del capítulo actual, y supone el núcleo de la metodología del aprendizaje automático. Generalmente, esta etapa consistirá, al igual que la etapa anterior, en la comparación de diferentes métodos de predicción con el fin de elegir aquel con mayor capacidad de generalización. Para la selección del método de clasificación podremos aplicar los métodos descritos en la sección 14.4.

Para un correcto desarrollo del modelo predictivo se deberá aplicar una estricta metodología de evaluación que incluya, tanto la etapa de selección/extracción de características, como la etapa de estimación del predictor. El capítulo 14 está dedicado en su mayor parte a estrategias y métricas para una correcta evaluación de los modelos predictivos.

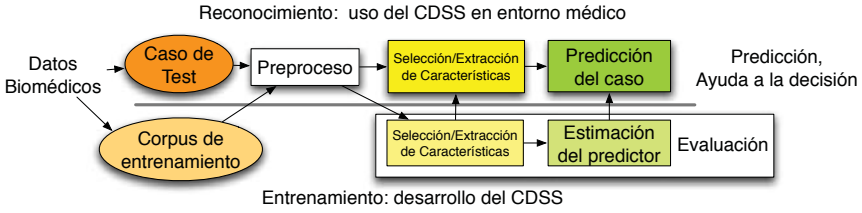


Figura 8.1: Etapas para el diseño de modelos predictivos.

Una vez obtenido el modelo de predicción el sistema puede implementarse para su uso dentro del CDSS en un entorno médico. En esta fase, la entrada será un nuevo caso de test, nunca visto por el modelo predictivo. Para que el caso sea compatible con el modelo predictivo, deberá realizarse el mismo preproceso que durante el entrenamiento y aplicar el método de selección/extracción de características seleccionado. Con el resultado de estos pasos podremos invocar la regla de decisión que implementa el modelo predictivo, obteniendo el resultado de decisión para el caso particular de estudio.

8.2. Problemas de aprendizaje automático

Tras el modelo general de decisión presentado en la sección 3.3, podemos estudiar algunos problemas tipo como la *clasificación, regresión* y *clustering* que han sido ampliamente tratados por la disciplina del aprendizaje automático.

Vapnik [108] describe el problema general de aprendizaje automático mediante tres elementos:

- Un generador produce muestras $x \in \mathcal{X}$ siguiendo una función de distribución de probabilidad $F(x)$.
- Un supervisor devuelve un valor $y \in \mathcal{Y}$ tras observar cada muestra x, siguiendo una función de distribución condicional $F(y|x)$.
- Un decisor que recibe como entrada un conjunto de pares (x_i, y_i) de acuerdo a la distribución conjunta $F(x, y) = F(x)F(y|x)$. Este decisor es capaz de implementar un conjunto de funciones $\hat{y} = f(x, \alpha), \alpha \in \Lambda$ para aproximar el valor y producido por el supervisor.

Las consecuencias de elegir $\hat{y} = f(x, \alpha)$ cuando el decisor observa x, pueden expresarse mediante una función de pérdida (o utilidad) $L(y, \hat{y})$, que conforme a la ecuación (3.37), asume un *riesgo condicional*

$$R(\hat{y}|x) = E_{y|x}[L(y, f(x, \alpha))], \tag{8.1}$$

que para variables continuas será:

$$R(\hat{y}|x) = \int L(y, f(x, \alpha))p(y|x)dy, \tag{8.2}$$

y para variables discretas:

$$R(\hat{y}|x) = \sum_{y \in \mathcal{Y}} L(y, f(x, \alpha))p(y|x). \tag{8.3}$$

Ante una observación x, el mínimo riesgo que asume el decisor viene dado por la opción que minimiza el riesgo condicional, por lo que la *regla de decisión* se define como

$$\hat{y}^* \leftarrow \arg\min_{\hat{y} \in \mathcal{Y}} R(\hat{y}|x). \tag{8.4}$$

En general, el decisor asumirá un *riesgo funcional* en su tarea igual al valor esperado de la pérdida de decidir $\hat{y}$ cuando el supervisor dice y, sobre los posibles valores conjuntos (x, y),

$$R(\alpha) = E_x[E_{y|x}[L(y, f(x, \alpha))]] = E_{x,y}[L(y, f(x, \alpha))]. \tag{8.5}$$

Por ejemplo, suponiendo que x e y toman valores continuos y condicionando sobre x, la ecuación (8.5) se puede expresar como

$$R(\alpha) = \int L(y, f(x, \alpha))p(x, y)dxdy = \int L(y, f(x, \alpha))p(x)p(y|x)dxdy. \tag{8.6}$$

De forma general, el objetivo del aprendizaje automático es encontrar la función cuyo parámetro α minimice el riesgo según la ecuación (8.5).

8.2.1. Clasificación

En un problema de clasificación, un decisor (llamado clasificador) que dispone de la información $x \in \mathcal{X}$ de un nuevo caso, se pregunta cual es la etiqueta y elegida por el supervisor de un conjunto $\mathcal{Y} = \{y_1, \ldots, y_C\}$ (conjunto de sucesos inciertos). Por lo tanto, para un problema de clasificación, el espacio de decisiones $\mathcal{D}$ es el mismo que el espacio $\mathcal{Y}$.

Un clasificador puede representarse mediante un conjunto de C *funciones discriminantes*

$$g = \{g_i : \mathcal{X} \to \mathbb{R}, i = 1, \ldots, C\}, \tag{8.7}$$

asociadas al conjunto de clases $\mathcal{Y}$, y que tras observar x de una nueva muestra, asigna la clase y_i como

$$\hat{y}_i \leftarrow \arg\max_i g_i(x), \tag{8.8}$$

$$\hat{y}_i \leftarrow \arg\min_i h_i(x), \tag{8.9}$$

$$h_i(x) \propto -g_i(x). \tag{8.10}$$

El clasificador óptimo α^* es aquel cuyas funciones discriminantes corresponden a las funciones de riesgo condicional (8.1),

$$h_i(x) = R(\hat{y}_i|x), i = 1, \ldots, C. \tag{8.11}$$

Regiones de decisión y fronteras de decisión

Como hemos visto en la ecuación (8.4), la regla de decisión asigna un valor $\hat{y}$ a cada observación x. Como consecuencia, el espacio muestral $\mathcal{X}$ queda dividido en C *regiones de decisión* disjuntas asignadas a cada posible valor de $y_i \in \mathcal{Y}$. Así pues, el decisor asigna la clase y_i a una observación x que cae en una región de decisión $\mathcal{R}_i$, esto es,

$$\mathcal{R}_i = \{x : g_i(x) > g_j(x), \forall j \neq i\}, i = 1, \ldots, C. \tag{8.12}$$

Las características de esas regiones de decisión dependerán de la topología que pueda representar el método de clasificación.

La *frontera de decisión* entre dos clases, y_i e y_j, es la frontera entre las regiones de decisión de ambas clases. La frontera de decisión entre estas clases corresponde al conjunto de puntos donde las funciones discriminantes correspondientes, $g_i(x)$ y $g_j(x)$, son iguales:

$$F_{y_i,y_j} = \{x : g_i(x) = g_j(x)\}, \tag{8.13}$$

A su vez, la frontera de decisión óptima entre las clases y_i e y_j es el conjunto de puntos x donde el riesgo condicional de elegir $\hat{y}_i$ o $\hat{y}_j$ es igual, es decir,

$$F^*_{y_i,y_j} = \{x : R[\hat{y}_i|x] = R[\hat{y}_j|x]\}. \tag{8.14}$$

Clasificador de Bayes

En un problema de clasificación es razonable pensar que el decisor desea reducir la probabilidad de errar en la predicción de cada nuevo caso, por lo que la función de pérdida 0-1 (ecu. 12.7) parece acertada para este propósito. Como la decisión estará basada en la información observada x, se puede expresar la decisión como una función $f(x, \alpha)$ dependiente de x y con parámetros α. Por tanto, se puede expresar la función de pérdida 0-1 como:

$$L(y, f(x, \alpha)) = \begin{cases} 0 & \text{si } y = f(x, \alpha) \\ 1 & \text{si } y \neq f(x, \alpha) \end{cases} \tag{8.15}$$

Para la clasificación de un nuevo caso observado x, el decisor tendrá una pérdida esperada que viene dada por la ecuación (8.1) y un riesgo global (8.5),

$$R(\alpha) = \int \sum_{y \in \mathcal{Y}} L(y, f(x, \alpha)) p(y|x) p(x) dx. \tag{8.16}$$

La función de pérdida 0-1 hace que únicamente los términos del sumatorio donde $y \neq f(x, \alpha)$ sean distintos de 0. Mediante el suceso complementario se puede expresar el riesgo como,

$$R(\alpha) = \int \sum_{y \neq f(x,\alpha)} p(y|x) p(x) dx = \int (1 - p(y|x)) p(x) dx. \tag{8.17}$$

Podemos ver que $1 - p(y|x)$ es la probabilidad de error del clasificador dada la muestra x, también llamada *probabilidad de error a posteriori*, $p(\text{error}|x)$, por lo que el riesgo $R(\alpha)$ será el error esperado del clasificador, $p(\text{error})$, también llamado probabilidad de error o error de generalización.

Por lo tanto, el clasificador óptimo o de mínimo riesgo, asigna la clase $\hat{y}$ que minimiza el error esperado condicionado a x. Esto es equivalente a asignar la clase que maximiza la probabilidad condicional,

$$\hat{y}^* \leftarrow \arg\min_{y \in \mathcal{Y}} p(error|x) \tag{8.18}$$

$$\hat{y}^* \leftarrow \arg\max_{y \in \mathcal{Y}} p(y|x). \tag{8.19}$$

A este clasificador se le denomina *clasificador de Bayes*, cuyas funciones discriminantes son

$$g_i = p(y_i|x), i = 1, \ldots, C, \tag{8.20}$$

y sus fronteras de decisión

$$p(y_i|x) = p(y_j|x), i = 1, \ldots, C-1; j = i+1, \ldots, C \tag{8.21}$$

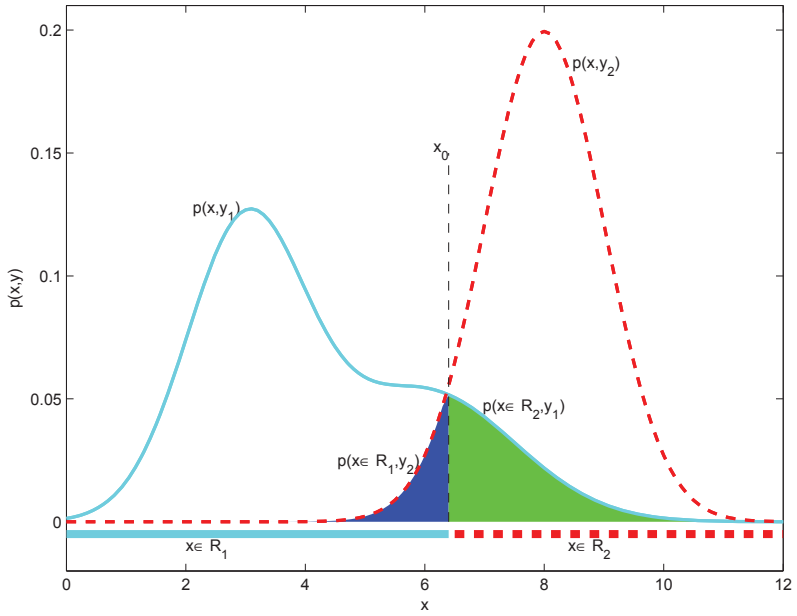


Figura 8.2: Clasificador de Bayes para dos clases y_1 e y_2 y observaciones unidimensinales x_0.

La figura 8.2 muestra un clasificador de Bayes para muestras de las que se observa una variable unidimensional $x \in \mathbb{R}$ en dos clases y_1 e y_2. El eje de abscisas representa el rango de interés del espacio de la variable x y el eje de ordenadas representa la probabilidad conjunta $p(x, y), y \in \{y_1, y_2\}$. La solución del problema define la región $x \in \mathcal{R}_1$ de decisión y_1 y la región $x \in \mathcal{R}_2$ de decisión y_2. La frontera de decisión entre ambas regiones es el punto x_0, donde $p(y_1|x) = p(y_2|x)$. Finalmente, podemos observar que la probabilidad de error $p(\text{error})$ es la suma de las áreas coloreadas en azul y verde que corresponden a la probabilidad de que una muestra de una clase caiga en la región de decisión asignada a la otra clase.

Función de pérdida asimétrica

El *clasificador de Bayes* (8.19), diseñado mediante la función de pérdida 0-1, supone una pérdida similar para cualquier tipo de error. Sin embargo, ciertas decisiones médicas pueden tener pérdidas asimétricas, como por ejemplo cuando se trata de diagnosticar un tumor como benigno o maligno. En este caso, decidir que el paciente tiene un tumor benigno cuando en realidad es maligno puede tener una pérdida asociada mayor que diagnosticarlo como maligno cuando realmente es benigno.

Para una clasificación en el espacio $\mathcal{Y} = y_1, \ldots, y_C$, la función de pérdida mediante la *tabla de pérdidas*, donde cada celda i, j define la pérdida provocada por elegir $\hat{y_j}$ cuando la clase correcta

es y_i:

$$L(y,\hat{y}) = \begin{array}{c|ccc} & \hat{y_1} & \cdots & \hat{y_C} \\ \hline y_1 & L_{1,1} & \cdots & L_{1,C} \\ \vdots & \vdots & \ddots & \vdots \\ y_C & L_{C,1} & \cdots & L_{C,C} \end{array} \tag{8.22}$$

En un problema de clasificación entre dos clases y_1, y_2 que está definido por la siguiente tabla de pérdidas asimétrica ($L_{1,2} \neq L_{2,1}$):

$$L(y,\hat{y}) = \begin{array}{c|cc} & \hat{y_1} & \hat{y_2} \\ \hline y_1 & 0 & L_{1,2} \\ y_2 & L_{2,1} & 0 \end{array} \tag{8.23}$$

El clasificador óptimo para este problema será:

$$\hat{y}^* = \begin{cases} y_1, & L_{2,1}p(x,y_2) < L_{1,2}p(x,y_1) \\ y_2, & L_{2,1}p(x,y_2) > L_{1,2}p(x,y_1), \end{cases}$$

Ejemplo 8.1 (Diagnóstico de tumores benignos y malignos)
Un grupo interdisciplinar de diagnóstico decide que el diagnóstico erróneo de un tumor maligno (m) supone una pérdida equivalente a cinco veces la pérdida de errar al diagnosticar un tumor benigno (b) como si fuera maligno.

Si definimos como unidad de pérdida aquella producida al diagnosticar incorrectamente un tumor benigno como maligno, la tabla de pérdida del problema queda como sigue:

$$L(y,\hat{y}) = \begin{array}{c|cc} & \hat{b} & \hat{m} \\ \hline y_b & 0 & 1 \\ y_m & 5 & 0 \end{array}$$

por lo que la regla de decisión será:

$$\hat{y}^* = \begin{cases} y_b, & 5p(x,y_m) < p(x,y_b) \\ y_m, & 5p(x,y_m) > p(x,y_b), \end{cases}$$

y la frontera de decisión será:

$$x_0 : 5p(y_m|x_0) = p(y_b|x_0).$$

La figura 8.3 representa el riesgo asumido al clasificar el tumor como benigno $R(\hat{y}_b)$ y al clasificar el tumor como maligno $R(\hat{y}_m)$. Podemos observar como el riesgo de clasificar como benigno cuando el tumor es maligno es muy alto debido a la función de pérdida asimétrica, lo que provoca que la frontera de decisión óptima x_0 se desplace hacia la derecha respecto a la frontera de decisión del clasificador de Bayes x_B, ampliando la región de decisión de y_m.

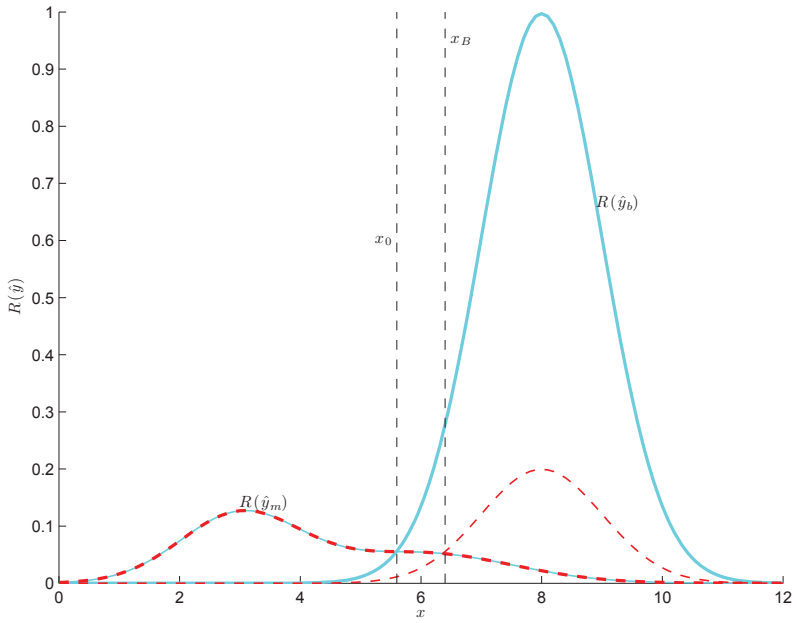


Figura 8.3: Clasificador con matriz de pérdida asimétrica.

8.2.2. Regresión

En un problema de regresión, un decisor que dispone de la información $\mathbf{x} \in \mathbb{R}^D$ de un nuevo caso, debe aproximar con $\hat{y} = \hat{f}(\mathbf{x})$ el valor $y \in \mathbb{R}$ elegida por el supervisor que responde a $y = f(\mathbf{x}) + \epsilon$. Si $y \in \mathbb{R}^{D'}$ entonces el problema se denomina regresión múltiple.

Una función de pérdida adecuada para el problema de regresión es la *función error cuadrático*,

$$L(y, \hat{f}(\mathbf{x})) = (y - \hat{f}(\mathbf{x}))^2, \tag{8.24}$$

por lo que el riesgo en la tarea de regresión será

$$R(\hat{f}(\mathbf{x})) = E_{x,y}[L(y, \hat{f}(\mathbf{x}))] = \int (y - \hat{f}(\mathbf{x}))^2 p(x, y) dx dy. \tag{8.25}$$

Como queremos minimizar el riesgo, podemos igualar su derivada respecto a $\hat{f}(x)$ a 0,

$$\frac{\partial R(\hat{f}(x))}{\partial \hat{f}(x)} = \frac{\partial}{\partial \hat{f}(x)} \int (y - \hat{f}(x))^2 p(x, y) dx dy = 0, \tag{8.26}$$

donde se ha fijado x al valor observado, por lo que si $R(\hat{f}(x))$ es mínimo, para cada x fijo, también cada término $\int (\hat{f}(x) - y)^2 p(x, y) dy$, debe ser mínimo,

$$\frac{\partial}{\partial \hat{f}(x)} \int (\hat{f}(x) - y)^2 p(x, y) dy = 2 \int (y - \hat{f}(x)) p(x, y) dy = 0, \tag{8.27}$$

por lo que

$$\hat{f}(x) = \frac{1}{p(x)} \int y p(x, y) dy = \int y p(y|x) dy = E_{y|x}[y], \tag{8.28}$$

por lo que la mejor solución $\hat{y} = \hat{f}(x)$, es el valor esperado de y condicionado a la observación de x.

8.2.3. Estimación de la densidad de probabilidad

En un problema de estimación de la densidad de probabilidad, el decisor desea conocer la densidad de probabilidad $p(x)$ de la cual procede un conjunto de muestras observadas $\mathcal{S}$. Para ello, asume que la distribución se encuentra representada en un conjunto de densidades $p(x|\alpha), \alpha \in \Lambda$.

Para este problema, la función de coste típicamente elegida es

$$L(p(x,\alpha)) = -\log p(x,\alpha). \tag{8.29}$$

8.3. El proceso de aprendizaje automático

Recordemos que ante una observación x la mejor decisión que se puede tomar es aquella que minimiza el riesgo condicional, tal como expresa la regla de decisión (8.4).

El riesgo esperado es una estimación de la pérdida esperada basada en el conocimiento de la distribución $p(y|x)$ de los sucesos inciertos y dada la observación x. Por lo tanto, cuanto mejor conozcamos esta distribución, mejor será el cálculo del riesgo y mejor la decisión tomada.

La dificultad viene dada por el desconocimiento de la distribución real de $p(y|x)$. Por lo tanto, debemos conseguir encontrar modelos que aproximen lo mejor posible dicha distribución. Precisamente, el proceso de aprendizaje automático permite aproximar, de forma implícita o explícita, dicha distribución mediante la incorporación de la información proporcionada por el conocimiento previo del problema y los experimentos realizados en relación con los sucesos inciertos.

8.3.1. Estimación por máxima verosimilitud

Para aproximar la función de probabilidad $p(y|x)$ dado un conjunto de observaciones independientes e idénticamente distribuidas (iid) $\mathcal{S} = \{\mathbf{s}_i\} = \{(\mathbf{x}_i, y_i)\},\ i = 1, \ldots, N$, con $\mathbf{x}_i \in \mathcal{X}^D$ e $y_i \in \mathbb{R}$, el método de máxima verosimilitud trata de maximizar la probabilidad con la que se obtendría el conjunto $\mathcal{S}$ si se generase con un modelo funcional $f(\mathcal{S}, \theta)$, siendo θ los parámetros asociados de dicho modelo. En otras palabras, la estimación por máxima verosimilitud consiste en estimar los parámetros θ de modo que la función $f(\mathcal{S}, \hat{\theta})$ alcance su valor máximo. Por lo tanto, la **función de verosimilitud** dependerá de los datos observados $\mathcal{S}$ y nos dará como resultado unos parámetros $\hat{\theta}$. Puesto que las variables observadas se asumen independientes la probabilidad conjunta será, por la ley multiplicativa (3.2), el producto de las probabilidades individuales:

$$\begin{aligned} \mathcal{L}(\theta|S) &= p(S|\theta) \\ &= \prod_{i=1}^{N} p(\mathbf{s}_i|\theta). \end{aligned} \tag{8.30}$$

Como se puede observar en la ecuación (8.30) maximizar la verosimilitud nos dará el *estimador máximo-verosímil* θ^* que será equivalente a maximizar el logaritmo de la verosimilitud:

$$\begin{aligned} \hat{\theta}^* &= \arg\max_{\theta} \mathcal{L}(\theta|S) \\ &= \arg\max_{\theta} \log \mathcal{L}(\theta|S) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \log p(\mathbf{s}_i|\theta). \end{aligned} \tag{8.31}$$

Para obtener $\hat{\theta}^*$ bastará con calcular las derivadas parciales de $\mathcal{L}(\mathcal{S}|\theta)$ respecto de los parámetros e igualar a cero en busca de las singularidades que hacen máxima la función de verosimilitud:

$$\frac{\partial \mathcal{L}(\theta|S)}{\partial \theta} = 0. \tag{8.32}$$

Una vez se han estimado los parámetros óptimos $\hat{\theta}^*$ del modelo, se podrá emplear para predecir los valores de observaciones futuras, $\mathbf{s}_{new}$, empleando $f(\mathbf{s}_{new}|\hat{\theta}^*)$.

Ejemplo 8.2 (Forma de un tumor de estirpe vascular observada por MRI))
Un hallazgo de interés en las imágenes de resonancia magnética cuando se estudian tumores de partes blandas suele ser la forma de la lesión, que indica la morfología externa de la misma. De forma categórica, podemos asumir que la forma de una lesión, puede tomar uno de los siguientes valores cualitativos: redonda, oval, lobulada, serpinginosa, fusiforme, cordón e irregular.

Un grupo multidisciplinar de diagnóstico por la imagen quiere caracterizar la forma de los tumores de estirpe vascular. Para ello, realiza un estudio aleatorio mediante el que obtiene un conjunto de 81 muestras cuya histologías fueron linfangioma, tumor glómico, hemangiopericitoma maligno, hemangioma o angioleiomioma.

Por el carácter cualitativo de la variable forma, x, deciden asumir una distribución multinomial con vector de parámetros $\mathbf{p}$ (la probabilidad de cada posible valor de la forma) y $L = 1$. El vector de probabilidades $\mathbf{p} = (p_1, \ldots, p_D)$ que define los parámetros de una distribución multinomial puede ser estimado mediante máxima verosimilitud sobre una muestra de N casos independientes $\mathcal{S} = (\mathbf{x_i}), i = 1, \ldots, N; \mathbf{x_i} \in \mathbb{N}^D, \sum_{d=1}^{D} x_i^{(d)} = L$, como

$$\begin{aligned}
\hat{\mathbf{p}} &= \arg\max_p p(\mathcal{S}|\mathbf{p}) && (8.33)\\
&= \arg\max_p \sum_{i=1}^{N} \log p(\mathbf{x_i}|\mathbf{p}) && (8.34)\\
&= \arg\max_p \sum_{i=1}^{N} \frac{L!}{\prod_{d=1}^{D} x^{(d)}!} \prod_{d=1}^{D} p_d^{x^{(d)}} && (8.35)\\
&= \frac{1}{NL} \sum_{i=1}^{N} \mathbf{x_i}. && (8.36)
\end{aligned}$$

De forma ilustrativa, un caso de forma oval estará representado por $(0, 1, 0, 0, 0, 0, 0)$.

La estimación por máxima verosimilitud de $\mathbf{p}$ puede calcularse mediante la ecuación (8.36) por conteo sobre el conjunto de muestras, obteniendo como resultado

$$\hat{\mathbf{p}} = (0{,}0370; 0{,}1235; 0{,}0988; 0{,}5679; 0{,}0370; 0; 0{,}1358).$$

Como crítica al método de máxima verosimilitud podemos observar en nuestros resultados que según la estimación basada en los 81 casos de nuestra base de datos, no es posible que un tumor de tipo vascular tenga forma de cordón, lo cual puede ser demasiado taxativo para cualquier resultado estadístico. Existen soluciones que corrigen estos resultados mediante *suavizado* [109], y aproximaciones que estiman la distribución de los posibles valores de los parámetros $\mathbf{p}$, en lugar de estimar un valor puntual, como veremos en el ejemplo 8.3.

Clasificador de máxima verosimilitud

La aproximación clásica del reconocimiento de formas describe el clasificador de Bayes (ver sección 8.2.1) en términos de probabilidades *a priori* y probabilidades *a posteriori*,

$$\begin{aligned}\hat{y}^* &= \arg\max_{y\in\mathcal{Y}} p(y|x)\\ &= \arg\max_{y\in\mathcal{Y}} p(y)p(x|y)\\ &= \arg\max_{y\in\mathcal{Y}} \log p(y) + \log p(x|y)\\ &= \arg\max_{y\in\mathcal{Y}} \log \hat{p}(y) + \log \hat{p}(x|y),\end{aligned}$$

donde $\hat{p}(y)$ y $\hat{p}(x|y)$ son estimados mediante máxima verosimilitud sobre el conjunto de muestras de entrenamiento $\mathcal{S} = (x_i, y_i), i = 1, \ldots, N$. Para cada posible clase en $\mathcal{Y} = y_1, \ldots, y_C$, la probabilidad a prior se puede estimar como $\hat{p}(y) = N_c/N$, y los parámetros de la función de densidad de probabilidad se calcula mediante la ecuación (8.32).

8.3.2. Inferencia bayesiana en problemas de decisión

El inconveniente de la aproximación por máxima verosimilitud es que el verdadero problema que se desea resolver no es $p(\mathcal{S}|\theta)$, es decir, la probabilidad de los datos dados los parámetros sino, precisamente, la probabilidad de los parámetros dados los datos, esto es, $p(\theta|\mathcal{S})$. La aproximación por inferencia bayesiana nos permiten resolver este problema a partir del teorema de Bayes (3.5).

Si se aplica este teorema al problema de hallar los parámetros de un modelo dados los datos $p(\theta|\mathcal{S})$ tendremos que

$$p(\theta|\mathcal{S}) = \frac{p(\mathcal{S}|\theta)p(\theta)}{p(\mathcal{S})}. \tag{8.37}$$

que es la expresión fundamental de la aproximación bayesiana donde $p(\mathcal{S}|\theta)$ es precisamente la **verosimilitud**. La expresión $p(\theta)$ es la **probabilidad a priori** y representa la información que tenemos de antemano. Por último, $p(\mathcal{S})$ es la verosimilitud marginal o **evidencia**:

$$p(\mathcal{S}) = \int p(\mathcal{S}|\theta)p(\theta)d\theta. \tag{8.38}$$

Con esta expresión podemos obtener $p(\theta|\mathcal{S})$, denominada **probabilidad a posteriori** de los parámetros. En resumen, tenemos que:

$$\text{posteriori} = \frac{\text{verosimilitud} \times \text{priori}}{\text{evidencia}}.$$

Una vez estimados los parámetros del modelo se deseará predecir el valor desconocido de una observación futura $\mathbf{s}_{new}$. En este punto conviene señalar que los parámetros θ no son fijos y únicos sino que, en la aproximación bayesiana, siguen una distribución de probabilidad. Por tanto, se deberán integrar las predicciones del modelo respecto a la distribución a posteriori de los parámetros. Esto es, dada la cantidad desconocida $\mathbf{s}_{new}$, obtenemos la **distribución predictiva final** para dicha cantidad del siguiente modo:

$$p(\mathbf{s}_{new}|\mathbf{s}_1, \ldots, \mathbf{s}_N) = \int p(\mathbf{s}_{new}|\theta)p(\theta|\mathbf{s}_1, \ldots, \mathbf{s}_N)d\theta. \tag{8.39}$$

En ocasiones, el cálculo de la integral de la distribución predictiva final de la ecuación (8.39) es computacionalmente compleja. Un aproximación muy empleada es encontrar un único valor con densidad de probabilidad a posteriori máxima $\hat{\theta}^*$. El uso de esta probabilidad a posteriori máxima (o MAP, *maximum a posteriori*) se suele describir como un método bayesiano, sin embargo

esta caracterización es poco acertada, excepto cuando la estimación del MAP se aproxima a la ecuación de la integral (8.39). Sin embargo, esta situación no suele ser habitual.

La distribución predictiva final puede tener otras formas en el caso de la regresión o la clasificación de observaciones. Cuando el objetivo es encontrar la distribución predictiva final, y_{new}, para un nuevo caso cuyas variables independientes han sido observadas $\mathbf{x}_{new}$, se emplea la siguiente expresión:

$$p(y_{new}|\mathbf{x}_{new}, (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)) = \\ p(y_{new}|\mathbf{x}_{new}, \mathcal{S}) = \int p(y_{new}|\mathbf{x}_{new}, \theta) p(\theta|\mathcal{S}) d\theta, \tag{8.40}$$

siendo $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. No se debe olvidar que, en este caso, la expresión de la verosimilitud para estos casos será:

$$\mathcal{L}(\theta|\mathcal{S}) = \prod_{n=1}^{N} p(y_n|x_n, \theta).$$

El cálculo de las expresiones indicadas suele ser bastante complejo. Analíticamente, para las expresiones que estiman la probabilidad a posteriori de los parámetros se puede lograr si se cumple una condición: que la probabilidad a priori de los parámetros $p(\theta)$ sea *conjugada*. Una definición formal de **probabilidad a priori conjugada** es la siguiente:

Definición Sea $\mathcal{P}$ una clase de distribuciones a priori de parámetros $p(\theta)$ y sea $\mathcal{F}$ una clase de distribuciones $p(y|\theta)$, entonces la clase $\mathcal{P}$ es **conjugada** de la clase $\mathcal{F}$ si

$$\forall p(y|\theta) \in \mathcal{F} \wedge p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}.$$

En otras palabras, si al multiplicar la distribución a priori por la verosimilitud va a resultar una distribución a posteriori (debidamente normalizada) que sea de la misma clase que la a priori. Este tipo de conjugadas nos permite obtener expresiones analíticas del cálculo de los parámetros. Sin embargo, no siempre el cálculo de las expresiones se podrán llevar a cabo de forma analítica. Cuando esto ocurre, se emplean métodos de muestreo de Monte Carlo.

Ejemplo 8.3 (Forma de un tumor de estirpe vascular (cont. ejemplo 8.2)) Supongamos ahora que disponemos de conocimiento a priori sobre la forma de los tumores de estirpe vascular, podemos entonces combinar dicho conocimiento con la información obtenida por experimentación mediante el Teorema de Bayes:

$$p(\mathbf{p}|\mathcal{S}) = \frac{p(\mathbf{p})p(\mathcal{S}|\mathbf{p})}{p(\mathcal{S})},$$

donde el conocimiento previo se expresa mediante $p(\mathbf{p})$, que es la distribución a priori de los parámetros que supondremos, que siguen una distribución de Dirichlet, por ser la conjugada de la multinomial. Los parámetros α de dicha distribución pueden interpretarse de forma intuitiva como el conteo de un conjunto de muestras anteriores producidas por la distribución multinomial de parámetros $\mathbf{p}$. En nuestro caso:

$$\alpha = (170, 500, 367, 2203, 162, 62, 536)$$

Por otra parte, podemos expresar el conteo para cada valor de $\mathbf{x}$ como $\sum_{i=1}^{N} \mathbf{x_n}$, que resume el experimento de N muestras del ejemplo 8.2:

$$\sum_{i=1}^{N} \mathbf{x_n} = (3, 10, 8, 46, 3, 0, 11).$$

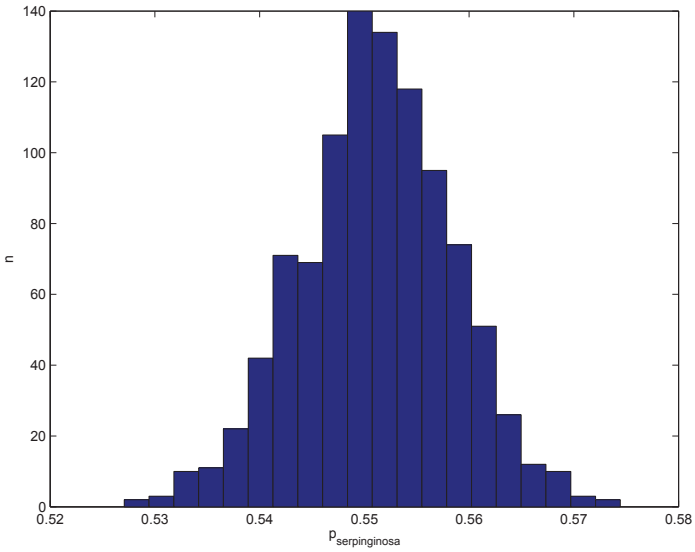


Figura 8.4: Muestreo de p(serpinginosa) tras la estimación de la distribución final $p(\mathbf{p}|\mathcal{S})$.

Por las propiedades de las distribuciones conjugadas, la distribución final (o distribución a posteriori) de $\mathbf{p}$ será una distribución de Dirichlet de parámetros $\alpha + \sum_{i=1}^{N} \mathbf{x_n}$, por lo que para nuestro problema,

$$\begin{aligned} p(\mathbf{p}|\mathcal{S}) &= Dir(\alpha + \sum_{i=1}^{N} \mathbf{x_n}) \\ &= Dir(173, 510, 375, 2249, 165, 62, 547). \end{aligned}$$

Para observar la distribución que toman los parámetros podemos muestrear, por ejemplo, mil observaciones y obtener el histograma de la figura 8.4 para la forma "serpinginosa", que vemos que tiene una forma normal centrada en 0.55. Del mismo modo, podríamos observar que el valor "cordón" ya no tiene una probabilidad 0 conforme calculábamos por máxima verosimilitud en el ejemplo 8.2, sino que puede tomar valores distribuidos en torno a 0,015.

8.4. Notas bibliográficas

Los libros sobre reconocimiento de formas y aprendizaje automático de Duda, Hart & Stork [92] y de Bishop [93] son fundamentales para conocer la disciplina y las técnicas de aprendizaje automático.

La llegada del *big data* provocó el despegue de la aplicación de las tecnologías del análisis masivo de datos biomédicos. Tal como muestra la edición del 2018 del *Big Data and Artificial Intelligence landscape* [110] y el *AI in Healthcare: Industry landscape* [111] muchas de estas tecnologías se han especializado en aplicaciones en las industrias de la salud y ciencias de la vida. Si bien la definición generalista de *big data* se refiere al tratamiento de grandes volúmenes de datos de alta velocidad, complejos y variables que requieren técnicas y tecnologías avanzadas para su captura, almacenamiento, distribución, gestión y análisis [112].

Sin embargo, la aplicación de *big data* a salud pone el acento en el uso del análisis de datos para extraer información y tomar decisiones mejor informadas [113]. *Big data* en salud no tiene por tanto la prioridad de manejar grandes conjuntos de datos o procedente de múltiples fuentes. El término se ha establecido como paraguas de los diferentes análisis computacionales que permiten obtener valor a partir del estudio de datos biomédicos. Por lo tanto, *big data* en salud se centra en el desarrollo y aplicación de técnicas computacionales para apoyar la investigación biomédica mediante el análisis de datos procedentes de fuentes de datos diversas, complejas, desorganizados, masivas, y multimodales, que pueden ser generados los investigadores, los hospitales, los centros de salud y dispositivos móviles en todo el mundo, incluyendo información clínica, fenotípica, genotípica, de comportamiento, de tratamiento, de exposición, molecular e imágenes y otras señales biomédica. Estos datos pueden ser utilizados para descubrir nuevos fármacos, estratificar pacientes, determinar las causas genéticas y ambientales de las enfermedades humanas, predecir pronósticos y supervivencia y para mejorar la gestión de los servicios sanitarios. Desde un punto de vista tecnológico, *big data* para salud se enfrenta a muchos desafíos: la gran y creciente cantidad de datos, la inherente variabilidad de la información, la complejidad de los procesos generadores de información, el carácter semi-estructurado de las fuentes de datos, la complejidad de las preguntas clínicas y de los modelos biomédicos. Además, es necesario validar y actualizar las soluciones computacionales para dar un servicio preciso a la clínica y a la investigación traslacional.

Tal es el interés, que informes como el emitido por Deloitte [114] confían la inteligencia artificial y la evidencia extraída del mundo real para desbloquear el valor de los datos de salud. Los datos del mundo real (RWD) proporcionan la información necesaria para que los investigadores puedan desarrollar una medicina más precisa y los médicos puedan predecir la respuesta de los pacientes a los tratamientos. Así pues, *big data* para salud se está convirtiendo en una infraestructura prioritaria para los servicios de salud y un activo empresarial fundamental.

Los ejemplos de mayor magnitud que encontramos en Europa están coordinados bajo el programa marco Horizonte 2020, la Agenda Digital de la Comisión Europea y la colaboración público/privada con la industria farmarcéutica. Específicamente, se han definido y están en desarrollo las temáticas "Big Data for Better Outcomes" del programa de trabajo IMI y "Big Data supporting Public Health policies" del programa de trabajo "Health, demographic change and well-being". La Comisión Europea está realizando una gran inversión en la búsqueda de nuevos paradigmas de gestión activa de la salud, incluyendo prevención primaria, gestión de crónicos pluripatológicos, ancianidad, enfermedades complejas y condiciones genéticas.

Capítulo 9

Funciones discriminantes, Redes Neuronales y Máquinas de vectores soporte

En este capítulo estudiaremos los fundamentos de las funciones discriminantes como solución al problema de clasificación. Relacionadas con las mismas, estudiaremos también las redes neuronales y las máquinas de vectores soporte. Para conocer la disciplina y las técnicas de aprendizaje automático es recomendable acercarse a los libros de Duda, Hart y Stork [92] y de Bishop [93].

9.1. Función discriminante lineal

El objetivo de los problemas de clasificación es asignar, a un objeto observado $\mathbf{x}$, una clase de entre $|\mathcal{Y}|$ clases posibles. Donde $\mathcal{Y}$ es el conjunto de todas las clases posibles para el problema y donde la clase a la que pertenece el objeto se anota como $y \in \mathcal{Y}$. Como se indica en la sección 8.2.1, el espacio muestral donde se representan las instancias observadas queda dividido en *regiones de decisión* separadas por las *fronteras de decisión*. El clasificador se representará mediante un conjunto de $|\mathcal{Y}|$ *funciones discriminantes*, como indica la ecuación (8.7). Cuando se aplica una función lineal al vector $\mathbf{x}$ se le denomina *función discriminante lineal*:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0, \tag{9.1}$$

donde $\mathbf{w}$ es un *vector de coeficientes*, también llamado *vector de pesos*, y w_0 es el *término independiente*.

Inicialmente se centrará la atención en el problema más sencillo que consiste en discriminar un objeto entre dos clases posibles y_1 e y_2, donde cada clase debería representarse mediante sendas funciones discriminantes lineales $g_1(\mathbf{x})$ y $g_2(\mathbf{x})$. La regla para clasificar los objetos $\mathbf{x}$ sería decidir la clase cuya función discriminante es mayor, esto es,

$$\hat{y}^* = \begin{cases} y_1 & \text{si } \; g_1(\mathbf{x}) > g_2(\mathbf{x}) \\ y_2 & \text{si } \; g_1(\mathbf{x}) < g_2(\mathbf{x}). \end{cases} \tag{9.2}$$

Sin embargo, ambas funciones pueden ser simplificadas en una única función cuando solo se consideran dos clases, ya que

$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= (\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x} + (w_{10} - w_{20}) \\ &= \mathbf{w}^T\mathbf{x} + w_0. \end{aligned} \tag{9.3}$$

En este caso la regla de clasificación será:

$$\hat{y}^* = \begin{cases} y_1 & \text{si } \ g(\mathbf{x}) > 0 \\ y_2 & \text{si } \ g(\mathbf{x}) < 0. \end{cases}$$

La frontera de decisión entre las dos clases corresponderá a los puntos donde la función discriminante es igual a cero, esto es $g(\mathbf{x}) = 0$

$$\mathcal{F} = \{\mathbf{x} : g(\mathbf{x}) = 0\}.$$

Cuando un objeto observado cae en la frontera de decisión el objeto $\mathbf{x}$ se puede asignar arbitrariamente a cualquiera de las dos clases o puede dejarse indefinido. La frontera de decisión $\mathcal{F}$ divide el espacio en dos regiones disjuntas asociadas a cada una de las clases: $\mathcal{R}_1$ a la clase y_1 y $\mathcal{R}_2$ a la clase y_2. Si $g(\mathbf{x})$ es una función discriminante lineal, esta frontera de decisión será un *hiperplano* separador.

A continuación se analizan algunas propiedades geométricas de las funciones discriminantes lineales. Supongamos que se toman dos puntos distintos, $\mathbf{x}_1$ y $\mathbf{x}_2$, que pertenecen a la frontera de decisión $\mathcal{F}$. Por definición,

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0.$$

Eliminando los términos independientes se obtiene $\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$. Este resultado nos muestra que el vector de pesos $\mathbf{w}$ es normal al hiperplano separador, ya que su producto escalar con el vector que definen los puntos $\mathbf{x}_1$ y $\mathbf{x}_2$ es nulo (ver figura 9.1). Por tanto, se deduce que el vector de pesos, $\mathbf{w}$, determina la orientación del hiperplano separador. Además, como $\mathbf{x}$ pertenece a la región $\mathcal{R}_1$ si $g(\mathbf{x}) > 0$, sabemos que el sentido del vector apunta a dicha región. La función $g(\mathbf{x})$ ofrece también una medida de la distancia de $\mathbf{x}$ al hiperplano $\mathcal{F}$. Para entender esto se puede expresar $\mathbf{x}$ como la suma de su proyección, $\mathbf{x}_p$, sobre el hiperplano $\mathcal{F}$ más el producto de la distancia r por el vector unitario de $\mathbf{w}$:

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{||\mathbf{w}||}. \tag{9.4}$$

Teniendo en cuenta que $g(\mathbf{x}_p) = 0$, si aplicamos la función discriminante a la ecuación (9.4) tenemos $g(\mathbf{x}) = r||\mathbf{w}||$, por lo tanto la distancia r será:

$$r = \frac{|g(\mathbf{x})|}{||\mathbf{w}||}, \tag{9.5}$$

ya que la distancia siempre debe ser $r \geq 0$.

A partir de la ecuación (9.5), se puede deducir la distancia del hiperplano $\mathcal{F}$ al origen de coordenadas. Como $g(\mathbf{0}) = w_0$, la distancia es:

$$r = \frac{w_0}{||\mathbf{w}||}. \tag{9.6}$$

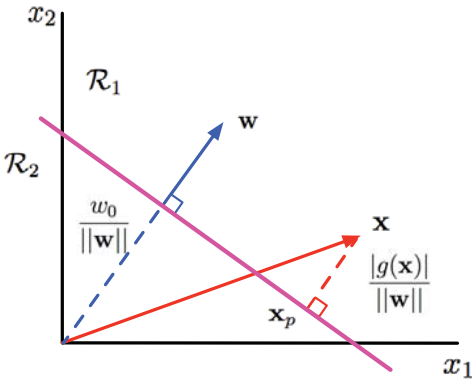


Figura 9.1: La geometría de una función discriminante lineal en dos dimensiones. La frontera de decisión, en color magenta, separa el espacio bidimensional en dos regiones disjuntas, $\mathcal{R}_1$ y $\mathcal{R}_2$. Esta frontera es normal al vector $\mathbf{w}$ y su desplazamiento viene definido por el parámetro w_0.

Por tanto, si $w_0 > 0$ el origen de coordenadas estará en el lado positivo de $\mathcal{F}$. Si $w_0 < 0$, el origen de coordenadas estará en el lado negativo del hiperplano separador. Finalmente, si $w_0 = 0$ el hiperplano $\mathcal{F}$ pasará por el origen de coordenadas. Ver figura 9.1.

9.1.1. Clasificación de múltiples clases

El problema de clasificación con múltiples clases aparece cuando el número de clases a discriminar es $|\mathcal{Y}| > 2$. En este caso la simplificación realizada en la ecuación (9.3) no puede llevarse a cabo. En su lugar, se empleará una función discriminante lineal $g_i(\mathbf{x})$ para caracterizar cada clase $y_i \in \mathcal{Y}$, $\forall i = 1, \ldots, |\mathcal{Y}|$. Bajo estas condiciones la regla de clasificación será

$$\hat{y}^* = \arg\max_i g_i(\mathbf{x}). \tag{9.7}$$

Como se comenta en la sección 8.2.1, esta regla es equivalente a escoger la clase y_i cuando $g_i(\mathbf{x}) > g_j(\mathbf{x})$, para todo $i \neq j$. La frontera de decisión entre cada par de clases y_i e y_j se compone de aquellos puntos que cumplen $g_i(\mathbf{x}) = g_j(\mathbf{x})$. Esta frontera de decisión $\mathcal{F}_{ij}$, es también un hiperplano separador definido como

$$(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + (w_{i0} - w_{0j}) = 0, \tag{9.8}$$

y tiene la misma forma que la frontera de decisión descrita para el problema de las dos clases y, por tanto, también tiene las mismas propiedades geométricas. El clasificador lineal para múltiples clases divide el espacio muestral en $|\mathcal{Y}|$ regiones disjuntas, conexas y convexas.

Hasta aquí hemos visto las funciones discriminantes lineales y sus propiedades para clasificar las observaciones $\mathbf{x}$ en base a la región a la que pertenece una vez aplicada la regla de clasificación. Sin embargo, necesitamos algoritmos capaces de obtener los valores adecuados para el vector de pesos de cada función discriminante $\mathbf{w}_i$, así como los valores de los términos independientes w_{i0}. A continuación se presentan algunos de los algoritmos más conocidos para el ajuste de estos parámetros.

9.1.2. Estimación de parámetros por mínimos cuadrados

El método de mínimos cuadrados se aplica habitualmente a problemas en los que se debe resolver un sistema de ecuaciones sobre-determinado, es decir, con más ecuaciones que incógnitas. La resolución exacta de estos sistemas de ecuaciones no suele ser posible. En su lugar, se busca minimizar una suma de cuadrados. El caso típico es el de la regresión lineal donde se tienen N observaciones, que corresponden con las N ecuaciones, y d variables independientes, que corresponden a las d incógnitas, siendo $N > d$. En la regresión lineal se trata de minimizar la suma de residuos al cuadrado.

En clasificación se puede emplear el método de los mínimos cuadrados para ajustar los parámetros de las funciones discriminantes lineales. Inicialmente, se considera la clasificación de dos clases para lo que únicamente se necesita una función discriminante lineal como muestra la ecuación (9.3). Partiendo de una muestra $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}$ para $i = 1, \ldots, N$, con $\mathbf{x}_i \in \mathbb{R}^D$ e $y_i \in \{+1, -1\}$, se puede expresar cada una de las observaciones como la combinación lineal de los parámetros de la función discriminante y las variables independientes:

$$\begin{aligned} y_1 &= w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_D x_{1D} + e_1 \\ y_2 &= w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_D x_{2D} + e_2 \\ &\ldots \\ y_N &= w_0 + w_1 x_{N1} + w_2 x_{N2} + \ldots + w_D x_{ND} + e_N \end{aligned} \tag{9.9}$$

y en forma matricial

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E}, \tag{9.10}$$

donde $\mathbf{X}$ es la matriz de observaciones aumentada[a] y el vector $\mathbf{Y}$ contiene las clases asociadas a dichas observaciones; la matriz $\mathbf{W}$ es la matriz de parámetros, incluyendo el término independiente w_0; por último, $\mathbf{E}$ es el vector de residuos, que es el error a minimizar. Concretamente,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1D} \\ 1 & x_{21} & x_{22} & \ldots & x_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{ND} \end{bmatrix};$$

$$\mathbf{W} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}; \qquad \mathbf{E} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}.$$

[a] La matriz $\mathbf{X}$ es aumentada porque incluye un vector columna de unos que se multiplica por la constante del término independiente.

Para obtener los valores apropiados de $\mathbf{W}$ que minimicen la expresión tenemos que $\mathbf{E} = \mathbf{Y} - \mathbf{XW}$ y, por la resolución de mínimos cuadrados, el problema se plantea como minimizar la expresión

$$E(\mathbf{W}) = ||\mathbf{E}||^2 = ||(\mathbf{Y} - \mathbf{XW})||^2 \tag{9.11}$$

derivando con respecto a $\mathbf{W}$, igualando a cero y reajustando la expresión se obtiene la siguiente solución para el cálculo de la matriz de pesos:

$$\mathbf{W} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{9.12}$$

Cuando se trabaja con múltiples clases, $|\mathcal{Y}| > 2$, es conveniente emplear una codificación 1-de-C que convierte la salida en un vector binario donde la posición de la clase a la que pertenece la observación toma el valor 1 y el resto toman el valor 0. En este caso, el vector de clases $\mathbf{Y}$ se convierte en una matriz donde la fila i es el vector codificado 1-de-C que representa la clase a la que pertenece la observación $\mathbf{x}_i$. El vector de pesos $\mathbf{W}$ también se convierte en una matriz de dimensiones $(D+1) \times |\mathcal{Y}|$, cuya fila k contiene el vector $\tilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k)$. El cálculo de la matriz de pesos es equivalente a la expresión de la ecuación (9.12).

Un problema importante de la estimación por mínimos cuadrados para modelos lineales es su falta de robustez frente a observaciones atípicas. En presencia de datos anómalos distribuidos asimétricamente la estimación por mínimos cuadrados puede ser sesgada e ineficiente.

9.2. Redes neuronales

Las redes neuronales artificiales (ANN, por sus siglas en inglés) [115, 116] son modelos muy empleados en las últimas décadas para clasificación y regresión. Estos sistemas se caracterizan por ser adaptativos, distribuidos y paralelos. El conocimiento que almacenan estos modelos, y que infieren a partir de observaciones, se distribuye a través de los pesos de las conexiones entre las unidades básicas de procesamiento. Las redes neuronales son capaces de generalizar a partir de observaciones ruidosas o incompletas. Existen multitud de tipos de redes neuronales artificiales, pero en esta sección nos centraremos únicamente en las redes conectadas totalmente hacia adelante (*feed-forward*), también conocidas como **perceptrón multicapa**. Este tipo de redes son capaces de ajustar cualquier tipo de función por compleja que sea. Todo depende del número de unidades básicas que se empleen para ello [117].

Las redes neuronales artificiales son modelos conexionistas que intentan emular el comportamiento de los sistemas nerviosos donde el elemento fundamental es la neurona. Del mismo modo, en las ANN el elemento fundamental serán las unidades básicas de procesamiento basadas en el **perceptrón** de Rosenblatt.

9.2.1. El perceptrón

El perceptrón de Rosenblatt [118] es un ejemplo de modelo discriminante lineal capaz de discriminar dos clases. Dadas unas observaciones $\mathbf{x}$ en un espacio D-dimensional, la expresión general del perceptrón es

$$y = f\Big(\sum_{d=1}^{D} w_d x_d + w_0\Big) \tag{9.13}$$

donde f es la función de activación que se verá a continuación; x_j es la entrada j-ésima de la neurona, equivalente a las dendritas en una neurona; w_j el peso de esa entrada en el proceso

de activación de la unidad, equivalente a la conexión entre neuronas, pudiendo ser excitadoras -pesos positivos- o inhibidoras -pesos negativos-; w_0 es el componente independiente, conocido como *bias*, que se correspondería con el umbral que desencadena la sinapsis eléctrica; la función de salida y sería equivalente a la salida de la neurona. En ocasiones, el bias se integra en el sumatorio de la expresión (9.13), siendo multiplicado por una entrada $x_0 = 1$, de modo que la expresión queda

$$\begin{aligned} y &= f\Big(\sum_{d=0}^{D} w_d x_d\Big) \\ &= f\Big(\mathbf{w}^{\mathrm{T}}\mathbf{x}\Big) \end{aligned} \tag{9.14}$$

La función de activación desempeña un papel destacado en la unidad perceptrón ya que determina y delimita los posibles valores de salida.

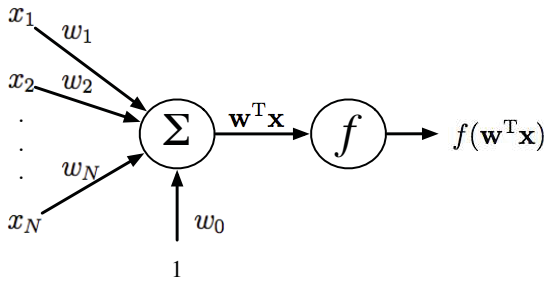


Figura 9.2: Ilustración del modelo perceptrón. La entrada se compone del vector $\mathbf{x}$ cuyas componentes se combinan linealmente con las componentes del vector de pesos $\mathbf{w}$, incluyendo el bias w_0 que se multiplica por 1. A esta integración también se la conoce como función de red. Posteriormente, se aplica la función de activación cuyo resultado será el valor de salida de la unidad.

El perceptrón presenta una limitación importante ya que únicamente puede discriminar problemas dicotómicos linealmente separables [119]. Sin embargo, si se conectan varios perceptrones entre sí se puede llegar a aliviar ambas limitaciones. En la sección 9.2.3 se verá un tipo de red neuronal artificial que conecta perceptrones para superar estos problemas: el perceptrón multicapa.

9.2.2. Funciones de activación

Existen múltiples funciones de activación aplicables a un perceptrón. Las funciones de activación más empleadas son:

Función escalón: es la forma más fácil de definir la activación de una unidad. Si la combinación lineal de las entradas y los pesos supera un cierto umbral establecido, la activación es 1. Si no lo supera, la activación es 0. La derivada de esta función es 0, por lo que no se puede emplear con algoritmos de aprendizaje basados en descenso por gradiente. Se comenta aquí por motivos históricos. La función escalón queda definida por la siguiente expresión:

$$f(x) = \begin{cases} 1 & \text{si } x >= \theta \\ 0 & \text{si } x < \theta \end{cases}$$

Función lineal o identidad: la función lineal o identidad responde a la expresión $f(x) = x$. Su derivada es $f'(x) = 1$.

Función logística: es una función sigmoide monótona creciente con límites asintóticos inferior, 0, y superior, 1. Al contrario que en la función escalón, la función logística es continua y derivable en todo el intervalo. Se define como:

$$f(x) = \frac{1}{1 + e^{-x}}$$

cuya derivada es:

$$f'(x) = f(x)(1 - f(x))$$

Función tangente hiperbólica: es una función sigmoide monótona creciente, continua y derivable con límites asintóticos inferior, -1, y superior 1. Se define como:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

cuya derivada es:

$$f'(x) = 1 - f(x)^2$$

Función softmax: la función softmax se emplea generalmente para obtener una interpretación probabilística en la salida de las redes neuronales ya que normaliza los valores de las unidades empleando la siguiente expresión:

$$f(x_i) = \frac{\exp\{x_i\}}{\sum_j \exp\{x_j\}}$$

Su derivada se expresa así:

$$\frac{\partial}{\partial x_j} f(x_i) = \begin{cases} f(x_i)\big(1 - f(x_i)\big) & \text{si } i = j \\ -f(x_i) f(x_j) & \text{si } i \neq j \end{cases}$$

Función lineal rectificada: la función lineal rectificada [120] es uno de los grandes avances en el desarrollo de las redes neuronales de principios de siglo, especialmente en la última generación de redes neuronales conocida como *deep learning*. Esta función de activación permite representación dispersas de los datos al saturar exactamente en 0, además cuando el valor de activación es mayor, que uno su derivada es exactamente 1 lo que evita un problema común en redes profundas donde el gradiente puede anularse al entrenar la red. Este problema se conoce en inglés como *vanishing gradient* [121]. Se define como:

$$f(x) = \text{máx}\big(0, x\big)$$

cuya derivada es:

$$f'(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Función lineal rectificada paramétrica: la función lineal rectificada [120] tiene una desventaja potencial durante la fase de entrenamiento debido a que el gradiente puede ser 0 cuando no se activa la unidad, lo que podría provocar que dicha unidad no se ajustase a los datos. Para evitar este problema surge la función lineal rectificada paramétrica que, compromete la saturación exacta a 0 mediante un parámetro α, a cambio de poder optimizar la unidad perceptrón durante el entrenamiento de la red. Se define como:

$$f(x) = \begin{cases} x & \text{si } x >= 0 \\ \alpha x & \text{si } x < 0 \end{cases}$$

cuya derivada es:

$$f'(x) = \begin{cases} 1 & \text{si } x > 0 \\ \alpha & \text{si } x < 0 \end{cases}$$

Función lineal rectificada exponencial: esta función propone rectificar la función lineal mediante una aproximación exponencial con el fin de aproximar la activación media de la unidad a 0, consiguiendo un entrenamiento más rápido a la vez que evita el problema de anulación del gradiente [122]. Su expresión es:

$$f(x) = \begin{cases} x & \text{si } x >= 0 \\ \alpha(\exp(x) - 1) & \text{si } x < 0 \end{cases}$$

cuya derivada es:

$$f'(x) = \begin{cases} 1 & \text{si } x >= 0 \\ \alpha + \alpha(\exp(x) + 1) & \text{si } x < 0 \end{cases}$$

9.2.3. El perceptrón multicapa

Existen multitud de arquitecturas de redes neuronales artificiales. Una de las más empleadas son las redes totalmente conectadas hacia adelante, también conocidas como **perceptrones multicapa**. En estos modelos se tienen numerosos elementos simples de procesamiento -perceptrones- conectados unos con otros y dispuestos en capas. Una *capa* es un conjunto de unidades cuyas entradas provienen de la misma fuente (la entrada **x** u otra capa de unidades), y cuyas salidas se dirigen al mismo destino (la salida **y** u otra capa). La **capa de entrada** está compuesta por unidades que reciben la información directamente del vector de observaciones **x**, transmitiendo esta información a la siguiente capa a través de las conexiones entre ellas. Cabe mencionar que, en muchas ocasiones, la capa de entrada no se considera realmente una capa. Esto sucede cuando se considera que las capas están conformadas por las conexiones entre unidades **w** en lugar de por las unidades. La **capa de salida** está formada por las unidades que devuelven la respuesta final de la red. Estas unidades pueden estar directamente conectadas con la capa de entrada o con una capa intermedia llamada **capa oculta**. La capa oculta puede estar conectada a su vez a otra capa oculta. El perceptrón multicapa funciona como una caja negra donde la entrada y la salida son conocidas, pero los valores intermedios que pasan por las capas ocultas no. Así pues, las unidades que forman parte de las capas ocultas no tienen contacto directo con la entrada ni con la salida. En la figura 9.5 se puede observar un esquema de red hacia adelante.

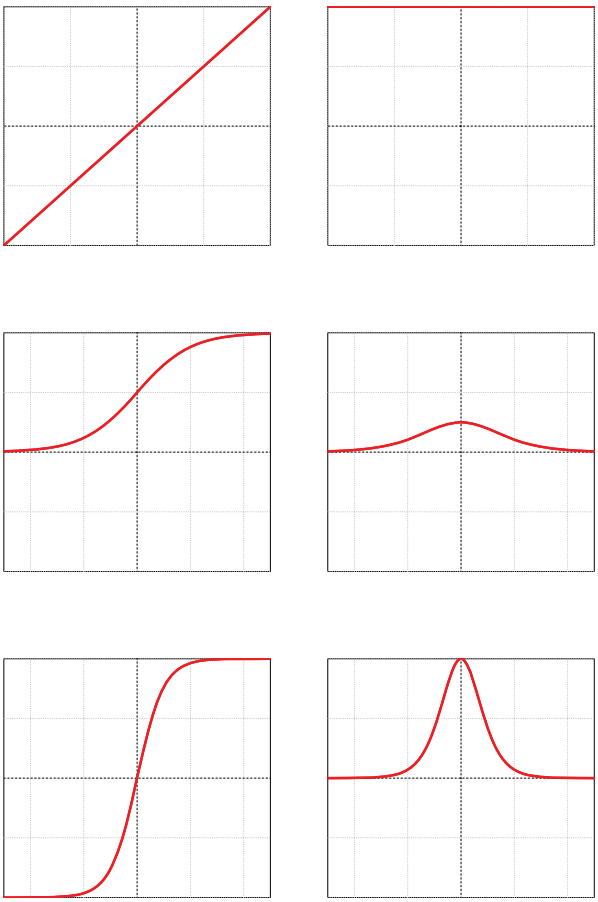


Figura 9.3: Ilustración de las funciones de activación (izquierda) y sus derivadas (derecha) para la función lineal o identidad (arriba), logística (centro) y tangente hiperbólica (abajo).

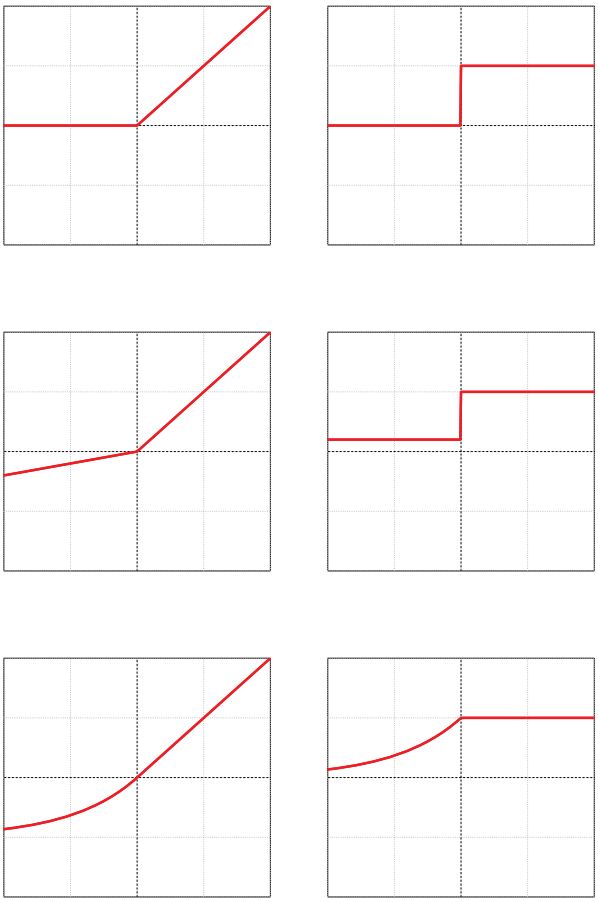


Figura 9.4: Ilustración de las funciones de activación (izquierda) y sus derivadas (derecha) para la función lineal rectificada (arriba), lineal rectificada paramétrica con $\alpha = 0{,}2$ (centro) y lineal rectificada exponencial (abajo).

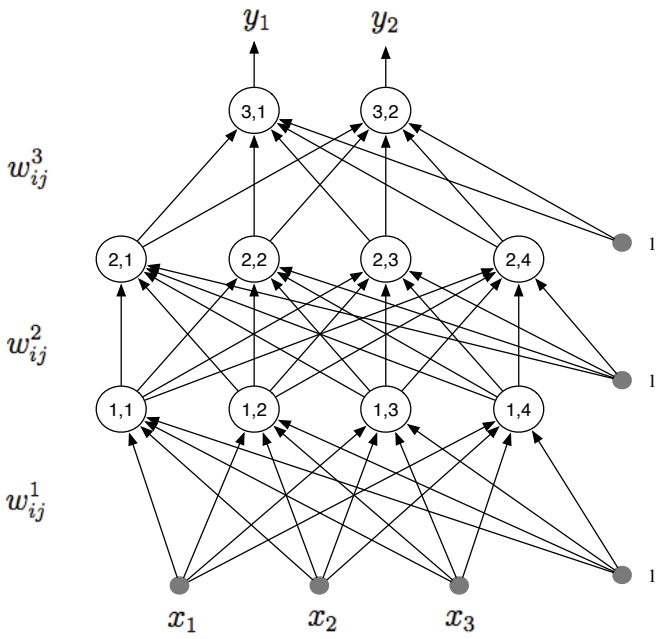


Figura 9.5: Ejemplo de arquitectura de un perceptrón multicapa con dos capas ocultas y una de salida. La entrada consta de un vector con tres variables $\mathbf{x} = (x_1, x_2, x_3)$, por tanto, son necesarias tres unidades de entrada. La salida consta de dos unidades, lo que equivale a dos clases para discriminar. Los pesos de las conexiones se definen como w_{ij}^k donde k hace referencia a la capa a la que llegan las conexiones, i se refiere a la unidad a la que llega la conexión y j se refiere a la unidad de la que sale la conexión. Los valores de bias están representados por un valor 1 y son considerados una conexión más.

En un problema de clasificación con $|\mathcal{Y}|$ clases a discriminar, la capa de salida de un perceptrón multicapa consta exactamente de $|\mathcal{Y}|$ unidades. Cada unidad de salida se corresponde con una función discriminante para la clase correspondiente. De este modo, como se vio con las funciones discriminantes lineales, el espacio muestral queda dividido en $|\mathcal{Y}|$ regiones. Las capacidades de un percetrón multicapa dependen directamente de la arquitectura que se emplee. La arquitectura se refiere a la descripción del número de unidades en cada capa, del número de capas en la red, de la función de activación de cada capa y de las conexiones entre las capas. Según la arquitectura de la red se podrá describir una topología más o menos compleja. Una red sin capas ocultas divide el espacio muestral en regiones disjuntas, conexas y convexas con fronteras lineales; si la red cuenta con una capa oculta, el espacio queda dividido también en regiones disjuntas, conexas, pero no convexas, a partir de fronteras lineales a trozos; una red con dos capas ocultas divide el espacio en regiones disjuntas, pero pueden ser no conexas y no convexas. Hasta el desarrollo del *deep learning*, no se empleaban más de dos capas ocultas porque se consideraba suficiente para ajustar cualquier función y, además, añadir más capas ocultas introducía el problema de la anulación del gradiente lo que impedía entrenar correctamente redes con más capas ocultas. Actualmente, con el desarrollo de nuevas estrategias como las de pre-aprendizaje, las nuevas funciones de activación rectificadas y el entrenamiento mediante *dropout* se ha conseguido llevar a cabo con éxito el entrenamiento de redes neuronales con más de dos capas ocultas, en lo que se considera la tercera generación de las redes neuronales después de la primera –la invención del perceptrón–, y la segunda –la invención del algoritmo de retropropagación del error para entrenar perceptrones multicapa–.

9.2.4. Algoritmos de aprendizaje

Hasta ahora se han visto las redes neuronales como modelos no lineales que ofrecen un vector salida $\mathbf{y}$ a partir de un vector de entrada $\mathbf{x}$ que se transmite hacia la salida mediante unas conexiones entre unidades con cuyos pesos se combinan linealmente y a las que se les aplica una función de activación. Sin embargo, no se ha descrito ningún algoritmo para ajustar los pesos de las conexiones $\mathbf{w}$ ni los bias w_0 de modo que aprendan a partir de las observaciones. Para ello, la aproximación es minimizar una función de error que dependa de los parámetros $\mathbf{w}$.

Así pues, dado el conjunto de observaciones $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, donde $\mathbf{x}_n$ son los vectores de características e $\mathbf{y}_n$ es el vector de salida que sigue una codificación 1-de-C, donde se activa a 1 únicamente el elemento del vector correspondiente a la clase a la que pertenece la observación, se debe minimizar la función de error

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{y}_n - \hat{\mathbf{y}}_n||^2, \tag{9.15}$$

donde $\hat{\mathbf{y}}_n = \mathbf{f}(\mathbf{x}_n, \mathbf{w})$, esto es, el valor de la red neuronal con parámetros $\mathbf{w}$ para la entrada $\mathbf{x}_n$.

Así pues, el ajuste de los pesos trata de hallar un vector $\mathbf{w}$ de manera que el error $E(\mathbf{w})$ se minimice. Para ello, los métodos de descenso por gradiente permiten ajustar los parámetros de manera iterativa hasta llegar a un mínimo local de la función de error $E(\mathbf{w})$. Durante el algoritmo de entrenamiento, los pesos de la red convergen gradualmente a valores tales que cada vector de entrada genere el vector de salida deseado.

El aprendizaje se basa en el algoritmo de retropropagación del error. Inicialmente, los pesos de las conexiones toman valores aleatorios pequeños $\mathbf{w}^{(0)}$. Después, estos valores se ajustan en el espacio paramétrico sucesivamente con el objetivo de reducir el error:

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} + \Delta\mathbf{w}^{(i)}, \tag{9.16}$$

donde i es el número de iteración y $\Delta\mathbf{w}^{(i)}$ es el incremento de los pesos. El cálculo de este incremento se calcula como:

$$\triangle\mathbf{w} = -\rho\frac{\partial E}{\partial w_{ij}} \tag{9.17}$$

siendo ρ el **factor de aprendizaje**. Se puede ver que este incremento depende de la derivada del error respecto de $\mathbf{w}$. Por esta razón, siendo $E(\mathbf{w})$ una función que depende de $\mathbf{f}(\mathbf{x}_d, \mathbf{w})$, es necesario que las funciones de activación de las unidades sean derivables. En el algoritmo de retropropagación del error estándar se emplea la regla delta generalizada. Para explicar esta regla, emplearemos una expresión escalar del error para la muestra n:

$$E_p = \frac{1}{2}\sum_j (y_{nj} - \hat{y}_{nj})^2 \tag{9.18}$$

donde y_{nj} es el valor deseado para la salida j-ésima cuando la entrada a la red es la observación $\mathbf{x}_n$ y donde $\hat{y}_{nj}$ es el valor que devuelve la red para esa misma muestra y la misma unidad de salida. Además, $E = \sum_n E_n$ es la medida total del error. Partiendo de un modelo perceptrón multicapa donde las unidades envían sus valores de salida a las capas siguiente y que tienen funciones de activación no lineales, monótonas no decrecientes y derivables, llamaremos función de red a la expresión

$$g_{nj} = \sum_i w_{ji}\hat{y}_{ni} \tag{9.19}$$

donde $\hat{y}_i = x_i$ si la unidad i es de entrada. De este modo, con la función de activación no lineal tendremos

$$\hat{y}_{nj} = f_j(g_{nj}) \tag{9.20}$$

donde g es diferenciable y no decreciente. Para obtener la generalización correcta de la regla delta se debe establecer

$$\Delta_n w_{ji} \propto -\frac{\partial E_n}{\partial w_{ji}} \tag{9.21}$$

De nuevo, aplicando la regla de la cadena tenemos,

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial g_{nj}}\frac{\partial g_{nj}}{\partial w_{ji}} \tag{9.22}$$

Por la ecuación (9.19), el segundo factor es

$$\frac{\partial g_{nj}}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}}\sum_k w_{jk}\hat{y}_{nk} = \hat{y}_{ni} \tag{9.23}$$

Ahora se define,

$$\delta_{nj} = -\frac{\partial E_n}{\partial g_{nj}} \tag{9.24}$$

Para calcular esta expresión se recurre de nuevo a la regla de la cadena:

$$\delta_{pj} = -\frac{\partial E_p}{\partial g_{pj}} = -\frac{\partial E_p}{\partial y_{pj}}\frac{\partial y_{pj}}{\partial g_{pj}} \tag{9.25}$$

Se calcula el segundo factor. Por la ecuación (9.20) tenemos

$$\frac{\partial \hat{y}_{nj}}{\partial g_{nj}} = f'_j(g_{nj}), \tag{9.26}$$

que es la derivada de la función no lineal g_j para la unidad j. Para calcular el primer factor, se deben considerar dos casos. Primero, se asume que la unidad y_j es una **unidad de salida**. En este caso, de la definición de E_p se sigue que

$$\frac{\partial E_n}{\partial \hat{y}_{nj}} = -(y_{nj} - \hat{y}_{nj}), \tag{9.27}$$

que es el mismo resultado obtenido para la regla delta estándar. Así, sustituyendo ambos factores en la ecuación (9.25) se obtiene la fórmula para corregir los pesos de la capa de salida:

$$\delta_{nj} = (y_{nj} - \hat{y}_{nj}) f'_j(g_{nj}). \tag{9.28}$$

Si u_j **no es unidad de salida**, se puede emplear la regla de la cadena para derivar el primer factor:

$$\begin{aligned}
\frac{\partial E_n}{\partial \hat{y}_{nj}} &= \sum_k \frac{\partial E_n}{\partial g_{nk}} \frac{\partial g_{nk}}{\partial y_{nj}} \\
&= \sum_k \frac{\partial E_n}{\partial g_{nk}} \frac{\partial}{\partial \hat{y}_{nj}} \sum_i w_{ki} \hat{y}_{ni} \\
&= \sum_k \frac{\partial E_n}{\partial g_{nk}} w_{kj} \\
&= -\sum_k \delta_{nk} w_{kj}.
\end{aligned} \tag{9.29}$$

Sustituyendo ambos factores en la ecuación (9.25), obtenemos

$$\delta_{nj} = f'_j(g_{nj}) \sum_k \delta_{nk} w_{kj}, \tag{9.30}$$

cuando la unidad u_j no es de salida. Las ecuaciones (9.28) y (9.30) nos proporcionan un procedimiento recursivo para calcular los valores δ para cada unidad de la red. Estos valores se emplean, a su vez, para calcular las variaciones de los pesos de las conexiones de la red neuronal:

$$\Delta_n w_{ji} = \rho \delta_{nj} \hat{y}_{ni}. \tag{9.31}$$

A este procedimiento se le llama **regla delta generalizada** y se aplica para entrenar perceptrones multicapa con conexiones hacia adelante con funciones de activación monótonas no decrecientes y derivables. En el algoritmo estándar de retropropagación del error la elección del factor de aprendizaje óptimo es uno de los problemas comunes a la hora de abordar tareas de aprendizaje. Generalmente, valores de aprendizaje pequeños convergen lentamente a la solución, mientras que para valores más altos, la convergencia es más rápida pero la oscilación mucho mayor, pudiendo llegar a situaciones de divergencia en las que la solución no se alcanza.

Otra versión del algoritmo es el de retropropagación con momento. La regla de aprendizaje de este algoritmo es:

$$\Delta \mathbf{w}^{(i+1)} = \rho \frac{\delta E}{\delta w} + \mu \Delta \mathbf{w}^{(i)},$$

siendo μ el factor *momento* que pondera el antiguo valor del peso de la conexión como parámetro para calcular el nuevo, introduciendo así un factor de inercia. La introducción de la inercia evita algunos problemas de oscilación en el algoritmo de retropropagación e incrementa la velocidad de aprendizaje de manera significativa.

El ajuste de todos los parámetros de una red neuronal es bastante tedioso. La elección de la arquitectura de red, el algoritmo de aprendizaje, sus parámetros y establecer un criterio de parada hacen que estos métodos necesiten una metodología robusta para validar los modelos. Habitualmente, se preparan una serie de redes neuronales con distinto número de capas ocultas y unidades en cada capa, cuyos parámetros toman valores aleatorios inicialmente. Posteriormente, se aplican distintos algoritmos de aprendizaje con distintos parámetros y se entrena la red con un conjunto de entrenamiento. Se establece un criterio de parada para no sobreajustar los modelos y, finalmente, se comparan los resultados de cada modelo frente a los datos de un conjunto de evaluación.

9.3. Máquinas de vectores soporte

Las máquinas de vectores soporte[b] (SVM, por sus siglas en inglés) se han convertido en un método muy popular para resolver problemas de clasificación y regresión. Son clasificadores no paramétricos basados en las funciones discriminantes lineales. Las SVM son clasificadores dicotómicos, ya que se limitan a discriminar 2 clases distintas. Para ello, las SVM tratan de hallar un hiperplano óptimo que separe en dos regiones el espacio muestral de modo que cada región pertenezca a una clase. Intuitivamente, se considera que el hiperplano es óptimo cuando la distancia entre los dos puntos más cercanos de cada clase al hiperplano es máxima. A este concepto, clave en las SVM, se le llamará *margen*. A continuación, se verá un caso linealmente separable que servirá para derivar el cálculo de los parámetros de la función discriminante lineal que definirá el hiperplano óptimo. Posteriormente, se extenderá la explicación a casos no linealmente separables y el uso de *kernels*.

9.3.1. Clases linealmente separables

Las SVM en su versión más sencilla son modelos lineales de la forma

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0.$$

Como ya se ha dicho, los modelos de SVM discriminarán únicamente dos clases. Así pues, el objetivo será ajustar un hiperplano separador dado un conjunto de observaciones cuyas clases son conocidas, $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, donde $\mathbf{x}_n \in \mathbb{R}^D$ son las observaciones y sus respectivas clases $y_n \in \{-1, 1\}$ son, por hipótesis, linealmente separables.

El hiperplano definido por un modelo discriminante lineal es una función tal que

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 > 0 \quad \text{si } y_n = 1$$
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 < 0 \quad \text{si } y_n = -1$$

Sin embargo, lo que se busca con las SVM es un hiperplano separador óptimo que maximice el margen entre las observaciones más cercanas de cada clase. Para ello, se requiere que el hiperplano

[b] A menudo se las conoce como *máquinas de soporte vectorial*, pero es una traducción errónea del inglés *support vector machines*.

$\mathcal{F} = \{\mathbf{x} : g(\mathbf{x}) = 0\}$ satisfaga las siguientes restricciones:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 \geq 1 \quad \text{si } y_n = 1 \tag{9.32}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 \leq -1 \quad \text{si } y_n = -1 \tag{9.33}$$

Estas condiciones se pueden resumir en una única condición sin más que multiplicar por y_n y restar 1 en ambos lados de las desigualdades:

$$y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0) - 1 \geq 0, \quad i = 1, \ldots, N. \tag{9.34}$$

Considérese un primer hiperplano $\mathcal{H}_1 = \{\mathbf{x} : g(\mathbf{x}) = 1\}$, formado por los puntos que cumplen la igualdad de la restricción (9.32) y un segundo hiperplano $\mathcal{H}_2 = \{\mathbf{x} : g(\mathbf{x}) = -1\}$, formado por los puntos que cumplen la igualdad de la restricción (9.33). La distancia al origen del hiperplano $\mathcal{H}_1$ será $|w_0 - 1|/||\mathbf{w}||$ y la del hiperplano $\mathcal{H}_2$ será $|w_0 + 1|/||\mathbf{w}||$. Ambos son paralelos al hiperplano separador, que determina la frontera de decisión, $\mathcal{F}$, cuya distancia al origen es $|w_0|/||\mathbf{w}||$. Así pues, no es difícil establecer que la distancia de los hiperplanos $\mathcal{H}_1$ y $\mathcal{H}_2$ a la frontera $\mathcal{F}$ es $r = 1/||\mathbf{w}||$. Por tanto, la distancia mínima entre dos clases, es decir, el margen será $m = 2/||\mathbf{w}||$ y no habrá ningún punto en la región definida entre los planos $\mathcal{H}_1$ y $\mathcal{H}_2$ (ver figura 9.6). Los puntos que cumplan la igualdad de la condición 9.34 son los llamados **vectores soporte** ya que, si se eliminasen, la solución final cambiaría totalmente. Por lo tanto, para optimizar el hiperplano separador $\mathcal{F}$ lo que se debe buscar es maximizar el margen, lo que equivale a minimizar $||\mathbf{w}||^2$, sujeto a la condición 9.34.

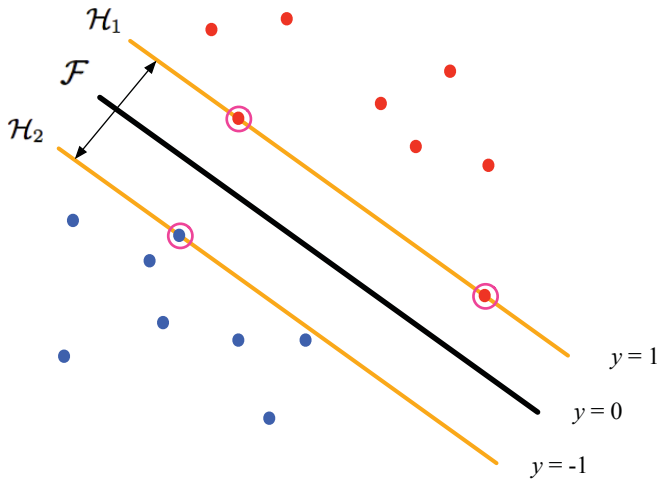


Figura 9.6: Ilustración del hiperplano separador en dos dimensiones y los hiperplanos $\mathcal{H}_1$ y $\mathcal{H}_2$ que definen el margen máximo. El margen es la distancia entre $\mathcal{H}_1$ y $\mathcal{H}_2$. Los puntos enmarcados por un círculo son los vectores soporte que definirán los parámetros del hiperplano separador $\mathcal{F}$.

Por conveniencia, el problema de optimización se resuelve mediante multiplicadores de Lagrange. En el método de Lagrange, las restricciones se multiplican por valores positivos o nulo

$\lambda_n \geq 0$ y se restan a la función que se desea optimizar. Así pues, la función lagrangiana a minimizar queda:

$$L_P(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} \lambda_n (y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0) - 1). \tag{9.35}$$

Ahora, se debe derivar la función $L_P(\mathbf{w}, w_0, \boldsymbol{\lambda})$ respecto a $\mathbf{w}$ y w_0, con lo que se obtienen las siguientes condiciones:

$$\mathbf{w} = \sum_{n=1}^{N} \lambda_n y_n (\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0), \tag{9.36}$$

$$\sum_{n=1}^{N} \lambda_n y_n = 0. \tag{9.37}$$

Si aplicamos estos resultados a la función $L_P(\mathbf{w}, w_0, \boldsymbol{\lambda})$ se obtiene una nueva lagrangiana conocida como *representación dual*:

$$L_D(\boldsymbol{\lambda}) = \sum_{n=1}^{N} \lambda_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m y_n y_m \mathbf{x}_n \mathbf{x}_m, \tag{9.38}$$

sujeta a las restricciones

$$\lambda_n, \lambda_m \geq 0, \quad n, m = 1, \ldots, N, \tag{9.39}$$

$$\sum_{n=1}^{N} \lambda_n y_n = 0. \tag{9.40}$$

La función dual L_D se maximiza derivando respecto a $\boldsymbol{\lambda}$ e igualando a cero. El problema de optimización satisface las condiciones de *Karush-Kuhn-Tucker*:

$$\lambda_n \geq 0, \tag{9.41}$$

$$y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) - 1 \geq 0, \tag{9.42}$$

$$\lambda_n(y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) - 1) = 0. \tag{9.43}$$

Cuando se cumplen las condiciones de *Karush-Kuhn-Tucker*, se garantiza encontrar una solución para problemas de optimización convexos como es el caso de las SVM. De este modo, dada la condición 9.43, para cada observación $\mathbf{x}_n$, o se anula el multiplicador de Lagrange, $\lambda_n = 0$, o bien $y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) = 1$. Esto implica que solo los puntos que cumplan $y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) = 1$ podrán tener un multiplicador de Lagrange asociado no nulo, $\lambda_n \neq 0$. Estos puntos serán los **vectores soporte** y pertenecerán a alguno de los hiperplanos que forman el margen, $\mathcal{H}_1$ o $\mathcal{H}_2$. El conjunto de vectores soporte se define como $\mathcal{VS} = \{\mathbf{x}_n : \lambda_n \neq 0\}$.

Finalmente, la solución óptima $(\mathbf{w}^*, w_0^*, \boldsymbol{\lambda}^*)$, una vez se tienen los valores óptimos de los multiplicadores de Lagrange, λ_n^*, queda

$$\mathbf{w}^* = \sum_{n=1}^{N} \lambda_n^* y_n \mathbf{x}_n = \sum_{\mathbf{x}_n \in \mathcal{VS}} \lambda_n^* y_n \mathbf{x}_n, \tag{9.44}$$

$$w_0^* = y_n - \mathbf{w}^{*\mathrm{T}}\mathbf{x}_n, \quad \forall \mathbf{x}_n \in \mathcal{VS}. \tag{9.45}$$

Una vez se han obtenido los valores óptimos de los parámetros del hiperplano separador se pueden clasificar nuevas observaciones empleando la siguiente función:

$$g(\mathbf{x}) = \sum_{\mathbf{x}_n \in \mathcal{VS}} \lambda_n^* y_n \mathbf{x}_n^{\mathrm{T}}\mathbf{x} + w_0. \tag{9.46}$$

Para clasificar nuevas observaciones a partir del modelo entrenado con los datos del conjunto $\mathcal{S}$, se evalúa el signo de la función $g(\mathbf{x})$, donde $\mathbf{x}$ es la observación a clasificar. Si $g(\mathbf{x}) > 0$, entonces $\mathbf{x}$ pertenecerá a la clase positiva y si $g(\mathbf{x}) < 0$, entonces pertenecerá a la clase negativa. Por último, si $g(\mathbf{x}) = 0$, entonces $\mathbf{x}$ pertenecerá a la frontera de decisión $\mathcal{F}$.

9.3.2. Clases solapadas

Hasta ahora se ha visto un caso linealmente separable, donde es posible encontrar un hiperplano que separe perfectamente todas las muestras de las dos clases. Ante este tipo de problemas, las SVM ofrecen modelos que separan exactamente los datos. Sin embargo, los problemas reales suelen presentar probabilidades condicionadas a la clase, $p(\mathbf{x}|y)$, que están solapadas. En estos casos no es posible hallar una separación lineal perfecta. Ante esta situación, la optimización anterior no encontraría una solución óptima. Por esta razón las SVM pueden ser modificadas para admitir que algunos puntos de entrenamiento puedan ser mal clasificados. Esto se puede conseguir relajando las restricciones (9.32) y (9.33). Para ello se introducen unas variables de holgura $\xi_n \geq 0$ con $n = 1, \ldots, N$ definidas como $\xi_n = |y_n - g(\mathbf{x}_n)|$. En la figura 9.7 se caracterizan estas variables. De este modo, las restricciones quedan redefinidas como:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 \geq 1 - \xi_n \quad \text{si } y_n = 1 \tag{9.47}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 \leq -1 + \xi_n \quad \text{si } y_n = -1 \tag{9.48}$$

Estas restricciones se pueden resumir en una única restricción al multiplicar por y_n en ambos lados de las desigualdades:

$$y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0) \geq 1 - \xi_n, \quad n = 1, \ldots, N. \tag{9.49}$$

Cada vez que se encuentre un error, la variable de holgura asociada a la observación mal clasificada será $\xi_n > 1$. Por lo tanto, $\sum_n \xi_n$ será una cota superior del número de errores de entrenamiento. Se debe considerar que, aunque las variables de holgura permiten trabajar con distribuciones solapadas, los datos anómalos[c] tienen una influencia importante ya que la cota superior de error crece linealmente con ξ_n.

Así pues, la función objetivo a optimizar para maximizar el margen teniendo en cuenta que las distribuciones de las clases pueden estar solapadas tendrá en cuenta la cota superior de error que estará controlada por un parámetro constante $C > 0$,

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N} \xi_n. \tag{9.50}$$

[c]En este caso concreto, por dato anómalo se entiende una observación que cae a una gran distancia de la frontera de decisión y en la región equivocada.

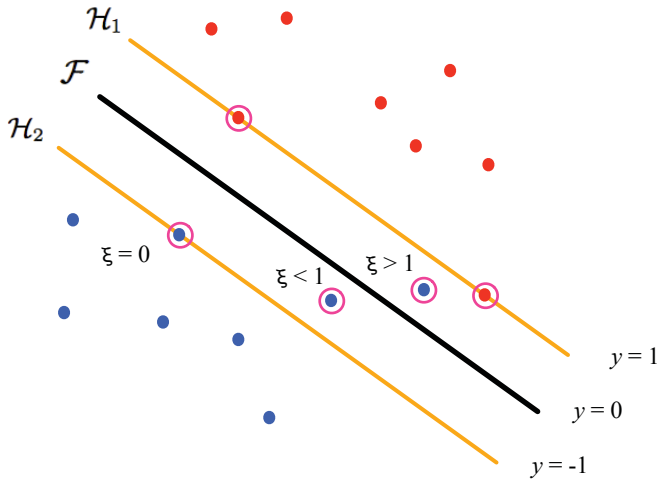


Figura 9.7: Ilustración de los valores de las variables de holgura $\xi_n \geq 0$. Si una observación se encuentra en la región correcta y en el margen o más allá de él, entonces $\xi_n = 0$. Si la observación cae en la región correcta, entre el margen y la frontera de decisión $\mathcal{F}$, entonces $\xi_n < 1$. Si la observación cae en la frontera de decisión, entonces $\xi_n = 1$. Si la observación cae en la región incorrecta, entonces $\xi_n > 1$. De ahí que $\sum_{n=1}^{N} \xi_n$ sea una cota superior del error de clasificación.

En el límite, cuando $C \to \infty$, el resultado será un modelo de SVM para el caso linealmente separable. Así pues, el objetivo para encontrar el hiperplano de separación óptimo se debe minimizar la expresión (9.50) sujeto a la restricción (9.49), junto a las variables de holgura $\xi_n \geq 0$. De nuevo, empleando multiplicadores de Lagrange tenemos

$$L_P(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n(y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0) - 1 + \xi_n) - \sum_{n=1}^{N}\mu_n\xi_n, \tag{9.51}$$

donde $\lambda_n \geq 0$ y $\mu_n \geq 0$ son los multiplicadores de Lagrange. En este caso, también se deben cumplir las condiciones de *Karush-Kuhn-Tucker*:

$$\begin{gathered}
\lambda_n \geq 0, \\
y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0) - 1 + \xi_n \geq 0, \\
\lambda_n(y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0) - 1 + \xi_n) = 0, \\
\mu_n \geq 0, \\
\xi_n \geq 0, \\
\mu_n\xi_n = 0.
\end{gathered}$$

Derivando L_P con respecto a $\mathbf{w}$, w_0 y $\boldsymbol{\xi}$ e igualando a cero y sustituyendo de manera análoga al caso linealmente separable se obtiene la función lagrangiana dual:

$$L_D(\boldsymbol{\lambda}) = \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{n=0}^{N}\sum_{m=0}^{N}\lambda_n\lambda_m y_n y_m \mathbf{x}_n\mathbf{x}_m. \tag{9.52}$$

La expresión es idéntica que para el caso linealmente separable, excepto por las restricciones. Si se deriva L_P respecto de ξ_n se obtiene que $\lambda_n = C - \mu_n$. Como $\lambda_n \geq 0$ y $\mu_n \geq 0$, se deduce que $\lambda_n \leq C$. Así pues, se debe minimizar 9.52 respecto a λ_n sujeto a las restricciones

$$0 \leq \lambda_n \leq C, \tag{9.53}$$

$$\sum_{n=1}^{N} \lambda_n y_n = 0, \quad n = 1, \ldots, N. \tag{9.54}$$

Las restricciones como la (9.53) se conocen como restricciones de caja (*box constraints*, en inglés). Finalmente, la solución para los parámetros es como antes

$$\mathbf{w}^* = \sum_{\mathbf{x}_n \in \mathcal{VS}} \lambda_n^* y_n \mathbf{x}_n, \tag{9.55}$$

$$w_0^* = y_n - \mathbf{w}^{*\mathrm{T}} \mathbf{x}_n, \quad \forall \mathbf{x}_n \in \mathcal{VS}. \tag{9.56}$$

La única diferencia con respecto al caso linealmente separable es que los multiplicadores de Lagrange tienen una cota superior en C. Esto quiere decir que el número de vectores soporte es, generalmente, mayor que para casos perfectamente separables mediante un hiperplano.

9.3.3. Kernels

Hasta ahora, incluso empleando márgenes con holgura, únicamente se han visto fronteras de decisión lineales definidas mediante hiperplanos. Sin embargo, existe la posibilidad de generalizar las fronteras de decisión a formas más complejas. Para ello se emplean métodos de *kernels*, que llevan a cabo una proyección de los datos de entrenamiento de su espacio original D-dimensional a un espacio M-dimensional donde $M \geq D$, de tal forma que los datos, que originalmente no son linealmente separables, puedan serlo en el nuevo espacio. Una de las grandes ventajas de las SVM es la posibilidad de emplear *kernels* para calcular las fronteras de decisión, lo que proporciona a estos métodos una gran flexibilidad para definir distintos tipos de fronteras.

Los modelos lineales pueden emplear una combinación lineal de los parámetros del modelo y una transformación de los datos, $\boldsymbol{\phi}(\mathbf{x})$, donde la función $\boldsymbol{\phi}$ aplica dicha transformación a los vectores originales para pasar de un espacio $\mathbb{R}^D$ a otro $\mathbb{R}^M$, esto es, $\boldsymbol{\phi} : \mathbb{R}^D \rightarrow \mathbb{R}^M$. De este modo, el modelo lineal puede ser

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) + w_0.$$

Es lógico pensar que si los datos se proyectan a un espacio de dimensiones mayores el coste computacional aumente en consecuencia. Sin embargo, lo métodos basados en *kernels* pueden evitar la proyección explícita de los datos. Como se ha visto, el vector $\mathbf{w}$ se puede expresar como una combinación lineal de todos o algunos de los datos de entrenamiento, en base a los vectores soporte escogidos, esto es, $\mathbf{w} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n$, donde $\alpha_n = \lambda_n y_n$. Así pues,

$$g(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n^{\mathrm{T}} \mathbf{x} + w_0. \tag{9.57}$$

Cuando se emplea la función $\boldsymbol{\phi}(\cdot)$ para transformar los datos, la expresión (9.57) toma la siguiente forma

$$g(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) + w_0. \tag{9.58}$$

A esta representación, empleando escalares α_n para la combinación lineal, se la conoce como representación dual. El producto $\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z})$, puede ser calculado de manera eficiente, sin necesidad de proyectar los datos, empleando una función *kernel*, que se define como:

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z}). \tag{9.59}$$

Al aplicar la función *kernel* en la expresión (9.58) se obtiene la expresión de la función discriminante en representación dual con *kernels*:

$$g(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) + w_0. \tag{9.60}$$

Para emplear funciones de *kernel* válidas se debe asegurar que el resultado del *kernel* es un producto escalar en el espacio muestral. Es decir,

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z}) = \sum_{m=1}^{M} \phi_m(\mathbf{x})\phi_m(\mathbf{z}).$$

Existe toda una aritmética para la construcción de *kernels* válidos si se tiene uno o más *kernels* de partida. Típicamente, los *kernels* más empleados son los siguientes:

- *Kernel* polinómico: $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + c)^d$.
- *Kernel* gaussiano: $k(\mathbf{x}, \mathbf{z}) = \exp(-\frac{1}{2\sigma^2}||\mathbf{x}-\mathbf{z}||^2)$. Donde $\sigma > 0$ es el parámetro que controla la anchura de la gaussiana.
- *Kernel* laplaciano: $k(\mathbf{x}, \mathbf{z}) = \exp(-\frac{1}{\sigma}|\mathbf{x}-\mathbf{z}|)$. Donde, del mismo modo, $\sigma > 0$ es el parámetro que controla la anchura de la laplaciana.

La combinación de un *kernel* adecuado y márgenes con holgura proporcionan a las SVM la capacidad de crear modelos a partir de casi cualquier tipo de problema con el que se encuentre. De ahí su éxito como método para clasificar y ajustar regresiones en los últimos años.

Ejemplo 9.1 (Caracterización de *kernels* válidos)
Si tenemos un espacio muestral bidimensional con $\mathbf{x} = (x_1, x_2)$ y queremos comprobar si un kernel $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z})^2$ es válido, siendo $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, todo lo que debemos hacer es desarrollar el producto escalar y comprobar que se puede descomponer en el producto escalar de las transformaciones de los vectores cuando se les aplica una función $\boldsymbol{\phi}(\cdot)$. Veamos,

$$\begin{aligned}
k(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^{\mathrm{T}}\mathbf{z})^2 \\
&= (x_1 z_1 + x_2 z_2)^2 \\
&= (x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2) \\
&= (x_1^2, \sqrt{2}x_1 x_2, x_2^2) \cdot \begin{pmatrix} z_1^2 \\ \sqrt{2}z_1 z_2 \\ z_2^2 \end{pmatrix} \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z}).
\end{aligned}$$

Siendo la función de transformación $\boldsymbol{\phi}(\mathbf{x}) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$. Es decir, la función de transformación es una aplicación $\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^3$, donde las funciones base se corresponden con $\phi_1(\mathbf{x}) = x_1^2$, $\phi_2(\mathbf{x}) = \sqrt{2}x_1 x_2$ y $\phi_3(\mathbf{x}) = x_2^2$.

9.4. Notas bibliográficas

El Aprendizaje Profundo, también conocido del inglés como *Deep Learning*, ha conseguido alcanzar niveles comparables a soluciones humanas en los últimos años. El aprendizaje profundo permite que los modelos computacionales compuestos de múltiples capas de procesamiento aprendan múltiples niveles de abstracción a partir de los datos [123]. Implementaciones de librerías de redes neurales profundas como TensorFlow y Theano, computación gráfica y en red y la gran cantidad de datos disponibles han permitido significativos avances en campos como la visión por computador, el reconocimiento del habla, el procesamiento del lenguaje natural, el reconocimiento de audio, el filtrado de redes sociales, la traducción automática, la bioinformática, el diseño de fármacos, la segmentación, clasificación y pronóstico de imágenes médicas, la inspección de materiales y los programas de juegos de mesa [124].

Capítulo 10

Modelos generativos

En los modelos generativos, la función discriminante se determina a partir de las probabilidades de la clase dado el objeto, $P(y|\mathbf{x})$. Para ello, los modelos generativos utilizan la estimación de las funciones de densidad de probabilidad condicionadas por la clase $p(\mathbf{x}|y)$ y las probabilidades de las clases a priori $P(y)$ para resolver el problema de clasificación. Es decir, la probabilidad a posteriori de la clase se genera a partir de la probabilidad a priori de la clase y de la probabilidad condicionada. Esto supone resolver un problema más complejo que el de la pura clasificación, pero a cambio podemos calcular las probabilidades a posteriori de las clases para los nuevos casos observados.

10.1. Clasificador de Bernoulli

El clasificador generativo de Bernoulli es un clasificador de Bayes (ver sección. 8.2.1) cuyas funciones de densidad de probabilidad de cada clase son distribuciones Bernoulli D-dimensionales, $\mathbf{x}|c \sim Be(\mathbf{p}_c)$ (3.2.1).

Por lo tanto, la regla de clasificación que define un clasificador de Bernoulli será elegir aquella clase cuya función discriminante sea máxima,

$$\hat{c}^* \leftarrow \arg\max_c g_c(\mathbf{x})$$

donde las funciones discrimantes son (8.21),

$$\begin{aligned} g_c(\mathbf{x}) &= p(c|\mathbf{x}) && (10.1)\\ &= \frac{p(c)p(\mathbf{x}|c)}{p(\mathbf{x})} && (10.2) \end{aligned}$$

Puesto que el denominador de la ecuación $p(\mathbf{x})$ es común a todas las clases, se deduce que la clase que maximice (10.2) será la misma clase que maximice al numerador, esto es,

$$\begin{aligned} \arg\max_c \left\{ \frac{p(c)p(\mathbf{x}|c)}{p(\mathbf{x})} \right\} &= \arg\max_c \Big\{ p(c)p(\mathbf{x}|c) \Big\} && (10.3)\\ &= \arg\max_c \Big\{ \log p(c) + \log p(\mathbf{x}|c) \Big\} && (10.4)\\ &= \arg\max_c \Big\{ \log p(c) + \sum_{d=1}^{D} x_d \log p_{cd} + (1 - x_d) \log 1 - p_{cd} \Big\} && (10.5) \end{aligned}$$

Siendo c cada una de las clases y d cada una de las variables. Se puede observar que la expresión final es una función lineal de $\mathbf{x}$, ya que $g_c(\mathbf{x})$ se puede expresar como

$$g_c(\mathbf{x}) = \mathbf{w}_c^{\mathrm{T}}\mathbf{x} + w_{c0}, \tag{10.6}$$

donde

$$\mathbf{w}_c = \log \mathbf{p}_c - \log \mathbf{1} - \mathbf{p}_c \tag{10.7}$$

$$w_{c0} = \log p(c) + \log \mathbf{1} - \mathbf{p}_c. \tag{10.8}$$

Si disponemos de $N = \sum_c N_c$ casos, podemos obtener la estimación de los parámetros de nuestro clasificador mediante la maximización de la función log-verosimilitud

$$\log \mathcal{L}(\Theta) = \sum_n \log p(c_n + \sum_d x_{nd} \log p_{cnd} + (1 - x_{nd}) \log 1 - p_{cnd},$$

como

$$\hat{p}(c) = \frac{N_c}{N}, \tag{10.9}$$

$$\hat{\mathbf{p}}_c = \frac{1}{N_c} \sum_{n_c} \mathbf{x}_n. \tag{10.10}$$

La estimación por máxima verosimilitud es asintóticamente insesgada. Esto significa que los parámetros estimados se aproximan al parámetro real cuando el número de casos disponible tiende a infinito. En términos prácticos, cuando el número de casos N es suficientemente grande. En caso contrario, esto es, cuando se dispone de un número de casos insuficiente el clasificador estimado por máxima verosimilitud corre el riesgo de sobreajustar los parámetros a las particularidades de los casos de entrenamiento y no generaliza lo suficiente. Este problema se agudiza cuando nuestra muestra no presenta todos los eventos posibles del espacio D-dimensional, lo que provoca que alguno de los parámetros sea nulo, provocando un aumento del error del clasificador debido a estimaciones que se anulan por esta causa.

Una solución óptima para evitar este problema es utilizar otros criterios de estimación como la inferencia bayesiana. Otra solución es utilizar métodos de regularización o suavizado. Esto quiere decir que el valor de los parámetros aprendidos se "suaviza" para evitar los problemas comentados anteriormente.

En distribuciones de Bernoulli podemos utilizar dos métodos de suavizado:

- Recortado simple. A partir de un umbral definido ϵ, redefinimos cada parámetro estimado $\hat{p}_{cd}$ como

$$\tilde{p}_{cd} = \begin{cases} \epsilon & , \text{si } \hat{p}_{cd} < \epsilon \\ 1 - \epsilon & , \text{si } \hat{p}_{cd} > 1 - \epsilon \\ \hat{p}_{cd} & \text{en otro caso.} \end{cases}$$

- Simulación de muestras. Añadimos una muestra artificial con todos los valores a 0 y otra con todos los valores a 1 en cada clase. Esto consigue que ninguno de los parámetros sea nulo. Este planteamiento equivale a modificar la estimación de $\hat{p}_c$ como

$$\hat{\mathbf{p}}_c = \frac{1}{N_c + 2}(\mathbf{1} + \sum_{n_c} \mathbf{x}_n).$$

10.2. Clasificador multinomial

De forma similar al clasificador de Bernoulli, podemos definir el clasificador multinomial cuando la observación responde a una variable aleatoria multinomial $\mathbf{X} = (X_1, \ldots, X_D), \sum_{d=1}^{D} X_d = L$ (ver sección 3.2.3) de parámetros $L, \mathbf{p}$, y por lo tanto las funciones de densidad de probabilidad de cada clase son distribuciones multinomiales $\mathbf{x}|c \sim Mult(L, \mathbf{p}_c)$ (3.2.3).

Por lo tanto, la regla de clasificación definida por un clasificador multinomial será elegir aquella clase cuya función discriminante sea máxima,

$$\hat{c}^* \leftarrow \arg\max_c g_c(\mathbf{x})$$

con funciones discriminantes lineales de la forma

$$\begin{aligned}
\arg\max_c \Big\{ g_c(\mathbf{x}) \Big\} &= \arg\max_c \Big\{ p(c|\mathbf{x}) \Big\} && (10.11)\\
&= \arg\max_c \Big\{ \log p(c) + \log L! - \sum_d \log x_d! + \sum_d x_d \log p_{cd} \Big\} && (10.12)\\
&= \arg\max_c \Big\{ \log p(c) + \sum_d x_d \log p_{cd} \Big\} && (10.13)\\
&= \arg\max_c \Big\{ w_{c0} + \mathbf{w}_c^{\mathrm{T}} \mathbf{x} \Big\} && (10.14)
\end{aligned}$$

donde

$$\begin{aligned}
w_{c0} &= \log p(c) && (10.15)\\
\mathbf{w}_c &= \log p_c. && (10.16)
\end{aligned}$$

El estimador máximo-verosímil para el clasificador multinomial con N muestras de entrenamiento es de nuevo compatible con el valor esperado para cada clase de la distribución multinomial:

$$\begin{aligned}
\hat{p}(c) &= \frac{N_c}{N} && (10.17)\\
\hat{\mathbf{p}}_c &= \frac{1}{\sum_{n_c} \sum_d \mathbf{x}_{nd}} \sum_{n_c} \mathbf{x}_n && (10.18)
\end{aligned}$$

Un uso típico del clasificador multinomial será la clasificación de textos mediante el conteo de la aparición de palabras clave en el documento. Sin embargo, no todos los documentos tendrán la misma longitud, por lo que en un documento de 4000 palabras, será normal encontrar más ocurrencias de una palabra que en un documento de 500 palabras, con lo cual la L_n de cada documento podrá ser diferente. Así pues, para hacer comparables los conteos $\mathbf{x_n}$ de los documentos, debemos normalizarlos por su L_n para hacerlos comparables. Para ello, podemos dividir el vector de conteos por L_n, y (opcionalmente) multiplicarlo por una constante, como la longitud media de los documentos $\bar{L}$, es decir, sustituir $\mathbf{x}_n$ por

$$\tilde{x}_n = \bar{L} \frac{\mathbf{x}_n}{L_n}$$

Una vez más, en caso de tener un conjunto amplio de sucesos aleatorios (D grande) y una muestra N limitada, la estimación mediante máxima verosimilitud puede dar valores nulos en algunos de los parámetros $\hat{\mathbf{p}}_c$. Esto hará que el cálculo de $p(\mathbf{x}|c)$ sea 0, lo que resulta poco operativo a la hora de resolver la decisión. Para solucionar este problema, podemos optar por por varias alternativas:

- Suavizado de Laplace. Sumaremos una pequeña cantidad $\epsilon > 0$ a todos los contadores $\sum_{n_c} \mathbf{x}_{nd}$ y volvemos a normalizar.
- Descuento absoluto. Se descuenta una cantidad fija b a todas las estimaciones no nulas, y se redistribuye entre los parámetros nulos (backing-off) o entre todos los parámetros (interpolación).

10.3. Clasificador gaussiano

El clasificador generativo gaussiano es también un clasificador de Bayes (sección 8.2.1) donde las funciones de densidad de probabilidad de cada clase son distribuciones gaussianas D-dimensionales, $\mathbf{x}|c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

De nuevo, la regla de decisión es la habitual en los clasificadores de Bayes, esto es, se escoge la clase que maximiza las funciones discriminantes

$$\hat{c}^* \leftarrow \arg\max_c g_c(\mathbf{x}),$$

En el caso de un clasificador gaussiano, como en los anteriores clasificadores de Bayes, las funciones discriminantes son equivalentes a obtener la probabilidad a posteriori de la clase, esto es, $g_c(\mathbf{x}) = p(c|\mathbf{x})$ y, por tanto, aplicando la regla de Bayes

$$g_c(\mathbf{x}) = \frac{p(c)p(\mathbf{x}|c)}{p(\mathbf{x})} \tag{10.19}$$

donde se asume que la probabilidad condicionada a la clase es una gaussiana multivariante $p(\mathbf{x}|c) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$

$$\equiv \quad p(c)p(\mathbf{x}|c) \tag{10.20}$$
$$\equiv \quad \log p(c) + \log p(\mathbf{x}|c) \tag{10.21}$$
$$\equiv \quad \log p(c) - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma_c\mathbf{x} - \mathbf{x}^{\mathrm{T}}\Sigma_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\log|\Sigma_c| - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\Sigma_c^{-1}\boldsymbol{\mu}_c. \tag{10.22}$$

Se puede observar que las funciones son cuadráticas, ya que $g_c(\mathbf{x})$ se puede expresar como

$$g_c(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{W}_c\mathbf{x} + \mathbf{w}_c^{\mathrm{T}}\mathbf{x} + w_{c0}, \tag{10.23}$$

donde

$$\mathbf{W}_c = -\frac{1}{2}\Sigma_c^{-1} \tag{10.24}$$
$$\mathbf{w}_c = \Sigma_c^{-1}\boldsymbol{\mu}_c \tag{10.25}$$
$$w_{c0} = \log p(c) - \frac{1}{2}\log|\Sigma_c| - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\Sigma_c^{-1}\boldsymbol{\mu}_c. \tag{10.26}$$

Sin embargo, cuando las matrices de covarianza son comunes a todas las clases, esto es, $\Sigma_c = \Sigma$, entonces el clasificador gaussiano es lineal ya que el parámetro $\mathbf{W}_c$ también es común y, por tanto, no aporta ninguna información para clasificar. Así pues, cuando las matrices de covarianza de las clases son comunes, la función discriminante queda

$$g_c(\mathbf{x}) = \mathbf{w}_c^{\mathrm{T}}\mathbf{x} + w_{c0}, \tag{10.27}$$

con

$$\mathbf{w}_c = \Sigma_c^{-1}\boldsymbol{\mu}_c \tag{10.28}$$

$$w_{c0} = \log p(c) - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\Sigma_c^{-1}\boldsymbol{\mu}_c. \tag{10.29}$$

Para estimar los parámetros de la distribución y de las probabilidades a priori se emplea el método de máxima verosimilitud. Si se dispone de un conjunto de datos $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, con $\mathbf{x}_i \in \mathbb{R}^D$ y siendo $y_i \in \mathcal{C}$ la clase a la que pertenece la observación i-ésima, entonces se puede maximizar la función log-verosimilitud $\log L(\Theta)$. Derivando $\log L(\Theta)$ con respecto a $\Theta = \{p(1), \ldots, p(C); \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C; \Sigma_1, \ldots, \Sigma_C\}$ e igualando a cero nos da las siguientes estimaciones:

$$\hat{p}(c) = \frac{N_c}{N} \tag{10.30}$$

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{n:c_n=c}\mathbf{x}_n \tag{10.31}$$

$$\hat{\Sigma}_c = \frac{1}{N_c}\sum_{n:c_n=c}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}, \tag{10.32}$$

donde $c = 1, \ldots, C$, N es el número total de observaciones y N_c es el número total de observaciones de la clase c.

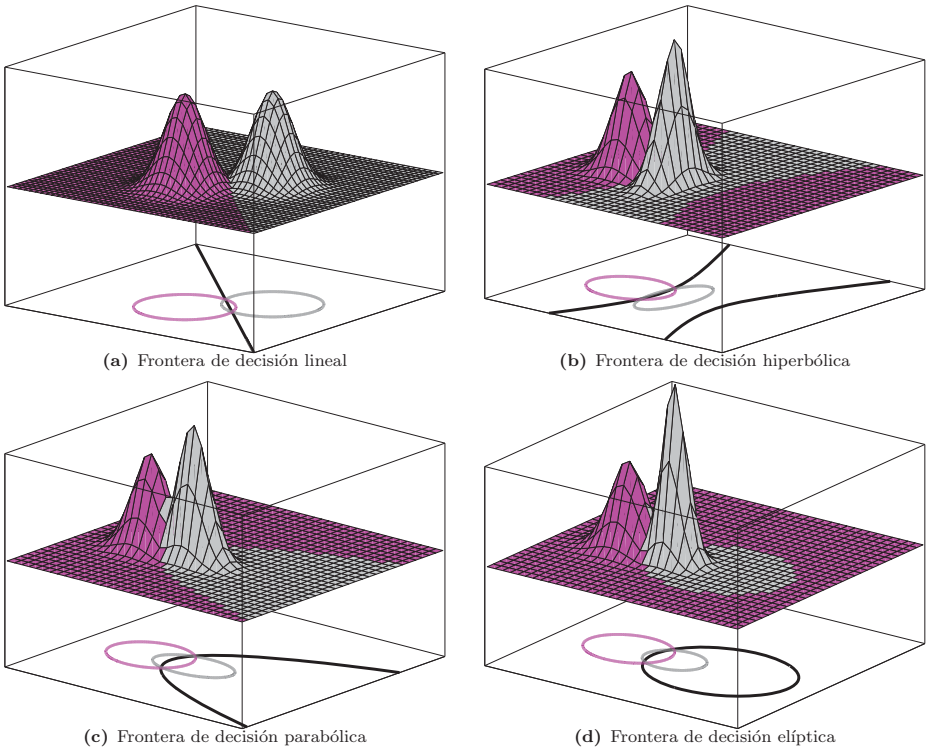
Como se ha visto, los clasificadores gaussianos permiten disponer de funciones discriminantes lineales y funciones discriminantes cuadráticas. Con las primeras se pueden describir fronteras mediante hiperplanos. Con las segundas pueden describir fronteras mediante curvas cónicas (ver figura 10.1).

Los clasificadores gaussianos disponen también de técnicas para suavizar los parámetros con el fin de evitar sobreajuste, especialmente cuando se disponen de pocas observaciones ya que las estimaciones de las matrices de covarianza en estas condiciones son poco robustas. Para solucionar este problema se disponen de, al menos, dos alternativas:

- Covarianza con umbral: establece un umbral θ para los coeficientes de correlación de modo que si se observa que $|\rho_{ij}| < 1 - \theta$ entonces se anula la covarianza correspondiente, esto es, $\sigma_{ij} = 0$. Si se estable un umbral $\theta = 0$ entonces se obtiene una matriz de covarianza diagonal.
- Regularización: se trata de realizar una combinación lineal convexa de la matriz de covarianza de la clase Σ_c, la matriz de covarianzas común Σ y la matriz de identidad $\boldsymbol{I}$:

  $$\hat{\Sigma}_c = \alpha_1\Sigma_c + \alpha_2\Sigma + \alpha_3\boldsymbol{I},$$

  donde $\alpha_1, \alpha_2, \alpha_3 \geq 0$ y $\alpha_1 + \alpha_2 + \alpha_3 = 1$.



(a) Frontera de decisión lineal

(b) Frontera de decisión hiperbólica

(c) Frontera de decisión parabólica

(d) Frontera de decisión elíptica

Figura 10.1: Ilustración de los posibles tipos de fronteras de decisión en dos dimensiones empleando un clasificador gaussiano para discriminar dos clases (gris y magenta).

Capítulo 11

Modelos discriminativos

En los modelos discriminativos, la función discriminante se determina a partir de las probabilidades de la clase dado el objeto, $P(y|\mathbf{x})$. Sin embargo, al contrario que en los modelos generativos vistos en el capítulo anterior, dicha probabilidad se calcula directamente y no a través de la descomposición mediante la regla de Bayes. Un caso típico de modelo discriminativo es la regresión logística, que es uno de los modelos más consolidados y estudiados dado que es un método que procede originalmente de la disciplina estadística. Para profundizar en estos métodos es recomendable acercarse a las obras de Hosmer y Lemeshow [125] sobre regresión logística y de McCullagh y Nelder [126] sobre modelos lineales generalizados.

11.1. Regresión logística

Cuando se emplea el vector original que representa a un objeto $\mathbf{x}$ un modelo discriminativo de regresión logística se limita a describir fronteras de tipo lineal. Sin embargo, existe la posibilidad de utilizar funciones de expansión sobre el vector $\mathbf{x}$ que nos permitirán hacer modelos no lineales que describan fronteras polinómicas de todo tipo. Una función de expansión se representa como $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$ donde cada componente $\phi_m(\mathbf{x})$ presenta una transformación sobre los datos originales. De ahora en adelante desarrollaremos la explicación empleando como entrada la expansión $\phi(\mathbf{x})$.

En un problema de clasificación de dos clases, definimos un modelo discriminante lineal generalizado sobre la expansión de $\phi(\mathbf{x})$ como el logaritmo de la razón entre las probabilidades de la clase $y = 1$ e $y = 0$, así pues

$$\log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) \tag{11.1}$$

Como $p(y = 1|\mathbf{x}) = 1 - p(y = 0|\mathbf{x})$, podemos obtener la probabilidad a posteriori de la clase $y = 1$ despejando la expresión (11.1)

$$p(y=1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))} \tag{11.2}$$

Esta función recibe el nombre de función logística. Esta función define una curva sigmoide con codominio $(0, 1)$. Es la misma función que se emplea como función de activación en la redes

neuronales artificiales. Es fácil observar que la probabilidad a posteriori de la clase $y = 0$ se obtiene con facilidad a partir de la igualdad anterior (11.2),

$$\begin{aligned} p(y=0|\mathbf{x}) &= 1 - p(y=0|\mathbf{x}) \\ &= \frac{\exp(-\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))}{1+\exp(-\mathbf{w}^{T}\phi(\mathbf{x}))} \\ &= \frac{1}{1+\exp(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))} \end{aligned} \tag{11.3}$$

Además de poder obtener la probabilidad a posteriori de cada clase, la regresión logística permite mediante test de hipótesis clásicos obtener la importancia relativa de cada variable de entrada, lo que nos aporta una información muy valiosa a la hora de seleccionar variables. La idea se fundamenta en la importancia de evitar modelos sobreajustados ya que la regresión logística es muy propensa a este problema cuando los datos son linealmente separables. Por lo tanto, lo que se busca es obtener un modelo parsimonioso, esto es, un modelo con pocos parámetros que sea capaz de explicar los datos. Estos modelos nos permiten también explicar la influencia de cada una de las variables de entrada sobre la variable de salida. Esta característica se explica a continuación.

Interpretación de las variables en regresión logística

Para explicar esta característica trabajaremos con una función de expansión identidad tal que $\phi(\mathbf{x}) = \mathbf{x}$. De este modo, podemos desarrollar la expresión (11.1) para analizar la interpretación de las variables independientes,

$$\log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = w_0 + w_1 x_1 + \ldots + w_D x_D \tag{11.4}$$

La influencia de la variable i-ésima x_i viene ponderada por el valor del parámetro asociado w_i. Para analizar la influencia de esta variable debemos suponer que el resto de variables tienen valores nulos. De este modo, podremos observar que

$$\frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \exp(w_i x_i) \tag{11.5}$$

es decir, el incremento en una unidad de la variable x_i incrementa la razón en $\exp(w_i)$. Por lo tanto, podemos deducir que si $w_i > 0$ entonces la variable x_i es un factor que aumenta la razón entre la probabilidad de la clase $y = 1$ y la clase $y = 0$. Dicho de otro modo, aumenta la probabilidad de la clase 1 y disminuye la probabilidad de la clase 0. Al contrario, si $w_i < 0$ entonces la variable x_i aumenta la probabilidad de la clase 0 y disminuye la probabilidad de la clase 1. Además, cuanto mayor sea la magnitud de w_i sabremos que la influencia en la razón de probabilidades será mayor. Este efecto, sin embargo, debe comprobarse con un test de hipótesis sobre si dicha influencia es realmente significativa (ver Hosmer y Lemeshow [125]).

Por último, se puede observar cómo el término independiente w_0 nos proporciona información acerca de las probabilidades a priori de las clases, dado que

$$\frac{p(y=1)}{p(y=0)} = \exp(w_0) \tag{11.6}$$

11.1.1. Estimación por máxima verosimilitud

Para estimar por máxima verosimilitud, lo primero que debemos hacer es definir la función verosimilitud. La verosimilitud para cada observación disponible $(\mathbf{x}_n, y_n)$ es la probabilidad a

posteriori $p(y = y_n|\mathbf{x}_n)$, enfatizando su dependencia del modelo con los parámetros $\mathbf{w}$, podemos expresarla como

$$p(y = y_n|\mathbf{x}, \mathbf{w}) = p(y = 1|\mathbf{x}_n)^{y_n}(1 - p(y = 1|\mathbf{x}_n))^{(1-y_n)} \tag{11.7}$$

recordemos que la expresión de la probabilidad a posteriori depende de los parámetros $\mathbf{w}$ en virtud de la expresión (11.2).

Dada una muestra de entrenamiento, $\mathcal{S} = (\mathbf{x}_n, y_n), n = 1, \ldots, N; \mathbf{x_n} \in \mathbb{R}^d, y_n \in \{0, 1\}$, la función de verosimilitud dados los datos puede escribirse como el producto de la probabilidad a posteriori de cada observación que se asume independiente e idénticamente distribuida al resto de observaciones,

$$\ell(\mathbf{w}|\mathcal{S}) = \prod_{n=1}^{N} p(y = 1|\mathbf{x}_n)^{y_n}(1 - p(y = 1|\mathbf{x}_n))^{(1-y_n)}, \tag{11.8}$$

Generalmente se emplea el logaritmo de la verosimilitud por razones computacionales, de este modo la log-verosimilitud queda

$$\mathcal{L}(\mathbf{w}|\mathcal{S}) = \log \ell(\mathbf{w}|\mathcal{S}) = \log p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{w}) = \sum_{n=1}^{N} y_n \log(p(y = 1|\mathbf{x}_n)) + (1 - y_n) \log(1 - p(y = 1|\mathbf{x}_n)). \tag{11.9}$$

Cuyo gradiente respecto a $\mathbf{w}$ es

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \sum_{n=1}^{N} \phi(\mathbf{x_n})(y_n - p(y = 1|\mathbf{x}_n)) \tag{11.10}$$

donde $p(y = 1|\mathbf{x}_n)$ viene dado por la expresión (11.2).

Al obtener una expresión no lineal, debido al uso de la función logística, no es posible calcular una solución cerrada para $\hat{\mathbf{w}}$, por lo que se debe aplicar un esquema iterativo de optimización. Se suele emplear para ello una función de error que es

$$E(\mathbf{w}) = -\mathcal{L}(\mathbf{w}|\mathcal{S}) \tag{11.11}$$

cuyo gradiente respecto de $\mathbf{w}$ es

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi(\mathbf{x_n})(p(y = 1|\mathbf{x}_n) - y_n) \tag{11.12}$$

La función de error es convexa, por lo que es posible minimizarla mediante el algoritmo iterativo de Newton-Raphson (ver sección E.1). La solución obtenida se denomina Iterative Reweighted Least Squares (IRLS), y en cada iteración el vector de parámetros $\mathbf{w}$ se actualiza de la siguiente forma

$$\mathbf{w}^{(t+1)} = (\boldsymbol{\phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(\mathbf{t})}, \tag{11.13}$$

donde (ϕ) es la matriz $N \times M$, donde cada fila i es $\phi(\mathbf{x_n})^{\mathrm{T}}$, $\mathbf{V}$ una matriz diagonal donde cada valor de la diagonal es la varianza de la distribución de Bernoulli con $V_{nn} = p(y = 1|x_n)(1-p(y = 1|x_n))$, y

$$\mathbf{z}^{(t)} = \boldsymbol{\phi}\mathbf{w}^{(t)} - \mathbf{V}^{-1}(\mathbf{p} - \mathbf{y}), \tag{11.14}$$

donde p es un vector de dimensión N donde el elemento n toma el valor $p(y_n|x_n)$.

Como ya se ha advertido, la estimación de modelos de regresión logística por máxima verosimilitud cae fácilmente en el sobreajuste del modelo al conjunto de entrenamiento $\mathcal{S}$; en efecto, para un problema linealmente separable, el rango de valores que puede tomar $\mathbf{w}$ permite encontrar una solución donde las probabilidades a posteriori de las muestras de entrenamiento para las clases etiquetadas sean 1, aunque los pesos tenga que llegar a valores en el infinito.

Ejemplo 11.1 (Predicción de metástasis en ganglios no-centinelas)
Los ganglios linfáticos axilares constituyen las estaciones de drenaje de la glándula mamaria. Esto ha justificado que las guías prácticas aplicables en cirugía del cáncer de mama recomienden la linfadenectomía axilar completa (ALND: axillary lymph node dissection) en pacientes con ganglios centinelas metastásicos para evitar la recaída axilar. Estudios recientes indican que entre el 30 % y el 70 % de las pacientes con ganglios centinela positivos no presentan ganglios no-centinela afectados, lo que podría evitar la linfadenectomía. Así pues, estamos interesados en desarrollar un modelo predictivo de la afección de otros ganglios linfáticos axilares cuando el ganglio centinela es metastásico. Este modelo permitiría decidir si es necesario realizar una linfadenectomía axilar completa o limitar el tratamiento a la extirpación de los ganglios centinela.

Supongamos que podemos adquirir durante la operación quirúrgica dos variables discriminantes (e.g. el tamaño de la metástasis en el ganglio centinela, x_1; y el tamaño tumoral, x_2) para decidir si existe algún ganglio axilar no-centinela afectado de metástasis, y.

Nos planteamos resolver el problema como un problema de clasificación. Para ello, utilizaremos un modelo lineal generalizado de regresión logística, sobre el que aplicar la siguiente regla de decisión

$$\text{Si } p(y=1|x_1,x_2) > 0{,}5 \text{ entonces } y=1$$

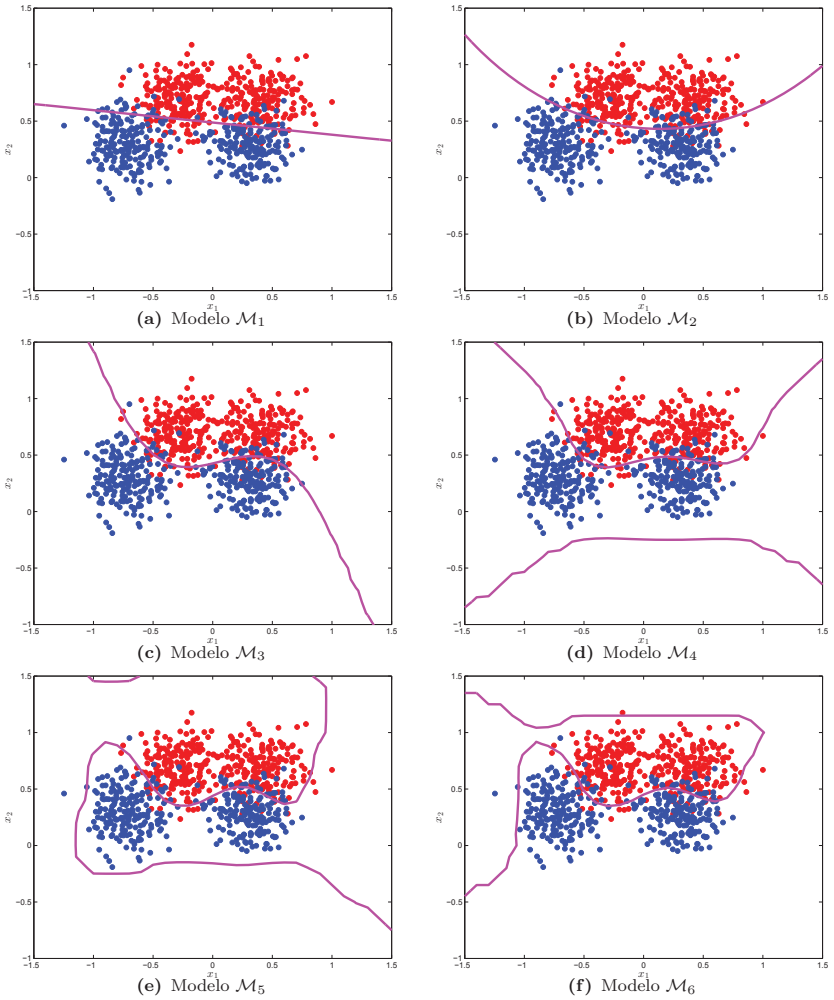
Para el entrenamiento del modelo se han recopilado 500 casos quirúrgicamente tratados, de los que se dispone la confirmación histológica de la existencia de metástasis en ganglios no-centinela[a].

Para observar el efecto de la complejidad del modelo en el aprendizaje se han entrenado seis modelos. El primer modelo $\mathcal{M}_1$ se basa en un polinomio de orden uno cuya función de expansión es $\phi(\mathbf{x}) = (1, 1, x_1, x_2)$, por lo que el vector de parámetros $\mathbf{w}$ tendrá tres componentes; el segundo modelo $\mathcal{M}_2$ se basa en un polinomio de orden dos, cuya función de expansión es $\phi(\mathbf{x}) = (1, 1, x_1, x_2, x_1^2, x_2^2)$; el tercer modelo $\mathcal{M}_3$ se basa en un polinomio de orden tres, cuya función de expansión es $\phi(\mathbf{x}) = (1, 1, x_1, x_2, x_1^2, x_2^2, x_1^3, x_2^3)$; siguiendo la estructura de estas funciones de expansión, el cuarto modelo $\mathcal{M}_4$, quinto modelo $\mathcal{M}_5$ y sexto modelo $\mathcal{M}_6$ se basan en polinomios de orden cuatro, cinco y seis, respectivamente. Estos modelos no capturan interacciones entre variables y requieren $G(D+1)$ componentes en el vector de parámetros $\mathbf{w}$, siendo D el número de dimensiones o variables independientes y G el grado del polinomio. En este caso los modelos cuentan con 3, 5, 7, 9, 11 y 13 parámetros respectivamente.

Pretendemos obtener una estimación puntual de $\mathbf{w}$ para cada modelo de regresión logística por máxima verosimilitud mediante IRLS. Para ello se emplean los datos de entrenamiento para ajustar los parámetros del modelo y se evalúa usando datos de evaluación independientes de los datos de entrenamiento. La figura 11.1 muestra las fronteras de decisión descritas por cada modelo.

En la siguiente sección compararemos las prestaciones de los modelos discriminativos de regresión logística estimados mediante máxima verosimilitud con sus equivalentes estimados mediante inferencia bayesiana.

[a]En este ejemplo utilizamos datos simulados y por lo tanto no relacionados a casos reales, específicamente, los datos pertenecen al corpus "rip" utilizado por Marc Girolami en el curso APG en la UPV.



(a) Modelo $\mathcal{M}_1$ **(b)** Modelo $\mathcal{M}_2$

(c) Modelo $\mathcal{M}_3$ **(d)** Modelo $\mathcal{M}_4$

(e) Modelo $\mathcal{M}_5$ **(f)** Modelo $\mathcal{M}_6$

Figura 11.1: Fronteras de decisión descritas por los modelos de regresión logística estimados con el algoritmo IRLS.

11.1.2. Regresión logística bayesiana

Como hemos visto, el estimador máximo verosímil nos permite calcular una estimación puntual que maximiza la verosimilitud de los parámetros $\mathbf{w}$ dados los datos observados. La estimación por inferencia bayesiana introduce los siguientes cambios conceptuales: en primer lugar, asume que los parámetros son una variable aleatoria y, por tanto, podemos obtener una distribución de dichos parámetros; en segundo lugar, asume que las probabilidades representan creencias subjetivas; en tercer lugar, estas creencias subjetivas se representan mediante una probabilidad a priori de los parámetros $p(\mathbf{w}|\alpha)$ y se modifican a partir de la observación de los datos. La probabilidad estimada de los parámetros se conoce como probabilidad a posteriori de los parámetros $\mathbf{w}$, dada la muestra de entrenamiento observada $\mathcal{S} = (\mathbf{X}, \mathbf{y})$,

$$p(\mathbf{w}|\mathbf{y},\mathbf{X},\alpha) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\alpha)}{p(\mathbf{y}|\mathbf{X},\alpha)}, \tag{11.15}$$

$$= \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\alpha)}{p(\mathbf{y}|\mathbf{X},\alpha)}, \tag{11.16}$$

donde

$$p(\mathbf{y}|\mathbf{X},\alpha) = \int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\alpha)d\mathbf{w},$$

y asumiendo una distribución inicial (a priori) de los parámetros $p(\mathbf{w}|\alpha) = N(0,\alpha^{-1}\mathbf{I})$. Además, también se emplea la función verosimilitud a partir del conjunto de las N muestras independientes e idénticamente distribuidas con la expresión:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \frac{\exp(\mathbf{w}^T\phi(\mathbf{x}))^{y_n}}{1+\exp(\mathbf{w}^T\phi(\mathbf{x}))}.$$

Uno de los problemas con el que nos encontramos es que la integral multidimensional no puede ser calculada analíticamente, por lo que necesitaremos calcularla mediante una aproximación determinista como la aproximación de Laplace o mediante una aproximación numérica, como los métodos de muestreo como Markov Chain Monte Carlo (MCMC). A continuación, el método de Laplace que aplicaremos propone aproximar la distribución a posteriori a una distribución gaussiana.

Estimación máximo a posteriori

Siguiendo la aproximación de Laplace, cuando el número de casos N es mucho mayor que el número de parámetros $\mathbf{w}$, la distribución final del modelo es aproximadamente gaussiana multivariante con valor medio $\mathbf{w}_{MAP}$ aquel que maximiza la distribución a posteriori y cuya matriz de covarianzas ($\mathbf{C}$) captura la curvatura de la distribución a posteriori

$$\mathbf{C} = -\Big(\frac{\partial^2}{\partial\mathbf{w}\mathbf{w}^{\mathrm{T}}}\log p(\mathbf{t},\mathbf{w}|\mathbf{X},\alpha)\Big)^{-1} \tag{11.17}$$

en el valor máximo $\mathbf{w}_{MAP}$. Por lo que se puede establecer que

$$p(\mathbf{w}|\mathbf{y},\mathbf{X},\alpha) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\alpha)}{p(\mathbf{y}|\mathbf{X},\alpha)} \approx \mathcal{N}(\mathbf{w}_{MAP},\mathbf{C}).$$

Así pues, podremos aproximar la distribución a posteriori de $\mathbf{w}$ si obtenemos el valor máximo $\mathbf{w}_{MAP}$ de dicha distribución y su curvatura $\mathbf{C}^{-1}$ en dicho punto. Como $p(\mathbf{y}|\mathbf{X},\alpha)$ no depende de los parámetros, el máximo de la distribución a posteriori $p(\mathbf{w}|\mathbf{y},\mathbf{X},\alpha)$ es el mismo que el obtenido por la verosimilitud conjunta $p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\alpha)$. Por lo tanto, la función a maximizar será el logaritmo de la verosimilitud conjunta,

$$\mathcal{L} = \log p(\mathbf{y},\mathbf{w}|\mathbf{X},\alpha) = \sum_{i=1}^{N} y_n\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - \log(1+\exp(w^{\mathrm{T}}\phi(\mathbf{x}_n))) - \frac{1}{\alpha}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \frac{D}{2}\log(2\pi\alpha^2),$$

cuya primera derivada es

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = \sum_{i=1}^{N} y_n\phi(\mathbf{x}_n) - \phi(\mathbf{x}_n)p(y=1|\mathbf{x}_n) - \frac{1}{\alpha}\mathbf{w} \tag{11.18}$$

$$= \phi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{p}) - \frac{1}{\alpha}\mathbf{w} \tag{11.19}$$

y cuya segunda derivada es

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}\mathbf{w}^{\mathrm{T}}} = \sum_{i=1}^{N} y_i \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^{\mathrm{T}} p(y=1|\mathbf{x})(1-p(y=1|\mathbf{x})) - \frac{1}{\alpha}\mathbf{I} \tag{11.20}$$

$$= -\phi(\mathbf{x})^{\mathrm{T}}\mathbf{V}\phi(\mathbf{x}) - \frac{1}{\alpha}\mathbf{I}, \tag{11.21}$$

donde $\mathbf{V}$ es una matriz diagonal conde $v_{ii} = p(y=1|\mathbf{x})(1-p(y=1|\mathbf{x}))$.

Con estos resultados ya nos es posible calcular la matriz de covarianzas $\mathbf{C}$ de la distribución aproximada y la curvatura $\mathbf{C}^{-1}$. Sin embargo, como tampoco es posible obtener una solución cerrada de $\mathbf{w}_{MAP}$, de nuevo se resuelve mediante el algoritmo iterativo de Newton-Raphson (ver sección E.1) como

$$\mathbf{w}_{MAP}^{(t+1)} = \mathbf{w}_{MAP}^{(t)} - \Big(\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}\partial \mathbf{w}^{\mathrm{T}}}\Big)^{-1}\frac{\partial \mathcal{L}}{\partial \mathbf{w}} \tag{11.22}$$

$$= \mathbf{w}_{MAP}^{(t)} + \big(\phi(\mathbf{x})^{\mathrm{T}}\mathbf{V}\phi(\mathbf{x}) - \frac{1}{\alpha}\mathbf{I}\big)^{-1}\big(\phi(\mathbf{x})^{\mathrm{T}}\mathbf{y} - \phi(\mathbf{x})^{\mathrm{T}}\mathbf{p} - \frac{1}{\alpha}\mathbf{w}\big) \tag{11.23}$$

$$= \big(\phi(\mathbf{x})^{\mathrm{T}}\mathbf{V}\phi(\mathbf{x}) - \frac{1}{\alpha}\mathbf{I}\big)^{-1}\phi(\mathbf{x})^{\mathrm{T}}\big(\mathbf{V}\phi(\mathbf{x})\mathbf{w} + \mathbf{t} - \mathbf{p}\big), \tag{11.24}$$

donde $\mathbf{p}$ es un vector de dimensión N con $p_i = p(y_i|x_i)$, y que se actualiza en cada iteración, junto con $\mathbf{V}$.

Distribución predictiva final

Hasta este momento se ha calculado la distribución de la probabilidad a posteriori de los parámetros. Pero falta saber cómo podemos usar esta distribución para llevar a cabo predicciones cuando se empleen datos nuevos. Para ello, una vez obtenida la aproximación de 11.15 como $\mathcal{N}(\mathbf{w}_{MAP}, \mathbf{C})$, podemos realizar predicciones mediante la distribución predictiva final, que se expresa como

$$p(y=1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{y}) = \int p(y=1|\mathbf{x}_{new}, \mathbf{w}) p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \alpha) d\mathbf{w},$$

donde $\mathbf{x}_{new}$ es el caso nuevo del que se desea conocer su probabilidad. Esta probabilidad debería estimarse computando la integral pero, de nuevo, esta integral no tiene una solución analítica conocida y solo podemos aproximarla por simulación o empleando la probabilidad Máxima A Posteriori (MAP). Para aproximar el valor de la probabilidad predictiva final por simulación (ver anexo E.3.4) se emplean los S modelos obtenidos mediante Monte Carlo,

$$p(y=1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{y}) = \frac{1}{S}\sum_{s=1}^{S} p(y=1|\mathbf{x_{new}}, \mathbf{w}_s) \tag{11.25}$$

$$= \frac{1}{S}\sum_{s=1}^{S} \frac{1}{1+\exp(-\mathbf{w}_s^T\phi(\mathbf{x}_{new}))}, \tag{11.26}$$

donde $\mathbf{w}_s \sim \mathcal{N}(\mathbf{w}_{MAP}, \mathbf{C})$.

Otra alternativa para aproximar la probabilidad predictiva final es asumir que la distribución final es apuntada (leptocúrtica) alrededor de $\mathbf{w}_{MAP}$, por lo que se considera despreciable la masa de probabilidad que rodea a este punto. De este modo, se aproxima dicha probabilidad como

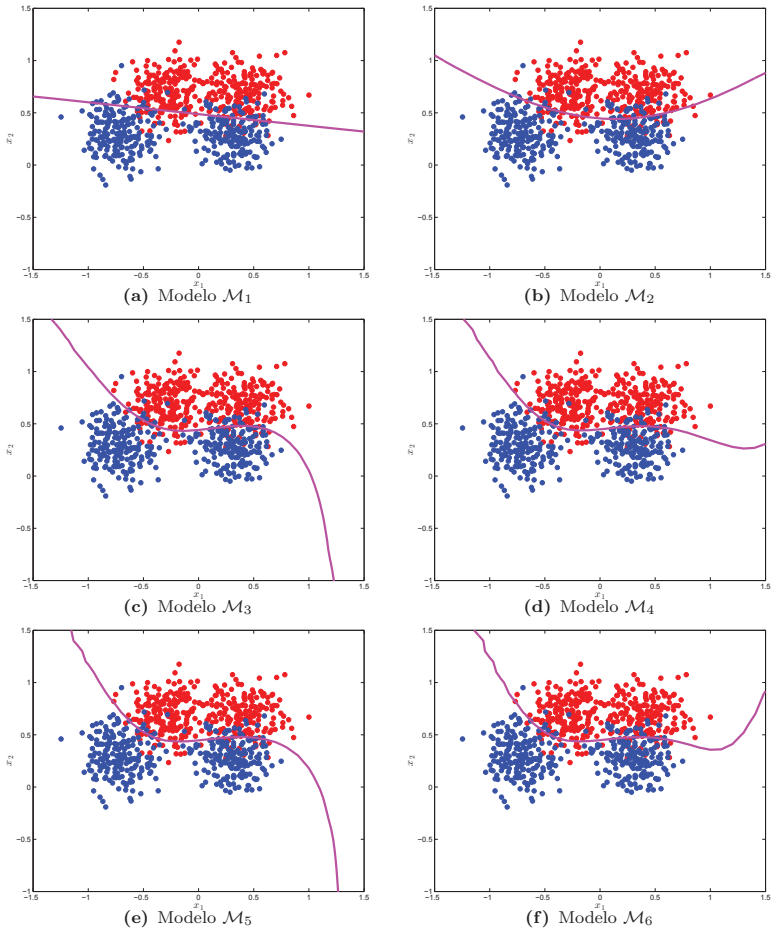
$$p(y=1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{y}) \approx p(y=1|\mathbf{x_{new}}, \mathbf{w_{MAP}}) \tag{11.27}$$

$$\approx \frac{1}{1+\exp(-\mathbf{w}_{MAP}^T\phi(\mathbf{x}_{new}))}, \tag{11.28}$$

Por lo que la regla de decisión queda como sigue: si $p(y = 1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{y}) > 0{,}5$, entonces $\mathbf{x}_{new}$ se asigna a la clase $y = 1$ y sino a la clase $y = 0$.

Ejemplo 11.2 (Predicción de metástasis en ganglios no-centinela (cont. ej. 11.1)) Seguimos intentando obtener modelos discriminativos de regresión logística para la clasificación del problema del ejemplo 11.1. En este caso, realizamos una estimación máximo a posteriori (MAP) de la distribución del vector de parámetros $\mathbf{w}$, asumiendo que siguen una Normal con una moda en $\mathbf{w}_{MAP}$ y asumiendo también que la distribución aproximada es suficientemente leptocúrtica. Al igual que antes, abordamos el aprendizaje de seis modelos de orden 1 a 6.

En este caso, tras obtener $\mathbf{w}_{MAP}$ mediante el algoritmo Newton-Raphson (ecuación 11.24, podemos realizar predicciones asumiendo que la distribución de $\mathbf{w}$ está representada por $\mathbf{w}_{MAP}$, por lo que aplicando 11.28, podemos obtener la probabilidad $p(y = 1|\mathbf{x}_{new}, \alpha, \mathbf{X}, \mathbf{y})$ para estimar las fronteras de decisión de cada modelo (figura 11.2).



(a) Modelo $\mathcal{M}_1$ **(b)** Modelo $\mathcal{M}_2$ **(c)** Modelo $\mathcal{M}_3$ **(d)** Modelo $\mathcal{M}_4$ **(e)** Modelo $\mathcal{M}_5$ **(f)** Modelo $\mathcal{M}_6$

Figura 11.2: Fronteras de decisión descritas por los modelos de regresión logística estimados con con inferencia bayesiana y estimación máximo a posteriori.

Estimación por simulación de Monte Carlo

El marco MCMC nos permite obtener una solución Bayesiana de la distribución de la probabilidad a posteriori de los parámetros en lugar de aproximarla mediante la aproximación de Laplace. El anexo E.3.4 describe la simulación de distribuciones a posteriori mediante el algoritmo Metrópolis y su generalización, el algoritmo de Metrópolis-Hastings.

Algoritmo 11.1 Algoritmo Metropolis para Regresión Logística

$\mathbf{w}^{(0)} \sim N(\mathbf{0}, \mathbf{I})$
for $t = 0, 1, 2, \ldots, m$ **do**
  $\mathbf{w}^* \sim N(\mathbf{w}^{(t)}, \epsilon^2\mathbf{I})$

$$\begin{aligned}\alpha(\mathbf{w}^*|\mathbf{w}^{(t)}) &= \text{mín}\{1, \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}^*)p(\mathbf{w}^*|\alpha)}{p(\mathbf{y}|\mathbf{X}, \mathbf{w}^{(t)})p(\mathbf{w}^{(t)}|\alpha)}\} \\ &= \text{mín}\{1, \frac{\exp(\phi(\mathbf{X})\mathbf{w}^*)^T\mathbf{t} - \sum_{i=1}^N \log(1+\exp\phi(\mathbf{x}_i)\mathbf{w}^*)p(\mathbf{w}^*|\mathbf{0}, \alpha^{-1}\mathbf{I})}{\exp(\phi(\mathbf{X})\mathbf{w}^{(\mathbf{t})})^T\mathbf{t} - \sum_{i=1}^N \log(1+\exp\phi(\mathbf{x}_i)\mathbf{w}^{(\mathbf{t})})p(\mathbf{w}^{(\mathbf{t})}|\mathbf{0}, \alpha^{-1}\mathbf{I})}\}\end{aligned}$$

  $u \sim Unif(0, 1)$
  if $\alpha(\mathbf{w}^*|\mathbf{w}^{(t)}) > u$ **then**
    $\mathbf{w}^{(t+1)} = \mathbf{w}^*$
  else
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)}$
  end if
end for

Específicamente, resolvemos la distribución final de $\mathbf{w}|\mathbf{y}, \mathbf{X}$ del modelo de regresión logística (ecuación 11.15) mediante el algoritmo Metrópolis 11.1. En el algoritmo asumimos un valor inicial de $\mathbf{w}$ muestreado a partir de una distribución normal tipificada y una distribución de transición $q(\theta|\theta^{(t)}) = N(\mathbf{w}^{(t)}, \epsilon^2\mathbf{I})$, siendo ϵ un factor de escala de la distribución esférica. Como distribución a priori de los parámetros se ha asumido $p(\mathbf{w}, \alpha) = N(\mathbf{0}, \alpha^{-1}\mathbf{I})$.

Para el cálculo de α, las constantes de normalización $p(\mathbf{y}|\mathbf{X}, \alpha)$ de numerador y denominador se cancelan, lo que resulta de gran interés práctico.

Finalmente, el numerador y el denominador de α se pueden implementar de forma robusta tomando la exponencial del logaritmo de $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$, por lo que

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \exp(\phi(\mathbf{X})\mathbf{w})^T\mathbf{t} - \sum_{i=1}^N \log(1+\exp\phi(\mathbf{x}_i)\mathbf{w})$$

Se suele considerar la cadena $\mathbf{w}^{(b+1)}, \ldots, \mathbf{w}^{(m)}$ como muestra representativa de $\mathbf{w}$, descartando los b primeros casos muestreados, para dar un margen de convergencia al algoritmo (p.e. 10 % del tamaño muestral).

Así pues, podemos aproximar mediante 11.24 el valor esperado de $\mathbf{w}$ como $\bar{\mathbf{w}} = \frac{1}{m-b}\sum_{i=b+1}^m w^{(i)}$. De forma similar se aproximaría la probabilidad a posteriori $p(y_{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{y})$ de una nueva muestra $\mathbf{x}_{new}$ (i.e. cualquier valor puntual de la distribución predictiva final) como $p(y_{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{y}) = \frac{1}{m-b}\sum_{i=b+1}^m p(y_{new}|\mathbf{x}_{new}, \mathbf{w}^{(i)})$.

Ejemplo 11.3 (Predicción de metástasis en ganglios no-centinelas (cont. ej. 11.1))
El modelo bayesiano de regresión logística aplicado en el ejemplo 11.2 asume una distribución

normal de los parámetros $\mathbf{w}$. Podemos resolver los modelos a posteriori para resolver el problema sin asumir esta suposición mediante simulación MCMC.

Aplicamos pues el algoritmo 11.1 a los seis modelos presentados en el ejemplo 11.1. Calculando la probabilidad a posteriori de las muestras de entrenamiento y test, obtenemos los resultados presentados en las dos últimas columnas de la tabla 11.1. Además, comparamos las prestaciones de todas las aproximaciones presentadas para el desarrollo de modelos discriminativos de regresión logística. Se puede observar cómo, en general, la versión de máxima verosimilitud (IRLS) ofrece peores resultados para aquellos modelos con mayor número de parámetros al generalizar peor y sobreajustar el modelo a los datos de entrenamiento. Esta conclusión se aprecia al observar cómo el error de entrenamiento es más bajo que el error de generalización. En cambio, las distintas versiones bayesianas ofrecen mejor generalización y sobreajustan menos a los datos de entrenamiento. Esto se debe a una propiedad de los modelos estimados con inferencia bayesiana como es la capacidad de regularizar de forma intrínseca gracias a la aplicación de hipótesis (subjetivas) de partida a través de la probabilidad a priori de los parámetros. Como desventaja, la simulación MCMC requiere mayor tiempo de cómputo ya que necesita muestrear un número suficientemente grande de casos (e.g. 250.000 en nuestros experimentos), que pueden suponer el cálculo de distribuciones complejas.

Tabla 11.1: Resultados de los modelos de regresión logística para el error del conjunto de entrenamiento, E(Tr), y el error de generalización, E(Gn), medido con el conjunto de evaluación con el modelo estimado con máxima verosimilitud mediante el algoritmo IRLS, mediante inferencia bayesiana con la aproximación de Laplace para la probabilidad a posteriori de los parámetros (LAP) y con aproximación Máximo A Posteriori (MAP) y por simulación (SIM) para la distribución predictiva final y, por último, mediante aproximación por MCMC para la distribución de la probabilidad a posteriori de los parámetros y la predictiva final.

	IRLS		LAP-MAP		LAP-SIM		MCMC	
Modelo	E(Tr)	E(Gn)	E(Tr)	E(Gn)	E(Tr)	E(Gn)	E(Tr)	E(Gn)
$\mathcal{M}_1$	12.2 %	11.9 %	12.0 %	11.9 %	12.0 %	11.9 %	12.0 %	11.9 %
$\mathcal{M}_2$	12.0 %	10.1 %	12.4 %	10.7 %	12.4 %	10.7 %	12.4 %	10.7 %
$\mathcal{M}_3$	9.8 %	9.3 %	10.6 %	8.8 %	10.6 %	8.8 %	10.6 %	8.8 %
$\mathcal{M}_4$	10.0 %	9.3 %	10.6 %	8.9 %	10.6 %	8.8 %	10.4 %	9.1 %
$\mathcal{M}_5$	8.8 %	9.3 %	10.6 %	8.8 %	10.6 %	8.8 %	10.4 %	8.8 %
$\mathcal{M}_6$	8.4 %	9.5 %	10.6 %	8.8 %	10.6 %	8.8 %	10.6 %	8.8 %

Capítulo 12

Modelos gráficos

Las redes bayesianas son modelos probabilísticos que pueden representar las relaciones de causalidad entre múltiples variables aleatorias mediante un Grafo Acíclico Dirigido (GAD). El modelado de problemas médicos reales, a partir de conocimiento experto, mediante redes bayesianas resulta intuitivo gracias a la representación gráfica de la causalidad. Además, las redes bayesianas permiten analizar las relaciones entre variables mediante la inspección del grafo, como por ejemplo la independencia condicional entre variables, lo que facilita el refinamiento del modelo mediante métodos formales.

Una vez modelado el problema, una red bayesiana puede ser utilizada para calcular las probabilidades conjuntas y marginales de las variables, así como la probabilidad condicional a la observación (llamada evidencia) de algunas de sus variables. La existencia de algoritmos eficientes para el cálculo de estas probabilidades hacen de las redes bayesianas una herramienta práctica para la ayuda a la decisión médica.

En esta sección nos centraremos en la aproximación a las redes bayesianas de variables discretas desarrollada en la Universidad de Aalborg y que resultó ser uno de los pilares para llevar esta metodología a su aplicación en problemas reales de decisión. Por último, también enumeraremos algunos métodos de aprendizaje automático de la estructura y de los parámetros de las redes bayesianas a partir de muestras de entrenamiento.

12.1. Causalidad y d-Separación

Una red bayesiana se compone de un conjunto de nodos y un conjunto de arcos dirigidos entre los nodos formando un GAD G.

Los nodos, i, (o vértices) de la red bayesiana (RB) representan a las variables aleatorias, X_i. Las variables de una RB pueden ser continuas o discretas. Por ejemplo, la variable discreta X_S podría tomar dos valores para indicar si un paciente fuma s o no fuma $\bar{s}$ y la variable X_B puede indicar si un paciente sufre un episodio de bronquitis b o no $\bar{b}$.

Los arcos (o aristas) dirigidos entre dos nodos pueden representar una relación de causalidad entre variables. Por ejemplo, $S \rightarrow B$ establece que fumar (X_S) es, en mayor o menor medida, causa de episodios de bronquitis (X_B). De forma gráfica, si hay un arco desde S hasta B diremos que S es *padre* de B y que B es *hijo* de S. Así pues, si sabemos que un paciente es o no fumador $X_S = e_S$, donde e_S puede tomar los valores $\{s, \bar{s}\}$ diremos que la evidencia de S influye en B y, por lo tanto, modifica la certeza (probabilidad) de que el paciente tenga bronquitis. De forma similar, al observar que un paciente tiene bronquitis, lo primero que se nos viene a la cabeza es plantearnos si el paciente fuma. Por lo tanto, a falta de evidencia de X_S, la observación de la evidencia de X_B modifica la creencia de X_S.

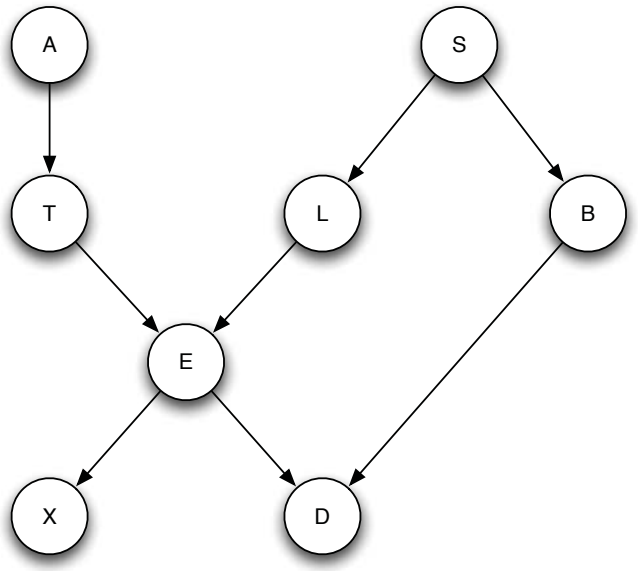


Figura 12.1: Red bayesiana Asia [127].

Ejemplo 12.1 (Red Asia de Lauritzen & Spiegelhalter [127])
Para ilustrar las explicaciones teóricas del capítulo utilizaremos la RB Asia (ver figura 12.1), introducida sin ánimo diagnóstico por Lauritzen & Spiegelhalter en [127].

Este modelo simplificado de diagnóstico establece que un paciente puede tener una enfermedad de pulmón (X_E) debido a tuberculosis (X_T) o a cáncer de pulmón (X_L). Haber viajado a Asia (X_A) recientemente puede ser causa de tuberculosis (X_T). Además, fumar (X_S) provoca cáncer (X_C) y bronquitis (X_B). A su vez, tener una enfermedad de pulmón (X_E) y tener bronquitis (X_B) son causas de disnea (X_D). Por último, una enfermedad de pulmón (X_E) influye en el resultado de la radiografía de pecho (X_X).

En los sucesivo denotaremos la variable X_A por el nombre de su nodo, A, sin que esto nos suponga una pérdida de generalización debida a la ambigüedad.

12.1.1. d-Separación

Como hemos visto, la evidencia de un nodo influye en la certeza de sus padres e hijos si estos no tienen evidencia observada. Además, esta actualización de la certeza de los nodos se extiende a lo largo de la red, modificando la certeza del resto de nodos. El aspecto más importante a estudiar en las redes bayesianas es, sin duda, cómo el cambio de la certeza de una variable influye en la certeza de otras variables.

Podemos estudiar el tipo de conexiones que se producen entre los nodos de una red para caracterizar si la evidencia de un nodo llega a influir o no en la creencia de otros. Las conexiones que pueden darse en una red bayesiana pueden resumirse en tres situaciones:

1. Conexión en serie.

   Tomemos del ejemplo 12.1 la conexión en serie entre los nodos $\{A, T, E\}$ (también llamada cola-cabeza o no cabeza-cabeza). El nodo A influye en T que, a su vez, influye en E. Por lo tanto, se podría pensar que la evidencia de A influye en E a través de T y, de forma similar, la evidencia de E influye en A. Sin embargo, si se observa la evidencia de T entonces la cosa cambia ya que T *bloquea* la conexión, haciendo que A y E se convierten en independientes condicionados a la observación de T. De este modo la evidencia de A ya no influye en E y viceversa. Se dice entonces que T d-separa a A y E.

2. Conexión divergente.

   Tomemos ahora la conexión divergente entre los nodos $\{L, S, B\}$, donde en el nodo intermedio S no hay arcos convergentes (cola-cola o no cabeza-no cabeza). La observación de la evidencia en B hará que modifiquemos la creencia de S, cambiando así la creencia de L. Sin embargo, si disponemos de la evidencia de S entonces la observación de B ya no influirá en la creencia de L ya que S *bloquea* la conexión, por lo que B y L son condicionalmente independientes dado S. Diremos que S d-separa a L y B.

3. Conexión convergente.

   Por último, como ejemplo del último tipo de conexión que puede darse en una red bayesiana, se toma la conexión entre los nodos $\{T, E, L\}$, donde todos los arcos convergen en el nodo intermedio E (cabeza-cabeza).

   Supongamos que disponemos de la evidencia T. Podemos pensar que E verá su creencia modificada. Sin embargo, L no tiene porque verse influida, ya que L y T son dos causas independientes. Pero si disponemos de la evidencia E, entonces el hecho de conocer T modificará la creencia de L. Esto se debe a que si se conoce la evidencia de la consecuencia E, entonces conocer una causa (T) nos modificará la creencia acerca de la influencia de la otra causa (L) sobre E.

   Ahora añadamos a nuestro estudio los descendientes de E: los nodos $\{X, D\}$. Si alguno de estos nodos es observado, entonces E modifica su creencia por lo que, una vez más, T y L dejan de tener un comportamiento independiente.

Los tres tipos de conexiones anteriores son las formas en las que una evidencia puede transmitirse por una variable. Por lo tanto, podemos generalizar los ejemplos en el *criterio de d-separación*.

Criterio de d-separación Dos conjuntos de variables $\mathbf{X_A}$ y $\mathbf{X_B}$ son independientes, dada la evidencia introducida en el conjunto de variables $\mathbf{X_Z}$ de una red bayesiana ($\mathbf{Z}$ d-separa $\mathbf{A}, \mathbf{B}$), si en todos los caminos no dirigidos entre los nodos de $\mathbf{A}$ y $\mathbf{B}$ existe un nodo N tal que:

1. N no es un nodo de aristas convergentes y N está en $\mathbf{Z}$.

2. N es un nodo de aristas convergentes y ni N ni sus descendientes están en $\mathbf{Z}$.

La figura 12.2 resume los tres casos de d-separación, donde las nubes representan grupos de nodos y des(N) representa el conjunto de nodos descendientes de N.

La d-separación de $\mathbf{A}$ y $\mathbf{B}$ por $\mathbf{Z}$, esto es, la **independencia condicional** de $\mathbf{A}$ y $\mathbf{B}$ dado $\mathbf{Z}$, nos asegura que, conocida la evidencia $\mathbf{Z}$, los cambios en la certeza de $\mathbf{A}$ no tienen impacto en la certeza de $\mathbf{B}$ y viceversa. Este hecho nos será de gran utilidad para simplificar el cálculo de actualización de la certeza de la red cuando introduzcamos una evidencia en la misma.

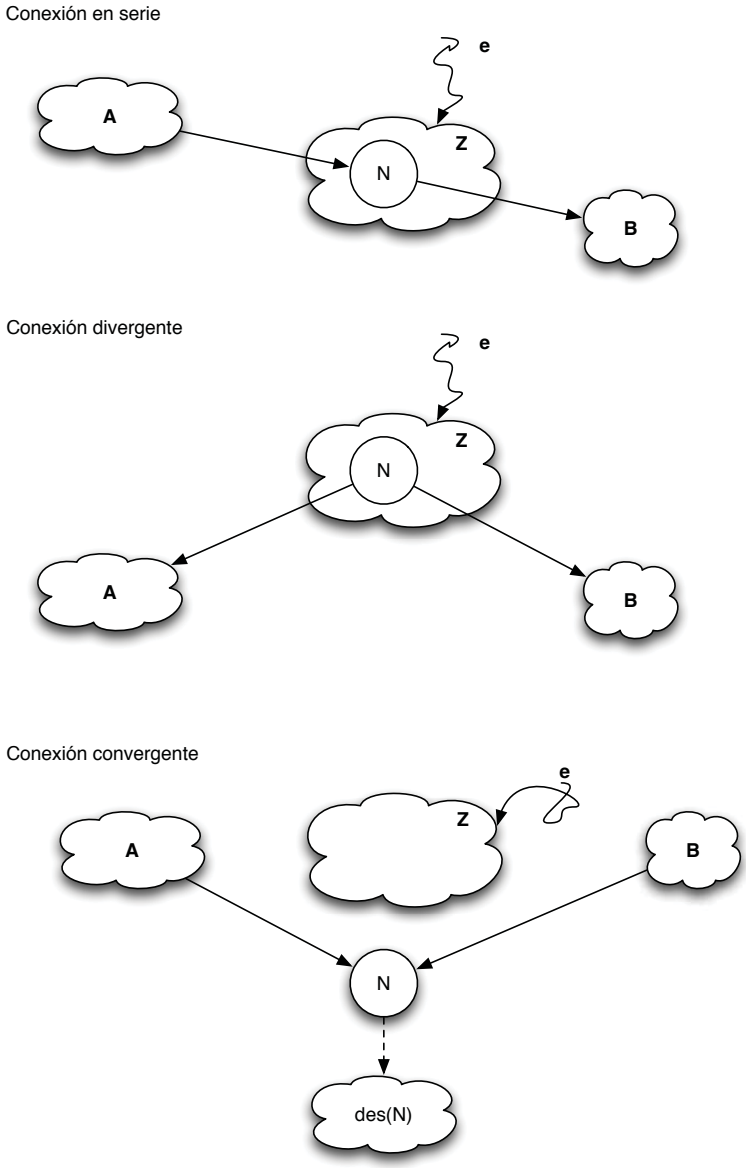


Figura 12.2: El grupo de nodos **Z**, de la cual se ha observado una evidencia e, *d-separa* **A** y **B**, si para toda ruta entre nodos de **A** y **B** existe un nodo N tal que i) no es convergente y N pertenece a **Z**; o ii) N es convergente (cabeza-cabeza) y ni N ni sus descendientes, $des(N)$, pertenecen a **Z**.

Ejemplo 12.2 (d-Separación en la red Asia (cont. ej. 12.1))
Queremos saber qué nodos de la red Asia (figura 12.1) son d-separables de B cuando disponemos de una evidencia $E = e$.

Evaluaremos el criterio de d-separación sucesivamente a cada nodo A, T, E, X, D, L, S de la red para comprobar si es d-separable de B dado $\mathbf{Z} = E$.

- ¿E d-separa A, B?

  La figura 12.3 analiza en profundidad la d-separación de estos nodos por E. Como vemos, podemos recorrer dos caminos para llegar desde A hasta B. El primero de ellos no está bloqueado. El segundo camino está bloqueado por D, ya que es un nodo cabeza-cabeza que no pertenece a $\mathbf{Z}$, ni tampoco sus descendientes. También está bloqueado por E, al ser cabeza-no cabeza y estar en $\mathbf{Z}$. Como el primer camino no está bloqueado, E no d-separa A, B.

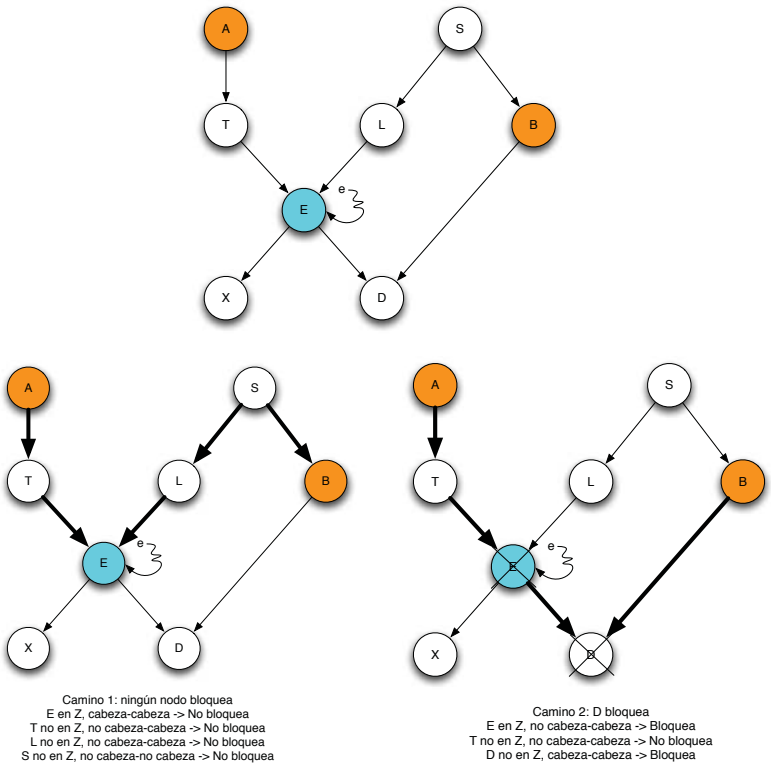


Figura 12.3: E no d-separa A, B. Los nodos en naranja son los testeados, los nodos en azul son los nodos observados, y el resto de nodos están en blanco. La cruz indica que el nodo *bloquea* el camino.

- ¿E d-separa T, B?

  La figura 12.4 analiza en profundidad la d-separación de estos nodos, que como vemos es bastante similar al caso anterior. Como vemos, podemos recorrer dos caminos para llegar

desde T hasta B. El primero de ellos no está bloqueado. El segundo camino está bloqueado por D, ya que es un nodo cabeza-cabeza que no pertenece a $\mathbf{Z}$, ni tampoco sus descendientes. También está bloqueado por E, al ser cabeza-no cabeza y estar en $\mathbf{Z}$. Como el primer camino no está bloqueado, Z no d-separa T, B.

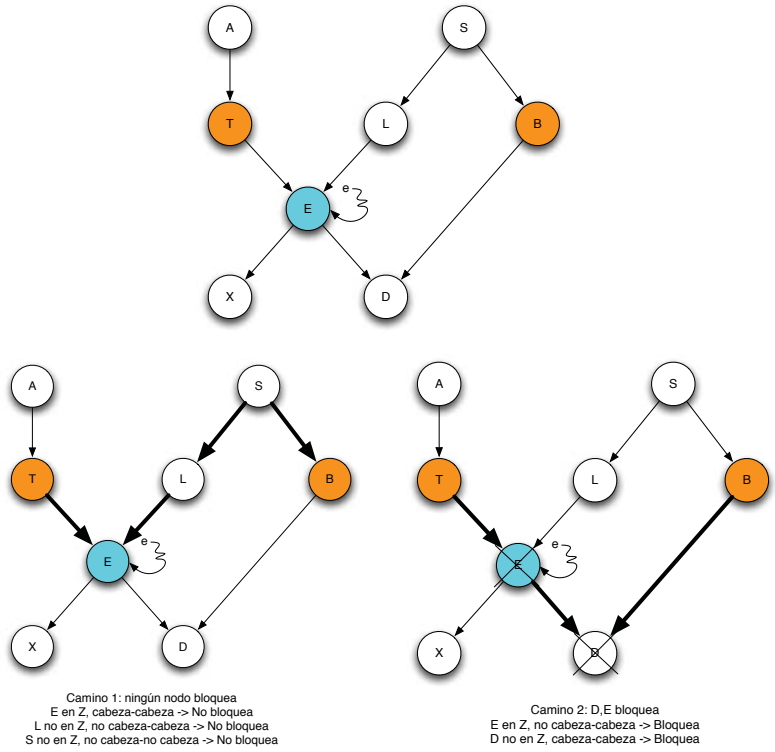


Figura 12.4: E no d-separa T, B. Los nodos en naranja son los testeados, los nodos en azul son los nodos observados, y el resto de nodos están en blanco. La cruz indica que el nodo sobre la que está *bloquea* el camino.

- ¿E d-separa X, B?

  La figura 12.5 analiza en profundidad la d-separación de estos nodos. Como vemos, podemos recorrer dos caminos para llegar desde X hasta B. Ambos están bloqueados por E, al ser un nodo de arcos no convergentes y estar en $\mathbf{Z}$. Además, el camino 2 también está bloqueado por D, ya que es un nodo cabeza-cabeza sin evidencia cuyos descendientes tampoco la contienen. Por lo tanto, E d-separa X, B.

- ¿E d-separa D, B?

  Como existe un arco directo entre B y D, los nodos no pueden ser d-separados.

- ¿E d-separa E, B?

  Como E pertenece a $\mathbf{Z}$ la cuestión no procede.

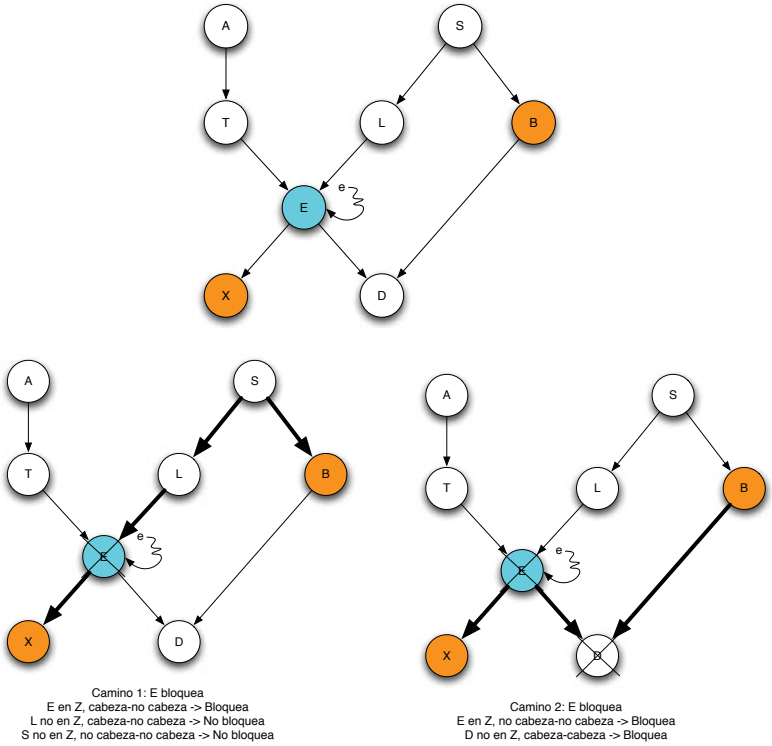


Figura 12.5: E d-separa X, B. Los nodos en naranja son los testeados, los nodos en azul son los nodos observados, y el resto de nodos están en blanco. La cruz indica que el nodo sobre la que está *bloquea* el camino.

- ¿E d-separa L, B?

  La figura 12.6 analiza en profundidad la d-separación de estos nodos. Como vemos, podemos recorrer dos caminos para llegar desde L hasta B. Como el primer camino no está bloqueado, E no d-separa L, B.

- ¿E d-separa S, B?

  Como existe un arco directo entre S y B, los nodos no pueden ser d-separados.

Por lo tanto, E únicamente d-separa B de X, por lo que si queremos calcular la creencia de B dada la evidencia observada en E, podremos prescindir de X, ya que no influye en el resultado.

12.2. Probabilidades condicionales

Para cuantificar la creencia de los estados que pueden tomar las variables de una red bayesiana se utiliza la teoría de la probabilidad, ya estudiada en la sección 3.1.

Utilizaremos la probabilidad condicional $p(D|E)$ para cuantificar la certeza de cada uno de los valores que puede tomar D condicionada a observar los valores de E, siendo E padre de D.

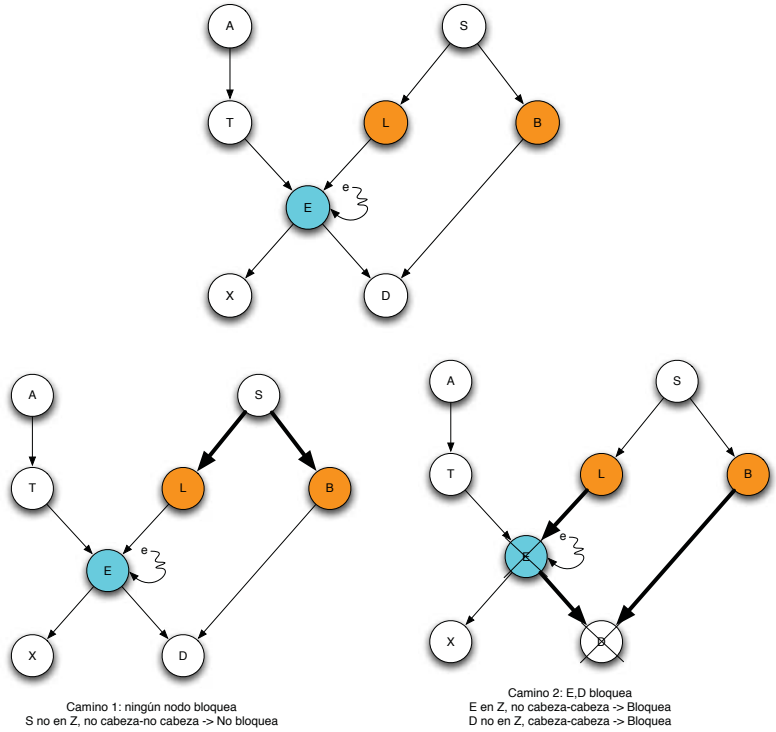


Figura 12.6: E no d-separa L, B. Los nodos en naranja son los testeados, los nodos en azul son los nodos observados, y el resto de nodos están en blanco. La cruz indica que el nodo sobre la que está *bloquea* el camino.

Como D tiene más padres, véase B, entonces E y B cooperan en la certeza de D, por lo que será más útil especificar $p(D|E,B)$. Generalizando, especificaremos $p(D|pa(D))$, para cada nodo, donde $pa(D)$ son los padres del nodo. Si un nodo no tiene padres, $pa(A) = \emptyset$, entonces su certeza vendrá expresada por la probabilidad incondicional, $p(A)$.

En RB de variables discretas las probabilidades condicionales se suelen especificar mediante tablas que contengan las probabilidades condicionales de los valores del hijo dadas las combinaciones de valores de los padres.

Ejemplo 12.3 (Creencia $p(D|E,B)$ en la red Asia (cont. ej. 12.1))
La probabilidad condicional $p(D|E,B)$ de la red bayesiana Asia que expresa con probabilidades la creencia de la variable *disnea* influida por E y B tiene la forma de la tabla 12.1.

En general, llamaremos *potencial*, y lo denotaremos como ϕ, a cualquier estructura que exprese la creencia de los valores de las variables mediante números reales positivos. Los potenciales suelen tener forma de tabla de contingencia, como el ejemplo que encontramos en la tabla 12.1. Es importante reflejar que los potenciales no necesariamente expresan las creencias como probabilidades normalizadas.

Ejemplo 12.4 (Potenciales de la red Asia (cont. ej. 12.1))
La tabla 12.2 define los potenciales de la red Asia.

Tabla 12.1: Probabilidad condicional p(D|E,B) de la red Asia.

	$E = e$		$E = \bar{e}$	
	$B = b$	$B = \bar{b}$	$B = b$	$B = \bar{b}$
$D = d$	0.9	0.7	0.8	0.1
$D = \bar{d}$	0.1	0.3	0.2	0.9

Tabla 12.2: Potenciales de la red Asia.

Potencial (ϕ)	Probabilidad condicional
ϕ_A	$p(A)$
ϕ_S	$p(S)$
ϕ_T	$p(T\|A)$
ϕ_L	$p(L\|S)$
ϕ_B	$p(B\|S)$
ϕ_E	$p(E\|T, L)$
ϕ_X	$p(X\|E)$
ϕ_D	$p(D\|E, B)$

12.3. Independencia condicional y definición de red bayesiana

En las conexiones serie y divergente que hemos estudiado en la sección 12.1.1, la observación de la evidencia en una variable Z bloquea la influencia que una variable A puede tener sobre otra variable B a través de la conexión. Este fenómeno se conoce como **independencia condicional** y puede expresarse en forma probabilística como

$$p(B|A,Z) = p(B|Z), \tag{12.1}$$

o de forma equivalente

$$p(A|B,Z) = p(A|Z), \tag{12.2}$$
$$p(A,B|Z) = p(A|Z)p(B|Z). \tag{12.3}$$
$$\tag{12.4}$$

En la conexión convergente, que también hemos estudiado en la sección 12.1.1, las variables A y B que comparten un hijo N son independientes cuando N no está observado, por lo que la observación de A no influye en B. Sin embargo, la observación de N hace que la observación de A influya en B, por lo que son dos variables *condicionalmente dependientes*. Cabe pensar que esto sucede porque, al conocer N y A, se puede saber cuánto del valor de N se debe a B.
Una red bayesiana de variables discretas se compone de:

- Un conjunto de variables (nodos) y arcos dirigidos, donde
- cada nodo tiene un conjunto finito de estados disjuntos, que forma
- una estructura en forma de GAD, donde
- cada nodo tiene asociado un potencial $\phi_A = p(A|pa(A))$.

Como vemos, esta definición no se refiere al concepto de causalidad. De hecho, no es imprescindible que exista dicha relación entre los nodos enlazados en la red. Formalmente, decimos que $\mathbf{X} = [X_1, \ldots, X_D]$ es una red bayesiana en base a G si para cada par de nodos A, B, sus variables aleatorias son condicionalmente independientes dado Z $p(A,B|Z) = p(A|Z)p(B|Z)$, donde Z d-separa A y B.

12.4. Factorización

Disponer de la distribución de probabilidad conjunta $p(\mathbf{X})$ cuando la dimensión D es grande es difícil y costoso. Ya que el GAD G de una RB define las independencias condicionales entre las variables, esta es una representación eficiente (además que intuitiva) de $p(\mathbf{X})$. Así pues, la distribución de probabilidad conjunta $p(\mathbf{X})$ se *factoriza* como el producto de probabilidades condicionales (o potenciales) asociadas a los nodos de la red bayesiana[a],

$$p(\mathbf{X}) = \prod_{i=1}^{D} p(X_i|pa(X_i)). \tag{12.5}$$

[a]La demostración de este enunciado es una aplicación directa de la d-separación, y puede encontrarse en la sección 1.4.3 de [128]

Ejemplo 12.5 (Factorización de la red Asia (cont. ej. 12.1))
Siguiendo el grafo de la figura 12.1 para la red Asia podemos factorizar la probabilidad conjunta como:

$$p(A,S,T,L,B,E,X,D) = p(A)p(S)p(T|A)p(L|S)p(B|S)p(E|T,L)p(X|E)p(D|B,E). \quad (12.6)$$

Teniendo acceso a la distribución de probabilidad conjunta $p(\mathbf{X}) = p(X_1,\ldots,X_D)$ podemos calcular la probabilidad de una variable $p(X_i)$ mediante la marginalización (3.3). El cálculo eficiente de esta marginalización es uno de los objetivos más importantes en redes bayesianas y el objeto de estudio de la sección 12.5.

Supongamos ahora que disponemos de una evidencia $\mathbf{e}$. Por ejemplo, podemos saber que L toma el valor $\hat{l}$. La probabilidad conjunta $p(\mathbf{X}, e)$ es la tabla resultante de poner a 0 todas aquellas posiciones donde $L \neq \hat{l}$. Como L toma los valores $\{l, \hat{l}\}$, podemos expresar la evidencia observada mediante el vector $\mathbf{e} = (\mathbf{1}, \mathbf{0})$, donde 1 indica que L toma el valor de la posición correspondiente. Con esta notación, podemos resolver $p(\mathbf{X}, e)$ como la multiplicación de $p(\mathbf{X})$ y $\mathbf{e}$, que factorizado

$$p(\mathbf{X}, e) = p(\mathbf{X}) \cdot \mathbf{e} = \prod_{\mathbf{i=1}}^{\mathbf{D}} \mathbf{p(X_i|pa(X_i))} \prod_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{e_j},$$

donde $\underline{\mathrm{e}}_j$ son cada uno de los n hallazgos observados.

Por marginalización (3.3) y aplicando el teorema de Bayes (3.5), podemos calcular la probabilidad de una variable X_i (o conjunto de variables) condicionada a la evidencia observada e:

$$p(X_i|e) = \frac{\sum_{j\neq i} p(\mathbf{X}, e)}{p(e)}.$$

12.5. Propagación de la creencia

Supongamos una red bayesiana de variables $\mathbf{X} = X_1,\ldots,X_D$ con una estructura G y potenciales ϕ. Los dos usos fundamentales que tiene una red bayesiana son i) calcular la probabilidad marginal $p(X_i)$ de una variable (o de un conjunto de variables) y ii) la probabilidad condicional $p(X_i|e)$ de una variable (o de un conjunto) condicionada a la evidencia e.

Ambas tareas pueden resolverse mediante la marginalización de la probabilidad conjunta $p(\mathbf{X})$ gracias a la factorización (12.5) de la red bayesiana. El coste temporal y espacial de la resolución, dependerá básicamente de la estrategia de marginalización que sigamos. Estudiaremos el algoritmo HUGIN[b], desarrollado por investigadores de la Universidad de Aalborg, que está considerado uno de los más eficientes para resolver esta marginalización.

Ejemplo 12.6 (Cálculo de $p(B)$ de la red Asia (cont. ej. 12.1))
Antes de desarrollar el algoritmo HUGIN, a modo de introducción, supongamos que queremos calcular $p(B)$ en la red Asia trabajando directamente con la marginalización de (12.6).

$$p(B) = \sum_{\mathbf{X}\backslash B} p(\mathbf{X}) = \sum_{\mathbf{X}\backslash B} p(A)p(S)p(T|A)p(L|S)p(B|S)p(E|T,L)p(X|E)p(D|B,E),$$

siendo $\mathbf{X}\backslash B$ el conjunto de nodos de la red distintos de B.

Utilizando la propiedad distributiva podemos evitar el cálculo de $p(\mathbf{X})$ mediante diferentes estrategias, por ejemplo

$$p(B) = \sum_A p(A) \sum_T p(T|A) \sum_S p(S)p(B|S) \sum_L p(L|S) \sum_E p(E|T,L) \sum_X p(X|E) \sum_D p(D|E,B),$$

[b] www.hugin.dk

implica que los sumatorios sucesivos sobre D, X, E y L sean 1. El potencial resultante del sumatorio sobre S tendrá una dimensión 1x2, que al multiplicarlo por $p(T|A)$ genera un potencial de dimensión 4x2, que es mayor que todos los potenciales asociados a la red bayesiana.

Como alternativa se puede plantear la estrategia

$$p(B) = \sum_S p(S)p(B|S) \sum_A p(A) \sum_T p(T|A) \sum_L p(L|S) \sum_E p(E|T,L) \sum_X p(X|E) \sum_D p(D|E,B),$$

qye implica que los sumatorios sucesivos D, X, E, L, T, A sean 1, y el potencial $p(S)p(B|S)$ es de dimensión 2x2, y por lo tanto no superior a la dimensión de los potenciales de la red.

En el ejemplo 12.6 hemos visto la importancia, en términos de coste espacial, de realizar la marginalización sucesiva (también llamada reducción de variables) en un orden adecuado. Los algoritmos de propagación de la creencia buscan una estrategia óptima de resolución.

12.5.1. Algoritmo HUGIN

El algoritmo HUGIN consta de dos fases:

1. Obtención del árbol de cliques.

   En esta fase se pretende conseguir una estructura en forma de árbol que permita un recorrido ordenado de los nodos respetando las dependencias expresadas por la red bayesiana.

2. Cálculo de probabilidades.

   Una vez tenemos el árbol de cliques, podemos comenzar con las operaciones sobre los potenciales (tablas de probabilidad condicional) para calcular las probabilidades marginales de nuestras variables o las probabilidades condicionales a una evidencia. Para ello, seguiremos el las rutas del árbol de cliques para resolver las probabilidades marginales de los nodos.

Ejemplo 12.7 (Aplicación de HUGIN a la red Asia (cont. ej. 12.1))
A continuación realizaremos la explicación del algoritmo HUGIN sobre la red Asia para evitar la complicación terminológica que supone una explicación genérica del mismo.

Obtención del árbol de cliques Seguiremos los siguientes pasos para obtener un árbol de cliques asociado a la red bayesiana.

- Obtención del grafo de dominios.

  El grafo de dominios de una red bayesiana es el grafo de arcos no-dirigidos que conecta cada nodo con sus padres, sus hijos y con los nodos con los que comparte hijos (arcos morales).

  Ejemplo 12.8 (Grafo de dominios de la red Asia (cont. ej. 12.7)
  La figura 12.7 presenta el grafo de dominios de la red Asia.

- Obtención del grafo triangular mediante rellenado de aristas.

  Un grafo triangular es aquel que no tiene ciclos con más de tres nodos. Esto se puede obtener mediante el *rellenado de aristas*. En el rellenado de aristas, elegimos un orden de eliminación de nodos, seleccionamos el primero y se añaden aristas hasta conseguir que el conjunto de nodos adyacentes del nodo sea completo[c]. Una vez hecho esto, se elimina el nodo y las aristas que lo conectan y se elige el siguiente nodo de la lista.

[c] un conjunto es completo si todos sus nodos están conectados a pares.

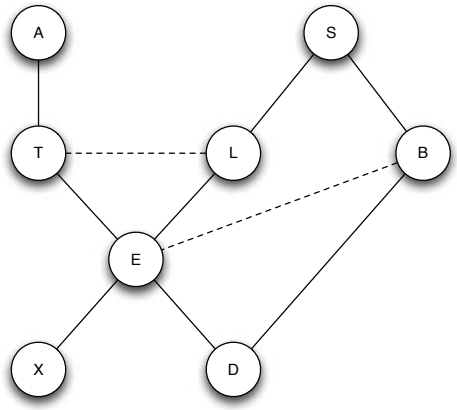


Figura 12.7: Grafo de dominios de la red Asia. Los arcos morales se representan mediante líneas discontinuas.

Ejemplo 12.9 (Grafo triangular de la red Asia (cont. ej. 12.8))
Tras aplicar el *rellenado de aristas* al grafo de dominios de la figura 12.7, con el orden de eliminación $\sigma = A, X, T, D, E, L, S, B$, obtenemos el grafo triangular de la figura 12.8.

- Identificación de los cliques del grafo triangular.

  Identificaremos ahora los cliques del grafo triangular obtenido del paso anterior. Un clique es un conjunto de nodos completo y maximal. Decimos que un conjunto es completo cuando todos sus nodos están conectados, y que es maximal cuando el conjunto es el mayor posible, y por lo tanto no está incluido en otros conjuntos completos.

  El coste computacional del algoritmo de propagación de la creencia dependerá principalmente del peso de los cliques, donde el peso se define como el producto de la cardinalidad de las variables incluidas en el clique. La selección del orden recorrido por el rellenado de aristas hará que el peso sea mayor o menor, y por lo tanto, resulta de interés una buena selección del orden seguido.

  Ejemplo 12.10 (Cliques la red Asia (cont. ej. 12.9))
  La tabla 12.3 enumera los cliques de la red Asia. Como todas las variables tienen dos posibles valores, los cliques de dos nodos tendrán un peso de 4, y los cliques de tres nodos tendrán un peso de 8.

- Construcción del árbol de cliques.

  Una vez tenemos identificados los cliques del grafo triangular, construiremos el árbol de cliques comenzando desde la raíz hacia las hojas. Como raíz estableceremos el clique de menor peso. A continuación se selecciona el clique libre cuya intersección con el árbol sea máxima, enlazándose con el clique enraizado con el que comparte mayor número de nodos. Se continua hasta que no quedan cliques libres.

  El conjunto de nodos que un hijo comparte con su padre se denomina separador. Cada arco entre un padre y un hijo tendrá un buzón de doble dirección anotado con el separador del hijo. El conjunto de nodos que un hijo no comparte con su padre se denomina residual.

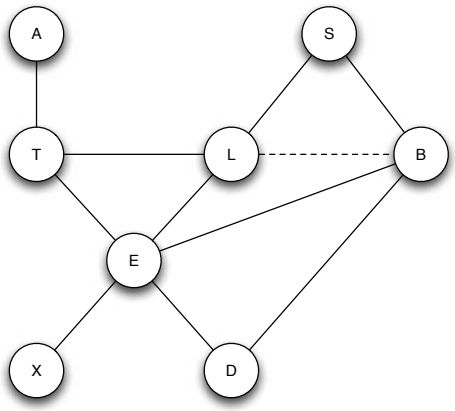


Figura 12.8: Grafo triangular de la red Asia. El arco añadido por rellenado de aristas en la etapa del nodo E tiene una trama discontinua.

Tabla 12.3: Cliques del grafo triangular de la red Asia.

Clique
A, T
T, L, E
S, L, B
L, E, B
E, B, D
E, X

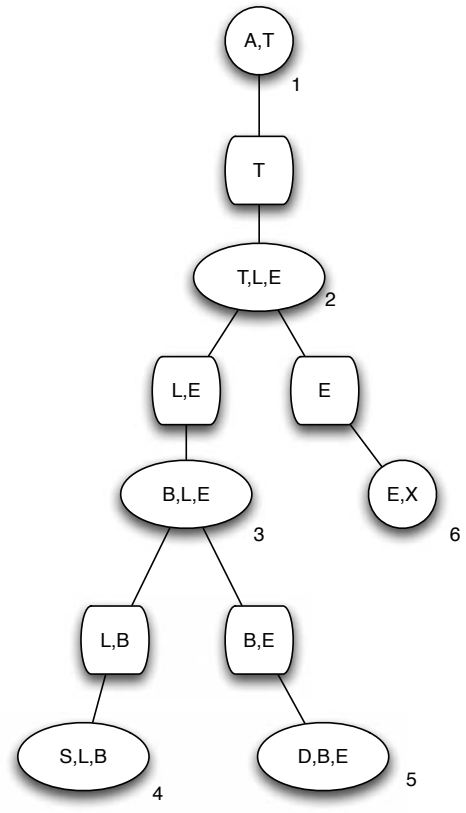


Figura 12.9: Árbol de cliques de la red Asia.

Ejemplo 12.11 (Árbol de cliques para la red Asia (cont. ej. 12.10))
La figura 12.9 presenta el árbol de cliques diseñado para la red Asia. Como ejemplo, el clique E, L, T tiene el nodo separador T con su padre A, T y los nodos residuales E, L.

Cálculo de probabilidades

- Cálculo de potenciales.

  Comenzaremos asignando el potencial ϕ de cada nodo a un solo clique. Entonces calcularemos el potencial Ψ de cada clique como el productorio de los potenciales ϕ que tiene asignados, siendo $\Psi = 1$ en caso de no tener ningún potencial ϕ asignado.

  Como resultado, la probabilidad conjunta $p(\mathbf{X})$ es la factorización de los potenciales de los cliques:

$$p(\mathbf{X}) = \prod_{i=1}^{C} \Psi_i,$$

  siendo C la cardinalidad del árbol de cliques. El mismo árbol servirá también para calcular probabilidades condicionadas a cualquier evidencia observada.

Tabla 12.4: .

Potencial de clique i (Ψ_i)
$\Psi_1 = \phi_A \phi_T = p(A)p(T\|A)$
$\Psi_2 = \phi_E = p(E\|T, L)$
$\Psi_3 = 1$
$\Psi_4 = \phi_S \phi_B \phi_L = p(S)p(B\|S)p(L\|S)$
$\Psi_5 = \phi_D = p(D\|E, B)$
$\Psi_6 = \phi_X = p(X\|E)$

Ejemplo 12.12 (Cálculo de potenciales Ψ de cada clique (cont. ej. 12.11)) Asignaremos los potenciales de la red Asia (table 12.2) a los cliques de la siguiente forma: ϕ_A, ϕ_T al clique 1, ϕ_E al clique 2, ϕ_L, ϕ_S, ϕ_B al clique 4, ϕ_D al clique 5 y ϕ_X al clique 6.

Por lo tanto, los potenciales Ψ de los 6 cliques, numerándolos según la figura 12.9 serán los presentados en la tabla 12.4.

- Fase de absorción de la evidencia.

  Si hemos observado una evidencia $E^* = e$ en un conjunto de nodos E^* podremos calcular la probabilidad condicional de las variables condicionada a dicha evidencia, p.e. $p(X_i|e)$.

  Para tener en cuenta esta condición, actualizaremos el potencial Psi_j de cada clique que contenga algún nodo con evidencia observada:

$$\Psi_j^* = \begin{cases} 0 & \text{si el valor de algún nodo del clique no es consistente con } e \\ \Psi_j & \text{en otro caso} \end{cases} \tag{12.7}$$

  No será necesario actualizar los cliques que no contienen nodos con obervaciones, por lo que sus potenciales no se modifican, $\Psi_k = \Psi_j, k \neq j$.

  Por lo tanto, la probabilidad conjunta condicionada a e será:

$$p(\mathbf{X}|\mathbf{e}) = \prod_{i=1}^{C} \Psi_i^*,$$

Ejemplo 12.13 (Absorción de evidencia $E = e$ en red Asia (cont. ej. 12.12)) Si hemos observado que la variable D toma el valor $\bar{d}$ y la variable X toma el valor x, disponemos de una evidencia e que afecta a las variables D, X. Para calcular las probabilidades condicionales a e, modificaremos los potenciales Ψ_5 como

$$\Psi_5^* = \begin{cases} 0 & D = d \\ \Psi_5 = p(D|E, B) & D = \bar{d}, \end{cases} \tag{12.8}$$

y Ψ_6 como

$$\Psi_6^* = \begin{cases} \Psi_6 = p(X|E) & X = x, \\ 0 & X = \bar{x}. \end{cases} \tag{12.9}$$

- Fase de propagación.

  Una vez definidos definitivamente los potenciales de los cliques, estamos en disposición de propagar la evidencia y creencia por el árbol para calcular las probabilidades de cada clique.

  La fase de propagación se realiza en dos etapas:

  - Etapa ascendente.
    Recorremos el árbol desde las hojas hacia la raíz todos los cliques. Para cada clique i con nodos G_i y padre j, calculamos la probabilidad $p(R_i|S_i)$ del residual de R_i condicionado a su separador S_i,

    $$m_i = \sum_{R_i} \Psi_i, \text{o si } S_i = \emptyset, m_i = sum_{G_i} \Psi_i \tag{12.10}$$

    $$p(R_i|S_i) = \frac{\Psi_i}{m_i} \tag{12.11}$$

    $$\tag{12.12}$$

    y propagamos la marginalización del potencial de i hacia j,

    $$\Psi_j = \Psi_j m_i.$$

    Ejemplo 12.14 (Fase ascendente de propagación (cont. ej. 12.12))
    La tabla 12.5 desarrolla la traza de la fase ascendente de propagación de la red Asia siguiendo el árbol de cliques de la figura 12.9.

  - Etapa descendente.
    Una vez calculadas las probabilidades del residual condicionado al separador de cada clique, podemos calcular la probabilidad conjunta de cada clique. Conocidas las conjuntas de cada clique, es inmediato calcular la probabilidad de cada variable por marginalización del resto del clique.
    En esta ocasión, recorreremos el árbol desde la raíz a las hojas. Como el clique raíz no tiene separador, ya tenemos calculada la probabilidad conjunta $p(G_1)$ de la etapa anterior. Para cada clique $i > 1$ con padre j, calcularemos la probabilidad conjunta de sus nodos G_i como

    $$p(S_i) = \sum_{G_j \setminus S_i} p(G_j), \tag{12.13}$$

    $$p(G_i) = p(R_i|S_i)p(S_i). \tag{12.14}$$

    Ejemplo 12.15 (Fase descendente de propagación (cont. ej. 12.14))
    La tabla 12.6 desarrolla la traza de la fase descendente de propagación de la red Asia siguiendo el árbol de cliques de la figura 12.9.

  Una vez acabada la fase descendente, podemos calcular la probabilidad de cada nodo $p(X_k)$ por marginalización del clique de menor cardinalidad que lo incluya:

  $$p(X_k) = \sum_{G_i \setminus X_k} p(G_i).$$

  Ejemplo 12.16 (Probabilidad de una variable en Asia (cont. ej. 12.15))
  Si quisiéramos conocer la probabilidad $p(E)$ de nuestra red Asia, simplemente tendríamos que marginalizar la probabilidad conjunta del clique más pequeño que contenga el nodo E, i.e. el clique 6,

  $$p(E) = \sum_{X} p(E, X).$$

Tabla 12.5: Traza de la fase ascendente de propagación para la red Asia.

Clique 6
$m_6 = \sum_X p(X\|E)$
$p(R_6\|S_6) = p(X\|E) = \frac{p(X\|E)}{\sum_X p(X\|E)}$
$\Psi_2 = p(E\|T,L) \sum_X p(X\|E)$
Clique 5
$m_5 = \sum_D p(D\|E,B)$
$p(R_5\|S_5) = p(D\|B,E) = \frac{p(D\|E,B)}{\sum_X p(D\|E,B)}$
$\Psi_3 = \sum_D p(D\|E,B)$
Clique 4
$m_4 = \sum_S p(S)p(B\|S)p(L\|S)$
$p(R_4\|S_4) = p(S\|L,B) = \frac{p(S)p(B\|S)p(L\|S)}{\sum_S p(S)p(B\|S)p(L\|S)}$
$\Psi_3 = \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
Clique 3
$m_3 = \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
$p(R_3\|S_3) = p(B\|L,E) = \frac{\sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}{\sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}$
$\Psi_2 = p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
Clique 2
$m_2 = \sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
$p(R_2\|S_2) = p(L,E\|T) = \frac{p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}{\sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}$
$\Psi_1 = p(A,T) = p(A)p(T\|A) \sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
Clique 1
$m_1 = \sum_{A,T} p(A)p(T\|A) \sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)$
$p(R_1\|S_1) = p(A,T) = \frac{p(A)p(T\|A) \sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}{\sum_{A,T} p(A)p(T\|A) \sum_{L,E} p(E\|T,L) \sum_X p(X\|E) \sum_B \sum_D p(D\|E,B) \sum_S p(S)p(B\|S)p(L\|S)}$

Tabla 12.6: Traza de la fase descendente de propagación para la red Asia.

Clique i	$p(S_i)$	$p(G_i)$
1	-	$p(A,T)$
2	$p(T) = \sum_A p(A,T)$	$p(E,L,T) = p(E,L\|T)p(T)$
3	$p(L,E) = \sum_T p(E,L,T)$	$p(B,L,E) = p(B\|L,E)p(L,E)$
4	$p(L,B) = \sum_E p(B,L,E)$	$p(S,L,B) = p(S\|L,B)p(L,B)$
5	$p(B,E) = \sum_L p(B,L,E)$	$p(D,B,E) = p(D\|B,E)p(B,E)$
6	$p(E) = \sum_{T,L} p(T,L,E)$	$p(E,X) = p(X\|E)p(E)$

En caso de observar una evidencia e la fase de propagación sería similar a la expuesta en los ejemplos anteriores, pero utilizando los potenciales Ψ^* actualizados durante la fase de absorción de la evidencia. Como resultado obtendríamos las probabilidades condicionales $p(X_k|e)$ de cada variable k a la evidencia e. Internamente, la diferencia consistiría en que las sucesivas multiplicaciones por 0 en las posiciones de los potenciales anularían las condiciones no consistentes con la evidencia del cálculo de probabilidades.

12.5.2. Simplificaciones

El algoritmo HUGIN es eficiente para calcular las probabilidades marginales de todos los nodos de la red bayesiana. Si nuestro interés reside en el cálculo de uno de los nodos, podemos aplicar dos simplificaciones que hagan la operación más eficiente:

- Simplificación por d-separación.

  Si estamos interesados en calcular la probabilidad de X_i condicionada a $E^* = e$, $p(X_i|e)$, los nodos que E^* d-separa de X_i no influyen en el cálculo, por lo que pueden eliminarse de la red.

  Ejemplo 12.17 (Simplificación por d-separación de $p(B|E = e)$ en Asia)
  Tal como vimos en el ejemplo 12.2, E d-separa X, B, por lo que podemos eliminar X de los cálculos de la probabilidad de B condicionada a observar evidencia en E, $p(B|E = e)$.

- Simplificación de nodos estériles.

  Un nodo es estéril si no recibe evidencia y todos sus hijos son estériles. Los nodos estériles no contribuyen a la probabilidad de los nodos no estériles, solo contribuyen a la probabilidad de sus descendientes.

  Ejemplo 12.18 (Simplificación de nodos estériles $p(B|E = e)$ en Asia)
  Si observamos $E = e$, los nodos X, D son estériles, por lo que no contribuirán al cálculo de probabilidades del resto de nodos, por lo que pueden ser eliminados del cálculo.

12.6. Modelado basado en conocimiento experto

Supongamos ahora que queremos diseñar una nueva red bayesiana para un problema médico específico. Podremos seguir tres aproximaciones:

- Modelado basado en conocimiento experto.
- Aprendizaje basado en casos.
- Combinada de conocimiento experto y aprendizaje basado en casos.

La tabla 12.7 comparar las pros y contras del modelado basado en conocimiento experto y el aprendizaje basado en casos.

Para el modelado de una red bayesiana mediante conocimiento experto seguiremos el siguiente procedimiento general:

1. Elección de variables.

   Deberemos elegir las variables que constituyen los nodos en la red bayesiana teniendo en cuenta la especifidad del problema médico a resolver. Debe tenerse en cuenta que la complejidad de la red depende, entre otros factores, del número de nodos, por lo que la selección de variables deberá ser lo más precisa posible.

Tabla 12.7:

	pros	contras
Experto	barato	sesgo psicológico, credibilidad, dificultad de cuantificar pontenciales
Apredizaje	directo (prospectivo), insesgado (prospectivo), rápido (retrospectivo), barato (retrospectivo),	tamaño de la base de datos, sesgo (retrospectivo), lento (prospectivo), caro (prospectivo)

2. Determinar los rangos de valores de las variables.

   Será necesario elegir un rango de valores que represente adecuadamente el conocimiento que aporta la variable al problema médico. En una red de variables discretas, debe tenerse en cuenta que el tamaño de los potenciales depende de la cardinalidad de las variables. No es obvio en variables cualitativas de carácter subjetivo comprobar que los valores elegidos son disjuntos y que abarcan todo el dominio de la variable.

3. Capturar la estructura de la red bayesiana.

   Teniendo las variables definidas, podemos pensar en las relaciones entre variables. Para variables con significado conceptual es fácil para las personas pensar en relaciones de causalidad, por lo que puede se de interés realizar una primera aproximación de la red mediante esta estrategia.

4. Refinar la estructura de la red bayesiana.

   Es un buen ejercicio intentar reducir el número de padres que un nodo recibe. Esto tiene dos razones: i) la dimensión de los potenciales de la red bayesiana y de la propagación de la evidencia dependen en gran medida del número de conexiones en la red; ii) nos puede permitir detectar redundancias debidas a diseños pobres de la red.

5. Comprobar la independencia condicional entre las variables del modelo.

   En este paso, al contrario que en el anterior, nos plantearemos si las independencias condicionales que se han generado por nuestro modelo son conceptualmente aceptables o no. En caso negativo, deberemos establecer nuevas relaciones entre variables, para subsanar la carencia conceptual de nuestro modelo.

6. Determinar las probabilidades condicionales.

   Como paso definitivo, deberemos especificar los potenciales de nuestra red bayesiana, que desde un punto de vista probabilístico se traduce en completar las tablas de probabilidades condicionales asociadas a cada nodo.

Como regla general aplicable a los puntos 1,2,3, deberemos tener en cuenta los niveles de evidencia científica que incorporan nuestras elecciones, según la MBE. Así pues, preferiremos aquellas variables, escalas y relaciones citadas en documentos con niveles de recomendación altos en las escalas de MBE (p.e. AHRQ). En definitiva daremos prioridad a las variables estudiadas mediante diseños multicéntricos (nivel A) y que han seguido una rigurosa metodología basada en ensayos aleatorios controlados ($\geq$ nivel B) y daremos menos prioridad a aquellos documentos que reflejen opiniones de comités de expertos (nivel E). Para la correcta interpretación de las fuentes de conocimiento médico, deberemos colaborar estrechamente con expertos, que podrán

guiar al diseñador en la relevancia de los elecciones para la resolución del problema médico a resolver.

Ejemplo 12.19 (Modelo emocional para terapia de depresión mayor)
La depresión mayor es una de las causas más comunes de discapacidad a corto y largo plazo en Europa. Además del sufrimiento de los propios pacientes, esta enfermedad implica un coste directo a los servicios de salud y un coste indirecto por la pérdida de productividad y la carga de cuidado que conlleva al entorno.

El uso de ordenadores para la terapia cognitivo-conductual para el tratamiento de la depresión mayor se apoya en la evidencia encontrada que indica que una terapia psicológica puede ser efectiva sin el contacto cara a cara, sin embargo, el valor de estos sistemas está limitado por la dificultad de mantener al paciente involucrado en la terapia durante largos periodos de tiempo. Esta limitación se intenta subsanar con paradigmas de la comunicación hombre-máquina basados en agentes virtuales. Específicamente, la interacción empática de estos avatares debería proporcionar una comunicación adecuada para trasmitir mensajes valiosos para la terapia del paciente siguiendo las guías establecidas por el terapeuta, pero adaptadas a cada momento.

Para conseguir una respuesta empática del avatar, es necesario conocer el estado afectivo del paciente. Para ello, nos planteamos realizar un modelo psicológico del paciente que permita predecir tanto el humor bajo un determinado ambiente y la emoción ante un evento a través de la evidencia observada mediante la monitorización de hábitos y expresiones del paciente y premisas del terapeuta sobre el estado clínico y las reacciones posibles del paciente.

Así pues, nos centraremos en el diseño de una red bayesiana para determinar el estado afectivo más probable del paciente durante la interacción con un agente virtual de asistencia a la terapia cognitivo-conductual de la depresión mayor.

Estado clínico. Siguiendo los objetivos de una terapia para la depresión mayor, podemos definir tres posibles estados del paciente:

1. Estable: cuando el paciente sigue su vida normal, sin recaídas.
2. Recaída posible: cuando hay signos iniciales de una posible recaída.
3. Situación de riesgo: cuando hay fuertes signos de recaída con posible situación de riesgo.

En un sistema de terapia continuada, nos plantearíamos un modelo predictivo para clasificar el estado clínico del paciente en estos tres posibles valores: estable, recaída y riesgo. Por ejemplo, este modelo predictivo podría basarse en la monitorización de los hábitos del paciente, p.e. patrones de sueño, actividad física y de alimentación; de expresiones del paciente, p.e. tono de voz; y de señales fisiológicas, p.e. conductividad de la piel, ritmo cardíaco, etc.

Asumiremos que el estado clínico influye en el estado afectivo del paciente, por lo que lo incluiremos como variable de nuestra red bayesiana.

El ambiente del paciente y los eventos que le afectan. Según el modelo determinista de evaluación propuesto por la teoría cognitiva de las emociones de Ortony, Clore and Collins (modelo OCC) [129], una persona está continuamente evaluando los eventos, situaciones y cosas que le rodean basándose en unos estándares, creencias y objetivos. Implementar un modelo de derivación de las variables de evaluación como [130] de un paciente con depresión mayor puede resultar complicado, por lo que optamos en primera aproximación por utilizar dos variables indicadoras. La primera recoge la percepción que el terapeuta tiene del ambiente del paciente, que podrá tomar dos valores: adecuado o inadecuado. La segunda indica la percepción que el paciente puede tener de un evento en el que está involucrado (típicamente durante la interacción con el agente virtual). Los valores que puede tomar un evento son: orden, buena noticia, mala

noticia, pregunta cognitiva, pregunta emocional, charla o silencio. El carácter de buena o mala noticia y pregunta cognitiva o emocional, puede ser preestablecida por el terapeuta, por lo que para cada evento se establece la percepción del evento mediante una tabla de correspondencia.

El modelo afectivo. Ell objetivo de nuestra red bayesiana es estimar el estado afectivo del paciente ante la observación de información como el estado clínico, el ambiente, el humor anterior del paciente, la percepción de un evento y la monitorización de hábitos y expresiones de la voz.

El estado afectivo de un paciente viene caracterizado por el humor y por la emoción que produce un evento. Comenzaremos por la caracterización del humor.

El humor. El humor es el estado emocional de media/larga duración, por lo que, temporalmente, es menos específico que la emoción (no se refiere a la reacción de un solo evento), y es más específico que la personalidad. Para caracterizar esta transitoriedad del humor en el paciente, tendremos en cuenta en nuestro modelo el estado de humor anterior, y el estado de humor actual. Según [131], el humor puede descomponerse en dos componentes ortogonales de menor nivel: la energía y la tensión.

Respecto a la Energía, una persona puede estar más energética o más cansada. En nuestro modelo, como estamos interesados en enfatizar la relación entre variables, hemos optado por discretizar la variable, por lo que tendremos dos posibles valores: energética, y cansada. Para establecer las relaciones entre variables, consideramos que la Energía está condicionada a la Energía anterior y al Estado clínico. En un Estado clínico estable, es relativamente probable que una persona continúe con la misma energía que en el momento anterior, con una ligera tendencia al incremento de su energía con el tiempo. Sin embargo, ante una posible recaída, podríamos pensar que la tendencia se invierte, y que es posible observar un cambio en la tendencia hacia el cansancio. La tabla 12.8 especifica las probabilidades condicionales de la variable Energía dadas las observaciones de Estado clínico y la Energía anterior.

Tabla 12.8: Probabilidades condicionales de la Energía.

Clinical State	Energy Past	energetic	tired
stable	energetic	.85	.15
stable	tired	.2	.8
possible	energetic	.5	.5
possible	tired	.1	.9
imminent	energetic	.2	.8
imminent	tired	.05	.95

Decimos que una persona está tensa o calmada, y podemos establecer una escala continua de Tensión entre estos dos polos. Una vez más discretizaremos la variable, por lo que la Tensión podrá tomar los valores tenso y calmado. Consideraremos que la Tensión está condicionada a la Tensión previa, el Estado clínico y al Ambiente. Así pues, un ambiente inadecuado, o una recaída aumentará la Tensión del paciente. La tabla 12.9 especifica la tabla de probabilidades condicionales de la variable a sus padres.

La emoción. La emoción es un fenómeno fisiológico que expresan la adaptación de un individuo a cierto evento. Lang [132] propuso caracterizar la emoción mediante dos dimensiones ortogonales de menor nivel: la atención (*arousal*) y la valencia (*valence*).

Tabla 12.9: Probabilidades condicionales de la Tensión

Clinical State	Tension Past	Environment	tense	calm
stable	tense	favorable	.7	.3
stable	tense	unfavorable	.8	.2
stable	calm	favorable	.1	.9
stable	calm	unfavorable	.3	.7
possible	tense	favorable	.6	.4
possible	tense	unfavorable	.9	.1
possible	calm	favorable	.3	.7
possible	calm	unfavorable	.6	.4
imminent	tense	favorable	.9	.1
imminent	tense	unfavorable	.95	.05
imminent	calm	favorable	.5	.5
imminent	calm	unfavorable	.8	.2

La atención (arousal, o excitación) es el estado fisiológico y psicológico de estar alerta, despierto o reactivo ante un evento. Discretizaremos la atención de un paciente en los valores atento y calmado. Podemos pensar que el humor actual (tanto la Energía como la Tensión) de una persona condiciona la atención de un paciente ante un evento; la tabla 12.10 recoge las probabilidades condicionales de la Atención a dichas variables.

La Valencia (valence o predisposición) establece la atracción intrínseca (placentera, o valencia positiva) o la aversión (no placentera, o valencia negativa) hacia un evento, objetivo o situación. En nuestro modelo, la Valencia está condicionada a la Tensión, pero no a la Energía del paciente; la tabla 12.11 especifica las relaciones condicionales de la Valencia.

Hábitos y expresiones del paciente El sistema de terapia cognitivo-conductual al que va dirigido el modelo afectivo que estamos diseñando está pensado para trabajar en el ambiente personal de un paciente, pudiendo monitorizar algunos de sus hábitos, como el Patrón del sueño, la Alimentación y el Ejercicio físico. Además, el interfaz principal del paciente con el sistema será la voz, por lo que también disponemos del tono de voz como indicador de las expresiones del paciente.

Hemos simplificado a valores binarios las cuatro variables, incluyendo un componente clínico es sus valores. Así pues, el Patrón del sueño será placentero o no placentero; la Alimentación será saludable o no saludable; y el Ejercicio físico será adecuado o inadecuado. Por su parte, solamente consideramos que el tono de voz tiende a ser placentero o enojado.

Las tres variables referentes a hábitos estarán condicionadas al humor, por lo que podemos establecer las probabilidades condicionales del Patrón del sueño, la Alimentación y el Ejercicio físico mediante las tablas 12.12, 12.13, y 12.14, respectivamente.

Por su parte, consideramos que el tono de voz viene directamente condicionado por la emoción del paciente, por lo que establecemos la tabla 12.15.

Tabla 12.10: Probabilidades condicionales de Arousal

Event configuration	Energy	Tension	calm	aroused
order	energetic	tense	0.1	0.9
order	energetic	calm	0.5	0.5
order	tired	tense	0.5	0.5
order	tired	calm	0.7	0.3
goodNew	energetic	tense	0.1	0.9
goodNew	energetic	calm	0.7	0.3
goodNew	tired	tense	0.6	0.4
goodNew	tired	calm	0.8	0.2
badNew	energetic	tense	0.1	0.9
badNew	energetic	calm	0.7	0.3
badNew	tired	tense	0.6	0.4
badNew	tired	calm	0.8	0.2
cognitiveQuestion	energetic	tense	0.7	0.3
cognitiveQuestion	energetic	calm	0.8	0.2
cognitiveQuestion	tired	tense	0.7	0.3
cognitiveQuestion	tired	calm	0.85	0.15
emotionalQuestion	energetic	tense	0.2	0.8
emotionalQuestion	energetic	calm	0.6	0.4
emotionalQuestion	tired	tense	0.5	0.5
emotionalQuestion	tired	calm	0.6	0.4
chat	energetic	tense	0.1	0.9
chat	energetic	calm	0.7	0.3
chat	tired	tense	0.6	0.4
chat	tired	calm	0.8	0.2
silence	energetic	tense	0.4	0.6
silence	energetic	calm	0.5	0.5
silence	tired	tense	0.5	0.5
silence	tired	calm	0.9	0.1

Tabla 12.11: Probabilidades condicionales de Valencia

Event configuration	Tension	pleasant	unpleasant
order	tense	0.5	0.5
order	calm	0.8	0.2
goodNew	tense	0.8	0.2
goodNew	calm	0.9	0.1
badNew	tense	0.2	0.8
badNew	calm	0.1	0.9
cognitiveQuestion	tense	0.8	0.2
cognitiveQuestion	calm	0.85	0.15
emotionalQuestion	tense	0.6	0.4
emotionalQuestion	calm	0.7	0.3
chat	tense	0.8	0.2
chat	calm	0.9	0.1
silence	tense	0.8	0.2
silence	calm	0.9	0.1

Tabla 12.12: Probabilidades condicionales de Patrón del sueño

Tension	Energy Past	pleasant	unpleasant
tense	energetic	.2	.8
tense	tired	.3	.7
calm	energetic	.7	.3
calm	tired	.3	.7

Tabla 12.13: Probabilidades condicionales de Patrón de alimentación

Tension	Energy Past	healthy	unhealthy
tense	energetic	.5	.5
tense	tired	.3	.7
calm	energetic	.8	.2
calm	tired	.2	.8

Tabla 12.14: Probabilidades condicionales de Actividad física

Tension	Energy Past	appropriate	inappropriate
tense	energetic	.4	.6
tense	tired	.2	.8
calm	energetic	.8	.2
calm	tired	.3	.7

Tabla 12.15: Probabilidades condicionales de Patrón de tono de voz

Arousal	Valence	anger	pleasure
calm	pleasant	.1	.9
calm	unpleasant	.7	.3
aroused	pleasant	.2	.8
aroused	unpleasant	.9	.1

Red bayesiana del estado afectivo del paciente La figura 12.10 muestra la red bayesiana obtenida de la relaciones específicas detectadas entre el estado Clínico (C), el Ambiente (En), el Evento (Ev), la Energía (E) y la Tensión (T) del humor, la Atención y la Valencia (V) de la emoción, el Patrón de sueño (S), el patrón Alimenticio (Ea), la Actividad Física (P) y el tono de voz (Vo) del paciente. El vector de variables observadas en un escenario típico de funcionamiento de la red bayesiana es: $\mathbf{Z} = \{C, En, Ev, PE, PT, Ea, S, P, Vo\}$. Además, estaremos interesado en obtener la probabilidad $p(X|\mathbf{e}_Z)$, para toda X en E, T, V, A, condicionada a la evidencia $\mathbf{e}_Z$ observada en $\mathbf{Z}$. Sin embargo, como hemos visto, la red bayesiana es suficientemente flexible para calcular la probabilidad esperada de cada variable dado un subconjunto de $\mathbf{Z}$. Esto puede ser interesante en caso de no disponer del estado de humor anterior o no se dispone de información actualizada de la monitorización de los hábitos del paciente.

Detección de nodos estériles. Teniendo en cuenta que en un escenario usual, los nodos del vector $\mathbf{Z}$ estarán observados, la red bayesiana de la figura 12.10 no tiene nodos estériles.

D-separación de los nodos no-observados. Analizamos la D-separación entre los nodos E, T, V, A, asumiendo que en el escenario normal observaremos la evidencia del vector $\mathbf{Z}$.

Los nodos E, T no son d-separable por $\mathbf{Z}$ porque son condicionalmente dependientes dado Ea, S, P or Vo, es decir, cuando Ea, S, P or Vo son observados. El mismo razonamiento se puede realizar para los pares E, A, T, V y V, A, por lo que ninguno de ellos es d-separable dado $\mathbf{Z}$. Sin embargo, todos los pares anteriores son condicionalmente independientes dado $\mathbf{Z}$. Por ejemplo, V, A son condicionalmente independientes dado $\mathbf{Z}$ porque la observación de Ev, y C bloque los caminos desde V hasta A. Así pues, la demostración de independencia condicional entre E, T y entre V, A es consistente con la descomposición del humor y la emoción en componentes ortogonales.

Simulaciones del modelo afectivo. Tomamos una serie de Pacientes Virtuales [133] con depresión mayor para simular el comportamiento de nuestra red bayesiana ante diferentes circunstancias. La tabla 12.16 define el conjunto de Pacientes Virtuales y las situaciones en als que se encuentra. Hemos planificado las simulaciones siguiendo una estrategia paso a paso, por lo que únicamente se modifica una variable desde las simulaciones VP2 a VP6, con el fin de detectar los cambios producidos por la variación del valor de una variable de $\mathbf{Z}$.

La tabla 12.17 describe los valores estimados por el modelo afectivo, calculados mediante la propagación de la evidencia observada en cada simulación de la tabla 12.16. Es importante darse cuenta como en la simulación VP3, la observación de los resultados de los hábitos produce un estado afectivo que podemos considerar como "no malo". Si estos hábitos no hubiesen sido observados (i.e. VP3.2), el humor previo tendría mucho peso al calcular el humor actual, por lo que el estado afectivo actual empeoraría. También es importante darse cuenta de lo fuerte que es el estado clínico en el modelo, por lo que una recaída lleva rápidamente la Energía del paciente al estado cansado, aumentando la probabilidad con el agravamiento de la recaída.

12.7. Aprendizaje basado en casos

El aprendizaje basado en casos de una red bayesiana pretende i) construir la estructura G de la red bayesiana, y ii) estimar el conjunto de parámetros (Θ) que controlan las relaciones entre los nodos (i.e. probabilidades condicionales) a partir de una muestra de entrenamiento $\mathcal{S} = (\mathbf{x_i}), i = 1, \ldots, N$, extraída aleatoriamente de una distribución conjunta de probabilidad desconocida $p_0(\mathbf{X}) = p(X_1, \ldots, X_D)$, de tal forma que cada caso i es el vector D-dimensional $\mathbf{x_i} = x_{i1}, \ldots, x_{iD}$.

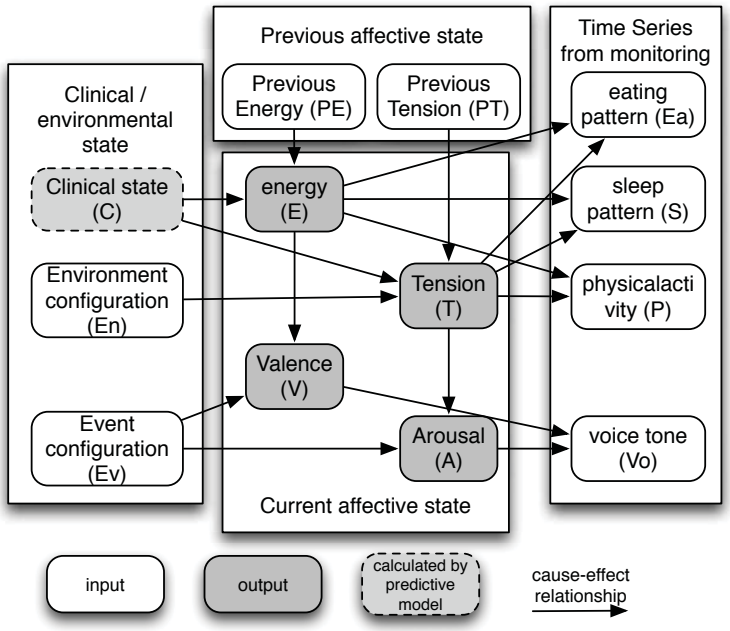


Figura 12.10: Red bayesiana del estado afectivo del paciente.

Tabla 12.16: Pacientes Virtuales y situaciones de estímulos. Las señales fisiológicas son enumeradas en el siguiente orden: patrón de alimentación, patrón del sueño, actividad física y tono de voz.

VP	C	En	Ev	Previous mood	Physiological signs
VP1	stable	favorable	badNew	energetic/calm	healthy, pleasant, appropriate, pleasure
VP2	stable	favorable	order	energetic/calm	healthy, pleasant, appropriate, pleasure
VP3	stable	favorable	order	tired/tense	healthy, pleasant, appropriate, pleasure
VP3.2	stable	favorable	order	tired/tense	na, na, na, na
VP4	stable	favorable	order	tired/tense	healthy, unpleasant, appropriate, anger
VP5	stable	unfavorable	order	tired/tense	healthy, unpleasant, appropriate, anger
VP6	possible	unfavorable	order	tired/tense	healthy, unpleasant, appropriate, anger
VP7	imminent	unfavorable	order	tired/tense	healthy, unpleasant, appropriate, anger

Tabla 12.17: Creencias del afecto del paciente, calculadas por la propagación de la evidencia observada en cada simulación en la figura 12.16

VP	PA mood	PA emotion
VP1	energetic: .992, calm: .991	calm: .826, unpleasant: 0.710
VP2	energetic: .992, calm: .993	calm: .540, pleasant: .944
VP3	energetic: .721, calm: .741	calm: .532, pleasant: .914
VP3.2	energetic: .781, tense: .813	arousal: .615, unpleasant: .791
VP4	energetic: .525, tense: .800	arousal: .709, unpleasant: .782
VP5	energetic: .525, tense: .872	arousal: .727, unpleasant: .800
VP6	tired: .675, tense: .941	arousal: .678, unpleasant: .821
VP7	tired: .815, tense: .972	arousal: .637, unpleasant: .831

Para la construcción de la estructura de la red bayesiana G existen algoritmos, casi siempre, basados en teoría de grafos y en descubrimiento de la independencia condicional que permiten encontrar una estructura compatible con la muestra de entrenamiento. Esta tarea queda fuera de los objetivos de este texto, y se remite al estudiante al curso de Redes bayesianas de Pedro Larrañaga [134] para una introducción al tema.

12.7.1. Aprendizaje de los parámetros de una red bayesiana

Supondremos una red bayesiana de la que sabemos su estructura G y de la cual pretendemos aprender sus parámetros Φ a partir de una muestra $\mathcal{S}$. En una red de variables discretas, donde cada variable $X_i, i = 1, \ldots, D$ puede tomar los valores $x_i^{(1)}, \ldots, x_i^{(r_i)}$, estamos interesados en aprender la probabilidad condicional $\phi_{ijk} = p(X_i = x_i^{(k)} | pa(X_i) = pa_i^{(j)}), k = 1, \ldots, r_i, j = 1, \ldots, q_i$, siendo q_i el número de posibles instancias diferentes de los padres de X_i, i.e. $q_i = \prod_{X_p \in pa(X_i)} r_p$. En definitiva, estamos interesados en aprender el valor de cada posición de los potenciales ϕ_i de cada nodo i de la red.

Ejemplo 12.20 (Parámetros a estimar de la red bayesiana Asia)
Supongamos la red de Asia (ej. 12.1), con los nodos numerados en el siguiente orden A, S, T, L, B, E, X, D y que por lo tanto identificaremos numéricamente como $i = 1, 2, 3, 4, 5, 6, 7, 8$ respectivamente. Supongamos también que el valor positivo de cada nodo, p.e. t para el nodo T, se numera con el valor $k = 1$, y el valor negativo, p.e. $\bar{t}$, con el valor $k = 2$.

Podemos identificar los parámetros necesarios para especificar el potencial ϕ_1 del nodo A como

$$\phi_A = \phi_1 = (\phi_{1-1}, \phi_{1-2}) = (p(X_A = a), p(X_A = \bar{a})),$$

los parámetros del potencial ϕ_T que depende de A ($r_3 = 2, q_3 = 2$) son

$$\begin{aligned} \phi_T = \phi_3 \quad &= (\phi_{311}, \phi_{321}, \phi_{312}, \phi_{322}) && (12.15)\\ &= (p(X_T = t | X_A = a), (p(X_T = t | X_A = \bar{a}), && (12.16)\\ &\quad p(X_T = \bar{t} | X_A = a), (p(X_T = \bar{t} | X_A = \bar{a})), && (12.17) \end{aligned}$$

y los parámetros del potencial ϕ_E que depende de T, L ($r_6 = 2, q_6 = 4$) son

$$\begin{aligned}
\phi_E = \phi_6 &= (\phi_{611}, \phi_{621}, \phi_{631}, \phi_{641}, \phi_{612}, \phi_{622}, \phi_{632}, \phi_{642}) && (12.18)\\
&= (p(X_E = e|X_T = t, X_L = l), (p(X_E = e|X_T = t, X_L = \bar{l}), && (12.19)\\
&\quad (p(X_e = e|X_T = \bar{t}, X_L = l), (p(X_E = e|X_T = \bar{t}, X_L = \bar{l}), && (12.20)\\
&\quad (p(X_E = \bar{e}|X_T = t, X_L = l), (p(X_E = \bar{e}|X_T = t, X_L = \bar{l}), && (12.21)\\
&\quad (p(X_E = \bar{e}|X_T = \bar{t}, X_L = l), (p(X_E = \bar{e}|X_T = \bar{t}, X_L = \bar{l}). && (12.22)
\end{aligned}$$

Independencia global de los parámetros. La verosimilitud de la muestra $\mathcal{S}$ dados los parámetros $\mathbf{\Phi}$,

$$L(\mathcal{S}, \mathbf{\Phi}) = p(\mathcal{S}|\mathbf{\Phi}) = p(\mathbf{x_1}, \ldots, \mathbf{x_N}|\mathbf{\Phi}) = \prod_{s=1}^{N} p(\mathbf{x_s}|\mathbf{\Phi}) = \prod_{s=1}^{N} p(x_{s1}, \ldots, x_{sD}|\mathbf{\Phi}),$$

puede escribirse por factorización de la probabilidad conjunta 12.5 como

$$\prod_{s=1}^{N}\prod_{i=1}^{D} p(x_{si}|pa(x_{si}, \mathbf{\Phi})) = \prod_{i=1}^{D}\prod_{s=1}^{N} p(x_{si}|pa(x_{si}, \mathbf{\Phi})),$$

que asumiendo independencia entre los potenciales

$$L(\mathcal{S}, \mathbf{\Phi}) = \prod_{i=1}^{D}\prod_{s=1}^{N} p(x_{si}|pa(x_{si}, \mathbf{\Phi})) = \prod_{i=1}^{D}\prod_{s=1}^{N} p(x_{si}|pa(x_{si}, \phi_{\mathbf{i}})) = \prod_{i=1}^{D} L(\mathcal{S}_i, \phi_{\mathbf{i}}),$$

donde con $\mathcal{S}_i$ denotamos las variables involucradas en el potencial ϕ_i.

Así pues, podemos estimar por máxima-verosimilitud los parámetros del potencial ϕ_i mediante las variables que influyen en X_i, independientemente del resto de variables.

Ejemplo 12.21 (Independencia global de parámetros en Asia (cont. ej. 12.20)) Tomaremos como ejemplo la variable E sin perdida de generalización al resto de variables. Gracias a la propiedad 12.7.1, la estimación de los parámetros del potencial ϕ_E solo depende de las variables T, L, E, por lo que únicamente se tomarán en consideración los valores que toman dichas variables en los casos de$\mathcal{S}$.

Independencia local de los parámetros. Reescribimos la verosimilitud mediante la agrupación de los casos de $\mathcal{S}_i$ por las combinaciones q_i de valores que toman los padres de X_i,

$$\begin{aligned}
L(\mathcal{S}, \mathbf{\Phi}) &= \prod_{i=1}^{D} L(\mathcal{S}_i, \phi_{\mathbf{i}}) = \prod_{i=1}^{D}\prod_{s=1}^{N} p(x_{si}|pa(x_{si}), \phi_{\mathbf{i}})) && (12.23)\\
&= \prod_{i=1}^{D}\prod_{j=1}^{q_i}\prod_{s=1}^{N_{ij}} p(x_{si}|pa(x_{si})^{(j)}, \phi_{\mathbf{ij}})) = \prod_{i=1}^{D}\prod_{j=1}^{q_i} L(\mathcal{S}_{ij}, \phi_{\mathbf{ij}}), && (12.24)
\end{aligned}$$

donde N_{ij} es el número de casos en $\mathcal{S}$ donde se observa la configuración $pa(x_i^{(j)})$.

Así pues, asumiendo independencia entre los parámetros de ϕ_i, la estimación de cada columna del potencial ϕ_i (ver tabla 12.2) es independiente del resto.

Ejemplo 12.22 (Independencia local de parámetros en Asia (cont. ej. 12.21)) Tomaremos de nuevo como ejemplo la variable E sin perdida de generalización al resto de variables. Gracias a la propiedad 12.24, el cálculo de los parámetros $\phi_{61k}, k = 1, 2$ solo se tendrán en cuenta los casos de la muestra donde $T = t, L = l$. De forma similar, $\phi_{63k}, k = 1, 2$ solo requiere observar las muestras donde $T = \bar{t}, L = l$.

Estimación por máxima verosimilitud Si definimos N_{ijk} como el número de casos que toma simultáneamente el valor $X_i = x_i^{(k)}$ y sus padres $pa(X_i)$ toman los valores $pa(x_i)^{(j)}$, es fácil demostrar que la estimación por máxima verosimilitud del parámetro ϕ_{ijk} es

$$\phi_{ijk} = p(X_i = x_i^{(k)} | pa(X_i) = pa(x_i)^{(j)}) = \frac{N_{ijk}}{N_{ij}}.$$

Cabe destacar que existen aproximaciones que solucionan la posible dispersión de las bases de datos, y que pueden dar conteos iguales a 0 para ciertas combinaciones de valores. Algunas de las soluciones están basadas en inferencia bayesiana, para incorporar conocimiento a priori sobre la distribución de los parámetros. Otras soluciones plantean suavizados sobre los potenciales.

También es importante comentar lo frecuente que es encontrar bases de datos con valores no observados en variables de algunos de sus casos. Si los datos no observados están dispersos y no resultan excesivos en comparación con el tamaño de la muestra, pueden ser de utilidad la aplicación de soluciones basadas en la imputación de la moda, o imputación de datos perdidos por el algoritmo Expectation-Maximization (EM).

12.8. Notas bibliográficas

Finn Jensen en [128] explica las bases, modelado, inferencia y aprendizaje de las redes bayesianas con variables discretas de forma clara y concisa. Pedro Larrañaga, en su curso de redes bayesianas [134], resuelve unos ejemplos muy aclaratorios del concepto de d-separación. Eva Millán, en su tesis doctoral, desarrolla un excelente capítulo [135] sobre los algoritmos de actualización de probabilidades en redes bayesianas. NETICA (norsys.com) permite modelar las redes bayesianas de forma rápida e intuitiva, además de comprobar su comportamiento ante la observación de evidencias.

Capítulo 13

Modelos basados en memoria

Los modelos vistos hasta ahora cuentan con unos parámetros que se ajustan en el proceso de entrenamiento en base a las observaciones disponibles. Posteriormente, pueden utilizarse para predecir nuevos casos sin necesidad de acceder a los datos observados anteriores. Estos son modelos que podemos llamar "sin memoria". Existen otro tipo de modelos que, al contrario que los modelos vistos en los capítulos anteriores, requieren el almacenamiento de los datos observados. En estos modelos basados en memoria la regla de clasificación depende de una métrica y no de los parámetros ajustados con los datos observados. El caso más conocido de modelo basado en memoria es el del vecino más próximo.

13.1. K-vecinos más próximos

Los modelos basados en el vecino más próximo (*nearest neighbour* y *K-nearest neighbour*) son modelos no paramétricos basados en distancias. Estos modelos asumen que el espacio muestral es un espacio métrico $\{X, d\}$, donde X es el conjunto de puntos u observaciones y d es una métrica o distancia, definida como $d : X \times X \to \mathbb{R}$. Además, una métrica debe cumplir las siguientes propiedades para todo $\mathbf{x}_i \in X$:

- No negativa: $d(\mathbf{x}_1, \mathbf{x}_2) \geq 0$. Si y solo si $\mathbf{x}_1 = \mathbf{x}_2$ entonces $d(\mathbf{x}_1, \mathbf{x}_2) = 0$.
- Simétrica: $d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_2, \mathbf{x}_1)$.
- Desigualdad triangular: $d(\mathbf{x}_1, \mathbf{x}_2) + d(\mathbf{x}_2, \mathbf{x}_3) \geq d(\mathbf{x}_1, \mathbf{x}_3)$.

Como sabemos, las observaciones pueden ser consideradas vectores de un espacio vectorial. Esta representación nos permite establecer un espacio métrico a partir del espacio vectorial empleando las métricas de la familia $L_p = d_p(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_i |x_{1i} - x_{2i}|^p\right)^{1/p}$. De este modo, las tres métricas más empleadas son:

- L_0 (o L_∞): $d(\mathbf{x}_1, \mathbf{x}_2) = \text{máx}_{1 \leq i \leq D} |(x_{1i} - x_{2i})|$.
- L_1: $d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{D} |(x_{1i} - x_{2i})|$.
- L_2: $d(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{i=1}^{D} (x_{1i} - x_{2i})^2\right)^{1/2}$.

Estas tres distancias son las más empleadas y comúnmente se conocen como distancia del ajedrez o de Cheryshev ($L0$), distancia de Manhattan ($L1$) y distancia euclídea ($L2$).

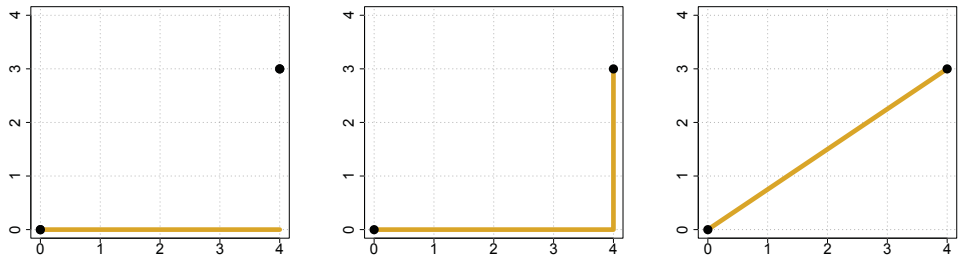


Figura 13.1: Ilustración de las métricas L0, L1 y L2, respectivamente. La distancia entre ambos puntos es 4 si se emplea la métrica L0, 7 si se emplea la métrica L1 y 5 si se emplea la métrica L2.

Si se dispone de un conjunto de observaciones $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, donde $\mathbf{x}_i \in X,\ i = 1, \ldots, N$ e y_i es la clase a la que pertenece a observación i-ésima, se puede establecer un espacio métrico $\{X, d\}$. Una vez establecido el espacio métrico, la clase a la que pertenece una nueva muestra $\mathbf{x}$ se calcula en base a la observación u observaciones más cercanas según la distancia $d(\cdot, \cdot)$. Esto es, si los puntos vecinos son de la clase y, entonces se asigna la clase y a la observación nueva $\mathbf{x}$. Este modo de proceder ha tenido bastante éxito debido a que es muy intuitivo.

13.1.1. Vecino más próximo

El vecino más próximo [136] es la regla más sencilla de estos métodos, ya que únicamente busca la clase del vecino más cercano en función de la métrica que se esté empleando. A los datos observados de una clase c se les denominan **prototipos** de la clase, P_c. La regla de decisión para una nueva observación $\mathbf{x}$ es

$$\mathbf{x} \in \hat{c} \Longleftrightarrow \exists \mathbf{z} \in P_{\hat{c}} : d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{z}') \forall \mathbf{z}' \in P_c, 1 \leq c \leq C, c \neq \hat{c}. \tag{13.1}$$

Es decir, a $\mathbf{x}$ se le asigna la misma clase a la que pertenece la observación $\mathbf{z}$ más cercana, bajo la métrica d. En el improbable caso de empate nos encontramos en una situación donde $d(\mathbf{x}, \mathbf{z}_1) = d(\mathbf{x}, \mathbf{z}_2)$ con $\mathbf{z}_1 \in c_1$ y $\mathbf{z}_2 \in c_2$. En este caso se escoge la clase que más representantes tiene. Esto es, se asigna la clase con mayor prevalencia o probabilidad *a priori*.

Las fronteras que describe el método del vecino más próximo quedan determinadas por el conjunto de puntos $\mathcal{S}$. De este modo, las funciones discriminantes son lineales a trozos ya que cada subconjunto de K vecinos describe una frontera lineal.

Una propiedad muy interesante del clasificador por el vecino más próximo es que, cuando el número de observaciones tiende a infinito, $N \to \infty$, el error del clasificador puede acotarse como:

$$P^* \leq P \leq P^* \Big(2 - \frac{C}{C-1} P^*\Big) \leq 2P^*,$$

donde P^* es el error teórico de Bayes, P es el error del clasificador y C es el número de clases.

13.1.2. K-vecinos más próximos

Se puede generalizar este concepto si, en lugar de tener en cuenta la etiqueta de clase del vecino más próximo, se tienen en cuenta las etiquetas de clase de los K vecinos más próximos [92].

De hecho, el vecino más próximo es el caso particular cuando K = 1. Si se tiene un conjunto de prototipos de cada clase P_c y siendo P_k el conjunto de los K vecinos más próximos bajo la métrica d, la regla de decisión de este tipo de clasificadores se define como

$$\mathbf{x} \in \hat{c} \Longleftrightarrow |P_k \cap P_{\hat{c}}| \geq |P_k \cap P_c|, 1 \leq c \leq C, c \neq \hat{c}. \tag{13.2}$$

Esta regla quiere decir que, una vez calculados los K vecinos más próximos, se contabilizan los representantes de cada clase y aquella clase que tenga mayor número de representantes entre los K vecinos más próximos será la etiqueta de clase que se asigne a la nueva observación $\mathbf{x}$. Generalmente, en caso de empate se desempata empleando la regla 1-NN.

Al igual que los modelos NN, los K-vecinos más próximos definen fronteras lineales a trozos (ver figura 13.2). Además, un clasificador K-NN tiende al error teórico de Bayes cuando se cumplen tres condiciones: $N \to \infty$, $\mathrm{K} \to \infty$ y ${}^{\mathrm{K}}/_{N} \to 0$. Por ejemplo, si se selecciona un valor $\mathrm{K} = \sqrt{N}$ se tiene garantías de alcanzar el error de Bayes si $N \to \infty$ [92]. Aunque este comportamiento teórico asintótico es inmejorable, depende en gran medida del número de observaciones disponibles. En este sentido, cuando se disponen de conjuntos de datos finitos resulta difícil garantizar dicho ideal.

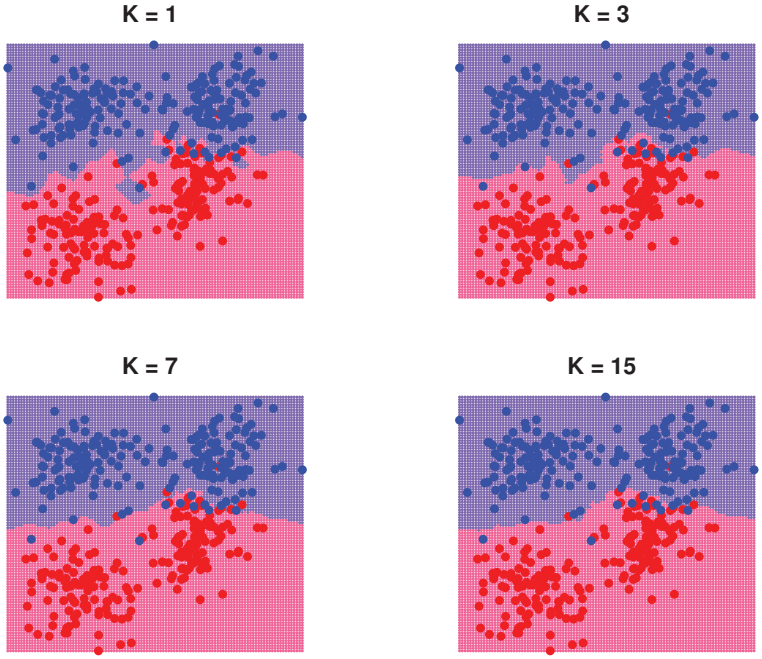


Figura 13.2: Ilustración de las fronteras y regiones definidas por modelos K-vecinos con distintos valores de K. Se puede observar que cuando K crece, las fronteras de decisión son menos abruptas.

Estos modelos no paramétricos se caracterizan por la necesidad de almacenar todos los prototipos etiquetados disponibles. Esto podría implicar unos elevados costes computacionales y de almacenamiento. Además, en la práctica nunca es posible saber por adelantado cuál es el mejor valor de K o cuál es la mejor métrica para cada problema, por lo que será necesario realizar un buen diseño experimental para acertar con estos valores.

13.2. K-vecinos probabilístico

El modelo de clasificación de los K-vecinos más próximos resulta atractivo por su sencillez, el amplio estudio teórico y los resultados empíricos recogidos en la literatura científica. Sin embargo, existen algunos inconvenientes con los K-vecinos:

1. Escoger el valor óptimo de K no es trivial. El método más empleado es usar una validación cruzada para seleccionar el mejor valor de K.

2. Las ventajas de su comportamiento asintótico depende de disponer de un número de datos suficientemente grande, cosa que no siempre es posible en problemas biomédicos reales.

3. Independientemente del valor de K, las predicciones del modelo no proporcionan una interpretación probabilística razonable. Generalmente, se adjudica una probabilidad que depende de la tasa de casos que pertenecen a la clase ganadora de entre los K vecinos más próximos. De este modo la probabilidad se discretiza en saltos de $1/K$. Este problema impide hacer un uso apropiado de los modelos de K vecinos más próximos en procesos de decisión.

En 2002 Holmes y Adams en [137] propusieron un clasificador de K-vecinos probabilístico que devuelve una probabilidad a posteriori de la clase y que a su vez es independiente del número de vecinos. Esto es posible gracias a la aplicación de inferencia bayesiana y al uso de múltiples modelos con diferentes valores de K. A continuación se explican los detalles del K-vecino probabilístico.

13.2.1. Modelo matemático

Dados unos datos observados $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, se define una función de verosimilitud como

$$p(\mathbf{y}|\mathbf{X}, k, \beta) = \prod_{n=1}^{N} \frac{\exp\left\{\frac{\beta}{k} \sum_{j\sim n|k} \delta(y_n, y_j)\right\}}{\sum_{c=1}^{C} \exp\left\{\frac{\beta}{k} \sum_{j\sim n|k} \delta(c, y_j)\right\}} \tag{13.3}$$

donde el vector $\mathbf{y}$ de dimensión $N \times 1$ representa la clase a la que pertenecen las observaciones, $\mathbf{X}$ es la matriz de las características de las observaciones de dimensión $N \times D$, k es el número de vecinos que se va a evaluar y β es una variable que pondera la intensidad de la asociación entre los vecinos. La expresión

$$\sum_{j\sim n|k} \delta(y_n, y_j) \tag{13.4}$$

representa el número de k vecinos más próximos de $\mathbf{x}_n$, bajo la métrica escogida, donde $\delta(\cdot, \cdot)$ es la función de Dirac ($\delta(a, b) = 1$ si $a = b$, en otro caso $\delta(a, b) = 0$).

Siguiendo los principios de la inferencia bayesiana, la estimación de la clase de una nueva observación $\mathbf{x}_{new}$ se estimará mediante la distribución predictiva final que marginaliza los parámetros del modelo, β y k. Como k es un parámetro discreto, se emplea el sumatorio en lugar de la integral:

$$p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{X}) = \sum_k \int p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{X}, k, \beta) p(k, \beta|\mathbf{y}, \mathbf{X}) d\beta \tag{13.5}$$

Se puede observar que esta expresión se compone de dos factores: la probabilidad de la clase dado el modelo $p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{X}, k, \beta)$ y la probabilidad a posteriori de los parámetros $p(k, \beta|\mathbf{y}, \mathbf{X})$. La primera de estas probabilidades tiene una expresión similar a la ecuación (13.3):

$$p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{X}, k, \beta) = \frac{\exp\left\{\frac{\beta}{k}\sum_{j\sim new|k}\delta(y_{new}, y_j)\right\}}{\sum_{c=1}^{C}\exp\left\{\frac{\beta}{k}\sum_{j\sim new|k}\delta(c, y_j)\right\}} \tag{13.6}$$

de este modo, la clase y_{new} más probable para la observación $\mathbf{x}_{new}$ vendrá dada por la clase más común entre los k vecinos más próximos, donde el parámetro β funciona como una variable de escalado.

Para la segunda, la probabilidad a posteriori de los parámetros, se estima mediante muestreo por Metropolis (ver apéndice E.3). Para ello, se asigna una distribución a priori para β y para k y se aplica el método de inferencia bayesiana que establece que la probabilidad a posteriori es proporcional a la verosimilitud por la probabilidad a priori, de modo que:

$$p(k, \beta|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, k, \beta)p(k, \beta) \tag{13.7}$$

La aproximación común [137, 138], propone adoptar una probabilidad uniforme discreta para el parámetro k y una probabilidad normal centrada en 0 y con una varianza pequeña para β. Además, como el parámetro β debe tomar valores positivos, cuando β toma valores aleatorios negativos se adopta la práctica de asignar $\beta = -\beta$. Así pues, al aplicar el algoritmo Metropolis se generan los parámetros k y β empleando las siguientes expresiones:

$$k_{n+1} = k_n \pm U[0, \ldots, K_{max}] \tag{13.8}$$

$$\beta_{n+1} = \beta_n + N(0, s^2) \tag{13.9}$$

siendo K_{max} un máximo número de vecinos definido por el ingeniero y s^2 una varianza generalmente menor que 1. Estos parámetros se aceptan con probabilidad

$$\alpha(\{k_{n+1}, \beta_{n+1}\}|\{k_n, \beta_n\}) = \text{mín}\left\{1, \frac{p(\mathbf{y}|\mathbf{X}, k_{n+1}, \beta_{n+1})}{p(\mathbf{y}|\mathbf{X}, k_n, \beta_n)}\right\} \tag{13.10}$$

Donde los factores del numerador y el denominador se estiman mediante la ecuación (13.3).

Tras generar mediante el algoritmo de Monte Carlo un conjunto de M pares de parámetros k y β, se puede estimar la probabilidad de una clase para una nueva observación aproximando la probabilidad predictiva final de la ecuación (13.5) mediante el cómputo de M modelos que empleen la ecuación (13.6):

$$p(y_{new}|\mathbf{x}_{new}, \mathbf{y}, \mathbf{X}, k, \beta) = \frac{1}{M}\sum_{k,\beta}\frac{\exp\left\{\frac{\beta}{k}\sum_{j\sim new|k}\delta(y_{new}, y_j)\right\}}{\sum_{c=1}^{C}\exp\left\{\frac{\beta}{k}\sum_{j\sim new|k}\delta(c, y_j)\right\}} \tag{13.11}$$

Este tipo de cálculo nos permite estimar una frontera de decisión basada en las probabilidades predictivas finales estimadas y que, por tanto, permiten establecer una frontera suave con gradiente como ilustra la figura 13.3.

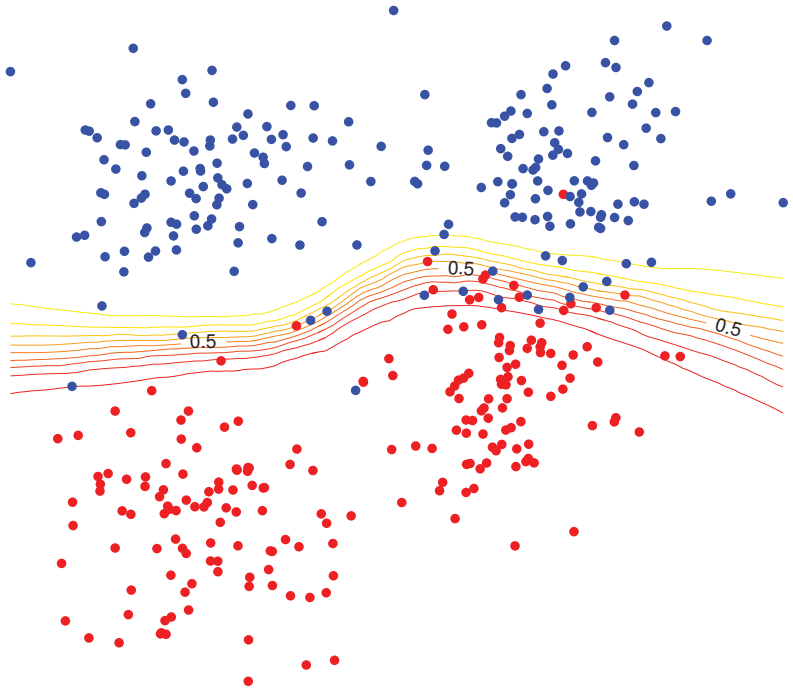


Figura 13.3: Ilustración de las fronteras y regiones definidas por el modelo K-vecinos más próximos probabilístico. La frontera de decisión se define como un gradiente de probabilidad donde la decisión de la clase se tomaría en la curva donde la probabilidad es 0,5.

Capítulo 14

Evaluación y selección de modelos de aprendizaje automático

Como hemos visto en el capítulo 8, el proceso de aprendizaje automático tiene como objetivo encontrar la solución que minimice el riesgo (8.5) de un decisor, es decir la pérdida esperada al afrontar nuevas situaciones del problema que resuelve.

Para obtener una solución al problema, empleando técnicas de aprendizaje automático, se dispone de conocimiento previo y de muestras procedentes de experimentos relacionados con el problema mediante los cuales se espera obtener un modelo lo más aproximado a la solución de mínimo riesgo y, por tanto, sea generalizable a nuevos casos. Cuando se abordan problemas de clasificación o regresión, cuyas funciones de pérdida son 0-1 (8.16) o el error cuadrático(8.25) respectivamente, el riesgo del modelo suele denominarse *error de generalización*. Por lo general, el error de generalización de un modelo estará influido por el sesgo y la varianza (o precisión) de las soluciones obtenidas por un algoritmo de aprendizaje, por lo que dicho error no llegará a ser el mínimo posible.

Conocer el riesgo (o el error de generalización) implica conocer la distribución real del problema, que suele estar oculta en los problemas de interés. Por lo tanto, necesitaremos realizar una estimación del riesgo lo más fiable posible mediante un conjunto de muestras limitado.

Antes de continuar, debemos considerar que nuestras estimaciones deben ser fiables. Todo estimador tiene un sesgo y una varianza, que definen su fiabilidad:

- Sesgo: $\epsilon - E(\hat{\epsilon})$
- Varianza: $E[(\hat{\epsilon} - E(\hat{\epsilon}))^2]$

Donde ϵ es el error de generalización, $\hat{\epsilon}$ es una estimación conseguida con un conjunto de muestras y $E[\hat{\epsilon}]$ es el valor esperado (media) de las estimaciones realizadas con diferentes conjuntos de muestras La figura 14.1 ilustra la fiabilidad de los estimadores en los términos de sesgo y varianza.

De forma práctica, durante el diseño de los decisores, podremos usar la estimación del riesgo para dos tareas: i) *la evaluación de modelos*, es decir, saber si los modelos son suficientemente buenos; ii) *la selección de modelos*, es decir, saber si un modelo es mejor que otro para la resolución del problema.

14.1. Descomposición del error de generalización

El error de generalización puede descomponerse en tres componentes, cuyo análisis nos dará pistas sobre la mejor estrategia de diseño de un modelo de decisión.

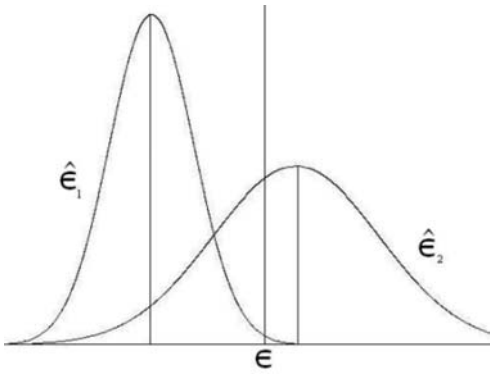


Figura 14.1: La fiabilidad de los estimadores $\hat{\epsilon}_1$ y $\hat{\epsilon}_2$ vendrá dada por su sesgo y la varianza. El sesgo será la distancia de la estimación media al valor real ϵ, la varianza será la de su distribución ante cambios de las muestras utilizadas durante la estimación.

La primera componente es el *error irreductible*, que corresponde al solapamiento entre clases que da lugar al error de Bayes en un clasificador y que corresponde al error ϵ en un problema de regresión $y = f(x) + \epsilon$. Como ya hemos estudiado, es el mínimo error posible y, por lo tanto, el objetivo del proceso de aprendizaje. Esta componente será cero si no existe solapamiento entre distribuciones, y distinto de cero si existen los casos $(\mathbf{x}, c_1), (\mathbf{x}, c_2)$ en $(\mathcal{X}, \mathcal{Y})$. El objetivo de un proceso de aprendizaje con funciones 0-1 o cuadrático es obtener este error irreductible.

La segunda componente es el *sesgo*, que indica la calidad media de las soluciones al problema que puede aportar el modelo. Corresponde a la discrepancia (distancia) entre el modelo medio estimado y el modelo de Bayes correspondiente a la distribución real. Indica la calidad media de las soluciones al problema que puede aportar en la familia de modelos elegida (p.e. LDA, QDA, 1-NN). Un sesgo alto indica que la familia elegida tiene una capacidad pobre de solucionar el problema de decisión. El sesgo tiende a ser mayor en familias de modelos simples (con pocos grados de libertad).

La tercera componente es la *varianza*, que corresponde a una medida de precisión de la solución al problema. Corresponde a la medida de variabilidad de los modelos estimados respecto al modelo medio estimado. Donde los modelos estimados pertenecen a una misma familia, y cada uno de ellos se ha estimado con diferentes muestras de la población. Una varianza alta indica que el modelo elegido puede variar fácilmente con pequeñas variaciones del proceso de aprendizaje, por lo que resultará una solución débil del problema. La varianza tiende a ser mayor en familias de modelos complejas (con muchos grados de libertad).

El sesgo y la varianza influyen en el error total de forma diferente dependiendo de la función de pérdida utilizada. Cuando la función de riesgo de un problema de regresión se define mediante el valor esperado de la función error cuadrático, el error final será la suma del error irreductible, el sesgo al cuadrado y la variabilidad [139].

Sin embargo, cuando abordamos un problema de clasificación, la influencia del sesgo y la varianza en el error final no es lineal y la interacción entre ambos hace que el sesgo asigne el signo al error producido por la variabilidad [92]. De esta forma, el valor absoluto del sesgo no será importante ya que únicamente su signo influye en el error de generalización. El valor absoluto de la varianza, por lo contrario, si que influye en el error por lo que es importante mantenerlo en valores bajos.

En general, podemos pensar que modelos con alto poder representativo tienen la suficiente flexibilidad para aproximar su solución a la solución óptima y, por lo tanto, tener un sesgo relativamente bajo. Sin embargo, esto se consigue a costa de aumentar su complejidad, lo que implica una mayor varianza, ya que los modelos complejos requieren el ajuste de un espacio paramétrico mayor que los modelos más simples. Como consecuencia, los modelos complejos requieren un número de muestras suficientemente alto, para llegar a reducir la varianza que las soluciones obtenidas pueden tener al utilizar conjuntos pequeños de muestras.
En cambio, los modelos simples requieren el ajuste de menor número de parámetros, por lo que su varianza será menor y por lo tanto requerirán menor número de muestras para su estimación. Sin embargo, la limitada flexibilidad de los modelos hará más difícil su aproximación a la solución óptima, por lo que serán modelos con un sesgo mayor.

La estimación de las componentes error irreductible, sesgo y varianza en problemas reales es difícil. Además, es el error de generalización el que necesitamos saber para evaluar y comparar los modelos. Sin embargo, con el análisis realizado, podemos decir que para el aprendizaje de modelos de clasificación será importante mantener ajustada la complejidad de los modelos, ya que es la varianza la que domina su error. Además, intentaremos aumentar en lo posible el uso de las muestras disponibles para entrenamiento para ajustar mejor el modelo a las características del problema.

Como conclusión del estudio de descomposición del error, podemos dar las siguientes recomendaciones:

- La estimación de las descomposición del error es difícil de llevar a cabo en problemas reales
- Sí que se puede realizar con simulaciones, o asumiendo distribuciones reales por conocimiento experto o complementario
- Duda & Hart deducen que en el error de generalización en clasificación influye:
  - La magnitud de la varianza
  - El signo del sesgo
- Por lo tanto es imprescindible controlar la varianza del modelo en valores pequeños
  - Se conseguirá manteniendo baja la complejidad de los modelos
  - Además, intentaremos aprovechar el máximo de muestras de entrenamiento posible para realizar un ajuste robusto de sus parámetros
  - Algunas aproximaciones buscan utilizar modelos complejos pero acotados mediante información a priori

14.2. Estimación del error de generalización

Recordemos que el objetivo de los problemas de aprendizaje es la minimización del riesgo, expresado mediante la ecuación (8.5). El cálculo del riesgo implica el conocimiento de la distribución del problema, pero esta suele ser desconocida. Podemos estimar el riesgo empírico mediante un conjunto de muestras $\mathcal{S} = \{(x_i, y_i)\}, i = 1, \ldots, N$, según la ecuación (3.11):

$$R_{\mathcal{S}}[\alpha] = \frac{1}{N} \sum_{i=1}^{N} L[y_i, f(x_i, \alpha)] \tag{14.1}$$

Así pues, para los problemas de clasificación y regresión, calcularemos la estimación del *error de generalización* mediante un conjunto de muestras pertenecientes a la distribución del problema.

Cuando utilizamos las mismas muestras de entrenamiento para estimar el error, utilizamos la aproximación por *resustitución* y obtenemos como resultado el llamado *error de entrenamiento*. Como resumen tendremos las siguientes características de nuestra estimación:

- Es una estimación optimista: subestimación del error
- Especialmente optimista en la evaluación de modelos complejos estimados por máxima verosimilitud
  - El modelo puede sobreajustarse a las muestras
  - La estimación basada en las mismas muestras resulta en un error irrealmente bajo

Ejemplo 14.1 (Clasificación de cáncer de mama por la morfología celular[a])
Supongamos que el clasificador de Bayes entre los caracteres *Benigno* (B) y *Maligno* (M) para cáncer de mama basado en la morfología de núcleos celulares observados mediante imágenes citológicas puede representarse mediante:

$$\begin{aligned} p(y=B) &= 0{,}5 \\ p(x|y=B) &= N(12{,}1465; 1{,}7805) \\ p(y=M) &= 0{,}5 \\ p(x|y=B) &= N(17{,}4628; 3{,}2040) \\ \hat{y}^* &\leftarrow \arg\max_{y\in\mathcal{Y}} p(y|x), \end{aligned}$$

siendo x el radio del núcleo celular (calculado como distancia media del centro a los puntos del perímetro), y las funciones de densidad de probabilidad condicional responden a distribuciones normales.

El error de generalización podría calcularse fácilmente ya que sabemos su distribución real[b] y en nuestro ejemplo obtendríamos $p(error) = 0{,}1349$.

Si, por el contrario, el modelo anterior fuera una estimación, no tendríamos la certeza que nuestra estimación fuera la correcta y, por lo tanto, lo que podríamos hacer es estimar el error de generalización mediante un conjunto $\mathcal{S} = \{(x_i, y_i)\}, i = 1, \ldots, N$, obteniendo el error empírico mediante la ecuación (14.1). Por ejemplo, para un conjunto i.i.d. de $N = 10^6$ muestras, el error empírico obtenido fue $p_{\mathcal{S}}(error) = 0{,}1352$[c].

Por lo que vemos, parece intuitivo utilizar el *conjunto de muestras de entrenamiento* disponible durante la preparación del modelo, $\mathcal{S}_T = \{(x_i, y_i)\}, i = 1, \ldots, N$ para estimar el error, obteniendo el llamado *error de entrenamiento*,

$$err_T = \frac{1}{N}\sum_{i=1}^{N} L[y_i, f(x_i, \alpha)]. \tag{14.2}$$

Sin embargo, cuando el número de muestras es limitado, esta estimación puede resultar optimista y, por lo tanto, no representar el rendimiento del predictor en muestras futuras. Esto es debido al sobreajuste del modelo al conjunto de entrenamiento (o *sobreentrenamiento*) que se produce principalmente en modelos complejos estimados por máxima verosimilitud.

[b] En este ejemplo hemos utilizado la integración numérica por trapecios sobre el rango $[0, 30]$ de la función $p(x, error) = min(p(x, y = B), p(x, y = M))$.

[c] La repetición del experimento supuso una desviación de la estimación respecto al error de generalización inferior a 10^{-3}.

Ejemplo 14.2 (Clasificación de cáncer de mama por la morf. celular (cont 14.1)) Utilizaremos ahora dos características, *textura* y *área*, extraídas de la morfología de los núcleos celulares para clasificar cáncer de mama. La textura se mide como la desviación estándar de los píxeles del núcleo frente a la media de la base de datos, por lo que ambas variables estarán en rangos de reales positivos.

Una vez más, queremos distinguir entre *Benigno* (B) y *Maligno* (M). Supongamos que el clasificador de Bayes responde a un clasificador gaussiano de matrices de covarianzas independientes con la siguiente configuración:

$$\begin{aligned} p(y=B) &= 0{,}6274 \\ \mu_{\mathbf{B}} &= \begin{pmatrix} 17{,}9148 \\ 462{,}7902 \end{pmatrix} \\ \mathbf{\Sigma_B} &= \begin{pmatrix} 15{,}9610 & -20{,}9526 \\ -20,9526 & 18033,0301 \end{pmatrix} \\ p(y=M) &= 0{,}3726 \\ \mu_{\mathbf{M}} &= \begin{pmatrix} 21{,}6049 \\ 978{,}3764 \end{pmatrix} \\ \mathbf{\Sigma_M} &= \begin{pmatrix} 14{,}2843 & 144{,}2469 \\ 144{,}2469 & 135378{,}3553 \end{pmatrix} \\ \hat{y}^* &\leftarrow \arg\max_{y \in \mathcal{Y}} p(y|x), \end{aligned}$$

y con error de generalización $p(error) = 0{,}0974$, si bien tanto el clasificador de Bayes como su error quedan desconocidos para el diseñador.

Supongamos que se dispone de una muestra de 200 casos para el entrenamiento de cuatro métodos de clasificación diferentes: clasificador gaussiano (con matrices de covarianza independientes por clase), clasificador gaussiano con matriz de covarianza común a todas las clases, red neuronal artificial (perceptrón de dos capas ocultas de 30 y 15 neuronas cada una) y clasificador por el vecino más próximo. Calculamos el error de entrenamiento con este mismo conjunto de 200 muestras de entrenamiento, esperando obtener una estimación del error de generalización.

Si este experimento lo repetimos un número moderado de veces (por ejemplo, 50 repeticiones) obtendremos los resultados parecidos a los diagramas caja-bigotes con la leyenda "entrenamiento" de la figura 14.2. Observamos que el clasificador gaussiano obtiene un error de entrenamiento $error_T$ de 0,095 de mediana, que se aproxima bastante el error de Bayes. El clasificador gaussiano con covarianza común tiene una mediana de 0,12, que podría indicarnos que tiene un rendimiento algo menor que el clasificador gaussiano con matrices independientes. La red neuronal tiene una mediana de 0,08 y su distribución se sitúa en los valores inferiores del rango 0-1. Esta red neuronal puede considerarse un modelo complejo para el problema a resolver, y además ha sido entrenada sin ningún método de regularización o de parada temprana, por lo que podemos sospechar que se ha producido sobreentrenamiento y que el error de generalización está subestimado al calcularlo como el error de entrenamiento. Por último, el clasificador del vecino más próximo, que es un clasificador local basado en diccionario de datos obtiene un error de entrenamiento igual a 0, debido a la propia metodología. Evidentemente, este error no es correcto, y deberemos disponer de métodos alternativos de estimación del error de generalización que sean más informativos.

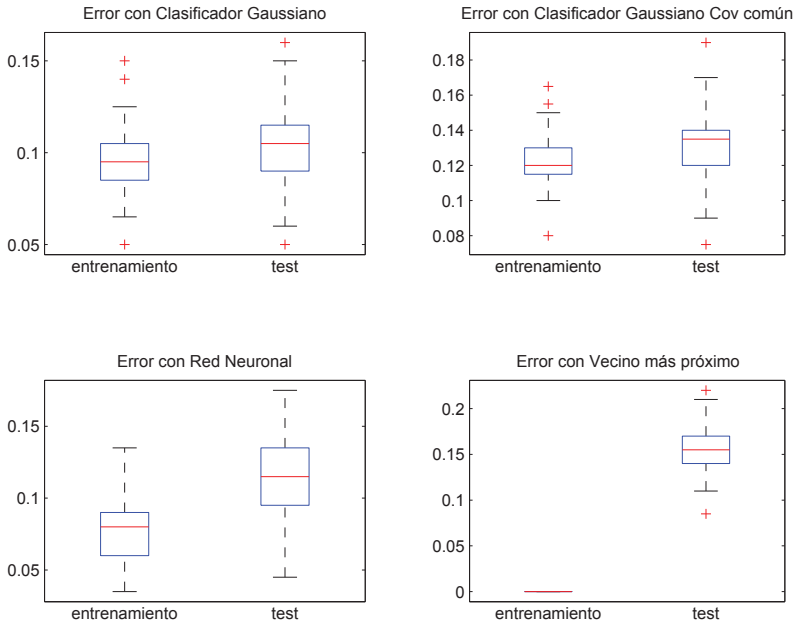


Figura 14.2: Errores de clasificación de cáncer de mama mediante Textura y Área.

Como hemos visto, el error de entrenamiento no nos asegura una estimación fiable del error de generalización. Sin embargo, nuestras tareas de evaluación y selección de modelos requieren disponer de una estimación útil. La solución obvia, pero no siempre factible, es estimar el error de generalización mediante nuevas muestras que no hayan sido utilizadas para el entrenamiento de los modelos. Así pues, definimos:

- el *conjunto de test*, $\mathcal{S}_t$, como el conjunto de muestras utilizado para evaluar el modelo de decisión mediante la ecuación (14.1), obteniendo el *error de test*;
- el *conjunto de validación*, $\mathcal{S}_v$, como el conjunto de muestras para seleccionar el modelo de decisión de un conjunto de modelos posible.

Ejemplo 14.3 (Clasificación de cáncer de mama por la morf. celular (cont 14.2)) La evaluación de los clasificadores entrenados en el ejemplo 14.2 mediante 200 casos nuevos de test obtendría unos resultados similares a los expresados por los diagramas caja-bigotes con la leyenda "test" de la figura 14.2. Podemos observar que el clasificador gaussiano obtiene un error de test de mediana 0,105, lo que supone una ligera sobreestimación del error de generalización, pero sigue siendo muy similar al error real y al estimado mediante el conjunto de entrenamiento. Cuando testeamos el clasificador gaussiano con matriz de covarianza común, la mediana obtenida es 0,135, lo que indica que este modelo tiene un rendimiento algo inferior que el anterior. La evaluación con test independiente de la red neuronal pasa ahora a distribuirse con una mediana de 0,115, que reafirma la subestimación del error calculado mediante las muestras de entrenamiento. Ahora conseguimos una estimación del error del clasificador del vecino más próximo, que resulta tener un rendimiento relativamente bajo respecto a los otros métodos, estando en valores entorno a 0,155.

La situación donde disponemos de un conjunto relativamente alto de muestras para dividirlas en tres bloques: *entrenamiento*, *validación* y *test* es idílica y poco realista actualmente en medicina. Por otra parte, como decíamos, nos interesa aumentar el conjunto de entrenamiento al máximo posible, por lo que buscaremos estrategias de entrenamiento, validación y evaluación que nos permitan aprender los modelos y calcular su error de generalización con un uso eficiente de las muestras disponibles.

14.2.1. Intervalo de confianza del error

En general, la estimación empírica del rendimiento de un clasificador mediante un valor puntual, por ejemplo el error de test, no es informativa si no está acompañado de un intervalo de confianza donde acotemos el rendimiento real, por ejemplo mediante el error de generalización. Concretamente, acompañar nuestros resultados con sus intervalos de confianza es de gran importancia cuando el número de casos de evaluación es pequeño, ya que la precisión de nuestra estimación es menor.

Podremos estimar el intervalo de confianza del error de generalización $p(error)$ de un clasificador mediante la evaluación con N muestras donde se han observado k errores, suponiendo que k tiene una distribución binomial (ver sección 3.2.2) y que $N \to \infty$, el error de test será $\widehat{err} = k/N$, por lo que se puede estimar que el error verdadero $p(error)$ estará contenido en el intervalo,

$$I_{95\,\%}(p(error)) = [\widehat{err} - zs, \widehat{err} + zs], \tag{14.3}$$

con un 95 % de probabilidad, siendo $z = 1{,}96$ y, el error estándar de la distribución binomial, $s = \sqrt{\widehat{err}(1-\widehat{err})/N}$ [d].

Ejemplo 14.4 (Cálculo del tamaño muestral de test)
Un grupo multidisciplinar en radiología está interesado en evaluar un clasificador basado en hallazgos de imagen para tumores de partes blandas del que esperan un $p(error) = 0{,}1$. Para planificar la recopilación de muestras, se plantean cuál es el número de muestras de test que necesitan para tener una confianza al 95 % de que el resultado estará dentro de un intervalo del $\pm i = \pm 0{,}03$ alrededor del error esperado.

Para ello, asumiendo los intervalos de confianza del modelo binomial de la ecuación 14.3, podemos calcular el número mínimo de muestras necesarias como

$$N = \lceil \frac{z^2 p(error)(1-p(error))}{i^2} \rceil. \tag{14.4}$$

Por lo que resuelven que el número de casos adquiridos para test debe ser como mínimo $N \geq 385$.

Kohavi en [141] propone una aproximación normal a la binomial y obtiene una alternativa más acertada al intervalo de confianza,

$$I^K_{95\,\%}(p(error)) = [\widehat{err} + \frac{(1-2\widehat{err})z^2}{2(N+z^2)} \pm zs], \tag{14.5}$$

donde

$$s = \sqrt{\frac{\widehat{err}(1-\widehat{err})}{N} + \frac{z^2}{[2(N+z^2)]^2}[1 - 4\widehat{err}(1-\widehat{err})(2+\frac{z^2}{N})]}.$$

La figura 14.3 muestra los intervalos de confianza obtenidos mediante la aproximación de Kohavi (14.5) variando el número N de muestras y el error de test $\widehat{err}$.

[d] Cuando N es pequeño, suele sumarse $0{,}5/N$ en ambos límites del intervalo.

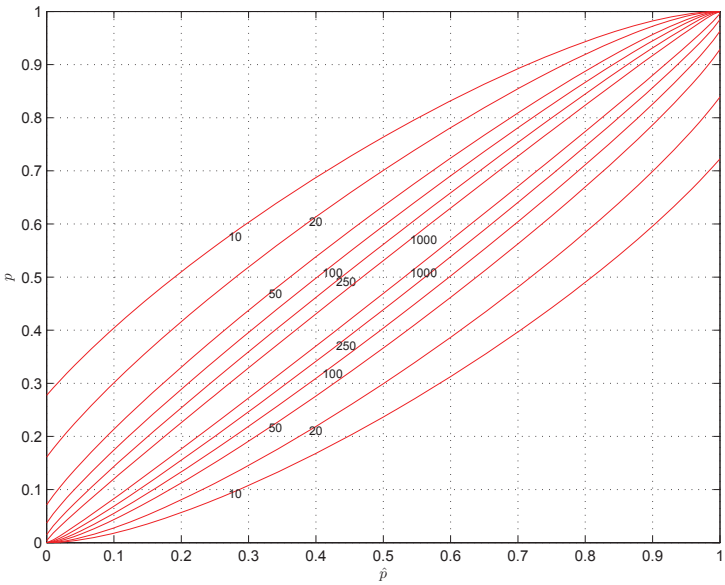


Figura 14.3: Intervalos de confianza al 95 % del error de generalización $p(error)$ para distintos valores de error de test $\widehat{err}$ y distintos tamaños N de la muestra.

Las aproximaciones anteriores son válidas cuando el tamaño de la muestra N es relativamente grande. La aproximación binomial asume que $N \to \infty$ y la propuesta de Kohavi se considera válida cuando $N\widehat{err}(1-\widehat{err}) \geq 5$ [142], por lo que para muestras pequeñas, o pequeños errores de test, estas propuestas no son las más adecuadas. Una alternativa útil cuando $10 \leq N \leq 200$ y $0 \leq k \leq \frac{1}{2}N$ es la estimación bayesiana de los intervalos de credibilidad [143, 144], que asumen una distribución binomial de los errores y ausencia de conocimiento del problema específico. Se define como

$$c_{95\,\%}(p(error)) = \big[\widehat{err} + \frac{2(N-2k)z\sqrt{0{,}5}}{2N(N+3)} \pm \sqrt{\frac{\widehat{err}(1-\widehat{err})}{N+2{,}5}}\big]. \tag{14.6}$$

Bajo esta aproximación, conforme aumenta la N, los intervalos de credibilidad se estrechan, tal como muestra la figura 14.4. Sin embargo, es interesante observar la corrección de las estimaciones de valores extremos, es decir, con resultados totalmente erróneos o perfectos, obtenidos con un número bajo de muestras.

Ejemplo 14.5 (Clasificación de cáncer de mama por la morf. celular (cont 14.3)) Supongamos ahora que se disponen de $N = 100$ muestras para realizar evaluar el clasificador obtenido en el ejemplo 14.3, y de ellas, $k = 10$ son errores de predicción.

El error de test será $\frac{10}{100} = 0{,}1$, y como el conjunto de test se considera pequeño, ya que $100 \cdot 0{,}1 \cdot 0{,}9 = 4{,}75 < 5$, calcularemos el intervalo de credibilidad del error de generalización mediante la ecuación (14.6), por lo que $c_{95\,\%}(p(error)) = [0{,}0527, 0{,}1688]$.

En resumen, si disponemos de un conjunto de N muestras suficientemente grande, podemos realizar el entrenamiento y selección de nuestro predictor y una estimación honesta de su error de generalización separando aleatoriamente las muestras en tres conjuntos: conjunto de entrenamiento, conjunto de validación y conjunto de test. Este método se denomina en inglés *holdout* o *partition*.

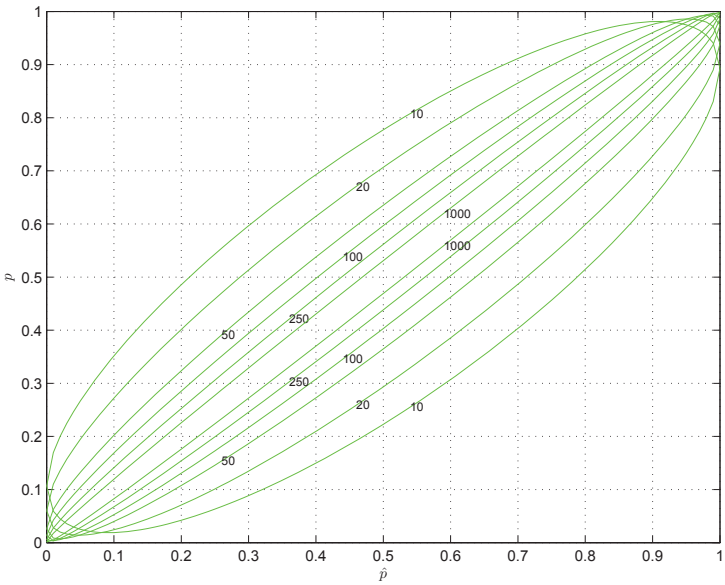


Figura 14.4: Intervalos de credibilidad al 95 % del error de generalización $p(error)$ para distintos valores de error de test $\widehat{err}$ y distintos tamaños N de la muestra.

Como resumen, si tenemos pocas muestras, *holdout* será pesimista, ya que solo una parte de los datos son usados para el entrenamiento del modelo, lo que perjudicará el modelo evaluado respecto al modelo que obtendríamos con todas las muestras.

Cuando repetimos k veces esta partición y promediamos los resultados el método se denomina *repeated holdout*. Por ejemplo, dejar 30 % de muestras aleatorias para test, y repetir 200 veces). Esta estimación poco sesgada (depende del ejecto del número de muestras usadas para entrenar en cada repetición), pero será muy variable por similitud de conjuntos de test. Cuando utilizamos esta estimación, no es correcto calcular la desviación estándar de la media muestral ya que las repeticiones no son independientes al compartir casos.

14.3. Estimación por remuestreo del error de generalización

Generalmente, dispondremos de un conjunto de muestras $\mathcal{S}$ de tamaño N para diseñar el modelo predictivo, incluyendo las etapas de entrenamiento, selección de modelos y evaluación del modelo final. Como hemos visto, es deseable utilizar el máximo número de muestras para ajustar el modelo, pero al mismo tiempo, una evaluación con un alto número de muestras hará más precisa la estimación del riesgo de nuestro modelo. Para conseguirlo, los experimentos de aprendizaje automático suelen utilizar técnicas de remuestreo, como *validación cruzada* o *bootstrap*, mediante las que reutilizaremos las muestras para entrenar los modelos de predicción y para estimar honestamente el error de generalización.

Debemos tener en cuenta que:

- Utilizar una muestra para testear un modelo que se ha entrenado con ella SESGA la estimación del error subestima el error

- Utilizar menos muestras de las disponibles SESGA la estimación, porque sobrestima el error al no conseguir el mejor modelo posible mediante las muestras disponibles
- Utilizar conjuntos similares de entrenamiento y test aumenta la VARIANZA de la estimación, ya que no sabremos que pasaría con conjuntos diferentes de muestras

14.3.1. Validación cruzada

En la validación cruzada con K bloques, el conjunto $\mathcal{S} = \{(x_i, y_i)\}$ de N muestras se divide en K conjuntos mutuamente excluyentes, $\mathcal{S}_1, \ldots, \mathcal{S}_K$, de aproximadamente el mismo tamaño y que se usarán como conjuntos de evaluación de forma iterativa. El predictor se entrena y evalúa K veces. En cada iteración $k \in 1, \ldots, K$, se entrena un modelo $f^{(-k)}$ con las muestras $\mathcal{S} \backslash \mathcal{S}_k$, y se evalúa con el conjuntos $\mathcal{S}_k$, obteniendo para cada muestra i del conjunto $\mathcal{S}_k$ el resultado $f^{(-k)}(x_i)$. El error de generalización puede estimarse mediante el *error de validación cruzada* como,

$$\widehat{err}_{CV} = \frac{1}{N} \sum_{i=1}^{N} L[y_i, f^{(-k)}(x_i)]. \tag{14.7}$$

Si el modelo predictivo es estable ante los cambios producidos por el borrado de casos en las particiones, entonces la estimación por validación cruzada no estará sesgada y la varianza del estimador será aproximadamente $\widehat{err}_{CV}(1 - \widehat{err}_{CV})/N$, por lo podremos pensar que hemos obtenido una estimación realista del rendimiento del modelo.

La elección del tamaño K de la partición suele ser decisión del diseñador. Cuando $K = N$, el método se denomina *leave-one-out* y es el que maximiza el número de muestras utilizadas para el entrenamiento del modelo de cada iteración. Por lo tanto, podremos pensar que la estimación por *leave-one-out* tendrá un sesgo pequeño. Sin embargo, como todos los modelos han sido entrenados con conjuntos muy similares, no estaremos seguros de la precisión de la estimación ante cambios en las muestras, por lo que el estimador puede tener una varianza considerable. Disminuyendo el valor K obtendremos conjuntos de evaluación más heterogéneos, bajando la varianza de la estimación. Sin embargo, el tamaño de los conjuntos de entrenamiento también disminuye, por lo que podremos obtener modelos más sesgados. El valor óptimo de K será aquel que obtenga un equilibrio entre el sesgo y la varianza del estimador. Mediante estudios empíricos se ha observado que $K = 10$ suele ser una buena elección para problemas reales de clasificación [141].

Como resumen podemos decir que:

- La estimación es poco sesgada, si se usa un K que deje bastantes muestras para entrenar
- La estimación es muy variable, porque se entrena con conjuntos similares, por lo que los modelos son similares (sin embargo los conjuntos de test son diferentes, lo que mejora)

Ejemplo 14.6 (Clasificación de cáncer de mama por la morf. celular (cont 14.2)) Entrenamos de nuevo los clasificadores gaussiano y red neuronal para solucionar la clasificación de cáncer de mama mediante la textura y el área de los núcleos celulares. Para mostrar la diferencia de comportamiento de las evaluaciones, dejaremos sobreentrenar la red neuronal. En esta ocasión únicamente disponemos de 100 muestras de entrenamiento y estamos interesados en observar la estabilidad de los modelos ante la variación de las muestras utilizadas.

La figura 14.5 muestra los resultados obtenidos en 50 repeticiones variando el número de particiones de la validación cruzada, desde 2 hasta N. En la figura observamos que ambos modelos no alcanzan la clasificación de Bayes, pero de forma comparativa, el clasificador gaussiano tiene menos sesgo y es más estable ante los cambios de K. También observamos una ligera sobreestimación de su error cuando se evalúa mediante menos de 10 particiones. El sobreentrenamiento

de la red neuronal se observa claramente con el aumento del error conforme aumenta K, lo que refleja la poca estabilidad del clasificador.

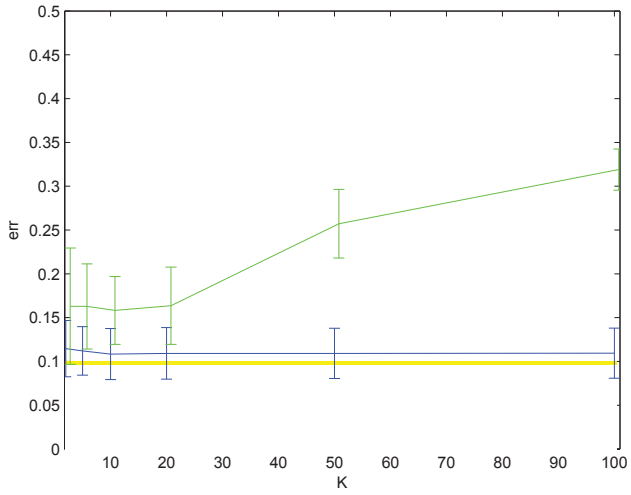


Figura 14.5: Evaluación mediante validación cruzada con 100 muestras y variando el número de particiones K de los clasificadores basados en un modelo gaussiano con matrices de covarianza independientes (azul) y la red neuronal 30x15 sobreentrenada (verde). Cada valor corresponde a la media de 50 evaluaciones, y se han añadido los intervalos de confianza. El banda amarilla corresponde al intervalo al 95 % del clasificador de Bayes.

14.3.2. Bootstrap

Otro método de estimación basado en el remuestreo es *bootstrap*. En *bootstrap* se realizan B iteraciones sobre un conjunto $\mathcal{S}$ de N muestras. En cada iteración b se creará un conjunto $\mathcal{S}_{(b)}$ de entrenamiento de N muestras, mediante muestreo con repetición siguiendo una distribución uniforme. El resto de muestras formarán el conjunto $\mathcal{S}_{(-b)}$ de test de la iteración b. La probabilidad que tiene una muestra de no aparecer en el conjunto de entrenamiento es de $(1 - \frac{1}{N})^N$. Cuando $N \to \infty$, entonces dicha probabilidad se aproxima a $e^{-1} \approx 0{,}368$. Esto implica que la probabilidad de aparecer en el conjunto de entrenamiento es aproximadamente 0,632. Para cada repetición b, calcularemos el error de entrenamiento $\widehat{err}_T^{(b)}$ y el error de test $\widehat{err}_t^{(-b)}$ mediante la ecuación (14.1) y calcularemos el estimador ,632-*bootstrap* como

$$\widehat{err}_{,632boot} = \frac{1}{B}\sum_{i=1}^{B} 0{,}632\widehat{err}_t^{(b)} + 0{,}368\widehat{err}_T^{(-b)}. \tag{14.8}$$

Bootstrap permite aproximar la varianza del error estimado a través de las repeticiones realizadas. La estimación del error por ,632-*bootstrap* compensa el sesgo que $\widehat{err}_t^{(b)}$tiene debido al uso del únicamente el 63,2 % de las muestras para entrenamiento. Para ello se pondera el error de test con el error de entrenamiento, o una estimación por *leave-one-out*. Para clasificadores locales, como el *vecino más próximo*, donde el error de entrenamiento subestima el error de generalización, *bootstrap* claramente sesga la estimación del error de generalización, sin embargo, para modelos

sencillos o regulados obtiene estimaciones insesgadas y, por lo tanto, fiable. Además, permite el cálculo de la varianza de la estimación.

Ejemplo 14.7 (Clasificación de cáncer de mama por la morf. celular (cont 14.6)) Cambiamos ahora de método de evaluación y utilizamos *bootstrap* variando el número de repeticiones B desde 2 hasta 100, manteniendo los métodos de clasificación, 50 repeticiones de los experimentos y disponiendo de 100 muestras para nuestros experimentos. En la figura 14.6 observamos que el error estimado para modelo gaussiano se acerca más al clasificador de Bayes que la solución evaluada mediante validación cruzada. Ya que el modelo gaussiano no es un modelo local, que podría hacernos sospechar de una subestimación de su error, podemos suponer que *bootstrap* obtiene una estimación de bajo sesgo. El comportamiento de la estimación del error de la red neuronal es el esperado de un modelo complejo con sobreentrenamiento.

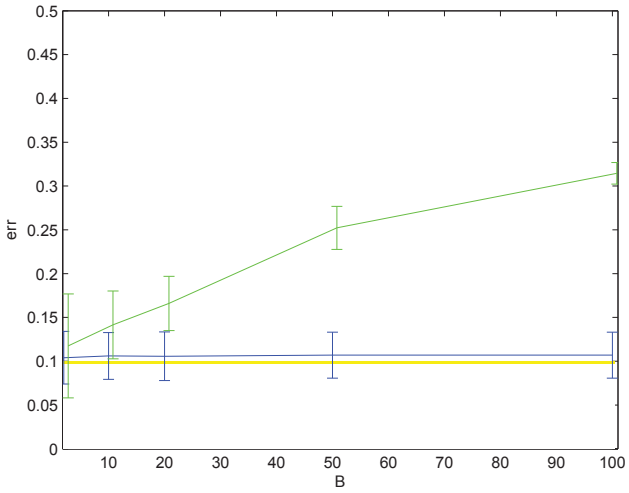


Figura 14.6: Evaluación mediante *bootstrap*, con 100 muestras y variando el número de repeticiones B, de los clasificadores basados en un modelo gaussiano con matrices de covarianza independientes (azul) y la red neuronal 30x15 sobreentrenada (verde). Cada valor corresponde a la media de 50 evaluaciones, y se han añadido los intervalos de confianza. El banda amarilla corresponde al intervalo al 95 % del clasificador de Bayes.

Ejemplo 14.8 (Valoración en screening del cáncer de mama) La mamografía es actualmente el método de screening más efectivo para cáncer de mama. Las tendencias actuales en ayuda a la decisión al screening de mama van por graduación BI-RADS según los hallazgos en mamografía para predecir la severidad (benignidad o malignidad) de las masas observadas.

Vamos a utilizar las variables *edad* y *densidad de la masa* de la base de datos "Mammographic Mass Data" [145] para entrenar y evaluar un clasificador gaussiano con matriz de covarianza común sobre la severidad de las lesiones: *benigna* o *maligna*. Trataremos las dos variables como si fueran continuas, si bien la variable *densidad de la masa* es ordinal, siendo 1 alta densidad, 2 media, 3 baja y el valor 4 indica que la masa contiene grasa. Para realizar nuestro estudio, disponemos de un total de 516 casos benignos y 445 casos malignos, pero más que obtener un clasificador de bajo error, nos interesa observar el comportamiento de los métodos de estimación

del error por remuestreo, validación cruzada y *bootstrap*. Por ello, hemos estimado el error de generalización mediante 500 muestras extraídas aleatoriamente y hemos repetido el proceso 500 veces, obteniendo un valor medio con el cual hemos calculado los intervalos de confianza del error de generalización que están representados mediante la banda amarilla de la figura 14.7. A continuación, hemos observado el error estimado mediante validación cruzada, variando K y mediante *bootstrap*, variando B, utilizando únicamente 300 muestras. Se han obtenido los resultados mostrados en la figura mediante las series azul y rojo, respectivamente. Observamos que en ambos casos tenemos resultados estables, con ligeras perturbaciones con valores muy bajos de K y B. Observamos que validación cruzada tiende a sobreestimar el error más que *bootstrap*, suponiendo que este último método funciona correctamente para modelos simples como este.

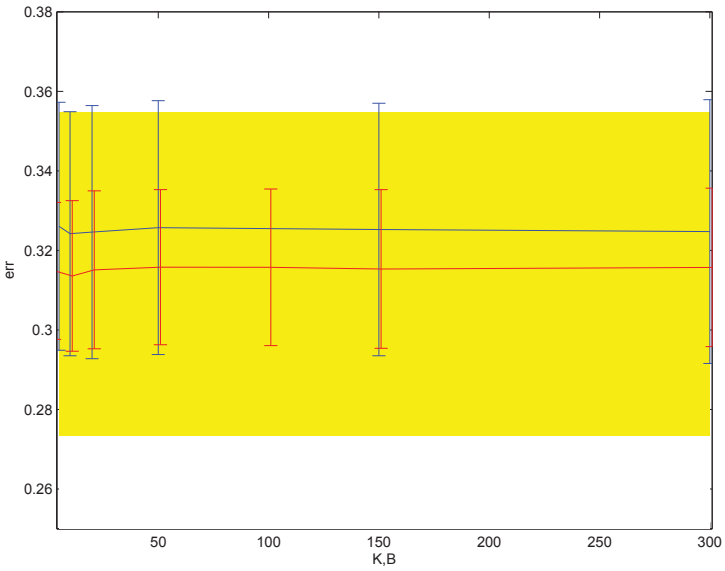


Figura 14.7: Evaluación por validación cruzada (azul) y *bootstrap* (rojo) de un clasificador gaussiano con matriz de covarianza común sobre la severidad de las lesiones: *benigna* o *maligna*, preparado con 300 muestras aleatorias (repitiendo el experimento 50 veces). La banda amarilla indica el intervalo al 95 % de confianza del error de generalización estimado con 500 muestras aleatorias, repetido 500 veces.

14.4. Selección de modelos

Suele ser común que durante el diseño del predictor probemos varios modelos, bien porque dispongamos de diversas metodologías, bien porque nuestra metodología requiera de la elección de algún parámetro por parte del diseñador.

Evidentemente, dado el conjunto de alternativas, queremos saber cuál de ellas es la mejor para nuestro propósito de obtener el predictor de menor riesgo. Así pues, en esta tarea, estamos más interesados en comparar los modelos entre sí que en hacer una estimación precisa del riesgo.

La primera alternativa ya la hemos visto en la sección 14.2 y consiste en estimar el error mediante un conjunto de validación (*error de validación*) distinto del conjunto de entrenamiento y del conjunto de test. Seleccionaremos aquel modelo con menor error de validación.

La segunda alternativa cuando queremos utilizar una estrategia de selección basada en validación pero no disponemos de suficientes muestras, es la *selección de modelos por error de validación con remuestreo*. En este caso, podemos utilizar los métodos de remuestreo estudiados en la sección 14.3 para estimar el error de validación. Sin embargo, este método de selección de modelos debe utilizarse adecuadamente cuando se necesita evaluar el modelo final y disponemos de muestras limitadas. Para obtener un resultado honesto, deberemos reservar un conjunto de test, con el que obtener una evaluación del modelo final, y utilizar el resto para entrenar y seleccionar el modelo mediante remuestreo. Si necesitamos aumentar las muestras de evaluación, será necesario anidar los bucles de remuestreo para la selección de modelos y para evaluación. Es habitual encontrar resultados publicados con el error del mejor modelo obtenido en la selección de modelos, en lugar del error de test, que sería el resultado honesto del experimento computacional.

La tercera alternativa consiste en el cálculo de indicadores basados en el comportamiento del modelo sobre el conjunto de entrenamiento y la complejidad del modelo. Como hemos visto, el optimismo al calcular el rendimiento de un modelo sobre el conjunto de entrenamiento suele incrementarse con la complejidad del modelo, por lo que un indicador que compense este optimismo es interesante para seleccionar un modelo con el sesgo y la varianza equilibrados. Un indicador bastante empleado es el indicador AIC (Akaike's information criterion),

$$AIC = -\frac{2}{N} loglik + 2\frac{d}{N},$$

donde el primer sumando de AIC hace referencia al rendimiento del modelo sobre el conjunto de entrenamiento, mientras que el segundo término indica la dimensión relativa del modelo. Este indicador está definido mediante la log-verosimilitud[e] ($loglik$) sobre el conjunto de entrenamiento, que para un problema de clasificación automática puede calcularse como $loglik = \sum_{i=1}^{N} \log p(y_i|x_i)$. La dimensión d, para modelos no regulados, será el número de parámetros a estimar en el modelo.

Ejemplo 14.9 (Clasificación de cáncer de mama por la morf. celular (cont 14.2)) Entrenamos de nuevo nuestros tres modelos para la clasificación de benigno/maligno con 100 muestras de morfología celular. Suponemos ahora que no disponemos de muestras de validación, y queremos establecer un ranking de preferencia entre los modelos, disponiendo del rendimiento de los modelos conjunto de entrenamiento, por lo que utilizamos el indicador AIC. La figura 14.8 muestra el error de entrenamiento ($\widehat{err}_T$), AIC y error de test ($\widehat{err}_t$) para los modelos gaussiano con matrices de covarianza independientes (cuadrático), gaussiano con matriz de covarianza común (lineal) y la red neuronal 30x15 (RN). Si bien la red neuronal consigue el mejor error de entrenamiento (mediana de 0.08), su indicador AIC se penaliza por el término de complejidad. El modelo cuadrático obtiene el mejor AIC, que efectivamente corresponde al modelo con menor error de test.

Existen otros indicadores similares, como BIC (*Bayesian information criterion*), definido a través de la comparación bayesiana de modelos [146], o la dimensión de Vapnik-Chervonenkis (*VC-dimension*).

Por otro lado, si se desea validar la existencia de diferencias significativas en el rendimiento de un modelo predictivo respecto a otro, o respecto a otros métodos clínicos, debemos diseñar un proyecto de estudio de valor añadido sostenido sobre un test estadístico, como los estudiados en el capítulo 19.

[e] Es importante observar que AIC no se define mediante la función de pérdida 0-1 u otras.

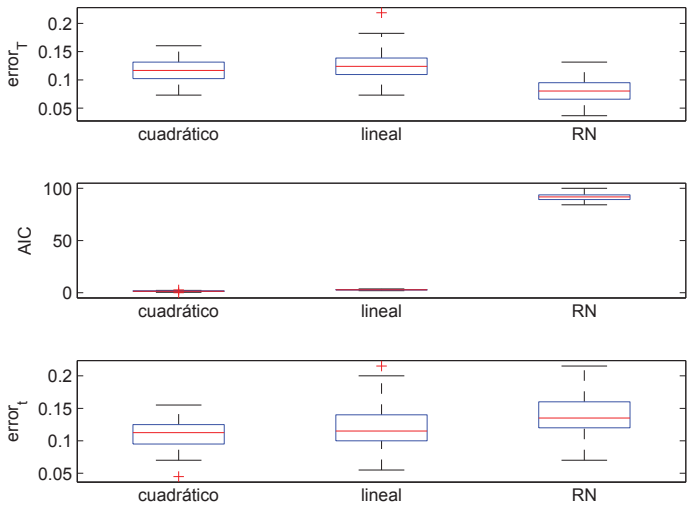


Figura 14.8: *AIC* en selección de modelos. Si bien la red neuronal tiene el error de entrenamiento más bajo, se ve penalizada por la complejidad del modelo. El modelo gaussiano de matrices de covarianza independientes obtiene el *AIC* más bajo, gracias a una log-verosimilitud mayor que el modelo con matriz de covarianza común, pese a tener más parámetros que estimar. La penalización sobre los modelos complejos puede llegar a resultar excesiva.

14.5. Métricas de evaluación

En la sección 3.3 definimos la función de pérdida (o utilidad) esperada (riesgo) como objetivo a optimizar por nuestros decisores. Estas funciones son las métricas de evaluación más generales que podemos utilizar para evaluar nuestros modelos predictivos.

En las secciones anteriores se ha utilizado el error de generalización a partir de la función de pérdida para el problema de clasificación automática. Sin embargo, en la sección 8.2.1 vimos que para ciertos problemas puede ser interesante observar (y decidir) el riesgo de tomar decisiones sobre problemas donde las equivocaciones tienen consecuencias distintas. Además, para problemas donde alguna de las clases es poco prevalente respecto al resto, el error de generalización puede ser poco informativo del rendimiento real del predictor. A continuación, revisaremos las métricas de evaluación más utilizadas para clasificadores.

Por simplicidad en la notación, pero sin pérdida de generalidad, supondremos un problema de clasificación de dos clases, $\{y_1, y_2\}$. Diremos que y_1 es la clase positiva, mientras que y_2 será la clase negativa. Si evaluamos un clasificador para estas dos clases con un conjunto de N casos, definiremos la *matriz de confusión* de la evaluación como:

	$\hat{y_1}$	$\hat{y_2}$
y_1	n_{11} (VP)	n_{12} (FN)
y_2	n_{21} (FP)	n_{22} (VN)

donde n_{11} es el número de muestras positivas clasificadas correctamente como positivas (Verdadero Positivo (VP)), y n_{21} es el número de muestras negativas clasificadas incorrectamente como positivas (Falso Positivo (FP)). De forma similar, n_{22} es el número de muestras negativas clasificadas como negativas (Verdadero Negativo (VN)), y n_{12} es el número de muestras positivas clasificadas incorrectamente como negativas (Falso Negativo (FN)). Los FP también suelen denominarse *errores de tipo I*, y los FN se denominan *errores de tipo II*.

Las métricas de evaluación típicamente utilizadas son el $\widehat{err}$ y el $\widehat{acc}$ (acierto o accuracy). El $\widehat{err}$, ya definido mediante la ecuación 14.1 como error empírico, que para nuestra notación para dos clases será

$$\widehat{err} = \frac{n_{12} + n_{21}}{N}. \tag{14.9}$$

De manera opuesta, el $\widehat{acc}$ se define como

$$\widehat{acc} = \frac{n_{11} + n_{22}}{N}. \tag{14.10}$$

Si fijamos como referencia la clase y_1, se define Recall (R) como

$$R_1 = \frac{n_{11}}{n_{11} + n_{12}}, \tag{14.11}$$

que nos indica la tasa de acierto del predictor para la clase y_1. También definimos Precision (P) como

$$P_1 = \frac{n_{11}}{n_{11} + n_{21}}, \tag{14.12}$$

para indicar la tasa de aciertos cuando se predice como resultado $\hat{y}_1$. De forma similar, podrían definirse el recall y la precisión para la clase 2. En medicina, se suele llamar *sensibilidad* al recall de la clase positiva (y_1 en nuestra notación), y se llama *especificidad* al recall de la clase negativa.

Basándonos en el recall y la precisión podemos definir otras métricas de evaluación alternativas, que enfatizan el comportamiento de los predictores en las clases del problema.

Definimos Geometric Mean of Recalls (GMOR) como

$$GMOR = \sqrt{R_1 R_2}, \tag{14.13}$$

para dos clases. De forma genérica, la GMOR se define como

$$GMOR = \sqrt[c]{\prod_{i=1}^{c} R_i}, \tag{14.14}$$

La GMOR resulta de gran interés cuando las clases del problema son de diferente prevalencia, por ejemplo, cuando estamos abordando problemas de *screening*, clasificaciones de diagnósticos raros, o clasificaciones de un diagnóstico frente a la agrupación de todos los demás diagnósticos. Como podemos observar en la figura 14.9, valores altos de accuracy (acierto) (acc), no tienen porque deberse a una alta tasa de acierto para ambas clases, sobretodo, cuando las muestras de una clase son muchas menos que las de la otra clase. Ante esta situación, el uso de GMOR puede ser mucho más realista que acc para informar de los resultados.

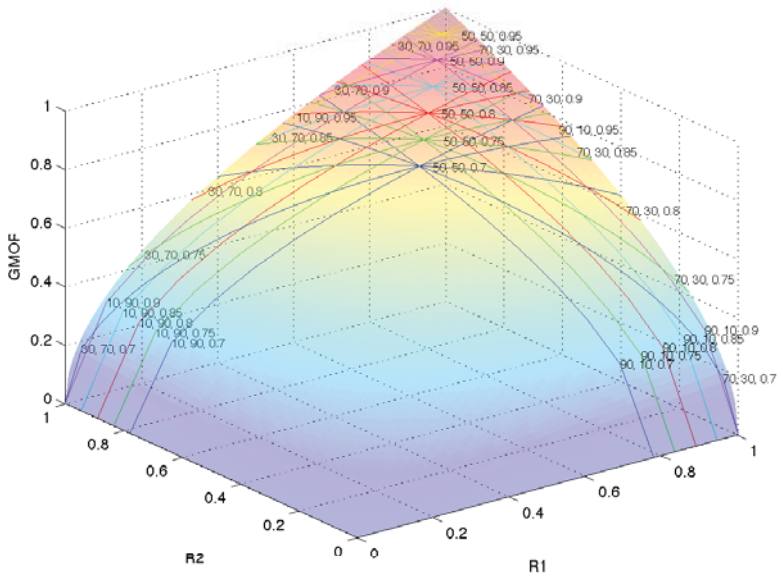


Figura 14.9: Representación en forma de superficie de la métrica de evaluación GMOR en función de los recall R_1 y R_2. Se ha representado mediante líneas azules los posibles valores GMOR que podría tener un clasificador con tasa de acierto 0.7 evaluado con 100 muestras de test distribuidas entre las clases y_1 e y_2 como [10,90], [30,70], [50,50], [70,30] y [90,10]. De forma similar se representa con líneas verdes los valores GMOR para un acierto de 0.75, rojas para 0.8, cyan para 0.85, magenta para 0.9 y amarillo para 0.95.

Similar a GMOR, se define el Balanced Accuracy Rate (BAR) como

$$BAR = \frac{R_1 + R_2}{2}. \tag{14.15}$$

En la figura 14.10 se observa que el BAR es menos pesimista que GMOR cuando uno de los recall es bajo y otro es alto, debido a su comportamiento lineal. Sin embargo, podemos observar que las líneas de posibles valores que puede tomar dado un conjunto de test es similar a GMOR. De forma similar a BAR, podemos definir Balanced Error Rate (BER) utilizando las tasas de error por clase en lugar de las tasas de acierto por clase.

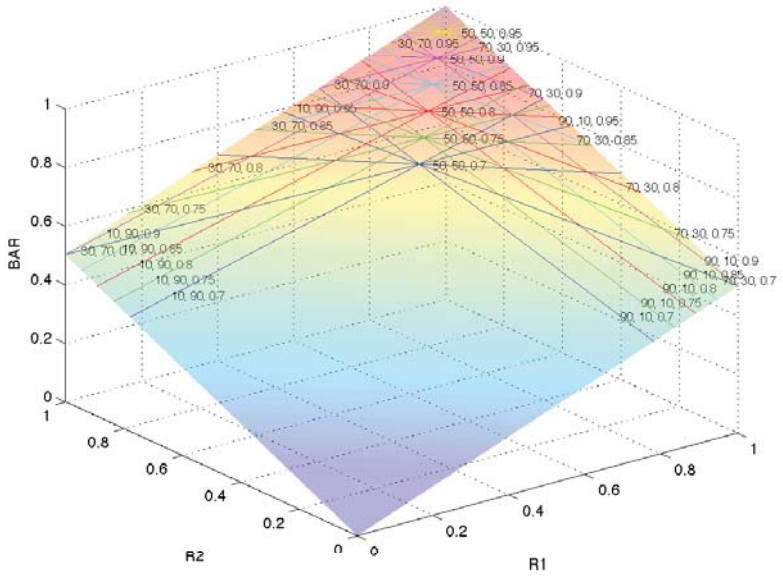


Figura 14.10: Representación en forma de superficie de la métrica de evaluación BAR en función de los recall R_1 y R_2. Se ha representado mediante líneas azules los posibles valores BAR que podría tener un clasificador con tasa de acierto 0.7 evaluado con 100 muestras de test distribuidas entre las clases y_1 e y_2 como [10,90], [30,70], [50,50], [70,30] y [90,10]. De forma similar se representa con líneas verdes los valores GMOR para un acierto de 0.75, rojas para 0.8, cian para 0.85, magenta para 0.9 y amarillo para 0.95.

Ejemplo 14.10 (Comparación de métricas para tumores hepáticos)
Veamos un caso extremo para entender la utilidad de las métricas propuestas. Se dispone de 100 pacientes, a 5 se les ha diagnosticado un tumor en el hígado a partir de unos marcadores hepáticos, el resto no tienen tumor. Deseamos desarrollar unos modelos de clasificación de tumores hepáticos a partir de datos cuyas clases están muy desbalanceadas, la clase positiva está infrarrepresentada en comparación a la clase negativa. Debemos pensar qué métrica es la más adecuada. Por ejemplo, partiendo de un clasificador trivial que prediga siempre que el paciente no tiene tumor hepático. Tomando como métrica el acierto, un clasificador trivial tendrá de media un 95 % de acierto. Sin embargo, su sensibilidad será nula. Es decir, nunca acertará cuándo un paciente tiene un tumor hepático, precisamente, el objeto del modelo de clasificación. La solución opuesta, predecir siempre que el paciente sufre un tumor hepático, es igualmente poco deseable porque tendría un acierto del 5 % y una especificidad nula. Se tendría que tratar a todos los pacientes con su consiguiente coste económico y disminución de calidad de vida. La solución adecuada es fijar una métrica alternativa como la GMOR o el BAR que tienen en cuenta tanto sensibilidad como especificidad con el fin de encontrar una solución equilibrada entre ambas clases.

Por último, definimos Geometric Mean of Recall and Precision (GMRP) para la clase positiva como

$$GMRP_1 = \sqrt{R_1 P_1}, \tag{14.16}$$

que viene definida por el recall y la precisión de la clase. Como vemos en la figura 14.11, podemos destacar que GMRP detecta las bajas tasas de acierto en la clase positiva cuando esta es poco prevalente, si bien a coste de penalizar su valor absoluto respecto al acc.

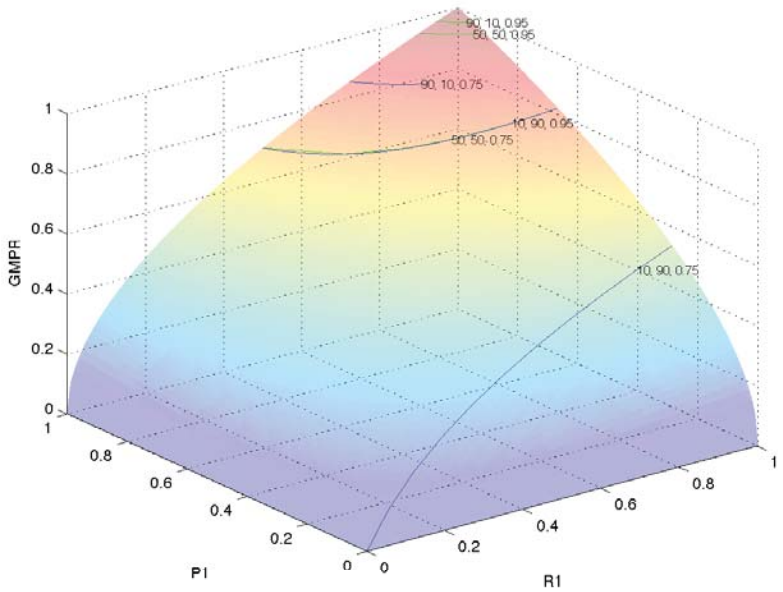


Figura 14.11: Representación en forma de superfície de la métrica de evaluación GMRP en función del recall R_1 y la precisión P_1 de la clase y_1. Se ha representado mediante líneas azules los posibles valores GMRP que podría tener un clasificador con tasa de acierto 0.75 evaluado con 100 muestras de test distribuidas entre las clases y_1 e y_2 como [10,90], [50,50], y [90,10]. De forma similar se representa con líneas verdes los valores GMRP para un acierto de 0.95.

14.5.1. Curva ROC: *Receiver Operating Characteristic*

Una de las métricas más empleadas a la hora de evaluar un modelo es la conocida curva ROC. La curva ROC recibe su nombre del inglés *Receiver Operating Characteristic* ya que originalmente se empleó en el análisis de señales de radar durante la segunda guerra mundial. La curva ROC representa gráficamente la sensibilidad de un modelo de clasificación de dos clases frente al valor complementario de la especificidad. Dicho de otro modo, compara la tasa de falsos positivos con la tasa de verdaderos positivos.

Si revisamos la *matriz de confusión*, la sensibilidad o tasa de verdaderos positivos representa el número de verdaderos positivos que el modelo ha clasificado correctamente. Esta información se representa en el eje Y del gráfico de la curva ROC. A su vez, la especificidad representa el número de verdaderos negativos que el modelo ha clasificado correctamente. Sin embargo, en el eje X de la curva ROC lo que representamos es 1-especificidad, es decir, la tasa de falsos positivos.

En la figura 14.12 podemos observar tres puntos que corresponden a tres modelos de clasificación distintos. El punto A representa un modelo prácticamente perfecto con una sensibilidad y especificidad del 100 %. El punto B representa un modelo con una sensibilidad del 90 % y una especificidad del 70 %. Al estar por encima de la diagonal representada por la función identidad mediante una línea discontinua se acepta como un modelo mejor que un modelo aleatorio. En cambio, el punto C representa un mal modelo de clasificación con una sensibilidad aceptable del 70 % pero con una especificidad muy baja del 20 %. Al estar bajo la línea diagonal se considera un modelo peor que el que conseguiríamos con un clasificador aleatorio. Otros puntos a tener en cuenta son: el punto (0,0), con sensibilidad nula y especificidad del 100 %, representa un modelo que siempre clasificaría las observaciones como pertenecientes a la clase negativa y por tanto siempre clasificaría correctamente los casos negativos, pero siempre fallaría al clasificar los casos positivos; el punto (1,1), que representa un modelo con 100 % de sensibilidad y especificidad nula, es el caso complementario en el que todos los casos positivos serían bien clasificados, pero ningún caso negativo se clasificaría correctamente; el punto (0,1) representaría el modelo perfecto como se explica con el punto A; por último, cualquier punto que caiga en la diagonal representaría un modelo aleatorio similar a clasificar lanzando una moneda al aire. Por ejemplo, el punto (0.5, 0,5) representa un modelo que clasificaría las clases positiva y negativa correctamente el 50 % de los casos. Otro ejemplo, el punto (0.9, 0.9) representaría un modelo que clasificaría correctamente el 90 % de los casos positivos, pero fallaría el 90 % de los casos negativos.

La curva ROC se crea a partir de los resultados de un modelo para un conjunto de observaciones que se emplean como validación o evaluación. El modelo debe clasificar cada observación y atribuir una probabilidad de pertenecer a la clase positiva. Ordenando las observaciones por su probabilidad estimada se pueden establecer puntos de corte donde los casos con probabilidad mayor que dicho punto de corte se consideran pertenecientes a la clase positiva y el resto se consideran pertenecientes a la clase negativa. Si se procede de esta manera de forma iterativa obtenemos n+1 puntos en la gráfica que formarán la curva ROC si se añaden los puntos (0,0) y (1,1), siendo n el número de observaciones para validar. Si la curva está por encima de la diagonal se considera que el modelo es mejor que el azar. La figura 14.13 representa varias curvas incluyendo la curva del azar que es la diagonal.

¿Pero cómo cuantificamos si el modelo es un poco mejor que el azar o mucho mejor que el azar? Para eso se dispone de una herramienta que resume la capacidad de clasificación del modelo: el área bajo la curva ROC, también conocida como AUC o estadístico C. Como se puede observar, el área bajo diagonal es un triángulo de área 0.5. A partir de aquí, el área de cualquier curva que esté por encima de la diagonal será mayor que 0.5. El área puede calcularse numéricamente con el método de los trapecios. Habitualmente, se considera que un modelo de clasificación con un área de 0.6 es un modelo pobre. A partir de ahí, un modelo con un AUC de 0.7 se considera un modelo aceptable, uno con un AUC de 0.8 se considera un buen modelo y uno con un AUC de 0.9 o mayor es considerado un modelo muy bueno o excelente. El AUC puede interpretarse probabilísticamente, ya que es equivalente a la probabilidad que tiene el modelo de clasificar una observación positiva escogida al azar por encima de una observación negativa escogida al azar. Es decir, si un modelo tiene un AUC del 90 %, esto quiere decir que, dadas una observación positiva aleatoria y una observación negativa aleatoria, en 9 de cada 10 casos la observación positiva tendrá una probabilidad de pertenecer a la clase positiva mayor que la observación negativa. Piense el lector qué ocurriría con un modelo cuyo AUC fuese del 10 %.

Cabe mencionar que el área bajo la curva ROC está íntimamente relacionada con el índice de Gini, ya que Gini + 1 = 2 AUC. En este caso, un AUC de 1 equivale a un índice de Gini de 1 y un AUC de 0.5 equivale a un índice de Gini de 0.

Finalmente, cuando se dispone de una curva ROC, es posible escoger el punto de corte del modelo que nos dará una combinación de sensibilidad y especificidad óptima. Este corte depende-

rá de la naturaleza del problema. Por ejemplo, es posible que en un programa de cribado sea más importante tener una alta especificidad sin comprometer especialmente la sensibilidad mientras que en un problema de ayuda al diagnóstico cobre mayor relevancia la sensibilidad en detrimento de la especificidad. Sin embargo, existen algunas estrategias matemáticas que permiten escoger una combinación de sensibilidad y especificidad equilibrado. Una de estas formas es emplear una métrica que combine ambas medidas o semejantes. Por ejemplo, el *F-score* permite encontrar el punto de corte óptimo para optimizar la sensibilidad y el valor predictivo positivo. Otra forma más gráfica para encontrar un punto de corte óptimo y equilibrado consiste en escoger el punto de la curva que corta a la recta que va del punto (0,1), que representaría un modelo perfecto, al punto (0.5, 0.5). Esta forma de escoger asume que la curva ROC es perfectamente convexa y, por tanto, ese punto de corte es el más cercano al punto (0,1). Esta forma es bastante conocida en la práctica médica, sin embargo es aconsejable establecer los criterios de selección del punto de corte óptimo antes de llevar a cabo la evaluación de los modelos para no cometer errores de decisión mediante la estrategia gráfica aquí explicada.

14.6. Notas bibliográficas

Hastie en [139] es una buena referencia para las estrategias de evaluación en aprendizaje automático. Para profundizar, algunos conceptos deben ser consultados en [92]. Los estudios de Kohavi fueron interesantes para caracterizar las metodologías de evaluación por remuestreo y la descomposición de las funciónes de error 0-1 [91, 141]. El artículo de Berrar [142] recoge una serie de recomendaciones y alternativas para estimar los intervalos de error que son de gran interés cuando nos encontramos con muestras de tamaño pequeño. En el artículo de Kim [147] se demuestra empíricamente las ventajas y desventajas de emplear los métodos de validación cruzada y *holdout* con repetición frente al *bootstrap*.

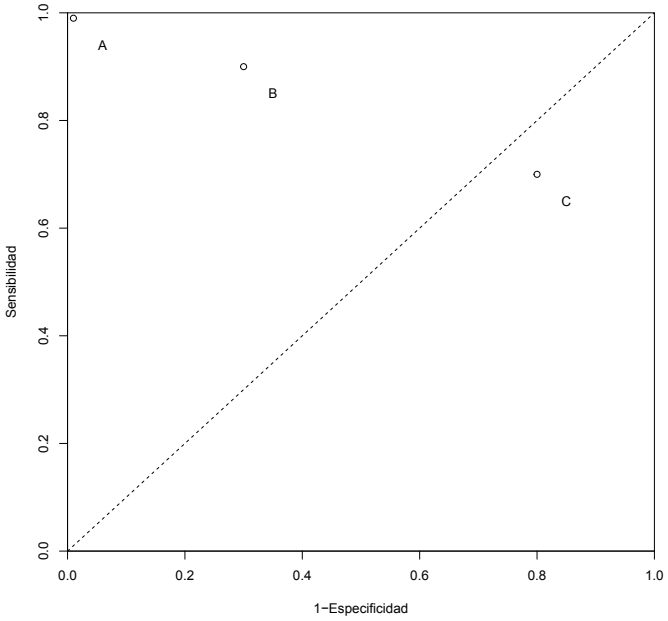


Figura 14.12: Ilustración de algunos ejemplos de modelos hipotéticos en el gráfico ROC.

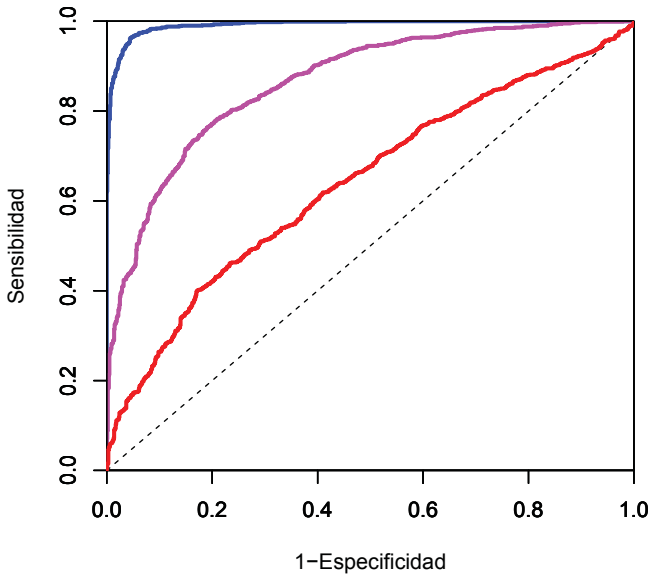


Figura 14.13: Tres curvas ROC representan los resultados de tres modelos de clasificación diferentes. El mejor modelo está representado por la curva azul y tiene un AUC de 0.99. El peor modelo está representado por la curva ROC y tiene un AUC de 0.61. La curva magenta representa un modelo con un AUC de 0.87.

Capítulo 15

Almacenes de datos y procesamiento analítico en línea

La Historia Clinica Electrónica (EHR) de una población es una de las mayores fuentes de información sobre la salud de sus ciudadanos, así como de situaciones que puedan comprometer la salud pública de una región. Efectivamente, una vez cubierta la asistencia médica de los pacientes, que representa el uso primario de un EHR, la información estructurada que contiene se puede emplear para usos secundarios como la predicción, prevención, planificación y gestión sanitaria.

En este capítulo, nos centraremos en la explotación de los datos históricos de salud mediante herramientas de Business Intelligence (BI) para apoyar programas de prevención de salud pública, planificar actuaciones sobre grupos de pacientes y gestionar recursos hospitalarios, entre otras utilidades. Dejaremos la predicción para el capítulo 8, donde veremos el uso de datos para la preparación de modelos predictivos.

15.1. Introducción

En un proyecto de BI se suelen estudiar *indicadores* o *medidas* sobre unos *hechos* agrupados bajo unas condiciones llamadas *dimensiones*. Así, por ejemplo, podemos estar interesados en saber el porcentaje (indicador) de camas ocupadas (hecho) en los hospitales de la Comunidad Valenciana (dimensión localización) a lo largo de un año (dimensión fecha). Algunos indicadores que se suelen definir en proyectos de sanidad son el consumo de tiras de glucemia capilar, la incidencia de tipos de cáncer, el número total de antibióticos prescritos, el porcentaje de bacterias resistentes a antibióticos[a] o el coste total. Las agrupaciones que se suelen definir vienen dadas por dimensiones temporales (tiempo), o geográficas (localización), pero también organizativas (servicios médicos), factores clínicos o poblacionales (diagnósticos o grupos poblacionales), o productos sanitarios. Además, las dimensiones suelen definir jerarquías de *niveles* para refinar en mayor o menor medida los agrupamientos. Así por ejemplo, la agrupación de los hechos a través de la dimensión tiempo puede realizarse por años, meses, quincenas, semanas, días, etcétera. En definitiva, los proyectos BI facilitan la navegación a través de *tablas de contingencia* sobre agregaciones o desagregaciones de las dimensiones del modelo dimensional.

De forma general, BI es el término que engloba al conjunto de herramientas para la explotación de datos existentes en una organización o empresa. Para facilitar la integración de los

[a]Un refinamiento de este indicador sería utilizar el CMI promedio de resistencia bacteriana a antibióticos: Concentración mínima inhibitoria (CMI) de un antibiótico se define como la mínima cantidad de antibiótico capaz de impedir el crecimiento bacteriano.

datos procedentes de múltiples fuentes de datos y su procesamiento analítico se suele preparar un almacén de datos (o *datawarehouse*) mediante plataformas que incluyen herramientas de i) *Extract*, *transform* and *load* (extracción, transformación y carga de datos) (ETL); ii) diseño de modelos multidimensionales; iii) *On-line Analytical Processing* (procesamiento analítico en línea) (OLAP); iv) obtención de resúmenes; y v) cuadros de mandos. Algunos textos incluyen las herramientas de *Data Mining* (minería de datos) (DM) como herramientas BI, pero resultan vagos en su descripción. En este capítulo trataremos las herramientas BI necesarias para integrar el procesamiento analítico en línea en cuadros de mando y dejaremos las herramientas de DM para el capítulo 8 donde se estudiarán bajo el contexto del aprendizaje automático.

Para guiar la exposición del tema, desde un punto de vista aplicado a medicina, desarrollaremos un ejemplo sobre vigilancia geográfica de la resistencia bacteriana y el uso de antibióticos. Complementaremos la exposición teórica del capítulo con una vista de implementación sobre Pentaho CE desarrollada en el anexo D.

Ejemplo 15.1 (Vigilancia de la resistencia bacteriana y el uso de antibióticos.) Las bacterias son los organismos más abundantes del planeta. Las bacterias patógenas pueden causar enfermedades infecciosas, como tuberculosis, cólera, sífilis, lepra, tifus, difteria, escarlatina, etcétera. Para combatir las infecciones bacterianas se utilizan los antibióticos, que inhiben la formación de la pared celular o detienen otros procesos de su ciclo de vida.

Al igual que el resto de organismos, aquellas bacterias con mutaciones que les permitan resistir la acción antibiótica serán las que crearán una descendencia resistente al antibiótico. Si bien la selección natural es propia de la evolución, se ha comprobado [148] que ciertos patrones sistemáticos de uso de los antibióticos aumentan el número de organismos resistentes a los antibióticos en una localización geográfica concreta. Específicamente, los siguientes factores pueden aumentar la resistencia antibiótica:

- Uso excesivo de antibióticos de amplio espectro (por ejemplo, las cefalosporinas de segunda y tercera generación, acelera en gran medida el desarrollo de resistencia a la meticilina).
- Los diagnósticos incorrectos.
- Las prescripciones innecesarias.
- El uso incorrecto de antibióticos por parte de los pacientes.
- El uso de los antibióticos como aditivos en la alimentación del ganado para aumentar el engorde.
- El uso intensivo de antibióticos en la agricultura.
- La introducción de antibióticos en limpiadores del hogar.

El control de la resistencia bacteriana y su relación con el uso de los antibióticos se considera actualmente un problema de salud pública mundial [149], y se aborda como tal por los servicios de salud, como en el proyecto ViResiST [b] en el Hospital Vega baja de Orihuela (Alicante) y sus centros asociados de atención primaria.

Un programa de salud pública para la vigilancia geográfica de la resistencia bacteriana y el uso de antibióticos debe considerar las siguientes cuestiones:

- Observar la evolución temporal de la resistencia de una bacteria determinada frente a un antibiótico determinado en una localización geográfica.

[b] www.viresist.org

- Observar la relación temporal entre el uso de los antibióticos y la resistencia de las bacterias en una localización geográfica.

A lo largo del capítulo, desarrollaremos un sistema de vigilancia basándonos en herramientas BI para dar soporte a un programa de salud pública que aborde estas preguntas. Durante el ejemplo, utilizaremos el contexto planteado por el proyecto ViResiST, pero los datos presentados son totalmente simulados y no responden en ningún caso a la realidad.

Más allá del modelo BI que diseñaremos en este capítulo, se podría diseñar un CDSS para la prescripción de antibióticos mediante modelos predictivos de la resistencia local esperada de las bacterias a antibióticos. Para ello, sería interesante complementar el desarrollo descrito en este capítulo con la metodología de diseño de modelos predictivos del capítulo 8.

Ejemplo 15.2 (Resistencia bacteriana y uso de antibióticos (cont. ej. 15.1.)) Basándonos en las tablas de Bretón en [150] sobre resistencias típicas de las bacterias a los antibióticos en cuatro hospitales valencianos formamos la tablas de aparición de bacterias (tabla 15.1) y su resistencia a antibióticos (tabla 15.2).

Tabla 15.1: Número de casos de infecciones bacterianas durante el periodo de estudio en cada hospital. Se incluyen los porcentajes de las bacterias (svi: *Streptococcus viridians*, sp: *Streptococcus pneumoniae* y sau: *Staphylococcus aureus*) condicionadas a cada hospital, así como los conteos y porcentajes marginales por hospital y bacteria.

Hospital	svi	sp	sau	TOTAL
H. General de Elche	262 (33 %)	263 (33 %)	267 (34 %)	792 (23.3 %)
H. de la Vega Baja	57 (26 %)	27 (12 %)	134 (62 %)	218 (6.4 %)
H. Dr. Peset	74 (22 %)	59 (18 %)	200 (60 %)	333 (9.8 %)
H. Clínico	294 (49 %)	92 (15 %)	216 (36 %)	602 (17.8 %)
H.General de Castellón	202 (14 %)	836 (58 %)	411 (28 %)	1449 (42.7 %)
TOTAL	889 (26.2 %)	1277 (37.6 %)	1228 (36.2 %)	3394

Simularemos un año de incidencias de antibióticos según las distribuciones de probabilidad expresadas por las tablas de aparición de bacterias (tabla 15.1) y su resistencia a antibióticos (tabla 15.2).

Según la revisión de Bretón en [150], se han observado aumentos de la resistencia de las bacterias de la tabla 15.1 a antibióticos según el resumen de la tabla 15.3.

Por simplicidad del ejemplo, simularemos únicamente el consumo de los antibióticos Penicilina (PEN), Amoxicilina (AMX), Ciprofloxacino (CIP) en las zonas de los hospitales del estudio según las tabla 15.4.

En nuestro ejemplo, supondremos que disponemos de un acceso de consulta al subconjunto del EHR que contiene los estudios de laboratorio de resistencia bacteriana a antibióticos y el uso de los antibióticos en los hospitales involucrados. La figura 15.1 representado por el modelo entidad-relación del subconjunto de la fuente de datos.

Tabla 15.2: Porcentaje (%) de bacterias con resistencia a antibióticos en los hospitales de estudio.

Bacteria	Hospital	PEN (%)	AMX (%)	CIP (%)
svi	H. General de Elche	40.9	39.5	45.0
	H. de la Vega Baja	0	0	31.3
	H. Dr. Peset	41.5	0	25.0
	H. Clínico	22.8	26.5	0
	H.General de Castellón	40.9	40.9	15.4
sp	H. General de Elche	59.2	53.7	61.8
	H. de la Vega Baja	70.0	0	0
	H. Dr. Peset	27.3	17.7	10.5
	H. Clínico	46.7	16.0	0
	H.General de Castellón	62.8	0	0
sau	H. General de Elche	97.4	98.7	10.2
	H. de la Vega Baja	91.8	91.4	2.9
	H. Dr. Peset	93.7	0	2.9
	H. Clínico	75.2	69.5	0
	H.General de Castellón	93.5	93.9	3.8

Tabla 15.3: Consumos de antibióticos que provocan resistencia de bacterias.

Consumo de	hace que la Bacteria	tienda a la resistencia a
-	*Streptococcus viridans* (svi)	β-lactámicos (v.g. penicilinas), macrólidos (v.g. eritromicina) y tetraciclinas
β-lactámicos, macrólidos	*Streptococcus pneumoniae* (sp)	eritromicina, β-lactámicos
penicilina, múltiples	*Staphylococcus aureus* (sau)	penicilina, múltiples

Tabla 15.4: Consumo simulado de antibióticos.

Hospital	Antibiótico	Unidades administradas
H. General de Elche	PEN	2640
	AMX	2532
	CIP	2789
H. de la Vega Baja	PEN	643
	AMX	682
	CIP	1350
H. Dr. Peset	PEN	1252
	AMX	758
	CIP	1310
H. Clínico	PEN	3568
	AMX	2591
	CIP	456
H.General de Castellón	PEN	9431
	AMX	7952
	CIP	2467

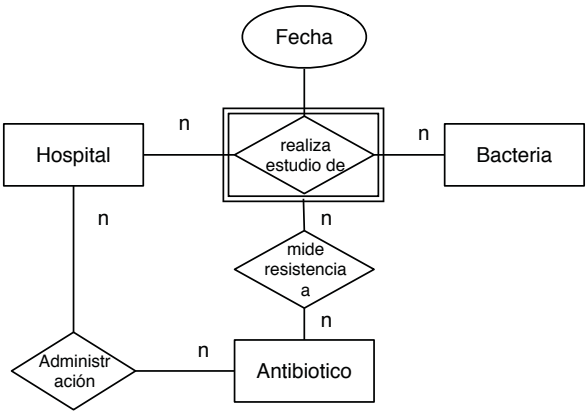


Figura 15.1: Modelo entidad-relación de la fuente principal de datos para el problema de resistencia bacteriana a antibióticos.

15.2. Modelo multidimensional

El uso de los datos en un proyecto BI va dirigido a calcular indicadores que resuman los hechos que se agrupan a través de las dimensiones. El modelo relacional utilizado por la mayoría de bases de datos no es adecuado como modelo conceptual para un almacén de datos. El modelo multidimensional es el más extendido entre las soluciones de BI. Este modelo organiza los datos en torno a los hechos que ocupan el centro de una estrella figurada rodeada de las dimensiones. Si las dimensiones constan de varios niveles a través de los cuales agregar los datos, en lugar de datos en estrella se denomina estrella jerárquica o copo de nieve si hay caminos alternativos.

Las plataformas BI suelen incorporar un editor de esquemas (o diseñador de cubos) que nos permitirá definir cada cubo a través de las medidas de los hechos de estudio y las dimensiones que definen las condiciones a través de las jerarquías de niveles.

La implementación lógica de un modelo multidimensional puede tener varias arquitecturas: ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP) y HOLAP (Hybrid OLAP). La arquitectura ROLAP utiliza modelos relacionales para implementar el modelo de estrella mediante tablas de hechos y de dimensiones. La principal ventaja de esta arquitectura es su flexibilidad ante los cambios. La arquitectura MOLAP enlaza mediante índices multidimensionales los hechos y medidas precalculadas para las agregaciones en tiempo de diseño, lo que permite una ejecución más rápida del procesamiento en línea. Las arquitecturas híbridas HOLAP almacenan la información detallada en estructuras relacionales ROLAP y las medidas precalculadas para agregaciones en cubos MOLAP.

En una arquitectura ROLAP, podemos implementar el modelo de estrella mediante:

- Tablas de dimensiones. Incluiremos una tabla por cada dimensión, que contenga una clave primaria y un campo por cada nivel de la dimensión, por ejemplo una tabla localización, podrá tener como registro "(1, Hospital General de Valencia, Valencia, Comunidad Valenciana)", donde el primer campo es la clave primaria y los siguientes son los valores de los niveles hospital, provincia y comunidad autónoma de la dimensión localización.

- Tablas de hechos. Incluiremos una tabla para cada hecho que contendrá las medidas del hecho y una clave ajena a cada dimensión de la estrella. Así, por ejemplo, la tabla de hechos "prescripción de medicamentos" tendrá un campo "coste" como medida, y los campos "localización", "medicamento", "fecha", "causa" corresponderán a las dimensiones de la estrella.

Visualmente es fácil imaginar un modelo con tres dimensiones en forma de cubo, cuyo volumen se trocea en pequeños cubos correspondientes a cada combinación de valores de cada dimensión, pudiéndose estos pequeños cubos agregarse subiendo de nivel en las jerarquías de las dimensiones o disgregarse, bajando de nivel en las dimensiones.

Generalmente, se denomina *datawarehouse* (o almacén de datos) al repositorio que contiene el conjunto de hechos y dimensiones definidos para un problema. Cada una de las estrellas que estudian un aspecto concreto del problema se denomina *datamart*, y pueden compartirse dimensiones entre ellos.

Ejemplo 15.3 (Modelo multidimensional (cont. ej. 15.2.))
Definiremos los hechos, indicadores y dimensiones de un modelo con dos estrellas jerarquizadas. En primer lugar podemos identificar los dos hechos relevantes de nuestro estudio: "resistencia de bacteria a antibiótico" y "uso de antibiótico". Definimos los indicadores para el modelo multidimensional enumerados en la tabla 15.5.

Tabla 15.5: Indicadores de resistencia de bacterias y uso de antibióticos.

Hechos	Indicadores
Resistencia de bacterias a antibióticos	Porcentaje de bacterias resistentes a antibióticos
Uso de antibiótico	Número de administraciones

La primera dimensión que definimos en el modelo es la **bacteria** causante de infecciones. Estudiaremos la resistencia de las bacterias enumeradas en la tabla 15.6. Es una lista limitada, pero suficiente para propósitos académicos. Por simplicidad, tampoco incluimos niveles de agregación de las bacterias (familias y características comunes), que podrían ser de interés para un estudio real.

Tabla 15.6: Bacterias incluidas en el modelo multidimensional.

svi	Streptococcus viridans
sp	Streptococcus pneumoniae
sau	Staphylococcus aureus

La segunda dimensión incluida en el modelo son los **antibióticos**, que pueden ser agrupados en familias. Restringiremos el estudio a los antibióticos incluidos en la tabla 15.7.

Tabla 15.7: Antibióticos incluidos en el modelo multidimensional.

Nomenclatura	Antibiótico	Familia
PEN	Penicilina	Penicilinas
AMX	Amoxicilina	Penicilinas
CIP	Ciprofloxacino	Quinolonas

La tercera dimensión del modelo será el **tiempo**, cuyo nivel más alto de detalle será el mes y el más bajo será el año.

La cuarta y última dimensión del modelo es el **lugar geográfico** de medida de resistencia de bacterias y administración de antibióticos. El nivel más bajo de detalle será el hospital donde se ha realizado la observación. Definiremos un nivel de menor detalle mediante agregación de hospitales en provincias. La tabla 15.8 enumera los hospitales incluidos en el ejemplo.

Tabla 15.8: Hospitales incluidos en el ejemplo, agregados por provincia.

Provincia	Hospital
Alicante	H. General de Elche
Alicante	H. de la Vega Baja
Valencia	H. Dr. Peset
Valencia	H. Clínico
Castellón	H.General de Castellón

Como resultado, obtenemos el datawarehouse con dos estrellas y cuatro dimensiones que presenta la figura 15.2.

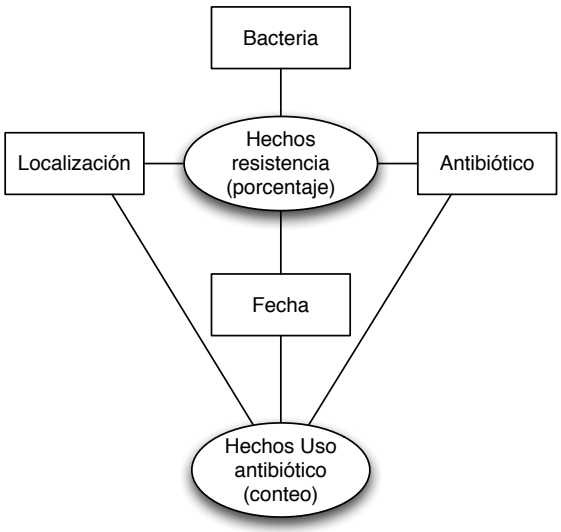


Figura 15.2: Datawarehouse de resistencia bacteriana a antibióticos.

Este proceso puede llevarse a cabo mediante el editor de esquemas de nuestra plataforma BI. Para ello, definimos el esquema del datawarehouse con dos estrellas (o cubos), uno de ellos con la medida de porcentaje de bacterias resistentes a antibióticos y el otro con el número de administraciones. Incorporaremos las dimensiones correspondientes a cada uno de los cubos, para lo que necesitaremos una única jerarquía en cada una de ellas. Dentro de la jerarquía de cada dimensión, se definirán los niveles en orden decreciente.

En paralelo al diseño del cubo, realizamos la implementación del modelo en una arquitectura ROLAP mediante una base de datos relacional con dos tablas de hechos (una para resistencia bacteriana y otra para uso de antibióticos) y tres tablas de dimensiones (antibiótico, bacteria y localización). La tabla de hechos "resistencia bacteriana" dispondrá de una clave primaria autonumerada, un campo indicador del resultado de test de resistencia, la fecha del test y tres

claves ajenas a las tablas de dimensiones. La tabla de hechos "uso de antibióticos" dispondrá de una clave primaria autonumerada, un campo fecha y claves ajenas a las tablas localización y antibiótico. Las tablas de dimensiones tendrán tantos campos como niveles tenga la dimensión, pudiendo ser la clave primaria el nivel de menor agregación o un autonumérico. Los niveles de la dimensión fecha pueden implementarse mediante funciones SQL de cálculo de años y meses sobre el campo fecha de las tablas de hechos, por lo que no es necesario una tabla explícita en la arquitectura ROLAP.

15.3. Carga y mantenimiento de un almacén de datos

Una vez diseñado el modelo multidimensional, debe poblarse de datos procedentes del EHR. Las suites ETL están generalmente formadas por un conjunto heterogéneo que facilitan principalmente

- la conexión con múltiples fuentes de datos para recuperar los registros;
- la manipulación de tablas y registros de datos, mediante transformaciones, uniones, generación de índices, etcétera;
- la creación de esquemas relacionales ROLAP de las estrellas, y metadatos para su mantenimiento y actualización (por ejemplo, *time stamping*);
- la planificación de cargas periódicas de datos que alimenten el almacén de datos.

No es demasiado común encontrar en el repositorio de información hospitalaria tablas de correspondencia entre los niveles de las dimensiones. Podremos completar nuestras fuentes de información creando tablas de correspondencia en bases de datos cercanas a nuestro almacén de datos que proporcionen la información de dominio necesaria para poder navegar por todos los niveles de nuestras dimensiones. Las herramientas ETL generalmente incluyen *wrappers* que permiten hacer conversiones entre ficheros de texto, hojas de cálculo, HTML, XML y tablas de bases de datos relacionales.

Ejemplo 15.4 (Creación de mapeo entre niveles de localización (cont. ej. 15.3.))
Para completar la información necesaria para el almacén de datos importaremos una matriz de correspondencia hospital-provincia desde una hoja de cálculo a nuestro sistema de información, consiguiendo el modelo entidad-relación ampliado de la figura 15.3. Para ello, será de utilidad utilizar las transformaciones que suministran las herramientas ETL de las plataformas de BI.

Una vez asegurado el acceso a las fuentes de datos, y que disponemos en ellas de la información para completar los datos de hechos y dimensiones, deberemos cargar los datos en el modelo multidimensional. Podremos integrar todas las fuentes de datos fácilmente mediante las herramientas de transformación y planificación de nuestro ETL.

Ejemplo 15.5 (Integración de datos (cont. ej. 15.4.))
En nuestro caso de estudio simulado supondremos que tenemos acceso al subconjunto de un sistema de información que almacena los estudios de resistencia bacteriana a antibióticos y el uso de los mismos según el modelo entidad-relación de la figura 15.3. Con la creación de tablas de correspondencia entre niveles de localización tenemos todos los datos necesarios para nuestro almacén de datos. La integración de los datos es viable y sencilla gracias a que el ETL puede establecer conexión con todas las fuentes para realizar las transformaciones necesarias para obtener los datos unificados.

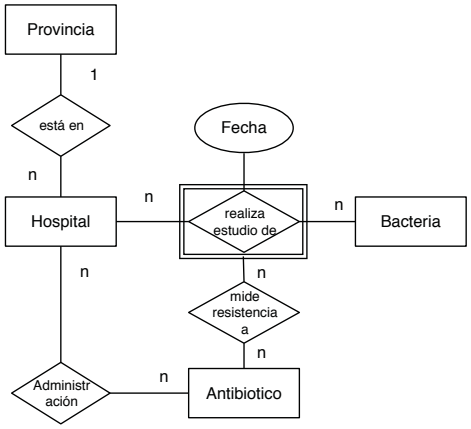


Figura 15.3: Modelo entidad relación ampliado de la fuente principal de datos para el problema de resistencia bacteriana a antibióticos.

Para una arquitectura ROLAP, se realizará la carga de datos mediante transformaciones desde el modelo relacional del sistema de información al modelo relacional de tablas de hechos y dimensiones que implementan las estrellas del datawarehouse. Además, esta carga debe ser planificada periódicamente para mantener el almacén de datos actualizado. Estas operaciones generalmente requieren un uso intensivo de los repositorios de información (por ejemplo EHR, sistemas de información externos o bases de datos departamentales), por lo que se suele planificar para momentos donde la actividad primaria del sistema de información es mínima o nula.

Ejemplo 15.6 (Carga del almacén de datos (cont. ej. 15.5.)
Las herramientas ETL permiten realizar la carga del almacén de datos a través de un interfaz gráfico para el diseñador del almacén o mediante la planificación de procesos en segundo plano.

Específicamente, cargaremos las tablas de la arquitectura ROLAP con datos de nuestro subconjunto simulado de EHR mediante la transformación del modelo entidad-relación de la figura 15.3 al modelo estrella jerárquica de la figura 15.2.

15.4. Procesamiento analítico en línea

El objetivo de la preparación del almacén de datos es explorar la información de nuestros datos históricos con el fin de encontrar tendencias que describan algún aspecto relevante de nuestro sistema de información sanitario. El núcleo de una plataforma BI para la exploración del almacén de datos es el servidor OLAP. El servidor OLAP se encarga de consultar el almacén de datos según el modelo multidimensional para generar tablas de contingencia con las medidas de los datos agregados a cierto nivel en cada dimensión.

El servidor OLAP ofrece los siguientes operadores de almacenes de datos (o de análisis):

- Drill: disgrega los datos a niveles más finos de detalle de una dimensión o varias.
- Roll: agrega los datos a niveles más gruesos de detalle de una dimensión o varias. Cuando Drill y Roll se ejecutan sobre dimensiones definidas en la primera consulta se denominan drill-down y roll-up; cuando son dimensiones diferentes, o hacen desaparecer dimensiones, se denominan drill-across y roll-across.

- Slice & Dice: este operador filtra los datos utilizados para calcular las medidas. El filtrado puede realizarse por valores de una o varias dimensiones.

- Pivot: reorientar las dimensiones en la tabla de contingencia.

Estos operadores actúan sobre el resultado obtenido por el servidor OLAP para ofrecer al usuario la vista de la estrella sobre la que trabaja. Por lo tanto, los operadores de análisis refinan la consulta realizada al abrir el cubo.

Además, los servidores OLAP ofrecen normalmente una representación gráfica de las tablas de contingencia obtenidas mediante la manipulación de los cubos.

Ejemplo 15.7 (Resistencia bacteriana por años y provincias (cont. ej. 15.6))
Una vez creado el almacén de datos de nuestro problema de resistencia bacteriana y uso de antibióticos, podemos aplicar los operadores de análisis para disgregar los datos por años y por provincias, manteniendo unidos todos los antibióticos y bacterias. Si aplicamos sobre la dimensión localización el operador pivot conseguiremos tres series temporales, una por provincia, del porcentaje de bacterias resistentes a los antibióticos. La figura 15.4 muestra las series temporales obtenidas, en forma de tabla de contingencia y como gráficas de líneas.

La navegación mediante las operaciones de agregación y filtrado nos harán fácil la revisión de los patrones más frecuentes del almacén de datos. Es posible guardar las vistas de análisis que se han obtenido durante la navegación por el almacén de datos. Además, las plataformas BI permiten componer cuadros de mando para integrar diferentes vistas de análisis en una pantalla única, con el objetivo de presentar el estado del proyecto de un solo vistazo.

Ejemplo 15.8 (Series temporales (cont. ej. 15.7))
La navegación por el cubo de resistencia bacteriana de antibióticos nos puede llevar a observar un patrón interesante en algún nivel de agregación. Por ejemplo, con datos simulados, hemos encontrado que en los meses de mayo y junio de 1998 se produjo un aumento de la resistencia de la bacteria *Staphylococcus aureus* (sau) a los antibióticos en la provincia de Alicante. La visualización conjunta mostrada en la figura 15.5 de la serie temporal de porcentaje de resistencia de antibióticos en los niveles de Provincia y Mes junto con la serie temporal de uso de Penicilina (PEN) durante el mismo periodo podría revelar una relación de causalidad entre este suceso y un aumento de la administración de penicilina durante marzo, abril y mayo del mismo año. Estas observaciones deberían venir acompañadas de un análisis estadístico con contraste de hipótesis para ratificar los hallazgos descriptivos encontrados.

15.5. Minería de flujos de datos

La salud y el bienestar son aspectos atemporales que involucran hábitos, monitorización y la prevención de eventos adversos de los ciudadanos durante su vida cotidiana. Los avances actuales en monitorización y comunicación de señales biomédicas e información ambiental del entorno del ciudadano hace cada vez más viable la gestión continua de sus condiciones vitales. Escenarios de la vida cotidiana de los ciudadanos, como el control de hábitos saludables de personas activas, la monitorización de ancianos, pasando por la terapia asistida de personas con depresión mayor, son algunos ejemplos de aplicaciones de la ayuda a la decisión que involucran grandes flujos de datos relacionados con la salud del paciente y que requieren una respuesta continua e inmediata.

La minería de flujos de datos (Mining data streams) es una subdisciplina de la minería de datos que estudia la forma de extraer estructuras de conocimiento desde modelos y patrones en flujos continuos de información. Esta disciplina se ha desarrollado en su aplicación a problemas financieros, transacciones bancarias, telecomunicaciones, seguridad informática, tecnología web,

datos multimedia, etcétera. Por ello, es fácil prever su aplicación tanto en laboratorios biotecnológicos y clínicos como en ambientes personales del ciudadano para el control de su salud y bienestar.

El requisito fundamental de la minería de flujos de datos es la rapidez de respuesta de sus algoritmos de decisión y un coste espacial asumible. Sin embargo, las aplicaciones de minería de flujos de datos requieren procesos complejos, como la detección de cambios por encima de niveles de ruido, el agrupamiento, la clasificación, la detección de patrones frecuentes y el análisis de series temporales. Por estas razones la minería de flujos de datos es una disciplina de alta intensidad científica y de gran proyección en el ámbito biomédico.

15.6. Notas bibliográficas

La explicación teórica del capítulo se ha basado fundamentalmente en el libro [151] de Orallo, Ramírez y Ferri y en la documentación técnica de Pentaho CE. La elaboración del ejercicio de resistencia bacteriana a antibióticos se basa en la información pública disponible del proyecto ViResist (http://www.viresist.org) y en la tesis de Gervas [149], si bien los datos son totalmente ficticios.

Gaber et al. en [152] realiza una buena revisión sistemática sobre la minería de flujos de datos. Las páginas web `http://www.csse.monash.edu.au/~mgaber/WResources.htm` y `http://wis.cs.ucla.edu/~hxwang/stream/bib.html` mantienen bibliografía actualizada sobre esta subdisciplina.

	Medidas		
	Porcentaje de Resistencia		
	Localizaciones		
Fechas	Alicante	Castellon	Valencia
1992	0,531	0,347	0,226
1993	0,577	0,362	0,333
1994	0,48	0,35	0,304
1995	0,517	0,366	0,224
1996	0,512	0,299	0,296
1997	0,558	0,317	0,27
1998	0,618	0,375	0,277
1999	0,523	0,339	0,298
2000	0,47	0,337	0,282

Figura 15.4: Tabla de contingencia y gráfica con las series temporales de resistencia bacteriana por años y provincias obtenido por el servidor OLAP sobre el cubo de resistencia bacteriana a antibióticos.

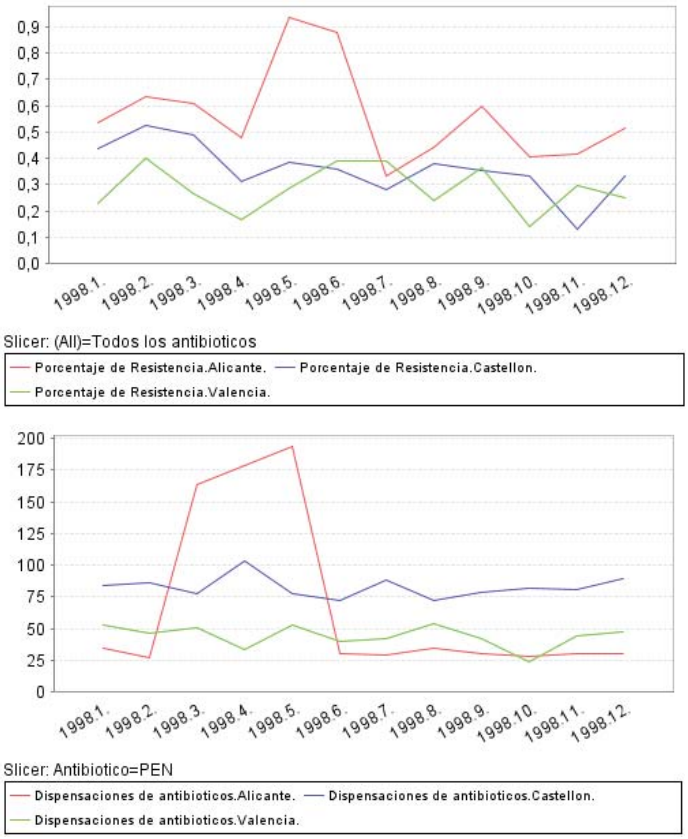


Figura 15.5: La visualización conjunta de la serie temporal de porcentaje de resistencia de antibióticos en los niveles de Provincia y Mes de 1998 junto con la serie temporal de uso de Penicilina (PEN) durante el mismo periodo podría revelar una relación de causalidad entre este suceso y un aumento de la administración de penicilina durante marzo, abril y mayo del mismo año.

Capítulo 16

Razonadores semánticos aplicados a medicina

La forma de resolver algunos problemas viene definida por un conjunto de reglas deterministas. La aplicación de estas reglas sobre una situación particular permite deducir conclusiones de forma lógica, y por lo tanto aplicar el conocimiento que se tiene sobre el problema a la situación particular.

Este tipo de resolución constituye una forma cómoda de resolver decisiones basándose en el conocimiento experto, si bien, no tiene en cuenta la incertidumbre intrínseca asociado a todo problema de decisión. En este capítulo haremos un breve repaso a la lógica formal para establecer los fundamentos sobre los que se sustentan los sistemas basados en reglas. Posteriormente, estudiaremos los sistemas basados en reglas y aplicaremos uno de ellos para resolver un sistema de medición del riesgo de un diabético de sufrir un evento adverso y un sistema de alertas para la prescripción de medicamentos.

Utilizamos el lenguaje Jess (www.jessrules.com), basado en CLIPS para resolver los ejemplos y ejercicios del capítulo, por lo que son directamente ejecutables en su motor de reglas. Los ejemplos utilizan estructuras estándar del lenguaje, que pueden ser fácilmente entendidas por el estudiante con el manual de Jess accesible desde la sección documentación de la página web del motor de reglas.

16.1. Introducción a la lógica simbólica

La lógica estudia los métodos de formalización del conocimiento. Este conocimiento se expresa mediante frases declarativas. Es decir, frases que expresan algo sobre el dominio de trabajo. Existen, al menos, dos niveles de abstracción según el grado de detalle que se quiera formalizar:

- Lógica proposicional o lógica de enunciados
- Lógica de predicados de primer orden

El elemento atómico de las proposiciones son las frases declarativas que constituyen por sí solos una unidad de comunicación de conocimientos y que llevan asociados un valor veritativo, esto es, pueden ser considerados Verdaderos o Falsos.

16.1.1. Fórmula bien formada

La lógica proposicional se compone de un alfabeto, compuesto a su vez de proposiciones y conectivas lógicas que son: la negación ($\neg$), la O lógica o disyunción ($\vee$), la Y lógica o conjunción ($\wedge$), la implicación ($\rightarrow$) y la coimplicación ($\leftrightarrow$).

Las gramáticas determinan qué combinaciones de símbolos son fórmulas bien formadas (fbf). Además, debe existir un mecanismo que nos permita asignar un valor veritativo (verdadero o falso) a una fórmula. En principio, cualquier variable proposicional es considerada una fórmula bien formada. La negación de una fórmula bien formada es, a su vez, una fórmula bien formada. Del mismo modo, la disyunción, conjunción, implicación y coimplicación de 2 fórmulas bien formadas son, a su vez, fórmulas bien formadas. Es decir, si p y q son fómulas bien formadas de un sistema, entonces $(p \vee q)$, $(p \wedge q)$, $(p \rightarrow q)$, $(p \leftrightarrow q)$ también son fórmulas bien formadas.

Ejemplo 16.1 (Fórmulas bien formadas)
Si tenemos que p, q, r y s son fórmulas bien formadas, entonces

$$\neg(p \vee q) \wedge r \rightarrow s$$

es, a su vez, una fórmula bien formada, pero

$$\neg \vee (p \wedge qr) \leftrightarrow s$$

no es una fórmula bien formada, pues concatena conectivas lógicas entre sí ($\neg\vee$) y fórmulas bien formadas sin una conexión entre ellas (qr).

16.1.2. Tablas de verdad

Las tablas de verdad son tablas matemáticas que muestran los valores veritativos o valores de verdad de las fórmulas bien formadas compuestas. A cada conectiva lógica básica se le suele asociar una tabla de verdad. Su uso se extendió a partir del *Tractatus Logico-Philosophicus* de Ludwig Wittgenstein. En la tabla 16.1 se muestra la tabla de verdad de la negación. La negación de una proposición toma el valor contrario de dicha proposición. La tabla 16.2 muestra la tabla de verdad de la disyunción. La disyunción solo es falsa si ambas proposiciones son falsas. En la tabla 16.3 se muestra los valores de verdad de la conjunción. La conjunción solo es verdadera si ambas proposiciones son verdaderas. En la tabla 16.4 se muestra la tabla de verdad de la implicación. La implicación está formada por un antecedente (que es p) y un consecuente (que es q). Si p es verdadero y q es falso, entonces el valor de la implicación es falso. En cualquier otro caso, el valor es verdadero. Por último, la tabla 16.5 muestra los valores de verdad de la coimplicación. La coimplicación (o doble implicación) es equivalente a decir que p implica q y q implica p, por lo tanto, solo cuando ambas proposiciones son verdaderas -o ambas son falsas- la coimplicación es verdadera.

p	**¬ p**
V	F
F	V

Tabla 16.1: Tabla de verdad de la negación.

p	**q**	**p** $\vee$ **q**
V	V	V
V	F	V
F	V	V
F	F	F

Tabla 16.2: Tabla de verdad de la disyunción.

p	**q**	**p** $\wedge$ **q**
V	V	V
V	F	F
F	V	F
F	F	F

Tabla 16.3: Tabla de verdad de la conjunción.

p	**q**	**p** $\rightarrow$ **q**
V	V	V
V	F	F
F	V	V
F	F	V

Tabla 16.4: Tabla de verdad de la implicación.

p	**q**	**p** $\leftrightarrow$ **q**
V	V	V
V	F	F
F	V	F
F	F	V

Tabla 16.5: Tabla de verdad de la coimplicación.

16.1.3. Equivalencias lógicas

Una equivalencia lógica ($\equiv$) se da cuando dos fórmulas bien formadas obtienen la misma tabla de verdad ante los mismos valores de verdad de las proposiciones atómicas. Un ejemplo, vemos que $p \rightarrow q$ es lógicamente equivalente a $\neg p \vee q$. Es lo que se llama supresión de la implicación. Esto se puede comprobar en las tabla de verdad que se muestran en la tabla 16.6. Existen una serie de equivalencias lógicas conocidas, algunas de las cuales se pueden encontrar en la tabla 16.7.

p	**q**	$\neg$ **p**	**p** $\rightarrow$ **q**	$\neg$ **p** $\vee$ **q**
V	V	F	V	V
V	F	F	F	F
F	V	V	V	V
F	F	F	V	V

Tabla 16.6: Las tablas de verdad de las fórmulas de la derecha son iguales, por lo tanto ambas fórmulas son lógicamente equivalentes.

Idempotencia	P$\vee$P $\equiv$ P	P$\wedge$P $\equiv$ P
Absorción	P$\vee$**V** $\equiv$ **V**	P$\wedge$**F** $\equiv$ **F**
	P$\vee$(P$\wedge$Q) $\equiv$ P	P$\wedge$(P$\vee$Q) $\equiv$ P
Elemento neutro	P$\vee$**F** $\equiv$ P	P$\wedge$**V** $\equiv$ P
Elemento complementario	P$\vee\neg$P $\equiv$ **V**	P$\wedge\neg$P $\equiv$ **F**
Propiedad conmutativa	P$\vee$Q $\equiv$ Q$\vee$P	P$\wedge$Q $\equiv$ Q$\wedge$P
Propiedad asociativa	P$\vee$(Q$\vee$R) $\equiv$ (P$\vee$Q)$\vee$R	P$\wedge$(Q$\wedge$R) $\equiv$ (P$\wedge$Q)$\wedge$R
Propiedad distributiva	P$\vee$(Q$\wedge$R) $\equiv$ (P$\wedge$Q) $\vee$ (P$\wedge$R)	P$\wedge$(Q$\vee$R) $\equiv$ (P$\vee$Q) $\wedge$ (P$\vee$R)
Leyes de De Morgan	$\neg$(P$\vee$Q) $\equiv$ $\neg$P $\wedge$ $\neg$Q	$\neg$(P$\wedge$Q) $\equiv$ $\neg$P $\vee$ $\neg$Q
Doble negación	$\neg\neg$P $\equiv$ P	
Supresión de la implicación	P$\rightarrow$Q $\equiv$ $\neg$P$\vee$Q	
Supresión de la coimplicación	P$\leftrightarrow$Q $\equiv$ (P$\rightarrow$Q) $\wedge$ (Q$\rightarrow$P)	

Tabla 16.7: Tabla de equivalencias lógicas más empleadas.

Cabe mencionar la importancia que tienen las leyes de De Morgan que nos facilita el convertir conjunciones en disyunciones y viceversa. Así como la equivalencia de la supresión de la implicación vista anteriormente y la propiedad distributiva en la conformación de Formas Clausales, relacionadas con la Forma Normal de Skolem, que se emplea comúnmente en los sistemas basados en reglas. Más adelante profundizaremos sobre ello.

16.1.4. Regla de inferencia

En el cálculo proposicional existe una serie de reglas de inferencia para deducir consecuencias dadas unas premisas que se suponen verdaderas. En concreto, una regla de inferencia que va a ser sumamente útil para los motores de razonamiento semántico es la conocida como *modus ponendo*

ponens o simplemente *modus ponens*, que en latín significa "el modo que afirma (el consecuente) afirmando (el antecedente)". Básicamente, si tenemos una implicación y el antecedente de esta implicación es verdadero, entonces podemos admitir como verdadero el consecuente

$$\begin{array}{l} P \Rightarrow Q \\ P \\ \hline \therefore \quad Q \end{array}$$

esta regla se puede leer como "si P implica Q y P es verdadera, entonces Q es verdadera". Una regla de este tipo se compondrá de un conjunto de premisas que constituyen el antecedente y un conjunto de conclusiones que constituyen el consecuente

$$P_1, P_2, \ldots, P_n \Rightarrow Q_1, Q_2, \ldots, Q_m$$

Generalmente, en los motores de razonamiento semántico actuales las comas que conectan las premisas del antecedente y las conclusiones del consecuente representan conjunciones. Un *hecho* se representa como una regla cuya premisa es un conjunto vacío. En los motores de inferencia deductiva actuales siempre se suele asumir que un hecho está representando algo del dominio de trabajo por lo que es, en efecto, verdadero.

Por lo tanto, de aquí en adelante consideraremos que un hecho es una proposición verdadera que nos informa de algo que ocurre en el dominio que estamos representando formalmente. Así, si representamos que un paciente tiene fiebre, es porque realmente dicho paciente tiene fiebre.

El conjunto de hechos y reglas junto con el *modus ponens* como silogismo es la base de deducción de todos los razonadores semánticos desarrollados y, por tanto, la base de representación del conocimiento de los sistemas de ayuda a la decisión de tipo III.

16.1.5. Lógica de primer orden

En la lógica de predicados de primer orden, los predicados equivalen a expresiones lingüísticas y son tratados como funciones cuyos argumentos son constantes o variables de individuos. Estas funciones están bajo el alcance de algún cuantificador de tipo *existencial* ($\exists$) o de tipo *universal* ($\forall$). El cuantificador universal quiere decir que los predicados se cumplen para todas las instancias de una variable y el cuantificador existencial quiere decir que existe alguna instancia que cumple el predicado.

El ejemplo clásico del silogismo de Aristóteles puede servir de ilustración. Así, tenemos una proposición de aridad 1 (que emplea una variable) como $H(x)$ y que asociamos al siguiente significado: "x es un hombre"; tenemos también la proposición $M(x)$ que significa "x es mortal"; por último, tenemos una constante p que hace referencia a "Platón":

Todos los hombres son mortales

Platón es un hombre

Luego Platón es mortal

Este ejemplo se puede formalizar mediante la lógica de predicados de primer orden como

$$\frac{\forall x H(x) \Rightarrow M(x) \quad H(p)}{\therefore M(p)}$$

Así pues, podemos formalizar la expresión lingüística "todos los hombres son mortales" como una regla y la proposición "Platón es un hombre" como un hecho, pudiendo deducir por *modus ponens* la consecuencia que asegura que "Platón es mortal", de nuevo como un hecho.

El uso de cuantificadores en la lógica de primer orden también conlleva una serie de leyes de suma importancia en los motores de inferencia deductiva o razonadores semánticos. Algunas de estas leyes se muestran a continuación. Entre ellas encontramos las más empleadas para obtener Formas Clausales que, como veremos enseguida, es la forma normal necesaria para implementar un sistema basado en reglas.

Descenso cuantificacional

$\forall x\ \mathrm{A}(x) \Rightarrow \exists x\ \mathrm{A}(x)$

Conmutatividad de cuantificadores

$\forall x \forall y\ \mathrm{A}(x,y) \equiv \forall y \forall x\ \mathrm{A}(x,y)$
$\exists x \exists y\ \mathrm{A}(x,y) \equiv \exists y \exists x\ \mathrm{A}(x,y)$
$\exists x \forall y\ \mathrm{A}(x,y) \Rightarrow \forall y \exists x\ \mathrm{A}(x,y)$
$\forall x \exists y\ \mathrm{A}(x,y) \not\Rightarrow \exists y \forall x\ \mathrm{A}(x,y)$

Gran distributividad

$\forall x\ (\mathrm{A}(x) \wedge \mathrm{B}(x)) \equiv (\forall x\ \mathrm{A}(x) \wedge \forall x\ \mathrm{B}(x))$
$\exists x\ (\mathrm{A}(x) \vee \mathrm{B}(x)) \equiv (\exists x\ \mathrm{A}(x) \vee \exists x\ \mathrm{B}(x))$
$\forall x\ (\mathrm{A}(x) \vee \mathrm{B}(x)) \not\equiv (\forall x\ \mathrm{A}(x) \vee \forall x\ \mathrm{B}(x))$
$\exists x\ (\mathrm{A}(x) \wedge \mathrm{B}(x)) \not\equiv (\exists x\ \mathrm{A}(x) \wedge \exists x\ \mathrm{B}(x))$

Leyes de De Morgan con cuantificadores

$\neg\exists x\ \mathrm{A}(x) \equiv \forall x\ \neg\mathrm{A}(x)$
$\neg\forall x\ \mathrm{A}(x) \equiv \exists x\ \neg\mathrm{A}(x)$
$\neg\forall x\ \neg\mathrm{A}(x) \equiv \exists x\ \mathrm{A}(x)$
$\neg\exists x\ \neg\mathrm{A}(x) \equiv \forall x\ \mathrm{A}(x)$

Una expresión en *Forma Clausal* se compone de una conjunción de predicados, donde los símbolos conjuntivos $\wedge$ se sustituyen por comas, y cuyas variables están cuantificadas universalmente de forma implícita. Así por ejemplo,

$$\underbrace{\mathrm{P}_1, \ldots, \mathrm{P}_n}_{\text{Cuerpo}} \Rightarrow \underbrace{\mathrm{C}_1, \ldots, \mathrm{C}_m}_{\text{Cabeza}}$$

En la parte izquierda de la Forma Clausal encontramos los antecedentes y recibe el nombre de *cuerpo*. En la parte derecha se encuentran las conclusiones y recibe el nombre de *cabeza*. Cuando solo existe cabeza y no hay cuerpo, esto es, cuando $n = 0$, entonces en lugar de una regla tendremos un hecho. Para pasar cualquier expresión lógica a Forma Normal y, por tanto, poder implementarla en un sistema basado en reglas se deben pasar por una serie de formas normales intermedias que se verán en la siguiente sección.

16.1.6. Formas normales

Forma Normal Prenexa

Una fórmula está en *Forma Normal Prenexa* (FNP) si está compuesta de un prefijo constituido de una cadena de cuantificadores y una fórmula sin cuantificadores o matriz de la fórmula. Es decir, una FNP es de la forma

$$Q_1 x_1 Q_2 x_2 \dots Q_n x_n M$$

donde cada Q_i es un cuantificador existencial o universal y donde M es la matriz o fórmula bien formada sin cuantificadores. Se puede demostrar que cualquier fórmula bien formada cerrada se puede transformar en una fórmula equivalente en FNP. Un fórmula bien formada es cerrada si no tiene variables libres, es decir, variables que no estén bajo el alcance de algún cuantificador. Si la fórmula no fuese cerrada se podría aplicar el cierre existencial. El *cierre existencial* consiste en añadir al principio de la fórmula un cuantificador existencial por cada variable. Para ello deberían seguirse los siguientes pasos:

1. Renombrar aquellas variables que tengan el mismo símbolo y distinto cuantificador
2. Eliminar las dobles implicaciones y las implicaciones
3. Conseguir que las negaciones solo afecten a fórmulas atómicas mediante la aplicación de las leyes de De Morgan con cuantificadores
4. Pasar los cuantificadores al principio de la fórmula aplicando las leyes conmutatividad y gran distributividad de cuantificadores

Forma Normal Conjuntiva

Una fórmula está en *Forma Normal Conjuntiva* (FNC) si se expresa como una conjunción finita de cláusulas donde cada cláusula es un literal o una disyunción de literales. La FNC no cuenta necesariamente con cuantificadores en la fórmula. Así, por ejemplo la fórmula bien formada

$$\neg \mathrm{P} \wedge (\mathrm{Q} \vee \mathrm{R}) \wedge (\mathrm{S} \vee \mathrm{T})$$

está en FNC, mientras que

$$(\mathrm{P} \wedge \mathrm{Q}) \vee \mathrm{R}$$

no está en FNC, aunque se puede expresar como tal sin más que aplicar la ley distributiva de la conjunción respecto a la disyunción y dejarla como $(\mathrm{P} \vee \mathrm{R}) \wedge (\mathrm{Q} \vee \mathrm{R})$.

Cuando la fórmula se compone de una disyunción finita de cláusulas donde cada cláusula es un literal o una conjunción de literales entonces se dice que la fórmula está en *Forma Normal Disyuntiva* (FND). Toda fórmula en FND puede transformarse a una FNC, y viceversa, sin más que aplicar las leyes de De Morgan y las propiedades distributivas.

Forma Normal de Skolem

Una fórmula bien formada cerrada está en *Forma Normal de Skolem* (FNS) si está en Forma Normal Conjuntiva Prenexa, es decir, en FNC y en FNP y, además, todos los cuantificadores

son universales. Para ello es necesario suprimir los cuantificadores existenciales mediante valores constantes. Es importante ver que la fórmula obtenida en FNS no será necesariamente lógicamente equivalente a la original, pero ambas fórmulas serán equisatisfacibles[a].

Para convertir una fómula cerrada en una FNS se deben buscar los cuantificadores existenciales comenzando por la izquierda. Después, si el cuantificador existencial está al principio de la fórmula, $\exists x\ \mathrm{P}(x)$, se sustituye la variable cuantificada existencialmente por una constante y se elimina el cuantificador, quedando $\mathrm{P}(a)$. Si el cuantificador existencial está precedido por un número de cuantificadores universales,

$$\forall x_1 \ldots \forall x_n \exists y\ \mathrm{P}(x_1, \ldots, x_n, y)$$

entonces se sustituye la variable cuantificada existencialmente por una función de las variables cuantificadas universalmente que le preceden, quedando

$$\forall x_1 \ldots \forall x_n\ \mathrm{P}(x_1, \ldots, x_n, f(x_1, \ldots, x_n))$$

Forma Clausal

Una fórmula cerrada que está en FNS puede convertirse fácilmente en una Forma Clausal sin más que suprimir los cuantificadores universales que están presentes de forma explícita y asumir que estarán presentes de forma implícita. Además, las conjunciones $\wedge$ se sustituyen por comas. Es decir, la fórmula en FNS

$$\forall x \forall y \forall z\ \mathrm{P}(x, y) \wedge \mathrm{Q}(z)$$

quedaría

$$\mathrm{P}(x, y), \mathrm{Q}(z)$$

en Forma Clausal, que como ya se ha dicho su implementación posterior en un sistema basado en reglas es directa. En el siguiente ejemplo pedagógico veremos cómo podríamos pasar desde el planteamiento de una expresión lingüística en un problema médico hasta la Forma Clausal asociada pasando por las formas normales intermedias.

Ejemplo 16.2 (Ejemplo de paso a Forma Clausal)
Supongamos que se nos plantea la siguiente expresión en un problema de oftalmología:

> "Todo paciente con presión intraocular alta o defectos periféricos del campo visual será examinado por un oftalmólogo experto en glaucoma y se le harán todas las pruebas que aporten información sobre el nervio óptico."

El primer paso es establecer unas proposiciones con semántica relacionada con el problema:

$\mathrm{P}(x) = x$ tiene presión intraocular alta.

$\mathrm{V}(x) = x$ tiene defectos periféricos del campo visual.

$\mathrm{O}(x, y) = x$ es examinado por y.

$\mathrm{G}(x) = x$ es un oftalmólogo experto en glaucoma.

$\mathrm{M}(x, y) = x$ pasa la prueba médica y.

$\mathrm{N}(x) = x$ aporta información sobre el nervio óptico.

[a] Dos fórmulas son equisatisfacibles si, y solo si, una fórmula es satisfacible entonces la otra también lo es. Una fórmula es satisfacible si es posible encontrar una interpretación o modelo que la haga verdadera.

De este modo, el problema expresado mediante lenguaje natural se puede expresar mediante lógica simbólica del siguiente modo:

$$\forall x\Big[\big(\mathrm{P}(x)\vee\mathrm{V}(x)\big)\Rightarrow\exists y\big(\mathrm{G}(y)\wedge\mathrm{O}(x,y)\big)\wedge\forall y\big(\mathrm{N}(y)\Rightarrow\mathrm{M}(x,y)\big)\Big]$$

El primer paso para dejar la expresión en Forma Clausal es renombrar las variables comunes a cuantificadores distintos, de este modo, podemos cambiar la variable y que es común a los cuantificadores $\exists$ y $\forall$ por una variable z en el último cuantificador sin modificar el significado de la expresión:

$$\forall x\Big[\big(\mathrm{P}(x)\vee\mathrm{V}(x)\big)\Rightarrow\exists y\big(\mathrm{G}(y)\wedge\mathrm{O}(x,y)\big)\wedge\forall z\big(\mathrm{N}(z)\Rightarrow\mathrm{M}(x,z)\big)\Big]$$

El segundo paso es eliminar las coimplicaciones y las implicaciones empleando las equivalencias apropiadas:

$$\forall x\Big[\neg\big(\mathrm{P}(x)\vee\mathrm{V}(x)\big)\vee\exists y\big(\mathrm{G}(y)\wedge\mathrm{O}(x,y)\big)\wedge\forall z\big(\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\Big]$$

Ahora, se aplican las leyes de De Morgan para conseguir que las negaciones solo afecten a fórmulas atómicas y no a cláusulas compuestas:

$$\forall x\Big[\big(\neg\mathrm{P}(x)\wedge\neg\mathrm{V}(x)\big)\vee\exists y\big(\mathrm{G}(y)\wedge\mathrm{O}(x,y)\big)\wedge\forall z\big(\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\Big]$$

El cuarto paso busca obtener la FNP al pasar todos los cuantificadores al principio de la fórmula obteniendo el prefijo y la matriz de la misma:

$$\forall x\exists y\forall z\Big[\big(\neg\mathrm{P}(x)\wedge\neg\mathrm{V}(x)\big)\vee\big(\mathrm{G}(y)\wedge\mathrm{O}(x,y)\big)\wedge\big(\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\Big]$$

El siguiente paso debe lograr obtener la Forma Normal Conjuntiva Prenexa aplicando la propiedad distributiva:

$$\begin{aligned}\forall x\exists y\forall z\big(&\neg\mathrm{P}(x)\vee\mathrm{G}(y)\big)\wedge\\&\big(\neg\mathrm{P}(x)\vee\mathrm{O}(x,y)\big)\wedge\\&\big(\neg\mathrm{P}(x)\vee\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\mathrm{G}(y)\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\mathrm{O}(x,y)\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\end{aligned}$$

El sexto paso es eliminar los cuantificadores existenciales para dejar la expresión en Forma Normal de Skolem. Para ello se sustituye la variable y del cuantificador existencial por una función de x que es el cuantificador que hay a la izquierda del cuantificador existencial, así:

$$\begin{aligned}\forall x\forall z\big(&\neg\mathrm{P}(x)\vee\mathrm{G}(f(x))\big)\wedge\\&\big(\neg\mathrm{P}(x)\vee\mathrm{O}(x,f(x))\big)\wedge\\&\big(\neg\mathrm{P}(x)\vee\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\mathrm{G}(f(x))\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\mathrm{O}(x,f(x))\big)\wedge\\&\big(\neg\mathrm{V}(x)\vee\neg\mathrm{N}(z)\vee\mathrm{M}(x,z)\big)\end{aligned}$$

Por último, se sustituyen las conjunciones por comas y se asumen como implícitos los cuantificadores universales eliminando su aparición explícita de la fórmula:

$$\big(\neg \mathrm{P}(x) \vee \mathrm{G}(f(x))\big), \big(\neg \mathrm{P}(x) \vee \mathrm{O}(x, f(x))\big), \big(\neg \mathrm{P}(x) \vee \neg \mathrm{N}(z) \vee \mathrm{M}(x, z)\big),$$
$$\big(\neg \mathrm{V}(x) \vee \mathrm{G}(f(x))\big), \big(\neg \mathrm{V}(x) \vee \mathrm{O}(x, f(x))\big), \big(\neg \mathrm{V}(x) \vee \neg \mathrm{N}(z) \vee \mathrm{M}(x, z)\big)$$

El objetivo pedagógico de este ejemplo es ver cómo es posible transformar una expresión lingüística en su versión lógica y cómo esta última se puede transformar en una Forma Clausal. Esto no significa que esta fuera la mejor forma de implementarlo en un Sistema Basado en Reglas, ya que cada posible regla deberá diseñarse conforme a las necesidades de cada momento y de cada usuario. De forma ilustrativa, la expresión anterior se ha convertido en un hecho con cabeza y sin cuerpo. Pero podría haberse convertido en un par de reglas con cuerpo y cabeza donde las expresiones de ambos estuvieran en Forma Clausal, por ejemplo

$$\mathrm{P}(x) \Rightarrow \mathrm{O}(x, f(x)), \mathrm{G}(f(x)), \mathrm{M}(x, z), \mathrm{N}(z)$$

$$\mathrm{V}(x) \Rightarrow \mathrm{O}(x, f(x)), \mathrm{G}(f(x)), \mathrm{M}(x, z), \mathrm{N}(z)$$

16.1.7. Unificación como método de razonamiento

Los sistemas basados en reglas (SBR) están formados por un *conjunto de reglas* (si-entonces), que representa el conocimiento de un problema, y por una *memoria de trabajo* o *base de hechos* que representa una situación particular del dominio del problema. Mediante la unificación de los hechos con los antecedentes de las reglas se puede llevar a cabo el razonamiento deductivo que nos permite obtener conclusiones. La unificación se puede definir como una sustitución de las variables de dos predicados distintos por otras variables que, aplicadas a ambos predicados, hacen que éstos sean iguales. Por ejemplo, la sustitución $\sigma = [x/f(z), y/z]$ unifica las expresiones $\mathrm{P}(x, f(z))$ y $\mathrm{P}(f(y), x)$ al hacerlas ambas iguales a $\mathrm{P}(f(z), f(z))$. En los sistemas basados en reglas existen, al menos, dos métodos de razonamiento que emplean el mecanismo de la unificación para llevar a cabo la inferencia deductiva. El primero es el encadenamiento hacia delante (*forward chaining*, en inglés) y el segundo es el encadenamiento hacia atrás (*backward chaining*).

En el encadenamiento hacia delante se parte de los hechos de la memoria de trabajo para unificar los antecedentes de las reglas de inferencia y, empleando el *modus ponendo ponens* se extraen nuevos hechos, que se interpreta como nuevo conocimiento, hasta llegar al objetivo deseado. Sistemas basados en reglas como CLIPS o Jess funcionan generalmente como sistemas con encadenamiento hacia delante.

En el encadenamiento hacia atrás se parte de los objetivos o hipótesis de trabajo y se recorren las reglas desde el consecuente al antecedente tratando de comprobar, por unificación, si existen hechos en la memoria de trabajo que confirmen el objetivo. El lenguaje de programación Prolog funciona mediante encadenamiento hacia atrás.

Las siguientes secciones están orientadas a sistemas basados en regla con encadenamiento hacia delante ya que los ejemplos están enfocados al uso de sistemas como CLIPS o Jess.

16.2. Sistemas basados en reglas

Como se ha adelantado, los sistemas basados en reglas están formados por un *conjunto de reglas* (si-entonces) y por una *memoria de trabajo*.

La memoria de trabajo contiene la información de la situación particular en forma de afirmaciones (llamados *hechos*). Cada regla del conjunto de reglas es una declaración *si antecedentes, entonces consecuente*, donde la satisfacción de todos los *antecedentes* por la memoria de trabajo,

provoca que la regla se accione. En un *sistema deductivo*, todas las reglas accionadas se disparan, por lo que todos los hechos expresados en los consecuentes de las reglas disparadas se incluyen en la memoria de trabajo. En un *sistema de reacción*, normalmente solo una regla se dispara al mismo tiempo, por lo que únicamente las acciones expresadas en su consecuente son ejecutadas. En los sistemas de reacción, un *procedimiento de resolución de conflictos* (p.e. una lista de prioridades) decide cuál de las reglas accionadas se dispara.

De forma resumida, vemos como podremos deducir nuevos hechos y realizar acciones aplicando las reglas a los hechos de la memoria de trabajo que el sistema tiene en un momento concreto.

Ejemplo 16.3 (Recursos para procedimientos sanitarios)
Queremos diseñar un sistema que recuerde la lista de recursos necesarios para realizar procedientos asociados a ginecología. El objetivo del sistema es comprobar que se dispone de todos los recursos, incluidos aquellos requeridos por otros recursos, antes de realizar el procedimiento, y así evitar preparatorios y encuentros fallidos con la paciente.

Una situación particular relacionada con nuestro sistema podría ser la realización de un procedimiento exudado vaginal a la paciente con identificador 234, que formaría parte de la memoria de trabajo del sistema.

Podemos incluir este hecho en la memoria de trabajo de un motor Jess mediante la siguiente línea:

```
(assert (procedimiento exudado-vaginal paciente234))
```

donde assert incluye el hecho (prcedimiento exudado-vaginal paciente234) en la memoria de trabajo[b].

Para realizar un exudado vaginal se requiere un hisopo, por lo que debemos reservar este recurso para la paciente 234. Para expresar esta condición, podríamos escribir una regla, del estilo *si se realiza el procedimiento exudado vaginal a la paciente 234, entonces asigna el recurso hisopo a la paciente 234*, que sería una de las reglas del conjunto de reglas del sistema. Fácilmente, podemos generalizar mediante variables esta regla para que contemple cualquier procedimiento, paciente y recurso: *Si se realiza el procedmiento ?p a la paciente ?id, y para realizar el procedimiento ?p se usa el recurso ?x, entonces asigna un recurso ?x a la paciente ?id.*

```
(defrule si-usa-material
  (procedimiento ?p ?id)
  (usa ?p ?x)
  =>
  (assert (recurso-asignado ?x ?id))
)
```

Al generalizar la regla, necesitamos un conjunto de hechos que nos digan que recursos se usan en cada procedimiento, por lo que debemos definir tres nuevos hechos con esta información en la memoria de trabajo:

```
(deffacts requisitos-exudados-vaginales
  (usa exudado-vaginal camilla)
  (usa exudado-vaginal hisopo)
  (usa exudado-vaginal tubo)
)
```

[b]Podemos apreciar la sintaxis prefija de las funciones (p.e. (+ 2 3) realiza la suma de los números 2 y 3)

Si ejecutamos nuestro sistema de reglas, encontraremos que tres nuevos hechos han sido incluidos en la memoria de trabajo, debido a la acción de la regla "si-usa-material" sobre el procedimiento exudado-vaginal en la paciente 234.

```
(recurso-asignado tubo paciente234)
(recurso-asignado hisopo paciente234)
(recurso-asignado camilla paciente234)
```

Podemos observar que los consecuentes de las reglas disparadas pueden incluir nuevos hechos en la memoria de trabajo, que provocan nuevas satisfacciones de reglas y por lo tanto nuevos disparos.

Cuando el encadenamiento sucesivo de reglas se realiza desde los antecedentes hacia los consecuentes, estamos aplicando un razonamiento hacia delante, y lo denominamos encadenamiento progresivo. Este razonamiento nos lleva desde los datos del problema (entrada) a la solución (salida), y es adecuado cuando se dispone de todos los hechos de entrada y cuando se sabe que la cantidad de posibles soluciones (conclusiones) es pequeña.

Ejemplo 16.4 (Recursos para procedimientos sanitarios (cont. ejemplo 16.3)) Profundizando en el problema, vemos que para usar el hisopo durante el exudado vaginal, se requiere un espéculo, entonces, también deberíamos reservar uno para la paciente. De la misma forma, un tubo, que necesitaremos para depositar la muestra, también requiere una pipeta y suero fisiológico para su correcta utilización en el exudado vaginal, por lo que también debemos reservar estos dos elementos. Observamos, pues, que tenemos una regla de requisitos entre recursos, que podemos expresar de forma genérica como: *si un recurso ?x se ha asignado a la paciente ?id y este recurso ?x requiere un otro recurso ?y, entonces debe asignarse el recurso ?y a la paciente ?id* y definir un conjunto de hechos con los requisitos particulares de los recursos de nuestro problema:

```
(deffacts requisitos-recursos
  (requiere hisopo especulo)
  (requiere tubo pipeta)
  (requiere frasco pipeta)
  (requiere tubo suero-fisiologico)
)

(defrule si-requiere-asigna
  (recurso-asignado ?x ?id)
  (requiere ?x ?y)
  =>
  (assert (recurso-asignado ?y ?id))
)
```

La ejecución del sistema de reglas ahora aplicará el encadenamiento progresivo para reservar suero fisilógico, pipeta por requerirse para usar el tubo y espéculo para usar el hisopo.

```
(recurso-asignado suero-fisiologico paciente234)
(recurso-asignado pipeta paciente234)
(recurso-asignado especulo paciente234)
```

Algunos motores de reglas también permiten comprobar el cumplimiento de hipótesis. Para ello, el sistema encadena hacia atrás (*encadenamiento regresivo*) las reglas hasta llegar a aceptar la hipótesis si las entrada lo corroboran, y rechazarla si no se dan las condiciones. Cuando el

problema a resolver es tal que un conjunto de entradas puede resultar en muchas conclusiones (alta amplitud de salida), entonces el encadenamiento regresivo es útil. Este encadenamiento también es de interés cuando no todas las entradas están disponibles al mismo tiempo, pero es suficiente saber si una conclusión es verdadera.

Algunos sistemas para el cálculo de riesgos se basan en reglas establecidas por los expertos por lo que resulta relativamente fácil extraer dicho conocimiento en forma de reglas para calcular el riesgo asociado a la situación del paciente.

Ejemplo 16.5 (Cálculo del riesgo de padecer un suceso adverso por diabéticos)
Se quiere diseñar un sistemas de alertas para diabéticos en su hogar, basado en el cálculo del riesgo de padecer un suceso adverso dado su estilo de vida y datos biomédicos de fácil obtención.

Así por ejemplo, podemos diseñar un sistema que recoja datos biomédicos como el tipo de diabetes, el hábito tabáquico, los niveles de colesterol LDL y HDL, triglicéridos, Presión Arterial Sistólica (PAS) y cantidad de hemoglobina glucosilada en la sangre (HbA1c), y convertir el conocimiento médico, como por ejemplo *Hiperglucemia: asociada a un aumento de la morbimortalidad por enfermedad cardiovascular. Por cada 1 % de disminución de la HbA1c, se reduce el riesgo de muerte asociada a diabetes un 14 %*, a una regla como

```
(defglobal ?*HbA1cMaxRecomendado* = 6)
(defrule hiperglucemia_HbA1c
  ?p <- (person {HbA1c < ?*HbA1cMaxRecomendado*})
  =>
  (bind ?*riesgo* (- ?*riesgo* (* (- ?*HbA1cMaxRecomendado* ?p.HbA1c) 0.14)))
  (printout t "Riesgo reducido por porcentaje de hemoglobina
glucosilada en sangre " ?*riesgo* crlf)
)
```

donde el índice riesgo, que acumulamos en una variable global, disminuye un 14 % por cada valor de HbA1c menor de 6.

Veremos a continuación el diseño de un CPOE que tiene en cuenta las contraindiaciones por interacción de medicamentos, alergias, enfermedades, y dosis. El sistema está basado en reglas hace uso de la prioridad para resolver conflictos entre reglas, y utiliza una estructura terminológica para resolver las interacciones entre medicamentos.

Ejemplo 16.6 (CPOE)
Se desea diseñar un CPOE que trabaje de forma permanente, admitiendo la entrada concurrente de prescripciones, y que actúe de filtro de las mismas, deteniendo aquellas que presenten alguna contraindicación. Las contraindicaciones de un medicamento pueden ser debidas a varios motivos:

1. La cantidad de dosis total administrada al paciente según su edad.
2. La interacción de alguno de sus componentes con los componentes de los medicamentos que actualmente toma el paciente.
3. La alergia de paciente a algún componente.
4. La incompatibilidad de algún componente del medicamento con los síntomas o enfermedades del paciente.

Tomaremos como caso de estudio la administración de paracetamol en una presentación de comprimidos de 500 mg, y para ello estudiamos las contraindicaciones y dosimetría que aparecen en su prospecto.

En primer lugar definimos las plantillas de hechos, ya que la complejidad del problema hace preferible utilizar estructuras que hechos ordenados.

La persona estará definida por una plantilla que incluya un identificador, la edad (para saber la dosis), y listas (campos multislot) de alergias, enfermedades y síntomas.

```
(deftemplate persona
  (slot ID)
  (slot edad)          ;;
  (multislot alergias)     ;;lista de alergias del paciente
  (multislot enfermedades)    ;;lista de alimentos actualmente ingeridos
  (multislot sintomas)     ;;lista de sintomas actuales del paciente
)
```

Mantendremos las dosis de los principios activos administradas durante el día a las personas en hechos de la siguiente plantilla:

```
(deftemplate persona-componenteActivo-dosis
  (slot persona)
  (slot componenteActivo)
  (slot dosis))
```

Los medicamentos estarán representados por un identificador, un componente activo, una lista de componentes, una presentación en forma de texto y una lista de indicaciones.

```
(deftemplate medicamento
  (slot ID)
  (slot componenteActivo)
  (multislot componentes)
  (multislot presentacion)
  (multislot indicacion)
)
```

En un sistema real, estos hechos deberían alimentarse a través de un Vademecum, en nuestro ejemplo definimos dos presentaciones de paracetamol.

```
(deffacts Vademucum "Medicamentos en la Farmacia"
  (medicamento (ID Mundogen500mgComprimidosEFG)
          (componenteActivo paracetamol)
       (componentes paracetamol almidon-pregelatimizado povidona acido-estearico)
       (presentacion comprimido 500)
       (indicacion fiebre dolor-muscular dolor-cabeza dolor-intensidad-leve
dolor-intensidad-moderada)
  )

  (medicamento (ID Termalgin650mgComprimidos)
          (componenteActivo paracetamol)
       (componentes paracetamol talco almido-maiz silice coloidal-anhidra
celulosa-microcristalina almidon-pregelatimizado povidona acido-estearico)
       (presentacion comprimido 650)
       (indicacion fiebre dolor-muscular dolor-cabeza dolor-intensidad-leve
dolor-intensidad-moderada)
  )
)
```


La prescripción de un medicamento va dirigida a una persona y tiene asociado un valor numérico que indica la dosis.

```
(deftemplate prescripcion
  (slot ID)
  (slot persona)
  (slot medicamento)
  (slot dosis)
)
```

y las acciones que el sistema reactivo realizará serán la detención de una prescripción de un medicamento a una persona, o la administración de una dosis de componente activo, que puede ser nueva para el paciente o la continuación del tratamiento.

```
(deffunction accion-detener (?pr ?p ?m ?t ?i)
  (printout t "Detener prescripcion " ?pr.ID " de " ?m " a " ?p " por " ?t " de "
?i "." crlf)
  (retract ?pr)
)
(deffunction accion-administrar-nuevo (?pr ?p ?m ?c ?d)
  (printout t "Administar prescripcion " ?pr.ID ": " ?d " dosis de " ?m " a "
?p "." crlf)
  (assert (persona-componenteActivo-dosis (persona ?p) (componenteActivo ?c)
(dosis ?d)))
  (retract ?pr)
)
(deffunction accion-administrar-continuacion (?pr ?p ?m ?c ?d ?pcd)
  (printout t "Administar prescripcion " ?pr.ID ": " ?d " dosis de " ?m " a "
?p "." crlf)
  (modify ?pcd (dosis (+ ?d ?pcd.dosis)))
  (retract ?pr)
)
```

Un componente tendrá que estar identificado de forma única, y especificar las dosis máximas para adultos y niños, además de las contraindicaciones por enfermedades, interacciones y síntomas.

```
(deftemplate componente
  (slot ID)
  (multislot es-un)
  (slot dosisMaximaAdultos)
  (slot dosisMaximaNinyos)
  (multislot contraindicacion-enfermedad)
  (multislot contraindicacion-interaccion)
  (multislot contraindicacion-sintoma)
)
```

Como vemos en la plantilla componente, hemos incluido un atributo es-un. Efectivamente, los componentes están clasificados en familias jerárquicas, como por ejemplo la aspirina, que es un antinflamatorio no esteroide, y por lo tanto es un analgésico. Por lo tanto, los componentes de nuestra memoria de trabajo pueden provenir de una terminología farmacológica para establecer

las relaciones es-un que permitan buscar interacciones entre componentes, directamente o por parentesco, por ejemplo, el paracetamol no debe interactuar con otros analgésicos, por lo tanto, si un paciente tiene como medicación actual ibuprofeno, no debe ser medicado con paracetamol.

En nuestro ejemplo especificaremos la terminología de los analgésicos y completaremos los atributos de paracetamol y unas contraindicaciones ficticias de síntomas incompatibles con los analgésicos.

```
(componente (ID amina)
    (es-un analgesico)
    (contraindicacion-sintoma empeoramiento dolorNinyo5Dias dolorAdulto10Dias
fiebre3Dias fiebre3Dias)) ;;fiebre3Dias para que se propague a paracetamol

(componente (ID antinflamatorio-no-esteroide) (es-un analgesico))
(componente (ID cannabinoide) (es-un analgesico))
(componente (ID opioide) (es-un analgesico))

(componente (ID fenacetina)(es-un amina))
(componente  (ID paracetamol) (es-un amina)
     (contraindicacion-interaccion analgesico)
     (contraindicacion-enfermedad higado renal cardiaco pulmonar anemia
alcoholismo-menor alcolismo-mayor embarazo lactancia)
     (contraindicacion-sintoma empeoramiento dolorNinyo5Dias dolorAdulto10Dias)
     (dosisMaximaAdultos 8)
     (dosisMaximaNinyos 5)
)

(componente (ID aspirina)  (es-un antinflamatorio-no-esteroide))
(componente (ID celecoxib)(es-un antinflamatorio-no-esteroide))
(componente (ID diclofenaco)(es-un  antinflamatorio-no-esteroide))
(componente (ID ibuprofeno)(es-un  antinflamatorio-no-esteroide))
(componente (ID ketoprofeno)(es-un  antinflamatorio-no-esteroide))
(componente (ID ketorolaco)(es-un  antinflamatorio-no-esteroide))
(componente (ID meloxicam)(es-un  antinflamatorio-no-esteroide))
(componente (ID naproxeno)(es-un  antinflamatorio-no-esteroide))
(componente (ID rofecoxib)(es-un  antinflamatorio-no-esteroide))
(componente (ID indometacina)(es-un  antinflamatorio-no-esteroide))

(componente (ID cannabis)(es-un cannabinoide))
(componente (ID tetrahidrocannabinol)(es-un cannabinoide))

(componente (ID alfentanilo)(es-un opioide))
(componente (ID carfentanilo)(es-un opioide))
(componente (ID buprenorfina)(es-un opioide) )
(componente (ID codeina)(es-un opioide) )
(componente (ID codeinona)(es-un opioide))
(componente (ID dextropropoxifeno)(es-un opioide) )
(componente (ID dihidrocodeina)(es-un opioide) )
(componente (ID beta-endorfina)(es-un opioide) )
(componente (ID fentanilo)(es-un opioide) )
(componente (ID heroina)(es-un opioide) )
```



```
(componente (ID hidrocodona)(es-un opioide))
(componente (ID hidromorfona)(es-un opioide))
(componente (ID metadona)(es-un opioide) )
(componente (ID morfina)(es-un opioide) )
(componente (ID morfinona)(es-un opioide))
(componente (ID oxicodona)(es-un opioide) )
(componente (ID oximorfona)(es-un opioide) )
(componente (ID meperidina)(es-un opioide) )
(componente (ID remifentanilo)(es-un opioide))
(componente (ID sufentanilo)(es-un opioide) )
(componente (ID tebaina)(es-un opioide) )
(componente (ID tramadol)(es-un opioide))
)
```

Para no administrar más dosis diaria de un medicamento a un paciente, declaramos dos reglas, una para comprobar la administración de su componente activo no se supera la dosis máxima diaria en adultos la otra para la misma condición en niños.

```
(defrule contraindicacion-dosis-adulto
  ?pr <- (prescripcion (persona ?p) (medicamento ?m)(dosis ?d))
  (persona (ID ?p) {edad > 11})
  (medicamento (ID ?m) (componenteActivo ?ca))
  (componente (ID ?ca) (dosisMaximaAdultos ?d1))
  (persona-componenteActivo-dosis (persona ?p) (componenteActivo ?ca) (dosis ?dh))
;;dosis historica que lleva el paciente
  (test (> (+ ?d ?dh) ?d1))
  =>
  (accion-detener ?pr ?p ?m dosisAcumulada (+ ?d ?dh))
)

(defrule contraindicacion-dosis-ninyo
  ?pr <- (prescripcion (persona ?p) (medicamento ?m)(dosis ?d))
  (persona (ID ?p) {edad <= 11})
  (medicamento (ID ?m) (componenteActivo ?ca))
  (componente (ID ?ca) (dosisMaximaNinyos ?d1))
  (persona-componenteActivo-dosis (persona ?p) (componenteActivo ?ca) (dosis ?dh))
;;dosis historica que lleva el paciente
  (test (> (+ ?d ?dh) ?d1))
  =>
  (accion-detener ?pr ?p ?m dosisAcumulada (+ ?d ?dh))
)
```

Para detener la prescripción si el paciente es alérgico a algún componente del medicamento, hemos de comprobar si alguno de ellos está en la lista de alergias del paciente. Esto puede realizarse comprobando que la intersección de ambas listas es distinto de cero, en cuyo caso, debe detenerse la prescripción.

```
(defrule contraindicacion-alergia-componente
  (persona (ID ?p) (alergias $?a))
  (medicamento  (ID ?m) (componentes $?c))
  ?pr <- (prescripcion (medicamento ?m) (persona ?p))
```

```
  (test (> (length$ (intersection$ ?a ?c)) 0))
  =>
  (accion-detener ?pr ?p ?m alergia (intersection$ ?a ?c))
)
```

Con esta regla, no estaríamos cubriendo la totalidad de alergias del paciente. Supongamos que el paciente es alérgico a todos los analgésicos, sabemos que el paracetamol es un analgésico, sin embargo, la lista de componentes del medicamento no contiene el término analgésico, sino paracetamol. Es necesario añadir a la lista de alergias del paciente todos los componentes que estén relacionados mediante "es-un" con todo término incluido en la lista.

```
(defrule extiende-alergias-terminologia
  (componente (ID ?z) (es-un ?x))
  ?p <- (persona (alergias $?a ?x $?b))
  (not (test (member$ ?z ?a)))
  (not (test (member$ ?z ?b)))
  =>
  (modify ?p (alergias ?a ?x ?b ?z))
;;  (printout t ?p.ID " " ?p.alergias crlf)
)
```

Igualmente con las contraindicaciones por interacción de un componente, sí especificado por un ancestro (p.e. amina), debe extenderse a todos sus descendientes (p.e. fenacetina y paracetamol), excepto si el componente a incluir en la lista de interacción es el propio componente.

```
(defrule extiende-interaccion-terminologia
  (componente (ID ?z) (es-un ?x))
  ?c <- (componente (ID ?id)(contraindicacion-interaccion $?a ?x $?b))
  (not (test (member$ ?z ?a)))
  (not (test (member$ ?z ?b)))
  (not (test (eq ?z ?id)))
  =>
  (modify ?c (contraindicacion-interaccion ?a ?x ?b ?z))
)
```

Y de forma similar, las posibles enfermedades y síntomas.

```
(defrule extiende-enfermedad-terminologia
  (componente (ID ?z) (es-un ?x))
  (componente (ID ?x)(contraindicacion-enfermedad $?a ?e $?b))
  ?c <- (componente (ID ?z)(contraindicacion-enfermedad $?e2))
  (not (test (member$ ?e ?e2)))
  =>
  (modify ?c (contraindicacion-enfermedad ?e2 ?e))
)

(defrule extiende-sintoma-terminologia
  (componente (ID ?z) (es-un ?x))
  (componente (ID ?x)(contraindicacion-sintoma $?a ?e $?b))
  ?c <- (componente (ID ?z)(contraindicacion-sintoma $?e2))
  (not (test (member$ ?e ?e2)))
  =>
```

```
  (modify ?c (contraindicacion-sintoma ?e2 ?e))
)
```

Lo que nos permite definir sin miedo a dejarnos interacciones, enfermedades o síntomas con las contraindicaciones, las siguientes reglas:

```
(defrule contraindicacion-componente-enfermedad
  (componente (ID ?c) (contraindicacion-enfermedad $?ce))
  (medicamento (ID ?m) (componentes $? ?c $?))
  ?pr <- (prescripcion (persona ?p) (medicamento ?m))
  (persona (ID ?p) (enfermedades $?e))
  (test (> (length$ (intersection$ ?e ?ce)) 0))
  =>
  (accion-detener ?pr ?p ?m enfermedad (intersection$ ?e ?ce))
)

(defrule contraindicacion-componente-sintoma
  (componente (ID ?c) (contraindicacion-sintoma $?cs))
  (medicamento (ID ?m) (componentes $? ?c $?))
  ?pr <- (prescripcion (persona ?p) (medicamento ?m))
  (persona (ID ?p) (sintomas $?s))
  (test (> (length$ (intersection$ ?s ?cs)) 0))
  =>
  (accion-detener ?pr ?p ?m sintoma (intersection$ ?s ?cs))
)

(defrule contraindicacion-componente-interaccion
  (persona-componenteActivo-dosis (persona ?p) (componenteActivo ?c2))
  ?pr <- (prescripcion (persona ?p) (medicamento ?m))
  (medicamento (ID ?m) (componentes $? ?c $?))
  (componente (ID ?c) (contraindicacion-interaccion $? ?c2 $?))
  =>
  (accion-detener ?pr ?p ?m interaccion ?c2)
)
```

Con estas reglas, el sistema puede detener las prescripciones que no deben llevarse a cabo por contraindicaciones. La acción de detener, implica a su vez eliminar el hecho prescripción de la memoria de trabajo. La resolución de conflictos por prioridad, nos permite implementar la administracion de medicamentos de forma elegante, ya que podemos definir reglas de administración de baja prioridad que se ejecuten sobre las prescripciones que no satisfacen las reglas de contraindicación. Así pues, definiremos una regla de continuación de prescripción, de baja prioridad (p.e. -99), que administrará y aumentará la dosis acumulada en el paciente del componente activo prescrito.

```
(defrule prescripcion-correcta-continuacion
  (declare (salience -99))
  ?pr <- (prescripcion (persona ?p) (medicamento ?m)(dosis ?d))
  (medicamento (ID ?m) (componenteActivo ?c))
  ?pcd <- (persona-componenteActivo-dosis (persona ?p) (componenteActivo ?c)
(dosis ?dh))
  =>
```

```
  (accion-administrar-continuacion ?pr ?p ?m ?c ?d ?pcd)
)
```

Finalmente, una regla de menor prioridad (-100) que administrará un medicamento no prescrito al paciente anteriormente, creando un hecho "persona-componenteActivo-dosis" en la memoria de trabajo para inicializar las dosis administradas al paciente.

```
(defrule prescripcion-correcta-nueva
  (declare (salience -100))
  ?pr <- (prescripcion (persona ?p) (medicamento ?m)(dosis ?d))
  (medicamento (ID ?m) (componenteActivo ?c))
  =>
  (accion-administrar-nuevo ?pr ?p ?m ?c ?d)
)
```

Ante un paciente como

```
(persona (ID Juan)
  (edad 32)
  (alergias pescado penicilina)
  (enfermedades renal)
  (sintomas fiebre)
)
```

se detendría una prescripción de paracetamol por enfermedad renal.

```
(prescripcion (ID Juan-20100505-1001)
  (persona Juan)
  (medicamento Mundogen500mgComprimidosEFG)
  (dosis 1)
)
```

A un paciente con alergia a las aminas, se detendría una prescripción de paracetamol, o a un paciente de 32 años se le permitiria tomar dosis sucesivas de paracetamol siempre que no superasen la dosis máxima acumulada para adultos.

Ejemplo 16.7 (Mycin)
Existen varios casos de uso de sistemas basados en reglas en el ámbito clínico. Uno de ellos es Mycin[c] es un sistema experto que fue adaptado para diagnóstico médico que realiza preguntas al usuario para llegar a conclusiones a través del encadenamiento.

16.3. Razonamiento semántico sobre ontologías

Una ontología, desde el punto de vista informático, es una formulación de un esquema conceptual de un dominio. Así pues, una ontología describe los conceptos de un dominio, sus individuos y las relaciones entre los mismos. La especificación formal de la ontología hace viable el razonamiento semántico sobre la misma, obteniendo implicaciones mediante los encadenamientos sobre las instancias y sus proiedades.

Web Ontology Language (OWL) es un lenguaje de marcas condificado en eXtensible Markup Language (XML) y construido sobre Resource Description Framework (RDF) para especificar

[c] `http://lazax.com/software/Mycin/mycin.html`, `http://www.cs.utexas.edu/users/novak/tmycin.html`

ontologías. OWL por si solo no es un lenguaje de reglas, por lo que se creó A Semantic Web Rule Language (SWRL) (`http://www.w3.org/Submission/SWRL`) para combinar OWL-Lite y OWL-DL con el lenguaje de reglas Rule Markup Language (RuleML). Como resultado, SWRL es un lenguaje de reglas del tipo antecedente-consecuente con capacidad inferencial sobre OWL-DL.

OWL y SWRL son lenguajes de especificación, por lo que necesitaremos un motor de razonamiento, como Jess y Pellet, para ejecutar o validar lo que describimos con ellos.

La herramienta Protégé-OWL [153] permite la creación y explotación de de ontologías mediante una intefaz gráfica para la creación de clases, propiedades e instancias. Además, Protégé permite la incorporación de plugins, para extender sus funcionalidades mediante el mismo interfaz de la herramienta. El plugin SWRLTab añade una nueva ventana a Protégé para la edición de reglas en SWRL que incluyan las clases, instancias y propiedades de la ontología del proyecto que editamos. Por si solo, el plugin SWRLTab no ejecuta un razonador semántico, sino que serán otros plugins, como SWRLJessTab, los que invoquen al razonador, convirtiendo el conocimiento OWL y las reglas SWRL a la base de conocimiento y reglas del razonador, ejecutando el razonador, y transfiriendo los hechos inferidos por el razonador como axiomas del modelo OWL.

Ejemplo 16.8 (CPOE en OWL+SWRL)
Implementaremos una pequeña parte de nuestro ejemplo 16.6 como ontología en OWL y reglas SWRL. Concretamente, únicamente extenderemos las alergias de una persona a componentes de medicamentos mediante la taxonomia de analgésicos. Utilizaremos para ello la herramienta protégé, que nos abstraerá de la sintaxis de los lenguajes, sin pérdida de control sobre nuestros modelos conceptuales.

Restringiremos nuestra ontología de dominio a dos clases: Componente y Persona, que serán los mismos conceptos de los templates en Jess del ejemplo 16.6 (ver figura 16.1).

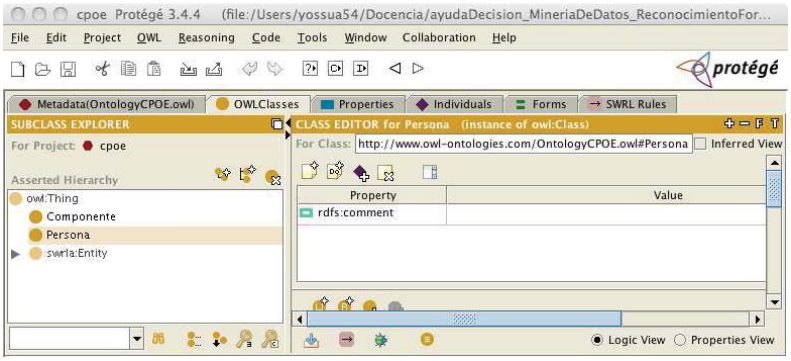


Figura 16.1: Clases Compnente y Persona en Protégé-OWL.

A continuación definiremos dos propiedades: es-un, con dominio y rango Componente; y alergia-a con dominio Persona y rango Componente (ver figura 16.2).

Poblaremos nuestra ontología con la taxonomia de analgésico descrita en la página 284 y con dos Personas: Juan con alergia a las aminas (ver figura 16.3) y Pedro con alergia al cannabis.

Siguiendo la taxonomia de los analgésicos, una persona con alergia a las aminas, tendrá alergia al paracetamol y a la fenacetina. Podremos inferir la lista completa de alergias de Juan y Pedro aplicando la siguiente regla lógica (ver figura 16.4):

```
Persona(?p) ∧ Componente(?c) ∧ alergica-a(?p, ?c) ∧ es-un(?c2, ?c)
→  alergica-a(?p, ?c2)
```

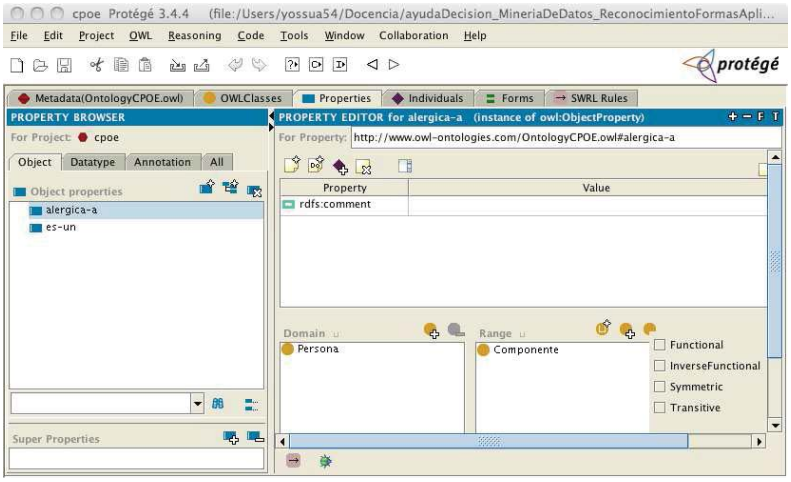


Figura 16.2: Propiedades es-un y alergia-a en Protégé-OWL.

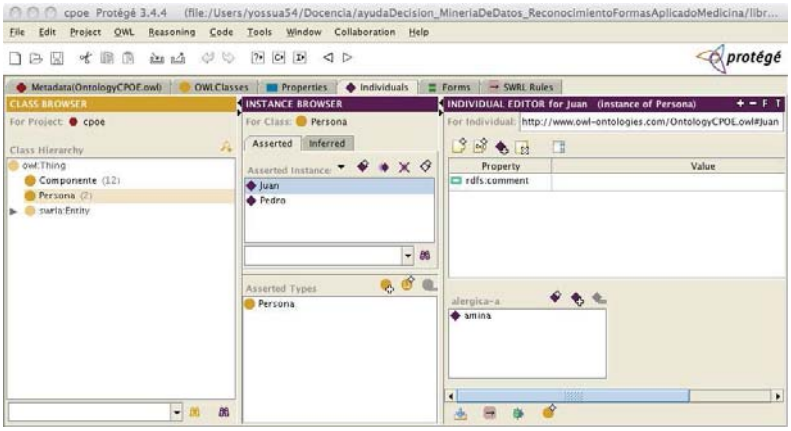


Figura 16.3: Instancia Juan de la clase Persona con alergia-a amina.

La ejecución de un razonador (ver figura 16.4), como Jess[d], nos dará como resultado dos nuevas instancias de la propiedad alergia-a asociadas a Juan: alergia-a paracetamol y alergia-a fenacetina (ver figura 16.5).

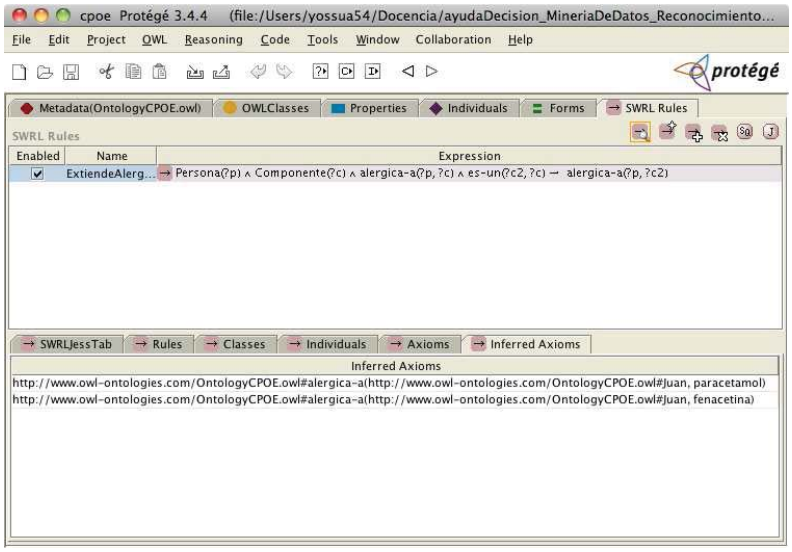


Figura 16.4: Especificación en SWRL y ejecución en Jess de la regla para extender las alergias con la taxonomia de los analgésicos.

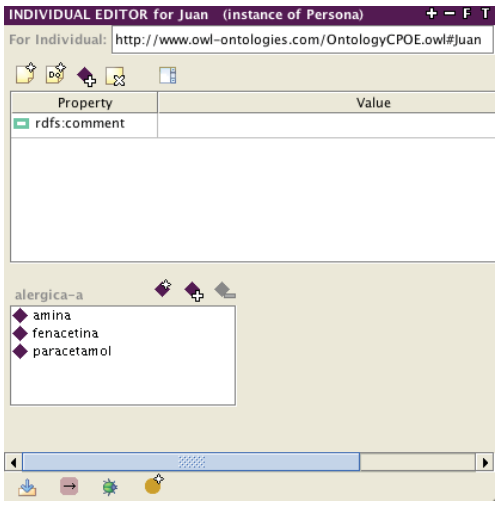


Figura 16.5: Resultado de la ejecución de la regla de inferencia para extender alergias a los analgésicos.

[d]Deberemos incluir jess.jar en el directorio *plugins/edu.stanford.smi.protegex.owl* para poder invovar al razonador desde Protégé

16.4. Lenguajes de guías clínicas

Un factor fundamental de éxito en la implantación de los CDSS es su ajuste al proceso asistencial del entorno donde actúa. Así pues, el conocimiento suministrado por los CDSS debe trasmitirse a los profesionales adecuados en tiempo y forma. Una herramienta interesante para adaptarse al proceso asistencial son las Guía de Práctica Clínicas (o simplemente Guías Clínicas). Una Guía Clínica es un documento que recoge directrices elaboradas sistemáticamente para asistir a clínicos y a pacientes en la toma de decisiones sobre la atención sanitaria adecuada para problemas clínicos específicos. En los últimos 25 años podemos encontrar ejemplos de sistemas de información que implementan Guías Clínicas. Entre otros mencionamos T-HELPER, DILEMMA, EON, Asgaard, PROforma, el proyecto europeo PRESTIGE, y el proyecto británico PRODIGY.

Sin duda los dos proyectos americanos más relevantes en la definición de estándares de Guías Clínicas son GLIF y SAGE. GLIF especifíca formalmente guías clínicas. Su versión 3 incorpora el estándar de mensajería de HL7 y con la intención de ser un estándar de Intercambio de Guías Clínicas. El proyecto SAGE incorpora estándares como HL7, junto con terminologías controladas ya existentes como LOINC y SNOMED sobre los conceptos desarrollados en GLIF. Más aún, investiga las relaciones entre estos estándares y los requisitos para completar el ciclo de vida completo de implementación de una guía clínica. En este sentido, presta gran atención al intercambio de conocimiento con la Historia Clínica Electrónica. Por desgracia, estos dos grandes proyectos no han tenido una continuidad, ni tampoco se ha dejado libres sus motores de ejecución.

En Europa, el lenguaje con mayor proyección ha sido PROforma. PROforma ha sido aplicado en herramientas clínicas como Tallys, para la asistencia de reuniones multidisciplinares de seguimiento de cáncer de mama en UK.

Un razonador semántico con niveles de evidencia científica y bibliografía asociada a las reglas que implementa, junto con un control del workflow puede implementar fácilmente un razonador semántico bastante flexible.

16.5. Notas bibliográficas

La explicación de los sistemas basados en reglas ha sido mayoritariamente extraída de los textos [154]. Los ejemplos han sido diseñados siguiendo los prospectos de medicamentos como Mundogen 500mg, guías de actuación contra la diabetes y manuales de toma de muestras en laboratorios clínicos [155]. Jess ofrece un manual para el programador suficiente para desarrollar todos los ejemplos del capítulo. La información de Protégé ha sido de gran ayuda para la explicación del ejemplo de OWL+SWRL.

Capítulo 17

Diseño de sistemas de ayuda a la decisión médica

El objetivo final de un CDSS es proporcionar conocimiento específico a los usuarios para tomar decisiones médicas asociadas a la salud de cada paciente. Como vimos en la sección 1.3, los CDSS pueden ser utilizados para uso primario o secundario como sistemas de alerta y/o recordatorio de situaciones de salud de pacientes concretos o poblaciones, ayuda al diagnóstico y al pronóstico, gestión de patologías crónicas, triaje, control de la calidad asistencial, gestión de costes temporales y recursos, planificación de riesgos, control de la calidad de biobancos multicéntricos, interpretación de señales biomédicas, búsqueda de marcadores discriminantes, búsqueda de casos similares, búsqueda de información bibliográfica relacionada con el paciente, búsqueda de patrones anómalos, simulaciones de escenarios, control de alertas poblacionales y aprendizaje basado en casos de personal sanitario.

La aproximación planteada en este libro para el diseño de un CDSS está basada en técnicas de Inteligencia Artificial (IA) y de Aprendizaje Automático (AA). Los capítulos 15, 16 y 8 profundizan en las tecnologías para desarrollar un modelo de conocimiento que resuelva el problema médico planteado al CDSS. En este capítulo analizaremos los elementos clave que rodean un modelo de conocimiento para hacer que un CDSS sea operativo en un entorno asistencial.

17.1. El modelo de conocimiento

Un símil de las aplicaciones informáticas con el lenguaje verbal situaría los CDSS en el nivel semántico, ya que requiere la "interpretación" del significado de los conceptos médicos para realizar la funcionalidad requerida. Por lo tanto, plantear un sistema genérico con características de CDSS, más allá de una aplicación *adhoc*, tendrá como núcleo un *modelo de conocimiento médico* que represente la semántica asociada al problema abordado. En este libro planteamos tres tecnologías complementarias para la construcción de modelos de conocimiento médico. En el capítulo 15 se estudia el uso de almacenes de datos y procesamiento en línea como base de conocimiento de un sistema de alertas poblacionales. En el capítulo 16 se estudian los motores de inferencia para implementar conocimiento médico en forma de reglas lógicas. Por último, el capítulo 8 estudia teoría de la decisión y aprendizaje automático como métodos supervisados de construcción de modelos probabilísticos para la predicción de diagnóstico, pronóstico y tratamiento de pacientes.

La implementación del núcleo de un CDSS genérico debe disponer de:

- Un motor de ejecución de modelos de conocimiento.

  El motor debe ser suficientemente versátil para ejecutar diferentes tipos de modelos de

conocimiento. Generalmente, los motores de ejecución de modelos de conocimiento estarán basados en un motor de inferencia (capítulo 16) para CDSS de nivel III y en motores de modelos predictivos (capítulo 8) para CDSS de nivel IV.

- *Plugins* de técnicas de modelos de conocimiento.

  Como se estudió en el capítulo 8, existen diversas técnicas de clasificación y regresión en los que se pueden basar los modelos predictivos: modelos lineales, cuadráticos, Support Vector Machines (SVM), Artificial Neural Networks (ANN), etcétera. Los *plugins* permiten incluir nuevas técnicas a un motor de ejecución de modelos, pudiendo actualizarse las capacidades del CDSS conforme es necesario.

- Lectura de ficheros de especificación de modelos de conocimiento.

  El esquema formado por el motor de ejecución y los *plugins* de técnicas de modelos de conocimiento se completa con la especificación de los modelos mediante ficheros formales legibles en tiempo de ejecución por el CDSS. Esto permite utilizar un mismo motor para múltiples problemas médicos y además actualizar los modelos una vez desplegados en el entorno médico.

- Identificación del problema médico independiente del modelo de conocimiento.

  Un sistema que identifique, de manera independiente, los problemas médicos a resolver y los diferentes modelos de conocimiento que lo resuelven permite 1) el uso versátil de los CDSS; 2) diferentes versiones de los modelos de conocimiento que resuelven un mismo problema, útil para auditorías posteriores; 3) la actualización de los modelos predictivos tanto de forma manual como automática. Por ejemplo, durante el estudio con resonancia magnética de un posible tumor cerebral, el radiólogo puede tener la hipótesis de que la masa estudiada es un glioblastoma. En tal caso, el radiólogo podría estar interesado en todos los modelos predictivos que clasifiquen glioblastoma con el resto de tipos de tumores. Sin embargo, si el radiólogo estuviera interesado en un diagnóstico diferencial entre glioblastoma y astrocitoma anaplástico, entonces el CDSS debería mostrar únicamente soluciones generadas por modelos de este problema dicotómico.

17.1.1. Metodología para el modelado del conocimiento médico

La consolidación de las tecnologías de ingeniería del conocimiento y aprendizaje automático durante las últimas décadas impulsó la creación de una metodología para el desarrollo de proyectos de minería de datos. Este modelo metodológico, conocido como CRISP-DM[a] [156], fue concebido en 1996 por los líderes del momento en el mercado de minería de datos y estaba enfocado principalmente a estandarizar el desarrollo de estas aplicaciones en la industria y las organizaciones comerciales. Aunque pretende ser una metodología genérica, los problemas de minería de datos biomédicos presentan algunos matices y peculiaridades que deberían tenerse en cuenta para poder encajar en dicho modelo metodológico. En esta sección se presentará, por tanto, una adaptación de dicha metodología al campo biomédico.

Cualquier metodología de minería de datos debería incorporar al menos las siguientes seis fases: i) análisis de la organización y el problema a resolver, ii) análisis de los datos, iii) preparación de los datos, iv) modelado, v) evaluación y vi) implantación de las soluciones. Además, las distintas fases del proceso de minería de datos deben ser interdependientes, de modo que el resultado de una fase alimente a la siguiente o realimenten a las anteriores con el fin de revisar

[a] CR*oss*-I*ndustry* S*tandard* P*rocess for* D*ata* M*ining*.

posibles errores. Este ciclo queda ilustrado en la figura 17.1, donde las flechas indican las dependencias más importantes entre fases y el círculo exterior simboliza el ciclo natural del proceso de minería de datos donde lo aprendido a lo largo de todo el proceso puede servir como experiencia para futuros proyectos de minería de datos.

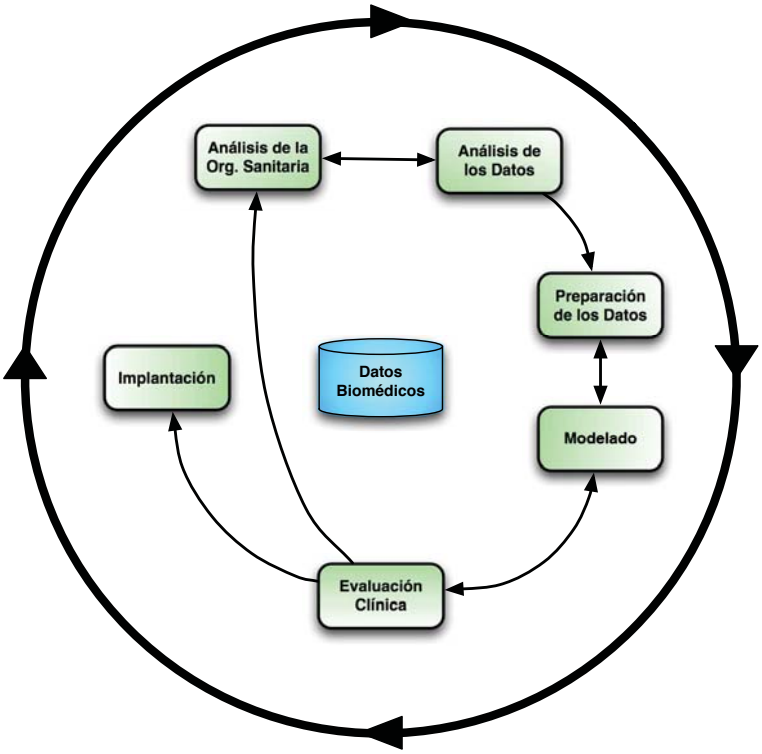


Figura 17.1: La minería de datos biomédicos puede basarse, con matices, en la metodología CRISP-DM [156] para proyectos de minería de datos.

Análisis de la organización sanitaria y el problema a resolver

El análisis de la organización sanitaria es quizás la fase más importante ya que el resto de fases dependen de una correcta comprensión de los objetivos del proyecto. Por lo tanto, es imprescindible determinar correctamente los objetivos del proyecto y asegurar que no se obtendrán resultados que respondan correctamente a la pregunta equivocada. Esto incluye detectar el tipo de problema ante el que nos encontramos: ayuda al diagnóstico, alertas médicas, gestión hospitalaria, planificación del tratamiento del paciente, sistemas de triaje automático, etcétera.

Una vez detectado el problema y los objetivos es muy importante valorar la situación de los recursos disponibles para hacer frente al proyecto. Esto incluye analizar tanto la disponibilidad del personal médico como del directivo, técnico y/o administrativo, así como la disponibilidad de recursos tecnológicos y de bases de datos, especialmente. Además, se debería incorporar un análisis de riesgos eventuales acompañado de una lista de soluciones potenciales ante dichos imprevistos.

Por último, para determinar si se han alcanzado los objetivos del proyecto de minería de datos biomédicos deben establecerse una serie de métricas. Por ejemplo, en un proyecto en el que se desee incorporar un sistema de ayuda al diagnóstico para glaucoma los objetivos podrían ser conseguir un modelo de clasificación automática con una especificidad y una sensibilidad superior al 95 % y al 90 %, respectivamente. Esta fase debería concluir con el desarrollo de un plan de proyecto que incluya los pasos a seguir mediante un calendario de tareas, hitos y entregables junto a un plan de contingencias.

Análisis de los datos

Como es obvio, la fase de análisis de los datos debe contar con un repositorio o colección inicial relativa al problema que se desea resolver. Esta fase permitirá al analista familiarizarse con los datos, identificando problemas de calidad, descubriendo características interesantes o detectando relaciones entre los mismos que permitan establecer hipótesis iniciales sobre la información subyacente.

Por lo tanto, esta fase incluye la recolección de un conjunto de datos iniciales. Esto no impide que a lo largo del proyecto se vayan obteniendo nuevos datos relativos al problema a resolver. Es de vital importancia poder establecer la disponibilidad de los mismos o el tiempo que llevará obtener un conjunto mínimo de datos con suficiente calidad, ya que el resto del proyecto puede depender por completo de esto. Una vez estén disponibles los datos, es conveniente llevar a cabo una exploración y un análisis estadístico descriptivo de los mismos: cantidad de casos disponibles, número de variables, frecuencias de aparición por clases y variables, distribuciones de las variables, análisis de correlación, visualización de los datos, detección de casos anómalos, test de hipótesis, etcétera.

Además, es muy habitual encontrar gran cantidad de datos perdidos, especialmente si los datos se han obtenido de distintos centros y en distintos periodos de tiempo. Por lo tanto, se deberían revisar las variables perdidas o datos con espacios en blanco. También se debería comprobar si todos los valores posibles se dan en una variable, verificar si los valores de los datos disponibles están dentro de los rangos establecidos, comprobar si hay una semántica asociada (por ejemplo, si los valores 0 y 1 se refieren a "no fumador" y "fumador", respectivamente), así como comprobar que no haya datos que entren en conflicto con el sentido común (por ejemplo, tener un caso de una paciente que haya dado a luz a los 85 años de edad).

Muchos de los inconvenientes que se encuentren en esta fase pueden resolverse en la siguiente una vez detectados y catalogados. Sin embargo, existe la posibilidad de observar problemas en los datos que pudieran obligarnos a volver a la fase anterior para replantear los objetivos y redefinir las cuestiones a resolver.

Preparación de los datos

La preparación de los datos incluye tareas como seleccionar, limpiar, transformar, integrar y formatear el conjunto de datos disponibles (ver capítulo 4). Esta fase es fundamental ya que un mal procesamiento de los datos garantizará un mal modelado posterior. Si se desea tener éxito en las siguientes etapas del proyecto es imprescindible que esta fase se haya realizado correctamente.

En la selección de datos se disponen de diversas aproximaciones para llevar a cabo la tarea. En principio, se deberían incorporar distintos criterios de inclusión y exclusión de casos si se fuesen a llevar a cabo análisis numéricos o estadísticos. Para esta tarea conviene contar con el apoyo de un experto relacionado con el problema[b]. La selección de datos también está intrínsecamente

[b]Es muy posible que esta colaboración deba estar clara en la primera fase para evitar problemas a estas alturas del proyecto.

relacionada con los métodos de selección de características para establecer qué variables pueden ser de mayor interés para el modelado posterior.

La limpieza de datos implica garantizar cierta calidad a los mismos. El conjunto de datos debería ser correcto, completo y consistente. Una forma directa de conseguir esto es rechazando aquellos casos que no cumplan los requisitos mínimos. Sin embargo, esto podría causar que el número de datos fuera insuficiente para la siguiente fase. Para evitar esto existen técnicas de imputación de datos perdidos, que pueden dotar al conjunto de cierta completitud y consistencia.

La transformación de los datos incluye tareas de tipificación, normalización, categorización y discretización de variables. Estas operaciones permiten que los datos estén preparados para poder ser empleados en la fase de modelización según las técnicas que se empleen. Además, se incluyen tareas de derivación de atributos donde se producen nuevas variables a partir de otras que estén contenidas en la base de datos. Por ejemplo, el índice de masa corporal se puede derivar a partir de los valores de altura y peso de un paciente o el equivalente esférico en oftalmología que se deriva de la potencia esférica (miopía o hipermetropía) y la potencia cilíndrica (astigmatismo). También es importante dar el formato apropiado a los datos para que puedan ser empleados posteriormente, esto incluye la codificación de datos (*dummy* o 1-de-C, por ejemplo). Por último, la integración de datos conlleva la combinación de múltiples tablas para crear nuevos registros o la agregación de nuevos valores a partir de sumarios de múltiples registros.

Tras esta fase de preparación de los datos se podrá llevar a cabo el modelado de los mismos en la etapa siguiente.

Modelado

Esta fase constituye probablemente el corazón del proyecto ya que los modelos a implantar serán probablemente útiles si se concluye con éxito. En esta etapa se seleccionan y se aplican distintas técnicas de modelado (ver capítulo 8) cuyos parámetros se calibran en función de los datos o el conocimiento disponible a estas alturas del proyecto. En algunos casos puede ser necesario volver a la etapa de preparación de los datos para adaptarlos a las necesidades de cada técnica.

En esta etapa deberá diseñarse un conjunto de experimentos para calibrar y evaluar cada uno de los modelos que se obtengan para lo que se seleccionará también una metodología de evaluación (ver capítulo 14). Básicamente, el analista deberá elegir entre una evaluación independiente, una validación cruzada o un método de remuestreo en función del número de casos disponible. Las pruebas empíricas permitirán evaluar la calidad y validez de los modelos en función de algún tipo de métrica. Por ejemplo, en problemas de clasificación es habitual medir la proporción de acierto o error de cada modelo.

Tras seleccionar el mejor tipo de modelo, el analista debería llevar a cabo un modelado completo con los datos disponibles respetando las condiciones y parámetros que mostraron mejor comportamiento durante el entrenamiento y la evaluación de los diferentes modelos. En este sentido es importante observar que el resultado que se obtenga mediante la metodología de modelado no será sino una estimación del error, o acierto, y la generalización del modelo final. Pero su verdadero desempeño solo podrá conocerse cuando se lleve a cabo una evaluación clínica independiente con datos obtenidos con posterioridad, esto es, en un entorno de uso real.

Evaluación clínica

Hasta esta fase y antes de proceder a la implantación de los modelos como solución de los problemas planteados al inicio del proyecto, el analista debe evaluar el modelo a conciencia y garantizar que alcanza los objetivos esperados. Las evaluaciones previas durante la fase de modelados daban cuenta de factores como la precisión o la generalización de los modelos. Sin embargo,

esta etapa debe valorar hasta qué punto se ajustan los modelos a los objetivos planteados inicialmente. Cabe mencionar que esta evaluación debe contar con el apoyo de expertos, generalmente médicos, que nos permitan cuantificar, en primer lugar, el desempeño o valor añadido de las soluciones proporcionadas y, en segundo lugar, la utilidad y facilidad de uso percibida por los usuarios.

El valor añadido de los modelos biomédicos puede llevarse a cabo mediante pruebas aleatorias controladas que, idealmente, se diseñaron en la fase inicial. Las pruebas aleatorias controladas se consideran ajenas a cualquier tipo de sesgo cognitivo, por lo que ofrecen una de las formas de evidencia científica más fiables. Por otro lado, la utilidad y facilidad de uso pueden medirse empleando cuestionarios Technology Acceptance Model (TAM) para medir la aceptación de la tecnología [157]. En cualquier caso este tipo de evaluaciones completas pueden ser costosas en tiempo y recursos y siempre existe la posibilidad de llevar a cabo un análisis de los resultados menos riguroso, aunque puede no ser aconsejable según el problema que se desee resolver.

Además, en caso de no alcanzar los objetivos propuestos, es conveniente analizar si existen tareas o factores importantes que se hayan podido obviar por cualquier razón y hayan repercutido en unos resultados por debajo de lo esperado. En función de las conclusiones de este análisis el analista o el líder del proyecto puede decidir dar por terminado el proyecto y pasar a la fase de implantación o, por el contrario, iniciar una nueva iteración del proyecto.

Implantación de las soluciones

El conocimiento adquirido a lo largo del proyecto y, normalmente, plasmado a través de los modelos desarrollados, debe ser organizado y presentado de forma que el usuario final, sea este un clínico o un directivo de la organización sanitaria, pueda usar la herramienta desarrollada de manera transparente en sus procesos de toma de decisiones cotidianos.

En función de los requerimientos, la implantación puede ser tan simple como generar un informe o tan compleja como implementar un servicio de minería de datos dinámico funcionando de manera distribuida entre diversos centros. La implantación deberá, por tanto, seguir un plan concreto que especifique una estrategia para facilitar la tarea. Además, es muy importante poder comunicar a los técnicos del centro sanitario todas las funcionalidades de la herramienta desarrollada, así como todas las acciones que deberán llevar a cabo ya que generalmente son ellos, y no los analistas, quienes terminan integrando las soluciones presentadas en los sistemas de información clínicos donde vayan a usarse.

También conviene diseñar un plan para monitorizar ciertos eventos de la aplicación de minería de datos que puedan considerarse importantes para llevar a cabo el seguimiento de la misma y proporcionar un servicio de mantenimiento a la organización sanitaria. Una estrategia de mantenimiento cuidadosamente diseñada y preparada evitará un uso incorrecto de los resultados del proyecto de minería de datos biomédicos.

Finalmente, se debe redactar un informe final con los resultados de todo el proyecto, donde se recojan las experiencias positivas y negativas del desarrollo y la implantación, las tareas desarrolladas, los resultados obtenidos y los errores y aciertos cometidos, incluyendo claves para la selección de técnicas de minería de datos biomédicos para poder emplearlas en situaciones futuras similares.

17.2. Verificación y evaluación de modelos de conocimiento

Es deseable que un modelo de conocimiento médico sea lo más eficaz posible para solucionar el problema médico a resolver, por lo que verificar que el modelo de conocimiento se ajusta

a lo esperado y evaluar su eficacia para resolver el problema médico será de gran interés. El capítulo 14 estudia la evaluación de modelos predictivos basados en AA y el capítulo 19 estudia la evaluación del efecto de un CDSS implantado en un servicio sanitario utilizando la metodología TAM para medir la eficiencia, eficacia y aceptabilidad del sistema.

Para modelos basados en evidencia médica y guías clínicas, las metodologías de verificación suelen reunir una serie de pasos que comienzan por la verificación funcional de la respuesta de la implementación dadas las condiciones que deben lanzar las reglas del sistema. La extensión que suele tener una guía clínica hace adecuada una metodología de testeo apoyada por un software que genere las combinaciones de valores en los rangos alrededor de los umbrales que lanzan las condiciones. Por ejemplo, la implementación de la guía de cuidados de salud para diabetes en adultos de la American Diabetes Association (ADA) abarcan un total de 75 páginas de reglas aproximadamente, lo que traducido a pruebas funcionales, corresponde a una magnitud de miles de combinaciones a testear. Acosta et al. en [158] utilizaron una aproximación de caja-negra durante un estudio de auditoría. Para ello reprodujeron sesiones de Cancer Multidisciplinary Meeting (MDM) para detectar discrepancias entre las recomendaciones del sistema y los acuerdos registrados por el panel de expertos. Groot et al. [159] utilizan la metodología *model checking* para validad las acciones de los médicos en comparación con el protocolo de acciones predefinidas a través de una guía clínica.

17.3. Credibilidad y evidencia médica referenciada

La credibilidad de un CDSS debe se siempre máxima debido a que su funcionalidad está directamente relacionada con la decisión de los médicos y la seguridad de los pacientes. Evidentemente, un CDSS tendrá gran credibilidad entre sus usuarios si realiza su tarea lo más correctamente posible. Sin embargo, los potenciales usuarios que no han podido utilizar el sistema durante un tiempo suficiente, no tendrán un criterio objetivo para confiar en el sistema sino es a través de referencias bibliográficas donde se demuestren los resultados de evaluación del CDSS sobre datos reales y donde se referencien claramente las fuentes de datos y conocimiento en las que se basa el sistema.

Una vez superada la credibilidad en el sistema, la siguiente credibilidad que debemos asegurar es la de cada decisión que realice el sistema ante un nuevo caso médico. Una vez más, el acierto del sistema será su mejor aliado, pero también al médico le gustará saber en base a que conocimiento se ha tomado la decisión. Generalmente, el médico confiará más en un sistema que 1) interpreta la decisión en base a los datos de entrada y que 2) muestra un intervalo de confianza de la predicción realizada.

Los CDSS basados en guías clínicas suelen asociar a cada respuesta las referencias a los trabajos clínicos que fueron usadas durante el diseño de la guía por consenso del comité editor. Cada una de las referencias tendrá asociada un nivel de calidad de la evidencia dependiente del método utilizado para generar la recomendación, tal como vimos en la sección 16.4. Otro indicador de idoneidad utilizado por Guyatt el al. [160] y Acosta et al. [158] es la solidez de la recomendación para el caso médico. A este respecto, una recomendación puede estar i) contraindicada para un caso médico, ii) los beneficios de la recomendación superar ampliamente los riesgos, o iii) los beneficios son similares que los riesgos.

17.4. Adaptación de CDSS a procesos asistenciales

El CDSS típico encontrado en la literatura científica suele ayudar a la resolución de un diagnóstico médico mediante un conjunto de observaciones puntuales del paciente. El tipo de

preguntas médicas abordadas suele ser 1) una respuesta dicotómica (por ejemplo, un diagnóstico diferencial entre glioblastoma y metástasis), 2) ratificar una hipótesis médica (p.e. el tumor es de tipo glioblastoma), 3) una clasificación multiclase (por ejemplo, entre más de dos posibles diagnósticos), 4) un pronóstico de supervivencia, o 5) un cálculo de dosis.

Este tipo de CDSS no requiere gran esfuerzo para su adaptación ya que apoya al médico en un punto específico del proceso asistencial. Sin embargo, suele ser difícil llegar a ver estos sistemas en producción en un ambiente asistencial. Algunas buenas prácticas en el diseño de estos CDSS para facilitar su adaptación son:

1. Detectar las preguntas médicas susceptibles de ser abordadas por CDSS y especificarlas de la forma lo más detallada posible.

2. Seleccionar aquellas preguntas de las que se dispone conocimiento, o datos biomédicos suficientes para abordarlo.

3. Buscar referencias bibliográficas y otros sistemas que aborden el problema médico a resolver con cualquier tipo de dato biomédico.

4. Estimar el coste que supondrá utilizar los datos biomédicos en un CDSS en producción.

5. Estimar el valor añadido que aporta el CDSS a la pregunta médica a resolver.

6. Diseñar la evaluación multicéntrica del sistema desde el principio del diseño del sistema.

Una circunstancia de particular interés donde los CDSS se hacen de particular interés en en los protocolos de actuación médica basados en consenso (p.e. método RAND/UCLA) donde un gran número de variables deben ser consultadas. Generalmente, estos protocolos llegan a tener una longitud tal que hace inviable su consulta durante la práctica clínica, por lo que un CDSS que implemente dicho protocolo puede dar valor al protocolo consenso. Un ejemplo de protocolo consenso RAND/UCLA son los "Estándares de Uso Apropiado de Cesáreas". Las tablas de consulta son 130 páginas, y un conjunto de variables nada despreciable. Sin embargo, un software que implemente el protocolo, recoja de la EHR la información clínica e interrogue únicamente con las variables necesarias al médico, puede llevar a la práctica clínica el protocolo difícilmente utilizable en papel.

Un escenario especialmente crítico para la implantación de un CDSS es el quirófano. Por ejemplo, los ganglios linfáticos axilares constituyen el principal drenaje de las glándulas del pecho. Esto justifica que las guías de práctica clínica actuales recomienden completar una linfadenectomía axilar a pacientes de cáncer e mama cuyos ganglios centinela contienen metástasis, con el fin de evitar recaídas. Sin embargo, estudios recientes muestras que entre el 30 % y el 70 % de las pacientes con metástasis en el centinela, no tienen metástasis en el resto de ganglios, por lo que la linfadenectomía podría ser evitada. Evitar una linfadenectomía cuando no es necesaria supone una reducción de los efectos adversos a la paciente y costes quirúrgicos y postquirúrgicos cuantificables en un rango entre 4kEUR a 25kEUR por paciente. Un CDSS que ayude a la decisión en el momento de la decisión quirúrgica puede ayudar a resolver este difícil problema. La implementación de un sistema tal, debe estár perfectamente validado, y debería aplicar las técnicas de teoría de la decisión vistas en el capítulo 8 sobre unas matrices bien estimadas de costes.

En algunos países, por ejemplo Reino Unido y Bélgica, el tratamiento del cáncer debe realizarse por consenso del equipo médico multidisciplinar del hospital. El formato para realizar dicha tarea son reuniones (MDM) de aproximadamente una hora donde todo el equipo médico se reune para decidir sobre aproximadamente treinta pacientes. El grupo principal de participantes es el de médicos senior, incluidos los oncólogos (oncología médica y oncología radioterápica),

el radiólogo, el histopatólogo y los cirujanos. Además también participan las enfermeras jefe y lo médicos residentes. Generalmente un médico residente destacado se encarga de preparar la reunión. La sesión se estructura por tipos de pacientes, por ejemplo en cáncer de mama se comienza por los pacientes en postoperatorio y se continúa con biopsiados, analizados con aguja fina jnto con Magnetic Resonance Imaging (MRI), solo MRI, etc. La mayoría de casos son rutinarios y no llevan complicación, por lo que los requisitos de un CDSS que apoye un MDM deben ser 1) ofrecer un interfaz intuitivo, adaptado a flujo de decisión del caso específico y que se puede seguir por todos los asistentes a la reunión, 2) ofrecer la información relevante del EHR del paciente 3) registrar todas las decisiones tomadas durante el MDM, 4) corroborar la decisión médica basada en las mejores guías de práctica clínica, 6) justificar las recomendaciones mediante evidencia médica, 7) ofrecer acceso a las herramientas predictivas disponibles en el mercado (por ejemplo, *adjuvant online!*, y 8) responder inmediatamente a las peticiones de los médicos. Para los casos complicados (generalmente uno en toda la sesión) pueden surgir discrepancias, y la recomendación de la guía puede ser de ayuda para hacer dudar al comité. Generalmente la discusión acabará en la solicitud de más datos biomédicos del paciente para apoyar la decisión en más informarción. Para dar una respuesta correcta para los casos difíciles, es importante que 1) la base de conocimiento del CDSS haya sido diseñada por un ingeniero de conocimiento médico (generalmente médico de formación, y experto en el campo específico de estudio), 2) la base de conocimiento esté totalmente actualizada con las tendencias mundiales de tratamiento en el campo, y 3) el CDSS sea operado por un experto que conozca las guías clínicas (preferentemente el ingeniero de conocimiento).

Cuando un CDSS va a trabajar sobre diferentes puntos del proceso asistencial (ver capítulo 2), es crucial que esté diseñado para poder adaptarse totalmente al proceso, para ello debe tener en cuenta que en un proceso asistencial participan múltiples actores y que se suelen realizar varias acciones simultáneas para un mismo paciente. Es crucial que la implementación de estos sistemas sean genéricas y permitan total flexibilidad ya que los procesos asistenciales pueden cambiar fácilmente y suele ser un hándicap la adaptación de los sistemas de información a los mismos. Para este caso de uso, es imprescindible la conexión de un CDSS al EHR y a las interfaces de los usuarios del sistema sanitario para que tenga alguna posibilidad de ser utilizado.

Además, si el sistema detecta deficiencias médicas o alertas en los pacientes, debe diseñarse de tal forma que sea poco invasivo con el desempeño asistencial, e incentivar la subsanación de los mismos, de forma directa o indirecta. Una forma directa de interacción poco invasiva podría ser marcar las circunstancias y registrarlas para próximas auditorías, un método indirecto sería utilizar las auditorías para mejorar el diseño de los sistemas de información y los CDSS para evitar el problema. En caso de detección de errores, por ejemplo en CPOE, se debe cuidar bien la tasa de falsos positivos que se generan, debidas a circunstancias no contempladas por el CDSS, siempre cuidando que la tasa de falsos negativos se mantenga en el mínimo posible.

El tratamiento continuado y sostenible de enfermos crónicos es en la actualidad uno de los focos principales de investigación en Independent Test (IT) para la Salud. Los costes y consumo de recursos de estos enfermos, junto con el aumento de la esperanza de vida y la inversión de la pirámide poblacional, llevan a los sistemas sanitarios a plantearse modelos donde el paciente se hace responsable proactivo de su salud y complementa los servicios de salud ofrecidos directamente por el sistema sanitario. Esto abre posibilidades a la incorporación de CDSS en los Personal Health System (PHS) que empoderen a los pacientes con las herramientas necesarias para una vigilancia activa de su salud. Generalmente un CDSS será en núcleo inteligente del PHS ofreciendo alertas y recomendaciones a los pacientes atendiendo a las premisas del médico, y mandará reportes resumidos puntuales e históricos a los servicios de salud. En la actualidad se trabaja en sistemas de este tipo para pacientes de diabetes tipos 1 y 2, y en personas con depresión mayor leve o moderada.

17.5. Interfaces de usuario en salud

Los interfaces en medicina deben facilitar la interacción del usuario con el sistema. Un interfaz para profesionales médicos deberá tener en cuenta la limitación de tiempo que tiene el usuario, y saber que su actividad principal no es la introducción de los datos en el sistema, sino su tarea clínica, que generalmente involucra la interacción directa con el paciente. La generación de alertas debe planificarse cuidadosamente ya que será fundamental para su aceptación que el sistema la ofrezca en el momento y formato adecuado, para no ser ignorada o incluso molesta.

Un CDSS deberá ser siempre que se pueda activo, y no esperar que el paciente acuda al sistema e introduzca los datos para obtener la información, ya que en ese caso será poco utilizado. Además, un sistema debe aportar conocimiento al profesional, si no este simplemente lo ignorará.

Históricamente, los interfaces de usuario médicos seguían un paradigma funcional, donde el profesional accedía a partir de menús o listas a las funciones del sistema relacionadas con su actividad. Esto requiere una navegación larga por menús y ventanas, además de requerir que el usuario recuerde las rutas. La tendencia actual del diseño de interfaces tiende a basarse en el proceso, por lo que el profesional debe tener accesible en su pantalla todo aquello que necesita en un momento dado a una distancia de un click. Estos interfaces recuerdan los cuadros de mando.

Microsoft Health ICT Resource Center desarrolló junto con diferentes profesionales de las tecnologías de la información (IT), y profesionales de la salud, tales como la Organización Mundial de la Salud (OMS), la Agencia Nacional de Seguridad del Paciente (NPSA), Institute for Safe Medication Practices (ISMP), el Servicio Nacional de Salud (NHS) una colaboración abierta[c] sobre Interfaces de Usuario en Salud, que publicó en junio del 2007. El propósito del estudio fue estandarizar los elementos de la pantalla para que los profesionales de atención médica puedan cambiar entre distintas aplicaciones e identificar rápidamente la información que necesitan con el propósito de aumentar la eficacia clínica y la mejora de la seguridad del paciente, mejorando la aceptación, fiabilidad y confianza en la aplicación. Como resultado, Microsoft proporciona una guía de diseño y un kit de herramientas de controles. La guía sirve al implementador para normalizar las interfaces clínicas de usuario, y al evaluador como guía de ergonomía y seguridad clínica de la aplicación.

La guía aborda distintas áreas, datos y terminología clínica, navegación, medicamentos, identificación del paciente y diseño. Respecto a las reglas que aplican más directamente a CDSS podemos destacar:

- La identificación del paciente debe estar destacada como cabecera permanente.
- Se recomienda el uso de tablas, cuyo número óptimo de columnas es dos, deben ser configurables.
- Las cabeceras de estados clínicos deben destacarse y siempre estar visibles. Las cabeceras deben agrupar la información de manera clara, y su término no debe repertirse en la información agrupada.
- Se recomienda usar terminologías médicas, como SNOMED-CT, pero sin mostrar etiquetas descriptivas poco entendibles por el usuario y destacando lo relevante.
- Deben evitarse los scrolls.
- Los gráficos deben tener una escala común, y la variable tiempo debe estar en el eje x.
- Los formularios de búsqueda deben ser autocompletables, y el orden no debe importar.

[c] `http://etdevents.connectingforhealth.nhs.uk`

- La etiqueta "Decision Support" en fondo gris indica que el sistema tiene todas las funcionalidades activas. La etiqueta en naranja indica que no está funcionando con todas las funcionalidades, y un símbolo de exclamación delante de la etiqueta indicará que el estado ha cambiado desde la última consulta.
- Un símbolo verde de "visto" indica que el servicio (por ejemplo, comprobación de alergias) está activo. Un aspa roja indicará lo contrario. Textos en gris indican que los servicios no están disponibles.
- Si no hay recomendaciones, debe indicarse.
- Si existen recomendaciones, deberán seguir un orden. Siempre deben mostrar la fecha, hora y prioridad. Las recomendaciones deberáan ofrecer el acceso a ampliación de la información.
- Las alertas de alta prioridad deben aparecer expandidas, con la cabecera en rojo y prevalecer sobre el resto.

Además, la guía detalla el uso de formas, fuentes, colores y focos de tabulación y ratón.

Además de guías de implementación, existen entidades certificadores de estándares de uso médico de interfaces para profesionales y usuarios.

17.6. Acceso a fuentes de datos heterogéneas

Es un hecho que la información médica relevante para la decisión médica sobre un paciente suele estar distribuida en múltiples fuentes de estructuras heterogéneas. Un CDSS no debe ser el encargado de resolver la integración de dichas fuentes de información, pero si que debe dar facilidades para poder utilizarlas. Los almacenes de datos y herramientas de procesamiento en línea suelen incluir módulos ETL (ver capítulo 15) que faciliten la selección de registros provenientes de diferentes fuentes de datos y la transformación de valores y registros a los requisitos del cubo de datos. De forma similar, existen múltiples herramientas para los desarrolladores que facilitan el mapeo de los datos para su explotación, tanto primaria como secundaria (por ejemplo, la herramienta LinkEHR accede y mapea diferentes fuentes a documentos estructurados mediante arquetipos de diferentes estándares[d].

17.7. Consistencia semántica

El tratamiento de la información de los CDSS se realiza a nivel semántico. Por lo tanto, se debe asegurar que los conceptos utilizados por el sistema mantengan siempre el mismo significado en todas las instancias. Esta consistencia adquiere especial relevancia en abstracciones temporales (por ejemplo, *recientemente*), en agregaciones, contextualizaciones de los conceptos o definición por composición de términos estandarizados. En definitiva, cuando un CDSS se compone de múltiples modelos de conocimiento es imprescindible la harmonización de los mismos.

17.8. Interoperabilidad de CDSS con EHR

La interoperabilidad semántica es la capacidad de compartir, agregar, analizar y comprender información ajena al sistema de manera automática, es decir, va más allá de comunicar datos o procesos entre sistemas.

[d] `http://www.linkehr.com`

Cada sistema implementa la persistencia, la lógica de negocio y la lógica de presentación que mejor se ajusta a su funcionalidad y contexto. En particular los CDSS tienen funcionalidades muy diferentes a la de los módulos generalistas de los Health Information System (HIS).

La esencia de la solución para obtener la interoperabilidad semántica es la estandarización que se compone de tres componentes: 1) la estructura de los datos (modelos de referencia), 2) la definición de los conceptos del dominio (arquetipos, y templates) y 3) el uso de vocabularios compartidos (terminologías).

Un modelo de referencia establece un modelo de datos común que define la sintaxis para la comunicación de información. Algunos ejemplos de modelo de referencia son UNE-EN 13606-Parte 1, HL7 v3 RIM, openEHR RM, CDISC Object Data Model, y Continuity of Care Record.

Una terminología es un vocabulario orientado al concepto (cada concepto solo tiene un significado, si bien puede haber ambigüedad entre términos). La construcción de una terminología se realiza mediante la recopilación de todos los conceptos de un dominio y su definición única. Las ontologías amplían esta definición mediante las relaciones entre los conceptos. Existen diferentes terminologías médicas, por ejemplo SNOMED-CT, LOINC, CIE-9/10, CPT, etc.

Un arquetipo es una estructura formal de representación de modelos clínicos detallados para ser procesables automáticamente. David Moner en su Taller UNE-EN 13606 establece la siguiente metáfora: "Los códigos son las palabras de un diccionario. El modelo de referencia es la gramática. Con palabras y una gramática podemos crear frases que tengan sentido o no. Los arquetipos definen aquello que tiene sentido". Los arquetipos dependerán del modelo de referencia utilizado, pudiéndose definir arquetipos y sus instancias en HL7-CDA, UNE-EN 13606, openEHR, etc.

17.8.1. Estándares de conocimiento

Al igual que los estándares de terminologías (p.e. SNOMED-CT) y los estándares de arquitectura (p.e. HL7-CDA) contribuyen para la interoperabilidad en las soluciones informáticas de los sistemas sanitarios, también son necesarios lenguajes de compartición de conocimiento médico que sean automáticamente ejecutables por los ordenadores.

El conocimiento médico puede ser expresado de varias formas, en forma de reglas lógicas, en forma de diagrama de flujo, en forma de modelo estadístico, etc. Para cada una de estas representaciones han aparecido diferentes estándares, unos directamente relacionados con la medicina, otros generalistas y utilizables en medicina de forma más o menos directa. A continuación veremos algunos ejemplos de estándares de conocimiento.

PMML: Predictive Model Markup Language

Predictive Model Markup Language (PMML) (`http://www.dmg.org/`) es un lenguaje de marcas basado en XML desarrollado por el Data Mining Group (DMG) para facilitar la compartición de modelos predictivos y de minería de datos entre aplicaciones. PMML es independiente del vendedor, por lo que facilita la utilización de herramientas diferentes para, por ejemplo, crear los modelos predictivos mediante un programa estadístico y posteriormente visualizarlos mediante un programa integrado en el escritorio médico o en un cuadro de mandos directivo.

PMML, al ser un esquema basado en XML, especifica el estándar mediante un *esquema XML* (i.e. documento XSD) que recoge la sintaxis que debe cumplir cualquier documento PMML. Este esquema XML puede obtenerse directamente desde la web del DMG.

La figura 17.2 describe marca PMML de un documento PMML versión 4.0, que típicamente contiene la especificación los modelos predictivos. Los componentes principales que contiene PMML son los siguientes:

- Header: contiene, entre otros datos, el copyright del modelo, su descripción, y un *timestamp* para especificar la fecha de creación del modelo.

- DataDiccionary: contiene la definición de todos los campos usados en el modelo, así como el tipo (optype) del campo: continuo, categórico u ordinal.
- DataTransformations: contiene transformación aplicada a los datos de entrada antes de aplicar el modelo. Algunas transformaciones son: normalización, discretización, mapeo de valores discretos, o la aplicación de una función.
- Model: contiene los modelos predictivos, como, por ejemplo, la descripción de una red neuronal. Los modelos que admite PMML se clasifican en AssociationModel, ClusteringModel, GeneralRegressionModel, MiningModel, NaiveBayesModel, NeuralNetwork, RegressionModel, RuleSetModel, SequenceModel, SupportVectorMachineModel, TextModel, TimeSeriesModel, y TreeModel.
- MiningSchema: contiene la especificación del uso (useType) de cada campo del modelo (p.e. active, predicted, and supplementary), identificándolos por su nombre (atributo name). También permite el tratamiento de *outliers*.
- Targets: permite el postproceso de las variables de predicción, mediante el escalado de variables continuas, o con valores por defecto en clasificaciones ante valores perdidos.
- Output: especifica los campos de salida del modelo predictivo.

Cada tipo de modelo se especifica mediante una estructura adaptada a las características de la técnica utilizada para crearlo. Por ejemplo, la figura 17.3 representa la estructura de un Árbol de Decisión (TreeModel) en un documento PMML-4.0. En un árbol de decisión, el objeto clave es el nodo (Node), que tiene una estructura recursiva mediante la cual se crea el árbol de decisión a través de reglas simples sobre una variable expresadas mediante la marca *SimplePredicate*.

Ejemplo 17.1 (Árbol de decisión Meningioma vs. No-Meningioma)
El diagnóstico de tumores cerebrales mediante espectroscopía de resonancia magnética es una tarea difícil. Esto es debido en gran medida a la dificultad de interpretación que presenta el espectro de resonancia magnética. Puede ser de interés un sistema de ayuda a la decisión médica para el diagnóstico de tumores cerebrales mediante espectroscopía de resonancia magnética nuclear. De hecho, existen varios prototipos científicos para dicha tarea, entre los que se encuentran CURIAM-BT, Interpret, eTUMOR-CADS, y HEALTHAGENTS.

Nos planteamos como ejemplo la discriminación del tipo de tumor meningioma frente al resto. Es una discriminación relativamente sencilla, ya que un radiólogo puede distinguir un meningioma fácilmente mediante MRI. Para ello, utilizando el algoritmo C4.5, se ha aprendido el árbol de decisión de la figura 17.4, que va realizando cortes sucesivos sobre los rangos de las variables para decidir si un caso es meningioma o no-meningioma. Por ejemplo, el primer nodo representa la separación en meningioma y no-meningioma sin utilizar ninguna variable predictora. Cuando utilizamos field12, y establecemos un corte en 0.8, obtenemos una primera separación de la muestra, que hace que el 94.67 % de los casos con un valor menor o igual a 0.8 caigan en el nodo de la izquierda, con lo cual, comenzamos a obtener una discriminación interesante de la muestra. La inclusión de más variables va mejorando el resultado, hasta cierto criterio de parada para evitar el sobreentrenamiento.

En PMML este diagrama se representa mediante un modelo predictivo de tipo TreeModel, del cual podemos ver el primer nodo y la primera separación en la figura 17.5. Vemos como el primer nodo presenta el conteo total de casos ($recordCount = 217$), e incluye dos nodos, de los cuales hemos expandido el primero de ellos. En este nodo hijo, aparece una marca SimplePredicate que contiene la regla $field12 \leq 0{,}801374$, con un conteo de 150 casos. El documento sigue con nodos recursivos que van desarrollando la especificación del árbol de decisión hasta llegar a las hojas, donde se clasifican definitivamente los casos de estudio.

Figura 17.2: Diagrama de la marca PMML de un documento PMML, que contiene la especificación de los modelos predictivos.

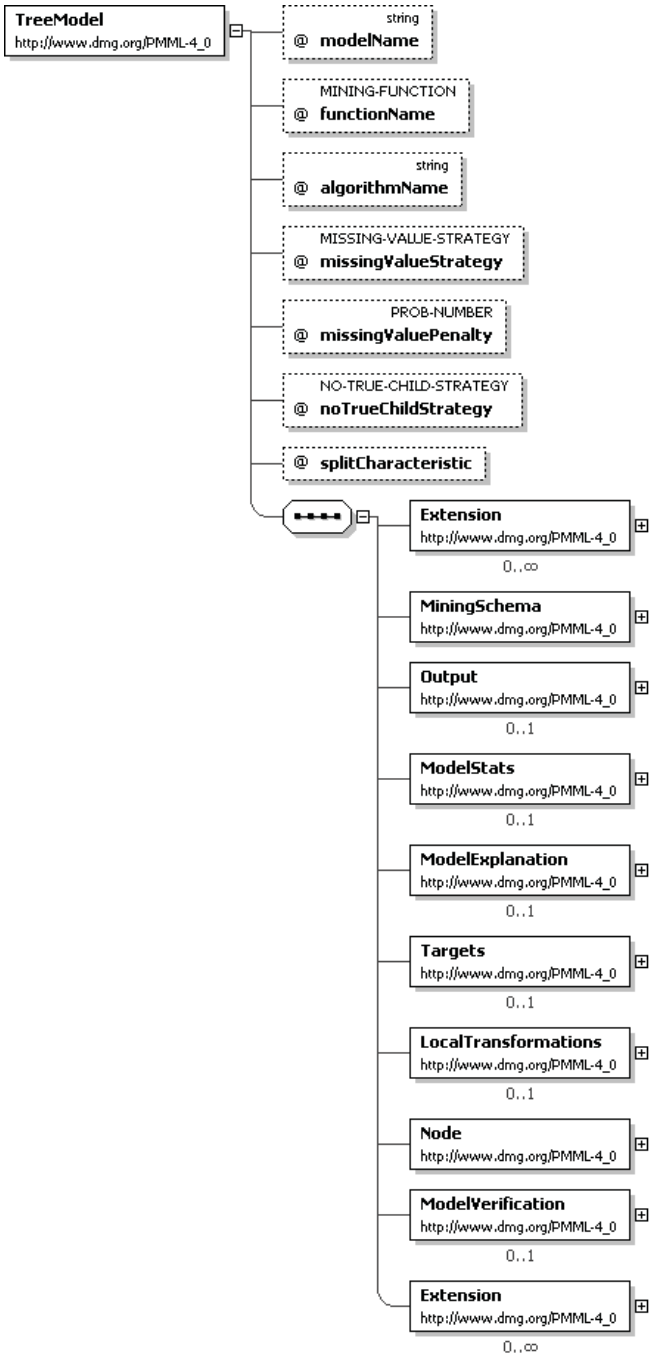


Figura 17.3: Diagrama de la marca TreeModel de un documento PMML, que contiene la estructura de un Árbol de Decisión.

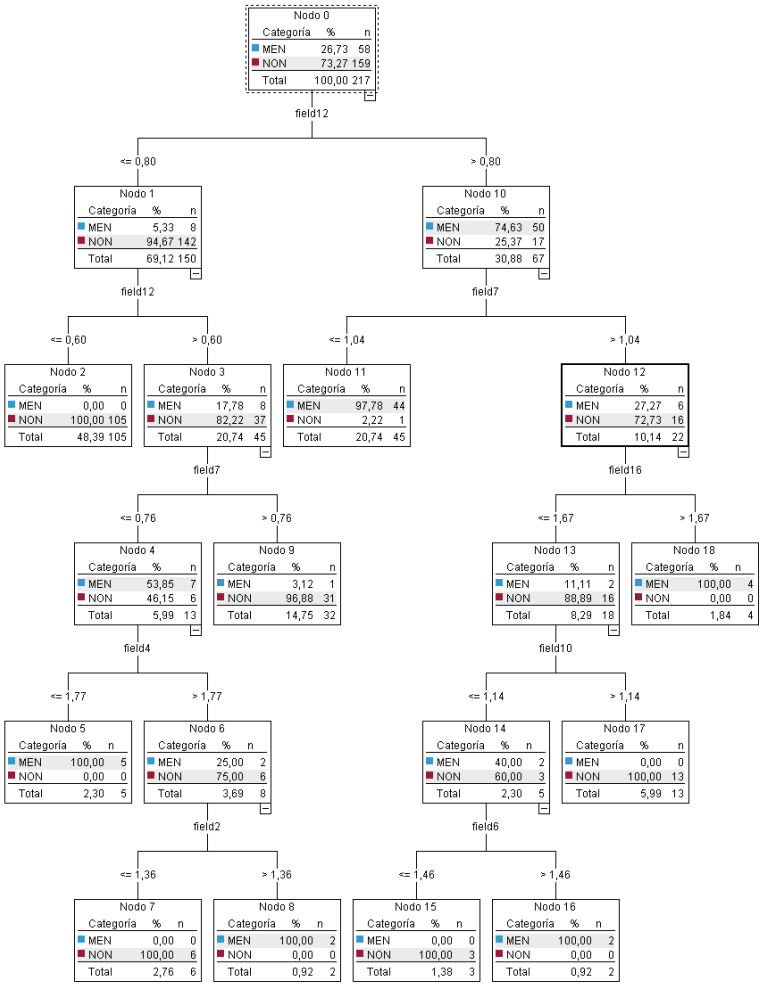


Figura 17.4: Árbol de Decisión para discriminar meningioma y no-meningioma.

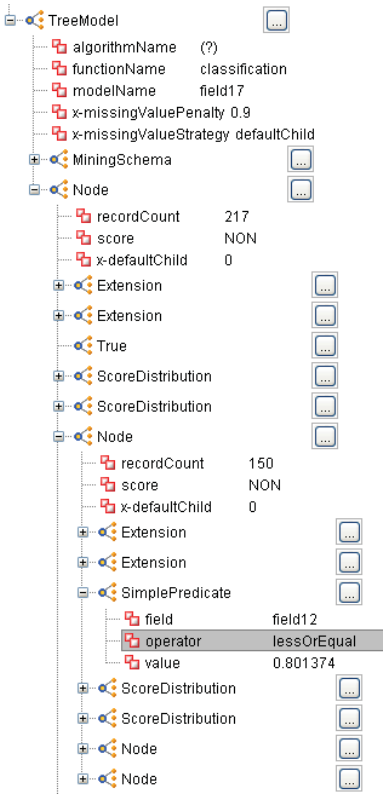


Figura 17.5: Parte del documento PMML representando el árbol de decisión de la figura 17.4.

17.9. Calidad del software

Un CDSS es un software que se utiliza para la asistencia sanitaria de los pacientes, por lo que debe cumplir con los más altos estándares de calidad. Actualmente, los requisitos de calidad en software de propósito médico están regulados a través de las directivas sobre dispositivos médicos. La sección 18.4 resume las directivas aplicables y cita las normas para su cumplimiento.

El estándar internacional IEC 62304, Medical device software - Software life cycle processes, especifica como requisitos generales 1) la gestión de la calidad del sistema, 2) la gestión de riesgos, y 3) la clasificación de seguridad del software. El estándar plantea un ciclo de vida de desarrollo apoyado por una documentación donde se permita trazar que 1) el plan de desarrollo y el análisis de requisitos se verifican en el testeo del sistema, 2) la arquitectura y el diseño detallado (por ejemplo, especificaciones funcionales y especificaciones técnicas se verifican en el test de integración y funcional y 3) que la especificación y desarrollo de módulos unitarios se verifican en los test unitarios. En definitiva, un ciclo de vida de software en forma de V bien documentado, apoyado por una matriz de trazabilidad que identifique cada elemento y su test, y la identificación de las responsabilidades y responsables de cada parte de software. También el mantenimieto del software requeriría una documentación similar continuando con el ciclo del software.

Se debe tener en cuenta que el componente más importante de un CDSS es el modelo de conocimiento médico. Este modelo de conocimiento puede implementarse incrustado en el software o como documentos formales leibles por el software. En cualquier caso, el proceso de construcción del modelo debe seguir una metodología bien definida (por ejemplo CRISP, ver sección 17.1.1) y disponer de una documentación y testeo similar a los aplicado a los módulos puramente de software del CDSS. La documentación del CDSS debe incluir las estimaciones del rendimiento que ofrece el modelo de conocimiento (publicadas generalmente en revistas científicas siguiendo las metodologías explicadas en el capítulo 14). Además, debe incluir las evaluaciones de aceptabilidad, eficacia y eficiencia que la implantación del CDSS (ver capítulo 19).

17.10. Calidad de datos

Los resultados de la investigación biomédica, incluido el desarrollo de CDSS, dependen de la toma de decisiones basadas en la información disponible. Los datos detrás de tal información son registrados por humanos o dispositivos basados en observaciones de hechos, en cualquier etapa del proceso de atención de la salud, y bajo un entorno o contexto. Sin embargo, tanto los humanos como los dispositivos están lejos de ser perfectos. Como resultado, pueden ocurrir errores, omisiones o cambios en los protocolos o prácticas durante la adquisición de datos en cualquiera de estas etapas del proceso de atención sanitaria o en cualquier contexto, lo que conduce a una información sanitaria poco fiable causada por una falta de calidad de datos.

Tal falta de calidad de datos es un asunto importante que conduce a decisiones equivocadas y procesos subóptimos. Esto es particularmente importante en la asistencia sanitaria, donde la fiabilidad de la información puede tener consecuencias directas en el proceso de atención de los pacientes. Además, una calidad de datos insuficiente puede perjudicar directamente los resultados de los estudios que reutilizan los datos, como los ensayos clínicos o, en el caso que nos ocupa, el desarrollo de CDSS. Muchos de los problemas de calidad de datos relacionados con la reutilización de la información clínica están relacionados con dos causas principales [161]: (1) las historias clínicas electrónicas originales están diseñadas para su principal objetivo de atención al paciente, sin tener en cuenta que la reutilización posterior de los datos puede requerir diferentes grados de calidad, y (2) las historias clínicas no están diseñadas para la prevención de problemas de calidad de datos. Por lo tanto, una evaluación de la calidad de datos es importante para estar al tanto de tales problemas para una reutilización adecuada de los datos, mejorar el valor de los datos y conducir a mejores decisiones.

El problema de calidad de datos ha sido estudiado durante años, especialmente en el ámbito industrial, basándose en la hipótesis de que los datos pueden considerarse un producto fabricado por las organizaciones. Aunque los datos biomédicos en la mayoría de los casos representan el estado del paciente, los datos en sí mismos son producidos por los profesionales de la salud, así como por los dispositivos. Bajo este supuesto, el Massachusetts Institute of Technology (MIT) lanzó en 1992 el programa Total Data Quality Management (TQM) [162], basado en las características de TQM introducidas a principios de la década de 1980 para la gestión de la calidad en la industria. Además, muchas otras propuestas de investigación y aseguramiento industrial de TDQM se han relacionado con la metodología de mejora de procesos Six Sigma de TQM [163–165]. Concretamente, el modelo 'DMAIC' puede ser utilizado para mejorar la calidad de datos y sus procesos relacionados, involucrando el siguiente ciclo de etapas: Definir, Medir, Analizar, Mejorar y Controlar.

Los protocolos de aseguramiento de calidad combinan actividades a diferentes niveles, desde el diseño del sistema de información, la formación de los usuarios en calidad de datos, hasta un control continuo de calidad de datos. Definir qué medir y cómo hacerlo es la base del aseguramiento de calidad de datos, siendo ellos los pasos iniciales para cualquier mejora de calidad de

datos. Existe un acuerdo general sobre la definición de calidad de datos en términos de idoneidad para el propósito [166, 167], y esto puede expresarse mediante las llamadas dimensiones calidad de datos. De este modo, se define como dimensiones a los atributos que representan un solo aspecto o construcción de la calidad de datos [168]. El trabajo de Wang and Strong (1996) [168] estableció un trabajo seminal hacia un marco conceptual para la evaluación de DQ considerando las dimensiones de DQ. Nos referimos al trabajo de [169] para una revisión exhaustiva de las metodologías de evaluación de DQ y su relación con las dimensiones. En el ámbito biomédico, encontramos múltiples clasificaciones de dimensiones de calidad de datos, como las propuestas en [170], [171], [172], o [173]. En la Tabla 17.1 recopilamos una propuesta de dimensiones de calidad de datos biomédicos que en la opinión de los autores cubren los problemas más importantes relacionados con el reuso de datos para el desarrollo de CDSS. En ella, incluimos dos factores de confusión potencialmente problemáticos en los datos generados entre múltiples fuentes (localizaciones geográficas, hospitales, proefsionales, etc.) y a lo largo del tiempo. En concreto, las diferencias en los protocolos, las poblaciones o incluso los sesgos inesperados, ya sea causados por los sistemas o por los seres humanos, pueden conducir a una heterogeneidad no deseada en los datos entre sus fuentes o a lo largo del tiempo. Esta variabilidad multifuente y temporal de los datos se reflejará en sus distribuciones estadísticas, en relación con los factores de confusión antes mencionados que, al final, representan un problema de calidad de datos (DQ) que debe ser abordado para una reutilización fiable de los datos. Estos problemas son considerados entonces en la dimensión de Estabilidad (nominación positiva de variabilidad), dimensión cuyos métodos de análisis e implicaciones se encuentran en el estado del arte [174–176].

Dimensión	Descripción
Completitud	Los valores de los datos están presentes
Consistencia	Los datos satisfacen restricciones (formato, rangos y valores permitidos, reglas de dominio, relaciones)
Corrección	Los valores son verosímiles, verdaderos o imparciales con respecto a su estado real
Unicidad	Los registros que representan a una misma entidad del mundo real no se replican
Temporalidad	Los datos se encuentran actualizados respecto a su estado real para la tarea en cuestión
Estabilidad temporal y espacial	Los conceptos y estadísticas inherentes a los datos son comparables entre fuentes (hospitales, profesionales, etc.) y a lo largo del tiempo
Relevancia	Los datos son útiles para su tarea
Contextualización	Los datos están anotados respecto al contexto de su adquisición, su significado y su semántica
Confianza	Los datos son confiables de acuerdo a la reputación de las partes involucradas en su adquisición

Tabla 17.1: Definiciones de dimensiones de calidad de datos propuestas por los autores

17.11. Notas bibliográficas

El estándar australiano HB307-2007 en [7] elaborado por la National Electronic Health Records Taskforce cataloga los Sistemas de Ayuda a la Decisión Médica según la elaboración de conocimiento sobre el paciente. Kawamoto [177] y García-Gómez [178] identifican funcionalidades deseables en los Sistemas de Ayuda a la Decisión Médica. Recomendamos el trabajo descrito en [179] por Sáez et al. como caso de estudio de diseño de un CDSS interoperable con Sistemas de Información Médica mediante el estándar HL7-CDA.

Capítulo 18

Implantación de sistemas de ayuda a la decisión médica

El objetivo inmediato de la implantación de un CDSS es facilitar su uso a los profesionales de la salud, gestores, proveedores de servicios, y pacientes cuando y donde necesitan realizar una decisión relacionada con salud. Evidentemente, el objetivo final de la implantación de un CDSS es la mejora de la calidad asistencial y de la salud individual y colectiva de las personas.

18.1. Adopción de los CDSS

El despliegue en un sistema sanitario de un CDSS debe centrarse en el problema a resolver para tener éxito teniendo en cuenta las características intrínsecas de los problemas de decisión médica; esto es, riesgo, complejidad y dinamismo, entre otros.

Estas características, si bien no identificadas explícitamente, llevan al foro *A Roadmap for National Action on Clinical Decision Support* del American Medical informatics Association (AMIA) a definir tres pilares fundamentales para la adopción de los CDSS en el ámbito médico [180]:

- El conocimiento disponible cuando es necesario

  El conocimiento del problema médico a resolver y la información específica del paciente deben necesariamente reducir el riesgo de la decisión a tomar, por lo que debe ser analizado durante el momento de tomar la decisión.

- Alta adopción y uso eficiente

  El uso sistemático de los sistemas basados en conocimiento reduce la variabilidad en la toma de decisiones, por lo que conllevan mayor eficiencia (temporal y económica) y eficacia (valor añadido al paciente) en el sistema sanitario.

- Mejora continua del conocimiento y de los métodos de ayuda a la decisión

  El acceso a las últimas tendencias de la evidencia científica, la auditoría y la realimentación de las consecuencias de las decisiones en la base de conocimiento lleva a la excelencia del sistema y por tanto a la convergencia hacia el objetivo primario de los CDSS.

Parece sin embargo sorprendente que las herramientas CDSS son una de las funcionalidades menos adoptadas en sistemas sanitarios, pese a ser las que potencialmente pueden aportar mayor valor añadido tanto al clínico como al paciente. Esto puede deberse a ciertas barreras que intentareamos identificar de forma extensiva.

18.1.1. Barreras a la adopción de los CDSS

La experiencia previa en implantaciones de CDSS en entornos médicos ha identificado una serie de barreras que pueden dificultar la incorporación de sistemas funcionalmente bien definidos. Estas barreras deben considerarse desde las primeras etapas del diseño del sistema y reservar los recursos necesarios durante la implantación A continuación se enumeran algunas barreras detectadas por [177, 181–186] entre otros:

- Retraso de la introducción de las Tecnologías de la Información y Comunicaciones (TIC) en el ámbito sanitario.

  Las TIC y la Ingeniería de Procesos están ampliamente integradas en sectores como la banca, la automoción, la energía o la logística. Sin embargo, es una realidad el retraso de estas tecnologías en el ámbito sanitario.

  Sistemas de información corporativos, cuadros de mando, sistemas de explotación de datos, que pueden encontrarse más o menos extendidos de forma transversal en grandes corporaciones u órganos directivos, han requerido más tiempo para ser introducidos en el ámbito sanitario. Si en los niveles gerenciales se observa el retraso en la incorporación de las TIC, todavía se hace más evidente cuando observamos el nivel clínico, donde el primer sistema a implantar, la EHR, puede considerarse completamente implantando únicamente en el 1.5 % de hospitales de U.S [185].

  Si que se observa una mayor implantación de las TIC en unidades históricamente asociadas con el procesamiento de señales biomédicas, como son los servicios de laboratorio y radiología, con un 75 % de implantación en EE.UU. [185].

- Metodología y variabilidad en la práctica médica.

  La práctica médica ha desarrollado a lo largo de la historia una metodología muy específica de trabajo, que viene avalada por los resultados obtenidos y que hace del profesional médico un decisor muy centrado en su experiencia. Esto, unido a la complejidad y riesgo de la tarea a resolver, incrementa la resistencia al cambio [186], y por lo tanto a la introducción de elementos externos a su práctica diaria.

  Sin embargo, se observa una variabilidad importante en la aceptabilidad de novedades dependiendo de las especialidades médicas. Sintchenko et al. en[181] observaron un claro aumento de aceptabilidad de los CDSS entre los médicos especialistas en cuidados intensivos respecto a los especialistas en enfermedades infecciosas. En su análisis apuntan hacia unos estilos diferentes en las especialidades, estando cuidados intensivos más predispuesta a buscar evidencias que puedan reducir el riesgo ante una decisión.

  De forma general podemos decir que la resistencia al cambio es un hecho cuando implica gran volumen de usuarios, lo que conlleva un abanico heterogéneo de percepciones. Concretamente, Jha [185] registra una resistencia por el personal médico cercana al 36 % de los casos encuestados, que resulta algo superior en hospitales con EHR que en hospitales donde no se han implantado.

- Dificultades de integración en el flujo de trabajo.

  La práctica clínica está migrando actualmente de la actuación por conocimiento tácito de los profesionales a la actuación en base a guías clínicas y procesos asistenciales integrados. A día de hoy se han hecho grandes esfuerzo en la definición de procesos asistenciales que identifican flujos de trabajo de los profesionales involucrados. Sin embargo, todavía se está en fase de adopción y evaluación en la mayoría de los casos, lo que resulta una barrera para la identificación de las etapas clave de decisión y los actores involucrados durante el diseño

de un CDSS específico para una pregunta médica. Además, la complejidad de los sistemas sanitarios hace que cada escenario de implantación tenga particularidades intrínsecas, por lo que se observa una clara heterogeneidad entre centros y entre servicios sanitarios.

- Características de los pacientes y situación clínica.

  Gravel, Graham et al. [182, 183] destacan las características de los pacientes y la situación clínica como dos barreras importantes para el uso de Sistemas de decisión compartida (Shared Decision-Making) (SDM). Esta barrera está realmente relacionada con la identificación correcta de la pregunta médica a resolver mediante los CDSS y el análisis del entorno de implantación, que deben ser cuidados desde las primeras etapas de desarrollo del sistema.

- Dificultad para acceder a la información relevante de los pacientes.

  Un problema inherente a la decisión es la falta de información procedente de los casos de estudio. Sin embargo, mucha de esta información podría estar registrada pero es realmente complicada de obtener por involucrar a multitud de personas. Por ejemplo, es sabido los grandes esfuerzos gubernamentales por involucrar a las mujeres en los programas de prevención secundaria (cribado) del cancer de mama. Otra información difícil de obtener es el seguimiento de los pacientes, tras un tratamiento, o episodio, que suele requerir mecanismos adecuados de registro, contacto con el paciente, etc; lo que conlleva un gran gasto de recursos. Además, los pacientes pueden cambiar de sistema sanitario, lo que conlleva el cese de seguimiento de dicho paciente.

- Dificultad para conseguir una muestra significativa para un ensayo clínico.

  La evalución de las tecnologías sanitarias es fundamental para avalar su implantación en la práctica clínica, incluso en base a las regulaciones sanitarias existentes [184]. Esta evaluación debe incluir la aceptación de la tecnología por parte de los profesionales involucrados en los procesos donde se incide, y demostrar el valor añadido en la salud de los pacientes, lo que implica realizar muestreos clínicos que involucren a profesionales y casos médicos (ver capítulo 19). Estos muestreos suelen ser complejos, económicamente costosos y temporalmente extensos, por lo que deben planificarse cuidadosamente y ser apoyados institucionalmente para que sean viables y rentables.

- EHR no implantada universalmente.

  Los CDSS adquieren su máxima utilidad cuando son interoperables con los sistemas de información del sistema sanitario (ver sección 1.2). Es evidente que la carencia de un EHR mínimo en el sistema sanitario limita las posibilidades de explotación de datos que puedan realizarse del mismo, si bien pueden suplirse con sistemas de laboratorio u otras soluciones *ad-hoc*.

  Jha subraya en una encuesta entre hospitales de EE.UU. que tan solo el 1.5 % de que respondieron disponen de un EHR global y un 7.6 % tiene sistemas básicos que implican al menos una unidad funcional.

- Bajo nivel de interoperabilidad semántica.

  La máxima utilidad de un CDSS puede obtenerse cuando interopera con el resto de sistemas de información y conocimiento del sistema sanitario al que da apoyo. Si bien en la actualidad se hacen grandes esfuerzos a través de programas nacionales e internacionales de estandarización de historias clínicas electrónicas, compartición de datos de laboratorio e imágenes médicas y definición de terminologías médicas es complejo encontrar un entorno interoperable más allá de pilotos o partes específicas del sistema sanitario. En la revisión

sistemática de Garg [184], el 47 % de los cien CDSS analizados estaban interconectados con la EHR o con el CPOE del sistema sanitario.

Dentro de esta barrera, podríamos incluir la falta de estructuración de la información clínica, cuyo ejemplo más claro es el registro del conocimiento médico en texto libre. Una solución a este limitante es la extracción automática de conocimiento médico mediante procesamiento de lenguaje natural [187], siendo esta una rama de estudio completa a estudiar en si misma.

- Restricciones de tiempo.

  La agenda diaria del profesional en medicina, y de los servicios sanitarios en general, suele estar sobresaturada, por lo que el profesional necesita máxima eficacia en su gestión del tiempo. Además, en ciertas circunstancias, la decisión debe tomarse lo antes posible. En este sentido, Friedman et al. [188] argumentan que la utilidad de la información médica percibida por los usuarios es una función de su relevancia, validación y el *esfuerzo realizado para encontrarla*.

  Sintchenko en [181] argumenta un incremento del tiempo necesario para llevar a cabo la tarea, aumentando de 113 a 245$s/caso$ de media en su estudio de 62 escenarios, por la falta de entrenamiento de los profesionales en los sistemas evaluados.

  También en la revisión sistemática de SDM realizada por Gravel, Graham et al. [182, 183] la gran mayoría de médicos encuestados indicaron la falta de tiempo como barrera a la hora de utilizar , lo que indica la clara necesidad de desarrollar metodologías adecuadas de implantación de CDSS en los entornos médicos.

  Por lo tanto, podemos asegurar que la accesibilidad, usabilidad, y adaptatividad son tres factores directamente relacionados con el éxito o fracaso de la implantación de un CDSS. Así pues, la integración eficiente de los sistemas en los procesos asistenciales y con la rutina del servicio sanitario, la formación y la asistencia a los profesionales son fundamentales para la implantación de los CDSS (ver sección 18.3.2). Como mejor solución, Garg observó en [184] que el valor añadido que los CDSS activos ofrecían a la eficiencia de los médicos era significativamente mayor que los sistemas pasivos (ver sección 1.2).

Además de estas barreras directamente aplicables a la problemática de los CDSS, Jha et. al [185] describe algunas barreras extra para la implantación de una EHR que pueden ser también aplicables a nuestros sistemas:

- Presupuesto inadecuado.

  En el estudio presentado por Jha, el 74 % de los hospitales consultados consideraró inadecuado el presupuesto invertido en el proyecto. Sin embargo, en hospitales con EHR, esta barrera se identifica como tal en un 60 % de los encuestados.

- Coste de mantenimiento.

  El 44 % de los hospitales consultados indicaron que el coste de mantenimiento es una barrera para decidir realizar una implantación del sistema, sin embargo esta barrera desciende al 30 % en aquellos hospitales donde tienen un EHR integrado.

- ROI indefinido.

  Otra barrera expresada por hospitales ante la implantación de EHR es la falta de una definición clara del retorno de la inversión. El 30 % de los hospitales encuestados que no tenían EHR observaron esta barrera, mientras que en hospitales con EHR es el porcentaje fue del 25 %.

Durante el diseño de un CDSS debe evaluarse el retorno de inversión, para dimensionar correctamente el sistema a desplegar. Por ejemplo, en un un análisis coste-beneficio Johnston et al. en [72] evaluan adecuado un CPOE con funcionalides de CDSS para un sistema ambulatorio con más de 10 médicos.

- Equipo informático inadecuado.

  Por último, el estudio de Jha identifica un 30 % de hospitales donde se considera que el personal informático para la implantación y mantenimiento es inadecuado, si bien el porcentaje desciende a un 20 % en los hospitales donde disponen de EHR.

18.1.2. Factores clave para la adopción de los CDSS

Una vez identificadas las barreras, podemos enumerar una serie de factores beneficiosos para llevar a cabo la implantación de un CDSS en un entorno clínico:

- Entender el problema médico a resolver como primer objetivo de la fase de diseño. Identificar e involucrar a los actores relevantes para el problema médico. Identificar y consensuar las preguntas a resolver, los sucesos inciertos y el espacio de decisiones. Definir las funciones de pérdida asociadas a cada decisión y simular los casos de uso más comunes (ver secciones 3.3 y 18.3.2) .
- Analizar el riesgo del proyecto (ver sección 18.2.4). Justificarlo en base a la evidencia científica [177] y las necesidades médicas.
- Estudiar el retorno de inversión proporcionado por el CDSS y compararlas con alternativas no computacionales o de menor complejidad. En favor de los sistemas computacionales, Kawamoto [177] encuentra diferencias significativas en el valor añadido ofrecido por estas soluciones frente a otras.
- Ofrecer sistemas inteligentes. Algo más que un sistema simple de guía (ver taxonomía en sección 1.2). Kawamoto [177] encontró que ofrecer recomendaciones daba un valor añadido significativamente más alto que ofrecer únicamente evaluaciones del proceso asistencial.
- Analizar las características de los pacientes y el entorno de decisión clínica para identificar limitaciones y facilitadores de la implantación.
- Integrar el CDSS con el proceso asistencial: dar la información adecuada en el momento adecuado. Kawamoto y Pearson identifican este factor como clave de la implantación [177, 186].
- Desarrollar sistemas activos. Los sistemas que inician la interacción con el usuario de forma autónoma aumentan el valor añadido a la eficiencia del decisor en un 25 % [177, 184, 186].
- Asegurar la accesibilidad, usabilidad, discreción y adaptatividad del CDSS mediante interfaces de usuario adecuadas [184], para reducir al mínimo la resistencia al cambio del personal médico.
- Asegurar la interoperabilidad con el resto de sistemas de información sanitarios.
- Asegurar la rapidez del sistema, brindar la información en tiempo real.
- Reservar los recursos de implantación (ver sección 18.3.2).
- Diseñar planes de formación y de asistencia para los decisores que usarán el CDSS.

- Identificar y reservar los recursos necesarios para demostrar la aceptabilidad, el valor añadido y el impacto obtenido. Difundir adecuadamente los resultados al sistema sanitario, la comunidad científica y divulgarlo entre los pacientes y población interesada.
- Diseñar la evaluación del sistema desde la perspectiva del valor añadido que ofrece el CDSS a la eficiencia al resolver la tarea médica y asegurar que el resultado hacia el paciente está en los niveles de eficacia esperados.
- Permitir la actualización y capacidad de evolución del CDSS.

18.2. Gestión de la innovación en CDSS

La implantación de un sistema de ayuda a la decisión médica es generalmente de tal envergadura que son las grandes instituciones públicas (ministerios, servicios de salud, u hospitales) o grupos de salud privados los potenciales compradores interesados en incorporarlos en sus procesos asistenciales integrados en los EHR.

Como hemos visto en los sucesivos capítulos, los CDSS se fundamentan en tecnologías innovadoras e incluso en tecnologías que se encuentran más allá del estado del arte de los actuales sistemas informáticos desplegados en el ámbito de la salud. Esta conlleva una dificultad clara para la especificación de sistema, así como su integración con los procesos asistenciales y sistemas actuales de las organizaciones.

Una institución que se plantee la incorporación de esta tecnología en su organización requiere una metodología de gestión de la innovación que le permita gestionar la inversión necesaria para conseguirla, respondiendo a un doble objetivo: 1) controlar el riesgo de la inversión en nueva tecnología, y 2) estimular a sus proveedores para que produzcan las mejoras en sus productos que respondan a los avances en dicha tecnología. Estos objetivos requiere una solución de compromiso entre ambos y la contratación pre-comercial puede llevarla a cabo, por lo que es una buena herramienta para la estrategia de gestión de la innovación en grandes instituciones de salud y por lo tanto para la evolución de los EHR actuales a CDSS corporativos.

18.2.1. Modelo actual

El modelo de contratación actual basado en el *desarrollo exclusivo* implica que un comprador público se reserva para su propio uso todos los resultados y beneficios del desarrollo (incluidos los derechos de propiedad intelectual). Esto hace que la propuesta de contratación tenga 1) un precio elevado, ya que las empresas que desarrollan los productos o servicios no puedan utilizarlos con otros posibles clientes, que suele conllevar que 2) no resulte atractiva cuando supone un alto riesgo debido la incorporación de nueva tecnología todavía en fase de consolidación.

Los compradores públicos europeos tienden a utilizar el desarrollo exclusivo, aunque sean uno más de los muchos usuarios potenciales de la solución desarrollada, por lo que dicha exclusividad no es imprescindible. Además, los compradores públicos suelen infravalorar los costes y los esfuerzos adicionales necesarios para cosechar los beneficios de los resultados, sobretodo cuando los desarrollos requieren mejoras tecnológicas. Como consecuencia, el desarrollo exclusivo obstaculiza la innovación, provocando una fragmentación del mercado mediante barreras que destruyen oportunidades de encontrar la comercialización de soluciones innovadoras.

18.2.2. Contratación precomercial: ámbito y definición

Algunas de las mejoras requeridas son tan exigentes desde el punto de vista tecnológico que o bien no existen aún soluciones comercialmente estables en el mercado, o bien las soluciones

existentes presentan deficiencias que exigen más innovación, generalmente aplicada al caso de uso específico demandado por el cliente. Así pues, no solo es necesaria la compra de productos y servicios comercialmente disponibles sino que también es necesaria la contratación de innovación de nuevas soluciones que mejoren lo que ofrece el mercado.

El ámbito de la la Contratación precomercial se incluye por lo tanto en la fase de gestión de la demanda dentro del ciclo de gestión de proyectos de una institución pública o privada. Por lo tanto, podemos verla como una etapa previa a la comercialización y complementaria a la gestión de los proyectos de implantación de productos o servicios en la organización.

El marco de la contratación precomercial se define como el planteamiento de la contratación de servicios de innovación distinto de aquellos según los que los beneficios pertenecen exclusivamente a la entidad adjudicadora para su utilización en el ejercicio de su propia actividad.

Una contratación precomercial puede cubrir actividades como la definición de soluciones competidoras, el desarrollo de prototipos o la implementación de un volumen limitado de primeros productos o servicios a medida con el fin de demostrar que el resultado se presta al despliegue masivo satisfaciendo normas aceptables de calidad. Sin embargo, la innovación no incluye actividades una producción o despliegue a gran escala por lo que no se plantea una integración completa de la solución que requiera el ajuste fino del desarrollo a los procesos existentes.

La característica fundamental de la contratación pre-comercial consiste en el reparto de riesgos y beneficios entre el adjudicatario y el proveedor, es decir, el comprador público no se reserva los resultados de la innovación para su propio uso en exclusiva sino que comparte con las empresas los beneficios de la innovación necesaria para desarrollar soluciones que superen las que hay disponibles en el mercado; como contrapartida también comparte los riesgos del proyecto con los proveedores. La contratación precomercial implica una competencia entre posibles proveedores de soluciones, que debe permitir al comprador 1) aprender sobre el problema y las soluciones posibles y 2) elegir las mejores soluciones que ofrece el mercado a los mejores precios. Una vez alcanzado el final de la fase de innovación, la contratación pública para la producción comercial puede incluir cualquier empresa que haya participado o no en la contratación precomercial.

Beneficio de la empresa

Los beneficios esperados por las empresas participantes en una contratación precomercial son:

- Reparto de beneficios de la innovación a través de la explotación de resultados
- Mejora del rendimiento de la innovación en relación con sus principales competidores (ventaja del pionero)
- Inversiones en innovación e investigación
- Anticipación de la demanda de nuevas soluciones que permite acortar el tiempo de llegada al mercado
- Reducción de costes de los proveedores pioneros para adoptar normas adoptadas posteriormente
- Mercado simétrico en la contratación pública para la producción comercial, basado en el conocimiento del problema y las posibles soluciones

Beneficio de las instituciones

Los beneficios esperados por las instituciones contratantes en una contratación precomercial son:

- Precios ajustados
- Elegir entre las mejores soluciones
- Mercado simétrico en la contratación pública para la producción comercial, basado en el conocimiento del problema y las posibles soluciones
- Mejora la calidad y la eficiencia de los servicios públicos a medio/largo plazo
- Asegura la competencia en la fase de comercialización (presión competitiva positiva)
- Aprendizaje mediante evaluación del rendimiento de los prototipos y de los productos de prueba en un auténtico entorno operativo comercial
- Influencia en los planes de trabajo de las empresas y en las futuras normas para ajustar los productos a sus necesidades
- Introducir nuevas soluciones más rápidamente
- La pronta participación en el proceso de innovación permite que los poderes públicos detecten antes posibles aspectos normativos
- La comprobación precoz de que la innovación responde a necesidades concretas de los servicios públicos
- Optimización del gasto destinado a innovación

Beneficio de la sociedad

Como beneficio general obtenido por la sociedad debido a la proliferación de este tipo de contratación, se podrían esperar los siguientes ítems:

- Reforzar la capacidad innovadora de la sociedad
- Estimular la innovación
- Promover el potencial de la contratación pública
- Mejorar el rendimiento de la innovación y la competitividad de la industria
- Solucionar escasez de inversiones en I+D+i
- Mejorar la capacidad de convertir las nuevas invenciones en nuevos productos y puestos de trabajo
- Promover normas abiertas de las soluciones

18.2.3. Contratación precomercial: posible implementación

Una implementación eficiente promovida por la Comisión Europea [189] consiste en distribuir la innovación en fases temporales asegurándose de que las empresas compiten para crear una serie de opciones evaluables tras cada fase. Cada fase, por lo tanto, consiste en 1) un desafío al mercado que se traduce en una invitación a varias empresas a desarrollar en competencia las mejores soluciones posibles para abordar el problema, y 2) la evaluación de los pros y los contras de cada solución. Como resultado, se produce un proceso de aprendizaje mutuo para los

compradores y las empresas que ayuda a determinar las necesidades funcionales de los problemas y el rendimiento y limitaciones de las soluciones.

Un posible esquema de ciclo de la innovación en fases aplicado a un problema especifico podría incluir:

- Fase 0. Investigación del problema
  - Input: creación de la necesidad, conocimiento de tecnología, conocimiento científico/-técnico
  - Output: idea del producto
  - Actores: institución pública, comités de expertos internos y externos independientes
  - Evaluación: consenso de idea del producto, robustez ante disensos
- Fase 1. Exploración de soluciones
  - Input: idea del producto
  - Output: catálogo de soluciones
  - Actores: múltiples empresas
  - Evaluación: selección de soluciones por comités internos y externos
- Fase 2. Prototipos
  - Input: soluciones seleccionadas
  - Output: prototipos exitosos, nuevas ideas, refinamiento de idea, refinamiento de soluciones
  - Actores: empresas de soluciones seleccionadas
  - Evaluación: selección de prototipos por comités internos y externos
- Fase 3. Desarrollo de primeros productos como pruebas de concepto
  - Input: prototipos seleccionados
  - Output: primeros productos limitados, refinamiento de idea, refinamiento de soluciones, estimación de costes de producción, estimación de problemas de integración, nuevas ideas
  - Actores: empresas de prototipos seleccionados (mínimo dos)
  - Evaluación: emisión de informes por comités de expertos internos y externos

Una vez finalizado el ciclo de innovación se pasaría a la contratación comercial, donde cualquier empresa que haya participado o no en la contratación precomercial puede participar, y donde la institución y las empresas conocen mejor la idea del producto demandado, las posibles soluciones tecnológicas, una estimación de costes de producción, de integración, posibles normas aplicables y estándares adoptables.

18.2.4. Casos de estudios

Uno de los retos que intentan cubrir los sistemas de ayuda a la decisión médica es asegurar la atención sanitaria universal de alta calidad y asequible que aborde las consecuencias del envejecimiento de la población mundial. Para conseguirlo son necesarios nuevos equipos que faciliten la prevención, el diagnóstico precoz, la optimación de los procesos asistenciales y la involucración de los ciudadados en el cuidado de su propia salud. Esto requiere una colaboración estrecha de las instituciones sanitarias y los proveedores capaces de producir innovación en las tecnologías de la información aplicadas a la salud. El marco de la contratación precomercial puede ser un instrumento estratégico para unificar los intereses de los actores involucrados en la obtención de soluciones dirigidas a la salud de los ciudadanos.

Caso de uso en CDSS: Plan Estratégico de Implantación de Ayuda a la Decisión Médica en los Procesos Asistenciales mediante ciclos de vida de contratación precomercial

La madurez de las tecnologías de la información aplicadas en salud hace pensar que, tras la historia clínica electrónica, el siguiente eslabón significativo es la ayuda a la decisión médica. Este eslabón, puede verse en el horizonte del corto-medio plazo, y su correcta incorporación en el uso asistencial pasa por un estudio centrado en el valor añadido que ofrece al profesional y al paciente. Para ello, se debe crear una estrategia de implantación interoperable CDSS en los servicios de salud basada en la gestión de la innovación, y por lo tanto siguiendo el ciclo de vida de Contratación Precomercial. Para iniciar dicha estrategia se ve adecuado definir y sostener un Plan Estratégico de Implantación de Ayuda a la Decisión Médica en los Procesos Asistenciales, donde puedan unirse instituciones sanitarias, universidades, empresas y grupos de interés en las TIC-Salud, para preparar el camino que permita al ámbito médico aprovecharse de la incorporación de los sistemas de ayuda a la decisión. Dicho plan estratégico serviría de embrión de las varias Fases 0 que cubrieran las funcionalidades de los sistemas de ayuda a la decisión identificadas en la sección 1.3. Los objetivos a cubrir por el Plan Estratégico de Implantación de Ayuda a la Decisión Médica en los Procesos Asistenciales son cinco:

1. Definición de pilotos pioneros en Ayuda a la Decisión Médica

   En este objetivo se buscarán preguntas médicas donde se estime un alto valor añadido por el uso de sistemas de ayuda a la decisión médica. Para ello, se involucrará a profesionales de la salud, se realizarán entrevistas a profesionales y se analizarán documentos científicos que permitan definir basados en la evidencia médica las funcionalidades de los sistemas de ayuda a la decisión. Los CDSS pueden ser utilizados como sistemas de alerta y/o recordatorio de situaciones de salud de pacientes concretos o epidemiológicos, ayuda a la prevención, diagnóstico, pronóstico, interpretación de señales biomédicas, planificación y/o control de terapias, calidad asistencial, riesgos, coste temporal, planificación de recursos, control de calidad en la adquisición de biobancos multicéntricos y búsqueda de patrones anómalos Los focos de estudio serán problemas de salud referentes a enfermos crónicos, cáncer, screening, hallazgos de imagen en señales radiológicas, tratamiento de procesos degenerativos, seguridad farmacológica, alergias, primaria, triaje, prescripción farmacológica, etc.

2. Mecanismos para la inclusión de los CDSS en procesos asistenciales

   Un factor fundamental de éxito en la implantación de los CDSS es su ajuste al proceso asistencial del entorno donde actúa. Así pues, el conocimiento suministrado por los CDSS debe trasmitirse a los profesionales adecuados en tiempo y forma. En este objetivo se estudiará la inclusión de los CDSS en los procesos de salud y como elementos de decisión objetiva en las guías clínicas.

3. Interoperabilidad semántica de CDSS

   Para llegar al máximo posible de despliegue de un CDSS, este debe ser interoperable con la historia clínica electrónica del paciente y otros sistemas de conocimiento del sistema sanitario. Vocabularios de terminología médica en forma de ontologías, como SNOMED, ICD9/10, LOINC, etc; han permitido la conceptualización de los diagnósticos, procesos, instrumental, y demás términos involucrados en el proceso asistencial. Es totalmente necesario que los sistemas de ayuda a la decisión interactúen de forma natural con la historia clínica electrónica. A la hora de leer datos, la solución viene dada por los estándares de estructura de información clínica, como HL7-CDA, ISO/EN 13606, o openEHR. A la hora de incorporar la ayuda a la decisión clínica como conocimiento biomédico, debe estudiarse cómo incorporar los estándares para compartir conocimiento, como PMML y RuleML a los repositorios de conocimiento médico.

4. Definición de biobancos digitales basados en el control de calidad

   Es una buena práctica que el diseño de CDSS esté motivado por un biobanco digital con datos biomédicos adquiridos por múltiples hospitales siguiendo un protocolo consenso y control de calidad. En el conjunto de datos recogidos en estos biobancos se incorpora la evidencia científica estudiada por los expertos internacionales que han participado en la adquisición de los datos, y que ha sido referenciada en las publicaciones científicas derivada de los mismos. Para que la evidencia científica incorporada en los CDSS sea de máxima utilidad, es interesante definir estrategias de recopilación de datos biomédicos multicéntricos basados en parámetros de calidad. En este objetivos se estudiarán las posibilidades técnicas basadas en la gestión de datos biomédicos, así como mecanismos automáticos de detección de datos anómalos, y auditoría de sistemas de información sanitarios para la generación de los biobancos digitales.

5. Aceptación de la tecnología, valor añadido e implantación

   La componente humana es la más importante en los sistemas de ayuda. Los sistemas de ayuda a la decisión están diseñados para ofrecer una ayuda objetiva al profesional de la salud, y con el objetivo de mejorar la atención del paciente. Las tecnologías que componen los sistemas de ayuda a la decisión deben mostrar sus resultados de forma útil a los profesionales involucrados en el flujo de decisiones. Por ello, la definición de mecanismos de evaluación que aseguren la aceptación y facilidad de uso de los sistemas en la práctica diaria asistencial resulta fundamental para la implantación de estas nuevas tecnologías. Como fin último, los sistemas de ayuda a la decisión, buscan la maximización de la calidad asistencial al paciente, por lo que el estudio del valor añadido sobre el mismo es totalmente necesario. Las metodologías de diseño de CDSS incluyen implícitamente el análisis de riesgos en su fase de desarrollo, por lo que es directa la extensión del mismo a los métodos de prueba aleatoria controlada utilizados generalmente en los estudios de valor añadido en clínica. Finalmente, es de especial relevancia el proceso utilizado de implantación de los sistemas de ayuda a la decisión en los ámbitos de salud. La implantación integral de programas de ayuda a la decisión clínica es abordada de forma integral en estudios internacionales, como los promovidos por HiMSS, donde se define, de forma tabular, los actores, recursos y sistemas involucrados en las actuaciones de mejora asistencial bajo la definición de indicadores evaluables. Este objetivo estará especialmente preocupado por los mecanismos de aceptación, valor añadido y las estrategias de implantación que aseguren una funcionalidad de calidad de los Sistemas de Ayuda a la Decisión Médica.

Tabla 18.1: Implantación de funcionalidades CDSS en hospitales de EE.UU. según la encuesta de Jha et. al en [185]. Los encuestados incluyeron sistemas electrónicos de farmacia, laboratorio y radiología, por lo que no pueden considerarse sistemas interoperables en el hospital.

Funcionalidad	Implantación en Hospital (%)
Guía clínica	17
Recordatorio clínico	23
Alerta a las alergias de medicamentos	46
Alerta a interacción entre medicamentos	45
Alerta a interacción de medicamentos y pruebas de laboratorio	34
Dosimetría asistida	31

18.3. Implantación de CDSS

18.3.1. Tasas de implantación de los CDSS

La implantación de sistemas informáticos con funcionalidades específicas de CDSS puede considerarse todavía embrional. La mayoría de los sistemas actualemente implantados están basados en guías clínicas electrónicas de tecnologías de nivel 1 o 3, o sistemas de alertas basados en niveles 2 y 3 (ver sección 1.2. Jha en [185], identifica en su encuesta a hospitales de EE.UU. las tasas de implantación de funcionalidades CDSS recogidas en la tabla 18.1

Como puede observarse, las tasas de implantación recogidas por Jha son extraordinarimente elevadas comparadas con la tasa de implantación de EHR (únicamente 1.5 % de los hospitales). Jha analiza que estas tasas pueden ser devidas a que los encuestados incluyeron funcionalidades CDSS proporcionadas por sistemas electrónicos de farmacia, radiología y laboratorio, pero no son realmente sistemas interoperables con el resto del hospital, y por lo tanto con la información historia clínica de los pacientes, por lo que podemos considerar que estas cifras están sobrestimadas.

18.3.2. Metodología HIMSS de implantación

Osheroff et al. en [190] proponen una metodología de implantación de Ayuda a la Decisión Médica guiada por las necesidades de la institución sanitaria y la medición del impacto de la implantación mediante indicadores de eficiencia y eficacia. La figura 18.1 establece las etapas de alto nivel para la implantación de un CDSS, incluyendo las posibles salidas que permitan comprobar el cumplimiento de las etapas de forma práctica.

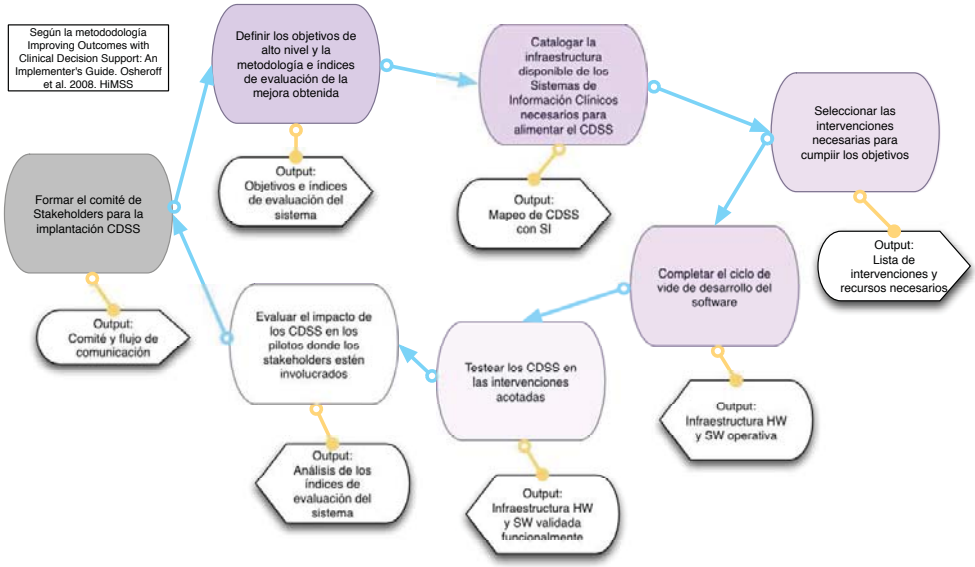


Figura 18.1: Etapas de alto nivel para la implantación de un CDSS y sus posibles salidas.

Los recursos humanos necesarios para realizar una implantación deben ser suficientes para asegurar un ajuste suave del CDSS al proceso asistencial y al perfil de los integrantes de la institución. Además, debe asegurarse el acceso a los siguientes grupos del sistema sanitario:

- Comité de Stakeholders. Direcciones de servicios médicos, direcciones del hospital y profesionales del hospital con conocimiento y capacidad de decisión en los centros a implementar. De aquí saldrán un conjunto de interlocutores directos con el director técnico y el implantador del CDSS. Definen y refinan los requisitos de alto nivel del sistema. Son informados de la evolución del desarrollo y deciden si se ha cumplido con el objetivo de la implantación.
- Interlocutores técnicos. Grupo de profesionales con conocimiento técnico que estarán involucrados en el uso del sistema o en tareas relacionadas con las entradas o salidas del mismo. Facilitan los requisitos técnicos del sistema.
- Ingeniero de conocimiento y/o documentalista. Son ingenieros de conocimiento del problema médico que aborda el CDSS por lo que facilitan las especificaciones funcionales al sistema.
- Data-manager. Conoce y entiende las fuentes de datos biomédicos a utilizar en la elaboración de los modelos de conocimiento.

18.4. CDSS como dispositivo médico

El ámbito de salud humana al que van dirigidos los CDSS y el software sanitario en general hacen que los niveles de seguridad aplicables a su funcionamiento y uso sean máximos. En Europa, estas consideraciones están recogidas por la directiva 93/42/CEE[a], modificada por la

[a]http://ec.europa.eu/enterprise/sectors/medical-devices/regulatory-framework

directiva 2007/47/CE del Parlamento Europeo y del Consejo de 5 de septiembre de 2007, y que ha entrado en vigor en marzo de 2010. En España, la directiva 2007/47/CE se ha implementado a través del Real Decreto 1591/2009, de 16 de octubre (BOE núm. 268, de 6 noviembre [RCL 2009, 2105]). Además, la Coordinación Europea de Organismos Notificados ha elaborado unas guías para promover una aproximación común a la directiva por parte de los desarrolladores y de los Organismos Notificados nacionales, que en el caso del software médico está recogida en la Recomendación NB-MED/2.2/Rec4[b]

Desde la modificación 2007/47/CE de directiva 93/42/CEE, cualquier programa informático utilizado solo o en combinación con finalidades específicas de diagnóstico y/o terapia destinado a ser utilizado en seres humanos con fines de diagnóstico, prevención, control, tratamiento o alivio de una enfermedad o diagnóstico, control, tratamiento, alivio o compensación de una lesión de una deficiencia, se considera *Producto Sanitario* (Artículo 1 de 93/42/CEE). Por lo tanto, un CDSS y el software sanitario en general, estan sujetos a la regulación específica de esta directiva comunitaria, por lo que los estados miembros de la UE adoptan las disposiciones necesarias para que los productos sólo puedan ser puestos en el mercado y/o ponerse en servicio si cumplen los requisitos establecidos en la presente Directiva cuando hayan sido debidamente suministrados, estén instalados y mantenidos adecuadamente y se utilicen con arreglo a su finalidad prevista (artículo 2 de la directiva 93/42/CEE).

Adcionalmente, un programa informático autónomo está considerado *producto sanitario activo*, y más específicamente, *productos sanitarios activos para diagnóstico* si está destinado a proporcionar información para la detección, el diagnóstico, el control o el tratamiento de estados fisiológicos, de estados de salud, de enfermedades o de malformaciones congénitas.

De forma general, los CDSS son *productos no invasivos*, y por lo tanto se incluyen en la *clase I* de productos sanitarios (Anexo IX de la directiva 93/42/CEE). Sin embargo, si los fines del software son el diagnóstico directo o la vigilancia de procesos fisiológicos vitales, entonces deben considerarse las condiciones especiales aplicables a productos activos para diagnóstico por lo que el software se incluiría en la *clase IIa*, y si la actividad supone un peligro inmediato para la vida, p.e. el funcionamiento cardíaco o SNC, entonces debe incluirse en la *clase IIb*. Además, como regla especial, los productos distinados específicamente al registro de imágenes radiográficas para diagnóstico se incluyen en la *clase IIa*.

Todos los productos sanitarios deben cumplir los requisitos esenciales establecidos en el Anexo I de la directiva 93/42/CEE que les sean aplicables habida cuenta de su finalidad prevista (Artículo 3 de la directiva 93/42/CEE). Así pues, los productos deberán diseñarse y fabricarse de forma tal ofrezcan las funcionalidades atribuidas por el fabricante durante el periodo previsto, y que su utilización no comprometa el estado clínico o la seguridad de los pacientes ni la seguridad y la salud de los usuarios. Los posibles riesgos asociados a la finalidad prevista deben ser aceptables en relación con el beneficio que proporcionen al paciente y compatibles con un nivel elevado de protección de la salud y de la seguridad. Esto implicará: i) la reducción del riesgo derivado de errores de utilización debidos a las características del producto y al entorno, y ii) tener en cuenta los conocimientos técnicos de los usuarios previstos. Los programas informáticos médicos deben ser validados con arreglo al estado actual de la técnica, teniendo en cuenta los principios de desarrollo del ciclo de vida, gestión de los riesgos, validación y verificación. En consonancia con esto, la NB-MED no recomienda el procedimiento de evaluación de la conformidad (CAP) típicamente usada para el resto de productos sanitarios, sino aplicar i) las metodologías de desarrollo asociadas a los ciclos de vida del software (requerimientos funcionales, planificación, control de riesgos, verificación y validación), ii) procedimientos de control de la documentación (trazabilidad del sistema) y gestión de la configuración del software, iii) gestión de la responsa-

[b]Esta versión de la recomendación (última a fecha de la edición de este libro) fue aceptada en noviembre de 2001, y por lo tanto no recoge las modificaciones de la directiva realizadas desde esta fecha.

bildad de los módulos desarrollados, y iv) control de compatibilidad de versiones y de hardware con el sistema.

Específicamente la evaluación de la conformidad de los productos de clase I deben seguir el procedimiento indicado por el Anexo VII de la directiva 93/42/CEE y efectuar, antes de la comercialización, la declaración CE de conformidad. Dicha declaración recogerá la descripción del producto, los resultados del análisis de riesgos, controles, evaluación preclínica y clínica, indicar los métodos de sustitución, y especialmente cumplir los requisitos esenciales especificados por el Anexo 1 de la directiva 93/42/CEE. Entre los requisitos esenciales a cumplir, el fabricante debe realizar una evaluación clínica mediante citación de publicaciones científicas, investigación clínica o una combinación de ambas. La investigación clínica deberá verificar las prestaciones y evaluar la seguridad del sistema, aplicando una metodología previamente planificada (p.e. RCT con número de casos precalculado) y anotación de acontecimiento adversos durante el ensayo (Anexo X de la directiva 93/42/CEE). Para los productos de clase I no será necesario acudir al organismo notificado del estado para obtener la evaluación de conformidad.

En la práctica existen guías y normas que implementan los procesos para cumplir los requisitos de la directiva 93/42/CEE. Las guías MEDDEV 2.5.1 y GHTF STED 2008 contienen indicaciones para la elaboración y el mantenimiento de los documentos técnicos incluidos en la documentación controlada para el marcado CE. Las normas ISO/IEC 62304 (Medical Device Software), ISO 14971 (Application of Risk Management to Medical Devices) e IEC 60601-1 (Medical Electrical Equipment, part 1) son las mejores guías para cumplir con los requerimientos exprexados en la directiva 83/42/CEE a partir de la modificación 2007/47/CE (la sección 17.9 enumera los elementos clave para el diseño de software médico según la norma IEC 62304). Por último, la norma armonizada ISO 13485 resulta de utilidad para demostrar el complimiento de los requisitos de gestión de la calidad de la directiva.

Otros mercados internacionales tienen sus propias normas que deben cumplir los dispositivos médicos para su comercialización En Estados Unidos de América la FDA regula los dispositivos médicos, en Canadá es HC, específicamente en China la SFDA, en Japón es JPSFB, en Australia la TGA, y en Sudamérica es MERCOSUR. Puede consultarse también la AHWP en Asia y GHTF para mercados globales occidentales.

18.5. Notas bibliográficas

El estándar australiano HB307-2007 en [7] elaborado por la National Electronic Health Records Taskforce cataloga los Sistemas de Ayuda a la Decisión Médica según la elaboración de conocimiento sobre el paciente. Kawamoto [177] y García-Gómez [178] identifican funcionalidades deseables en los Sistemas de Ayuda a la Decisión Médica.

La implantación integral de programas de ayuda a la decisión clínica es abordada de en [190], donde se define, de forma tabular, los actores, recursos y sistemas involucrados en las actuaciones de mejora asistencial bajo la definición de indicadores evaluables.

Bermejo coordina el monográfico [191] sobre seguridad en la información en entornos sanitarios, abordando los fundamentos, aspectos éticos y legales, requisitos de seguridad, su gestión y tecnologías que la implementan.

En 2018 la FDA ha aprobado la comercialización del primer CDSS como dispositivo médico usando inteligencia artificial para detectar retinopatía diabética [192]. La FDA redujo el riesgo limitando la indicación de uso para examinar a los adultos que no tienen síntomas visuales durante una retinopatía mayor que la leve, para remitirlos a un especialista en atención ocular [65].

Capítulo 19

Evaluación de ayuda a la decisión médica

En 1996, Miller estableció que la evaluación de los CDSS debe tener en cuenta *si el usuario junto con el sistema es mejor para realizar una tarea específica que el usuario sin asistencia* [193]. Desde entonces, se han propuesto diferentes métricas de evaluación, teniendo en cuenta la eficacia del CDSS para resolver la tarea, pero también el efecto que tiene sobre la eficiencia del proceso asistencial, la medición de tiempos, y de forma subjetiva, la aceptatibilidad y la percepción de los profesionales médico y los pacientes y el alcance de la implantación del sistema [194].

Para tener constancia del valor añadido que aporta un CDSS, será necesario comparar su efecto con la forma anterior de resolver la tarea. Una comparación honesta debe realizarse mediante el experimento adecuado, siendo la Prueba controlada aleatoria (RCT) el *gold standard* en investigación médica. Si bien algunos autores como Liu [195] piensan que el RCT no es la metodología adecuada para evular los HIS, ponen como excepción a los CDSS por estar directamente relacionados con el proceso de decisión médica sobre el paciente. Por lo tanto, actualmente, podemos considerar los RCT el procedimiento actualmente más aceptado para evaluación de CDSS.

En la literatura podemos encontrar un número significativo de evaluaciones de CDSS para analizar las tendencias actuales de su rendimiento en la práctica clínica. En general, los problemas médicos donde se han llegado a evaluar más CDSS coinciden con los problemas de salud más prevalentes en la población: diabetes, enfermedades cardiovasculares, afecciones respiratorias, cancer, depresión [184], sobretodo para la planificación y cribado de las enfermedades. Sintchenko en [181] observó un descenso de la variabilidad de las prescripciones de antibióticos que realizaban los médicos respecto a un panel de expertos mediante un CDSS basado en el indicador de riesgo de infección pulmonar (CPIS) en pacientes con riesgo de neumonía asociada a la ventilación automática. Tomando la decisión del panel de expertos como *gold stantard*, el CDSS redujo significativamente la variabilidad de las soluciones en comparación con el caso control y respecto al uso de guías clínicas validadas de prescripción de antibióticos. Garg no observó ningún CDSS de ayuda al diagnóstico que demostrara mayor eficacia en el resultado del paciente respecto a la práctica médica sin el sistema [184]. Esta observación ha llevado a centrar las evaluaciones actuales de los CDSS en la mejora de la eficiencia obtenida en la resolución de la tarea y menos en la mejora de los resultados de diagnóstico directamente [194].

Las funcionalidades más evaluadas, por ser las de mayor implantación, han sido los sistemas de alertas y prescripción de medicamentos [184, 186]. En el estudio de Garg[184], el 64 % de los CDSS analizados aportaron valor añadido a la eficiencia de los decisores médicos. Por funcionalidades, el 40 % de los sistemas de ayuda al diagnóstico aportaron valor añadido a la tarea, así como el 76 % de los sistemas recordatorios, el 62 % de los sistemas de planificación de

enfermedades, el 66 % de los sistemas de alertas y prescripción. Por otra parte, Pearson [186] observa una tasa de éxito mayor en sistemas implantados en hospitales que en ambulatorios. Sintchenko et al. compararon el uso de CDSS con el uso de únicamente guías de antibióticos y el uso de resultados de laboratorio, observando una mejora en el acierto del 11 % de media. Ramnarayan en [196] evaluó el CDSS ISABEL para diagnósticos y planificación de procedimientos diagnósticos en pediatría mediante un test pareado donde cada médico diagnosticaba sin usar el CDSS y una segunda vez utilizando la ayuda del sistema.

En este capítulo revisaremos las métricas de evaluación, y los test estadísticos de contraste de hipótesis, aplicables al análisis de resultados de las Prueba controlada aleatoria (RCT). Veremos también conceptos relacionados con la evaluación de los CDSS, como son metodologías de evaluación de la aceptabilidad, de la calidad del software, y normas de software sanitario. Acabaremos, como en los capítulos anteriores, con una revisión bibliográfica basada en casos de uso de los temas estudiados en el capítulo.

19.1. Métricas de evaluación

19.1.1. Métricas sobre la eficacia de la intervención

Las métricas típicas de evaluación de la eficacia de un CDSS con respuestas discretas son el error (ecuación 14.9), el acierto (ecuación 14.10), y para respuestas continuas el error cuadrático (ecuación 8.25); si bien las métricas más generalistas son el riesgo (ecuación 14.1) y la utilidad esperada. Como vemos, estas métricas coinciden con las estudiadas en la sección 14.5. Efectivamente, un médico o el tándem médico+CDSS son en definitiva predictores que dan su respuesta ante la observación de un nuevo caso.

En la sección 14.5 también definimos métricas que tienen en cuenta los verdaderos y falsos positivos y negativos, como son la *sensibilidad* y la *especifidad*, el *recall* (ecuación 14.11), la *precisión* (ecuación 14.12) y el Area Under the Curve (AUC) de la Receiver operating characteristic (ROC). También estudiamos la utilidad de métricas como GMOR (ecuación 14.13), BAR (ecuación 14.15) y GMRP (ecuación 14.16) ante circunstancias de desbalanceo en el número de muestras de test por clase.

19.1.2. Métricas sobre la eficiencia de la intervención

Muchos CDSS inciden directamente sobre la planificación de procesos asistenciales, por lo que es deseable medir la mejora obtenida gracias a la intervención. Esta mejora está directamente relacionada con la eficiencia de los sistemas sanitarios y por lo tanto en la provisión universal de servicios de salud, por lo que inciden en la calidad asistencial ofrecida a la población.

En este sentido, cuando medimos la eficiencia, no es la tasa de acierto de un modelo predictivo lo que nos interesa, sino métricas como la capacidad asistencial, el consumo de recursos, el coste económico, la tasa de errores médicos, la variabilidad inter e intracentro (o servicio) lo que nos da una visión del efecto del CDSS en el entorno donde interviene.

Algunos de las métricas de eficiencia utilizadas por los evaluadores de CDSS son:

- Número de citas intermedias (o recitaciones) por paciente
- Número de hospitalizaciones por paciente
- Número de reingresos por paciente
- Recursos requeridos (p.e. camas, sillones de tratamiento)

- Capacidad asistencial del servicio
- Coste económico del proceso asistencial
- Tasa de errores en la prescripción
- Número de eventos adversos por paciente
- Tasa de procedimientos repetidos
- Variabilidad asistencial inter e intracentro (o servicio)
- Tiempos de espera de los pacientes
- Tasa de involucración (y abandono) en programas de cribado
- Calidad de vida (desviación respecto a estándares poblacionales)
- Calidad de los registros de biobancos

Berner en [194] recoge la opinión de varios autores sobre lo adecuado de usar las métricas de eficiencia para evaluar sistemas de ayuda al diagnóstico, ya que reflejan más adecuadamente que la métricas de eficacia el impacto producido en el proceso asistencial aplicado sobre el paciente.

19.1.3. Métricas temporales de la intervención

El consumo de tiempo del médico que requiere la decisión es otro de los aspectos interesantes a medir cuando esta se apoya en un CDSS. Como premisa, un profesional quiere un sistema de ayuda a la decisión que le permita resolver los casos convencionales en 30 segundos, y le ayude a profundizar de forma analítica y comparativa en casos complejos.

Por lo tanto, no suele ser el tiempo la variable a minimizar de forma absoluta, pero si que debe tenerse en cuenta para mejorar la usabilidad de los interfaces de usuario, y para medir el posible impacto en la planificación de los recursos del sistema sanitario.

Sintchenko [181] observó en la prescripción de antibióticos en pacientes con neumonía asociada a la ventilación automática un aumento de 113 segundos a 245 segundos provocado por el CDSS. Los autores del estudio identificaron como posible causa la falta de entrenamiento del personal sanitario con el sistema, y no queda claro si la reducción de variabilidad obtenida compensaba el tiempo extra de intervención.

La medición del tiempo empleado puede extenderse a nivel de proceso asistencial, por lo que podría considerarse a si mismo una métrica de evaluación de la eficiencia.

19.1.4. Otras métricas de evaluación

Otras métricas utilizadas en los estudios pueden referirse a la *percepción* que tiene el personal sanitario y/o de los pacientes del efecto sobre la intervención del CDSS.

Holbrook et al. [197] midieron, como indicador complementario, el optimismo de los pacientes crónicos de diabetes II después del ensayo. Para ello, preguntaron a los pacientes sobre el efecto del sistema en el control de su diabetes, sobre la relación con sus proveedores de servicios sanitarios y sobre la calidad asistencial.

Sintchenko en [181] midió la confianza de los médicos al utilizar los CDSS, sin observar mejoras significativas en sus resultados. En escenarios como la confirmación diagnóstica del cáncer de mama durante el cribado se puede observar que la tasa de citas intermedias por paciente puede considerarse elevada, siendo posiblemente síntoma de incertidumbre en la decisión diagnóstica,

mostrando la necesidad de la adaptición de programadas de ayuda a la decisión que mejoren el proceso.

Algunos directivos de Sistemas de Información en Salud consideran que la evaluación de los sistemas informáticos no debe realizarse mediante métricas de eficiencia sanitaria, ya que consideran estas métricas influidas por múltiples factores. Algunas iniciativas como European Institute for Health Records (EUROREC)[a] han definido indicadores de calidad de HIS que incluyen funcionalidades de alertas médicas, sin embargo no incluyen muchas otras de las funcionalidades de los CDSS (ver sección 1.3).

19.1.5. Métricas compuestas

Los estudios centrados en la mejora de la calidad asistencial puede que necesiten métricas diferentes al acierto diagnóstico, la precisión pronóstica, o el riesgo. La calidad asistencial puede medirse mediante diferentes indicadores, generalmente calculados mediante la composición de varias observaciones.

Holbrook et al. [197] definieron una métrica compuesta por ocho factores de riesgo (presión sanguínea, colesterol, hemoglobina glucosilada, peso, dieta, problemas en el hígado, hábito tabáquico) para crónicos de diabetes tipo II para medir la calidad del cuidado de la enfermedad.

Sintchenko en [181] propone una métrica compuesta por la tasa de adopción del CDSS y la eficacia obtenida al usarlo.

Ramnarayan en [196] propone una métrica para medir conjuntamente la eficacia y la eficiencia de la ayuda de un CDSS que sugiere un conjunto de diagnósticos llamado plan. La métrica de Ramnarayan está compuesta por un factor de calidad diagnóstica (DQS: diagnostic quality score) y un factor de calidad del gestión clínica (MQS: management quality score). Ambos factores son calculados como funciones *de cada diagnóstico sugerido por el plan*, *la especifidad del plan* y la *sensibilidad del plan*. Sin embargo, DQS pondera cada diagnóstico sugerido en función de su verosimilitud vistas las características clínicas (eficacia) y la relevancia clínica del diagnóstico; miesntras que MQS pondera cada diagnóstico en función de lo apropiado que es para el proceso dignóstico y lo seguro que resulta la situación clínica para el paciente.

19.2. Contraste de hipótesis

Para poder saber si el efecto de un CDSS supone una mejora en la tarea intervenida se debe comparar con la práctica médica habitual mediante una métrica de evaluación. Un contraste de hipótesis nos permitirá tener la evidencia estadística del efecto del CDSS en el ámbito médico.

Una *hipótesis* estadística es la asunción sobre una o varias poblaciones. Esta hipótesis puede ser cierta o falsa. Un decisor puede aceptar o rechazar la hipótesis mediante un *contraste de hipótesis* basándose en la información extraída de muestras $\mathcal{S}$ de la población.

Se suele formular la hipótesis con intención de rechazarla, por lo que se llama *hipótesis nula* y se designa como H_0. Rechazar la hipótesis nula implica aceptar la *hipótesis alternativa*, H_1. Podemos tomar como hipótesis nula que la diferencia entre el valor que toma una variable (p.e. la métrica de evaluación) en dos poblaciones (una intervenida y otra control) es debida al azar, y por lo tanto no hay diferencia ($\theta = \theta_0$); la hipótesis alternativa podría ser que hay diferencia ($\theta \neq \theta_0$, constraste bilateral), que $\theta > \theta_0$ (contraste lateral izquierdo), o que $theta < theta_0$ (contraste lateral derecho).

Si un decisor rechaza la hipótesis nula H_0 cuando es verdadera, entonces se dice que comete un *error de tipo I* (típicamente un falso positivo), siendo la probabilidad de esta situación

[a] http://www.eurorec.org

$\alpha = p(\text{rechazar } H_0|H_0 \text{ verdadera})$, y se define como *nivel de significación* del contraste (llamándose *nivel de confianza* a $1-\alpha$). Si un decisor acepta H_0 cuando es falsa, entonces comete un *error de tipo II* (típicamente un falso negativo), siendo la probabilidad de esta circunstancia $\beta = p(\text{aceptar } H_0|H_0 \text{ falsa})$. Además, a la probabilidad de aceptar H_1 cuando es verdadera $p(\text{aceptar } H_1|H_1 \text{ verdadera}) = 1-\beta$ se le llama em potencia de contraste. La tabla 19.1 resume estas situaciones.

Tabla 19.1: Situaciones en un contraste de hipótesis.

	Aceptar H_0	Rechazar H_0
H_0 cierta	Correcto	Error tipo I (α)
H_0 falsa	Error tipo II (β)	Correcto

Normalmente se espera que la probabilidad α de cometer errores de tipo I sea inferior a 0,05, aunque también se usan otros niveles como $\alpha < 0{,}1$ o $\alpha < 0{,}01$ para adoptar niveles de significación más relajados o estrictas. Por otro lado, la probabilidad β de errores de tipo II se espera que sea inferior a 0,02 o a 0,01, lo que implica una potencia de constraste de 0,08 o 0,09, respectivamente.

Se debe elegir un estadístico T[b] cuya distribución de probabilidad $F(T|\theta)$ está relacionada con la hipótesis en estudio y es conocida. El contraste de hipótesis será la función:

$$\Phi(\mathcal{S}) = \begin{cases} 1 & T(\mathcal{S}) \in \Omega \\ 0 & T(\mathcal{S}) \notin \Omega, \end{cases} \tag{19.1}$$

donde $\mathcal{S} = 1$ indica que debemos rechazar la hipótesis nula H_0 (y aceptar H_1), mientras que $\mathcal{S} = 0$ indica que no hay evidencias para rechazar H_0. Ω es la región de rechazo, y se debe elegir de tal manera que la probabilidad de que $t(\mathcal{S})$ caiga es su interior sea baja si H_0 es verdadera, concretamente su probabilidad se establece al nivel de significación α del contraste.

Así por ejemplo, si establecemos el nivel de significación $\alpha < 0{,}05$ para un contraste bilateral de hipótesis basándonos en un estadístico con distribución gaussiana, entonces la probabilidad de la región de rechazo $F(\Omega) = 0{,}05$, debe distribuirse en ambas colas de la distribución, $\Omega = (-\infty, t_-] \cup [t_+, \infty)$, siendo $F(T < t_-) = 0{,}025$ y $F(T > t_+) = 0{,}025$ y por lo $t_- = -1{,}96$ y $t_+ = 1{,}96$. Por lo tanto, la hipótesis nula H_0 se rechaza si el valor del estadístico T es inferior a $-1{,}96$ o superior a 1,96. Si el contraste fuera lateral derecho, entonces $F(\Omega) = F(T > t_+) = 0{,}05$, por lo que la hipótesis nula es rechazada si $T > 1{,}65$.

En caso de realizar múltiples test sobre nuestra muestra debe tenerse en cuenta la posibilidad de obtener rechazos de hipótesis nula simplemente por repetición del test. Debe por lo tanto aplicarse una corrección por múltiple test para evitar conclusiones incorrectas. Bonferroni y False Discovery Rate (FDR) son algunas técnicas para este propósito.

19.2.1. Contrastes de dos distribuciones normales independientes

Vamos a comparar los valores que toman dos poblaciones gaussianas, $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$, de las que se han obtenido de forma independiente dos muestras $\mathcal{S}_1$ y $\mathcal{S}_2$ de tamaños N_1 y N_2. En las evaluaciones, típicamente, cada una de las muestras corresponde a un

[b] Un estadístico (muestral) es un valor en función de la muestra.

grupo de pacientes *intervención* sobre el que se aplica un CDSS y un grupo *control*, sobre el que se realiza la práctica habitual sin CDSS.

Como decimos, nos centraremos en distribuciones normales, lo que requiere la comprobación de dicha hipótesis mediante métodos gráficos (p.e. histogramas, diagramas de cajas o gráficos de normalidad) como mediante test estadísticos (test de Kolmogorov-Smirnov, test de Shapiro-Wilks). Si estos test no se cumplen, deberemos usar pruebas no paramétricas en lugar de los métodos expuestos aquí.

Así pues, estudiaremos una serie de test relacionados con la diferencia entre las medias de las poblaciones,

$$H_0 : \mu_1 - \mu_2 = 0, \tag{19.2}$$

asumiendo ciertas consideraciones sobre las varianzas.

Contrastes de medias con varianzas conocidas

Estamos interesados en estudiar la diferencia entre las medias de las dos distribuciones (ecuación 19.2), y asumimos conocidas las varianzas poblacionales, por lo que el estadístico

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim N(0,1), \tag{19.3}$$

que en el contraste bilateral con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 \neq 0, \tag{19.4}$$

si $|Z| > z_{1-\alpha/2}$, entonces rechazamos H_0; y en un contraste lateral derecho con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 > 0, \tag{19.5}$$

si $Z > z_{1-\alpha}$, entonces rechazamos H_0, siendo z_γ el cuantil γ de la distribución $N(0,1)$.

Comparación de medias homocedáticas

Seguimos interesados en comparar las medias de dos distribuciones (ecuación 19.2), pero ahora desconocemos sus varianzas, aunque sabemos que son iguales, por lo que utilizaremos el siguiente estadístico:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim t_{N_1+N_2-2}, \tag{19.6}$$

donde $\hat{S}^2$ es la cuasivarianza muestral ponderada de $\hat{S}_1^2$ y $\hat{S}_2^2$ es,

$$\hat{S}^2 = \frac{(N_1-1)\hat{S}_1^2 + (N_2-1)\hat{S}_2^2}{N_1+N_2-2}$$

y $t_{N_1+N_2-2,1-\alpha}$ es la distribución t de student con $N_1 + N_2 - 2$ grados de libertad.

En el contraste bilateral con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 \neq 0, \tag{19.7}$$

si $|T| > t_{N_1+N_2-2,1-\alpha/2}$, entonces rechazamos H_0; y en un contraste lateral derecho con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 > 0, \tag{19.8}$$

si $T > t_{N_1+N_2-2,1-\alpha}$, entonces rechazamos H_0, siendo $t_{df,\gamma}$ el cuantil γ de la distribución t de student con *df* grados de libertad.

Comparación de medias no homocedásticas

Ahora seguimos interesados en comparar las medias de dos distribuciones (ecuación 19.2), pero ahora desconocemos sus varianzas, y además son significativamente diferentes, por lo que debemos modificar el estadístico de la siguiente forma:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{N_2}}} \sim t_f, \tag{19.9}$$

donde t_f es la distribución t de student con f grados de libertad,

$$f = \frac{(\frac{\hat{S}_1^2}{N_1} + \frac{\hat{S}_2^2}{N_2})^2}{\frac{1}{N_1+1}(\frac{\hat{S}_1^2}{N_1})^2 + \frac{1}{N_2+1}(\frac{\hat{S}_2^2}{N_2})^2} - 2.$$

De nuevo, en el contraste bilateral con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 \neq 0, \tag{19.10}$$

si $|T| > t_{f,1-\alpha/2}$, entonces rechazamos H_0; y en un contraste lateral derecho con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 > 0, \tag{19.11}$$

si $T > t_{f,1-\alpha}$, entonces rechazamos H_0, siendo $t_{f,\gamma}$ el cuantil γ de la distribución t de student con f grados de libertad.

Comparación de medias ponderadas no homocedásticas

Una forma de tener en cuenta la agrupación de pacientes en factores naturales (p.e. pertenecer a un mismo centro de salud, ser diagnosticados por un mismo médico, pertenecer a la misma familia, pertenecer al mismo grupo social, etc) es realizar el test estadístico sobre los resultados de la evaluación sobre los grupos en lugar de realizarlo directamente sobre los pacientes individuales. De forma general, los grupos no tienen que estar compuestos por el mismo número de pacientes, por lo que no sería correcto asignar el mismo peso a cada grupo en el estudio, tal como lo haría el cálculo de una media aritmética. Podemos tener en cuenta que los grupos tienen diferente número de pacientes mediante el cálculo de medias ponderadas, y realizar un test estadístico de comparación de dichas medias que exprese esta circunstancia.

Sea X_{1i}, una muestra i al que se ha asignado un peso w_i procedende del conjunto $\mathcal{X}_1$ de N_1 elementos, y sea X_{2j} una muestra j con peso w'_j procedente del conjunto $\mathcal{X}_2$ de N_2 elementos, independiente de $\mathcal{X}_1$.

Asumimos que todas las muestras $X_{1i}, i = 1, \ldots, N_1$ siguen una distribución normal con media μ_1 y que las muestras $X_{2j}, j = 1, \ldots, N_2$ siguen una distribución con media μ_2,

$$X_{1i} \sim N(\mu_1, \frac{\alpha_1}{w_i}), i = 1, \ldots, N_1 \tag{19.12}$$

$$X_{2j} \sim N(\mu_2, \frac{\alpha_2}{w'_j}), j = 1, \ldots, N_2, \tag{19.13}$$

y que $\sum_{i=1}^{N_1} w_i = 1$ y $\sum_{j=1}^{N_2} w_j = 1$; y calculamos las medias ponderadas de cada grupo como

$$\bar{X}_1 = \sum_{i=1}^{N_1} w_i X_{1i} \tag{19.14}$$

$$\bar{X}_2 = \sum_{j=1}^{N_2} w'_j X_{2j}. \tag{19.15}$$

Al igual que en 19.2, podemos estudiar si las medias μ_1 y μ_2 son iguales, asumiendo la siguiente hipótesis nula:

$$H_0 : \mu_1 - \mu_2 = 0, \tag{19.16}$$

desconocidas α_1 y α_2. Para ello, planteamos necesitaremos el estadístico:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\hat{\alpha}_1 + \hat{\alpha}_2}} \sim t_u, \tag{19.17}$$

donde $\hat{\alpha}_1 = \frac{S_1}{N_1 - 1}$ y $\hat{\alpha}_2 = \frac{S_2}{N_2 - 1}$, siendo $S_1 = \sum_{i=1}^{N_1} w_i (X_{1i} - \bar{X}_1)^2$ y $S_2 = \sum_{j=1}^{N_2} w'_j (X_{2j} - \bar{X}_2)^2$; y donde t_u es la distribución t de student con u grados de libertad,

$$u = \frac{(\hat{\alpha}_1 + \hat{\alpha}_2)^2}{\frac{\hat{\alpha}_1^2}{N_1 - 1} + \frac{\hat{\alpha}_2^2}{N_2 - 1}}.$$

En el contraste bilateral con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 \neq 0, \tag{19.18}$$

si $|T| > t_{u,1-\alpha/2}$, entonces rechazamos H_0; y en un contraste lateral derecho con hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 > 0, \tag{19.19}$$

si $T > t_{u,1-\alpha}$, entonces rechazamos H_0, siendo $t_{u,\gamma}$ el cuantil γ de la distribución t de student con u grados de libertad.

19.2.2. Contrastes sobre la diferencia de proporciones

Las métricas de evaluación suelen ser variables derivadas de respuestas dicotómicas (distribuciones de Bernoulli). Por ejemplo, el accuracy (acierto) (acc) es la proporción de muestras donde el clasificador acierta respecto al total de muestras evaluadas. El número de aciertos en cada muestra tendrá una distribución binomial, por lo que la proporción será de un modo aproximado un distribución normal.

$$\hat{P}_1 \sim N(p_1, \frac{p_1(1-p_1)}{N_1}) \tag{19.20}$$

$$\hat{P}_2 \sim N(p_2, \frac{p_2(1-p_2)}{N_2}) \tag{19.21}$$

$$\tag{19.22}$$

Nos interesa comprobar si las proporciones p_1 y p_2 son iguales, por lo que definimos la hipótesis nula como

$$H_0 : p_1 - p_2 = 0, \tag{19.23}$$

y planteamos el estadístico

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{N_1} + \frac{\hat{p_2}(1-\hat{p_2})}{N_2}}} \sim N(0,1) \tag{19.24}$$

que en el contraste bilateral con hipótesis alternativa

$$H_1 : p_1 - p_2 \neq 0, \tag{19.25}$$

si $|Z| > z_{1-\alpha/2}$, entonces rechazamos H_0; y en un contraste lateral derecho con hipótesis alternativa

$$H_1 : p_1 - p_2 > 0, \tag{19.26}$$

si $Z > z_{1-\alpha}$, entonces rechazamos H_0, siendo z_γ el cuantil γ de la distribución $N(0,1)$.

19.2.3. Otros contrastes de hipótesis

Debe elegirse el contraste de hipótesis adecuado dependiendo del estudio y de la métrica de evaluación. La tabla 19.2 puede ser de utilidad para elegir el test estadístico a utilizar dependiendo del objetivo y del tipo de métrica.

19.2.4. Estimación del tamaño muestral

Cuando se realiza la evaluación de un sistema de ayuda a la decisión de forma descriptiva (y no comparativa), podemos calcular el tamaño muestral de nuestro experimento mediante los métodos presentados en 14.2.1.

Para la evaluación comparativa de un sistema de ayuda a la decisión (u otra intervención) se debe estimar el tamaño de la muestral necesario para realizar un contraste de hipótesis con cierto nivel de significación (α) y potencia ($1-\beta$). A continuación veremos como calcularlo para la comparación de medias y la comparación de proporciones[c].

Estimación del tamaño muestral para la comparación de proporciones

En el contraste sobre la diferencia de proporciones hemos supuesto que la estimación de la diferencia de proporciones $\hat{p}_1 - \hat{p}_2$ tiene un error estandar $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}$, por lo que el intervalo en el que confiamos al $1-\alpha$ que esté contenido el valor real de la diferencia $p_1 - p_2$ es

$$p_1 - p_2 \in \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} SE.$$

Asumiendo que se quiere realizar un estudio con $N_1 + N_2 = N$ muestras, que esperamos unas proporciones $\tilde{p}_1, \tilde{p}_2$ y teniendo en cuenta el nivel de significancia α y la potencia $1-\beta$, se puede calcular N como

$$N = \frac{2(z_{\alpha/2}\sqrt{2\tilde{p}_m(1-\tilde{p}_m)} + z_\beta\sqrt{\tilde{p}_1(1-\tilde{p}_1) + \tilde{p}_2(1-\tilde{p}_2)})^2}{(\tilde{p}_1 - \tilde{p}_2)^2}, \tag{19.27}$$

donde $\tilde{p}_m = \frac{\tilde{p}_1 + \tilde{p}_2}{2}$.

Ejemplo 19.1 (Tamaño muestral en la evaluación de CDSS en talasemias) Se quiere comparar la eficacia obtenida mediante el uso de un CDSS en el diagnóstico de anemias talasémicas con el procedimiento diagnóstico habitual. Estudios preliminares indican que la eficacia del procedimiento habitual es de $\tilde{p}_H = 0{,}75$, mientras que se espera aumentar hasta $\tilde{p}_C = 0{,}9$ mediante el CDSS. ¿Cuál es el número de muestras necesarias para realizar un contraste de comparación de proporciones con un nivel de confianza $\alpha = 0{,}05$ y una potencia $1-\beta = 0{,}9$?

La aplicación directa de 19.27 indica que se debe realizar una captura de $N = 266$ casos para la comparar efecto del CDSS con el procedimiento habitual.

Estimación del tamaño muestral para la comparación de medias

De forma similar a la comparación de proporciones, podemos estimar el tamaño muestral $N = N_1 + N_2$ necesario para realizar una comparación de dos medias $\tilde{\mu}_1, \tilde{\mu}_2$ de varianza igual y conocida σ^2 teniendo en cuenta el nivel de significancia α y la potencia $1-\beta$ como

$$N = \frac{4\sigma^2(z_{\alpha/2} + z_\beta)^2}{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}. \tag{19.28}$$

[c]Podemos encontrar unas calculadoras de los métodos a continuación explicados en `http://www.rad.jhmi.edu/jeng/javarad/samplesize`

Tabla 19.2: Test estadístico adecuado dependiendo del objetivo y del tipo de métrica.

	Tipo de métrica			
Objetivo	Medida (Gaussiana)	Medida (No Gaussiana)	Rangos, puntuaciones y binomial (2 posibles valores)	Tiempo de supervivencia
Comparación de muestra con población	t-test (1 muestra)	Wilcoxon	χ^2, Binomial	
Comparación de dos muestras no pareadas	t-test (no pareado)	Mann-Whitney	Fisher's, χ^2, Binomial	Log-rank, Mantel-Haenszel
Comparación de dos muestras pareadas	t-test (pareado)	Wilcoxon	McNemar's	Regresión de riesgos proporcionales condicionales
Comparación de tres o más muestras no pareadas	análisis de varianza (one-way anova)	Kruskal-Wallis	χ^2	Regresión de riesgos proporcionales condicionales
Comparación de tres o más muestras pareadas	análisis de varianza (repeated-measures anova)	Friedman	Cochrane Q	Regresión de riesgos proporcionales condicionales
Comparación de muestras no pareadas	t-test (no pareado)	Mann-Whitney	Fisher's, χ^2, Binomial	Log-rank, Mantel-Haenszel
Comparación de muestras pareadas	t-test (pareado)	Wilcoxon	McNemar's	Regresión de riesgos proporcionales condicionales
Asociación entre dos variables	Pearson correlation	Spearman correlation	Contingency coefficientes	
Valor predictivo desde otra variable medida	Regresión lineal simple o regresión no-lineal	regresión no-paramétrica	Regresión logística simple	Regresión de riesgos proporcionales condicionales
Valor predictivo desde otra variable medida	Regresión lineal múltiple o regresión no-lineal múltiple		Regresión logística múltiple	Regresión de riesgos proporcionales condicionales

Los efectos de correlación entre las muestras (p.e. por encontrarse en un mismo grupo poblacional o un mismo hospital) y la contaminación durante el experimento (p.e. provocado por el efecto indirecto de la intervención en los pacientes de control), suelen corregirse mediante el agrupamiento de pacientes y el aumento del tamaño muestral. Veremos estas correcciones en la sección 19.3.1.

19.3. Prueba Controlada Aleatoria (RCT)

RCT (del inglés *Randomized Controlled Trial*) es un tipo de experimentación científica ampliamente utilizado para comprobar la eficacia y eficiencia de tecnologías, servicios, y tratamientos médicos. En una RCT los individuos son asignados aleatoriamente[d] a un grupo *intervención*, sobre el aplica la tecnología experimental, o a un grupo *control* sobre el que se sigue aplicando la metodología habitual.

Etapas de un RCT

Una RCT consta de tres etapas sucesivas:

1. Diseño del estudio.

   Contribuye en un 30 %-40 % a la validez y fiabilidad de la prueba, y consiste de las siguientes tareas:

   - Configuración del protocolo

     Donde se elabora el documento marco del estudio, y se recoge el diseño del mismo, que será utilizado durante su desarrollo.
   - Definición de las métricas de evaluación

     Elegimos la métrica primaria y las secundarias a estudiar, con el fin de ofrecer la máxima evidencia del valor añadido del estudio.
   - Definición de los estudios de inclusión y exclusión de pacientes

     Deben definirse los pacientes que se beneficiarán de la intervención, atendiendo a la protección de pacientes definida por la ICH Guideline on Good Clinical Practice (ICH E6).
   - Plan del análisis estadístico y tamaño de la muestra

     En esta tarea debe especificarse, para cada métrica de evaluación, el estudio estadístico a realizar. Para comparar los valores obtenidos para cada métrica de evaluación en los grupos *intervención* y *control* elegiremos un contraste de hipótesis adecuado según el tipo de variable aleatoria que sea la métrica (ver sección 19.2).

     Por motivos éticos y económicos, debe minimizarse el tamaño de la muestra utilizada en la prueba asegurando el nivel de significancia y la potencia del estudio, por lo que deberemos estimar el tamaño idóneo de la misma, como estudiamos en la sección 19.2.4.
   - Diseño de formularios para el registro de casos

     Debe diseñarse el formulario y los campos que contendrá cada caso, para ello se seguirá el protocolo del estudio.

[d] Si bien el método de aleatoriedad completa es el básico, existen diferentes alternativas para realizar esta asignación con el fin de corregir desbalanceos entre grupos o por superar restricciones éticas (ver Wikipedia: Randomized Controlled Trial).

- Diseño logístico de la realización del estudio

  Deben reservarse los recursos y tiempo necesarios para realizar el estudio, así como prever posibles contingencias y dotar de mecanismos para solucionarlas.

2. Realización del estudio

   La realización del estudio supone entre el 50 % y el 60 % de la calidad del mismo. Durante la realización se deben abordar las siguientes tareas:

   - Aleatorización de casos

     Esta tarea es la encargada de asignar a cada paciente un grupo de estudio, y puede seguir diferentes estrategias: aleatorización completa, permutación de bloques (para formar el balanceo de clases), aleatorización adaptativa por la covarianza, etc.

   - Enmascaramiento

     Cuando evaluamos CDSS no tiene mucho sentido hablar de test ciegos o doble ciegos, ya que para el paciente es trasparente el método por el que decide el médico, y el médico sabe si está utilizando el CDSS o no; en cualquier caso, debe minimizarse en lo posible la contaminación entre los grupos *intervención* y *control*.

   - Diseño del repositorio de información del ensayo

     Es necesario un repositorio digital más o menos sofisticados para registrar la información obtenida durante la prueba. Cada vez más se buscan soluciones integradas con la historia clínica electrónica de los pacientes para evitar duplicados y registrar eventos asociados a los ensayos de forma autónoma.

   - Monitorización del estudio

     Debe asegurarse el desarrollo, registro y recopilación de los datos siguiendo rigurosamente el protocolo del estudio para aumentar la calidad del mismo al máximo. En la medida de lo posible, la monitorización debe realizarla personal externo al grupo investigador de la intervención.

3. Análisis estadístico e informe de resultados

   Esta última etapa supone entre el 10 % y el 20 % de la calidad del estudio, ya que depende en gran medida de la planificación del análisis estadístico y del desarrollo del estudio como entradas de la etapa.

   En análisis estadístico seguirá el plan desarrollado durante el diseño del estudio. El informe y sus conclusiones debe publicarse en publicaciones científicas del dominio de la aplicación y/o de ayuda a la decisión médica. Deben reflejarse los detalles necesarios para reproducir el estudio y deben presentarse las consideraciones éticas derivadas del mismo.

Ejemplo 19.2 (RCT de un CDSS para tumores cerebrales)
El equipo multidisciplinar de ayuda a la decisión de un Sistema de Salud desea evaluar el valor añadido que aporta un CDSS a la eficacia del diagnóstico radiológico de tumores cerebrales mediante espectroscopía de resonancia magnética.

Por lo tanto, se plantea comprobar mediante un RCT que un radiólogo que dispone del CDSS acierta más en el diagnóstico que un radiólogo sin el CDSS. El sistema de salud en el que han implantado el sistema dispone de un número suficiente de radiólogos, pero ha reservado recursos para involucrar a 34 de ellos en el estudio.

Tras la etapa de desarrollo del estudio, se obtuvo del resultado del estudio que resume la tabla 19.3[e] por radiólogo.

[e]Basado en Kerry SM, Bland JM, BMJ,1998;316:54

Tabla 19.3: RCT para la evaluación comparativa del valor añadido de un CDSS. Se resume el resutado del ensayo por radiólogo, como porcentaje de pacientes donde se acierta el diagnóstico (en comparación con la histopatología).

Grupo intervención		Grupo control	
Aciertos	Total	Aciertos	Total
20	20	7	7
7	7	33	37
15	16	32	38
28	31	23	28
18	20	16	20
21	24	15	19
6	7	7	9
5	6	19	25
25	30	90	120
53	66	64	88
4	5	15	22
33	43	52	76
32	43	14	21
16	23	83	126
44	64	14	22
4	6	21	34
10	18	4	10
341	429	509	702

Si acc_i (de intervención) a la proporción de aciertos mediante el uso del CDSS y acc_c (de control) a la proporción de aciertos de un caso sin el uso del CDSS, podemos enunciar las hipótesis nula y alternativa del constraste sobre la diferencia de proporciones (sección 19.2.2) como:

$$H_0 : acc_i = acc_c H_1 : acc_i > acc_c,$$

que como vemos en H_1 es un *contraste lateral derecho* ya que el interés está en comprobar si el uso de CDSS mejora el acierto del radiólogo.

Con la muestra obtenida y asumiendo independencia entre los casos, a través del test de diferencia de proporciones 19.2.2, podemos decir que la diferencia de 0,07 entre las proporciones de acierto al usar CDSS y no usarlo es significativa con un nivel $\alpha = 0{,}004$.

19.3.1. Limitaciones de las pruebas controladas aleatorias

En la práctica, las condiciones de los entornos reales llevan a dos efectos no deseados en las pruebas: la *correlación* y la *contaminación* de las muestras [198].

Una limitación de las pruebas controladas aleatorias sobre pacientes individuales es no tener en cuenta que estos pacientes pueden estar agrupados por factores comunes como pertenecer a la misma familia, ir al mismo hospital, o, sobretodo, ser atendido por el mismo médico. No tener en cuenta este agrupamiento natural de pacientes, puede llevar a conclusiones erróneas debido a la *correlación* entre los pacientes y a la *contaminación* que se produce en el grupo de pacientes control a través del beneficio indirecto de la intervención. Para solucionar este problema, es necesario incorporar los agrupamientos naturales de pacientes en el análisis de las pruebas controladas aleatorias [198].

La *contaminación* es la extensión del efecto de la intervención de un grupo a otro. Puede ocurrir cuando el grupo control está expuesto a la intervención o el grupo intervención está expuesto a tratamiento control. El medio de transmisión puede ser el propio médico involucrado en el experimento, ya que puede aplicar la experiencia obtenida durante la intervención en los pacientes control. También los propios pacientes pueden ser el medio de transmisión en ciertos ensayos por estar próximos físicamente o relacionados entre sí. La consecuencia de la contaminación es la atenuación de la intervención y por lo tanto una tendencia a obtener un resultado del ensayo falsamente negativo. Las soluciones a la contaminación pueden ser dos:

- Aumentar el tamaño muestral del experimento en un factor $\frac{1}{(1-contamination)^2}$, siendo *contamination* la proporción del efecto atenuado; lo que, por lo tanto, reducirá el efecto relativo de la contaminación en la comparación.
- Separar los sujetos control de los sujetos intervención, que elimina el efecto de la contaminación.

La *correlación* entre los pacientes de un agrupamiento puede deberse a la pertenencia a grupos poblacionales comunes (edad, sexo, localización, etc). Además, pacientes visitados por el mismo médico pueden recibir tratamientos más parecidos que pacientes visitados por diferentes médicos. Por último, la correlación puede deberse a transmisiones directas entre pacientes, que pueden ser debidas a genotipos familiares, o hábitos o comportamientos comunes. A su vez, los médicos de un mismo servicio u hospital también estarán influidos por las mismas condiciones del entorno, por lo que su comportamiento también podrá tener una correlación positiva.

Si existe correlación entre pacientes de un agrupamiento pero no se consideran en el estudio, pueden darse las siguientes circunstancias:

- Si los agrupamientos se producen entre los pacientes del grupo de intervención y del grupo control (circunstancia más común), el resultado de la prueba tenderá a ocultar la posible

diferencia significativa entre los grupos de estudio, aumentando la varianza de cada grupo innecesariamente, por lo tanto a un resultado falsamente negativo del ensayo. Si en el experimento se tienen en cuenta estos agrupamientos, se podrá disminuir la varianza, obteniendo estimaciones más precisas de los estadísticos de comparación.

- Si los agrupamientos tienen pacientes control o pacientes intervención pero no de ambos, entonces el resultado puede llevar a resultados falsamente positivos en el ensayo, ya que la diferencia entre agrupamientos contribuye a la diferencia entre el grupo intervención y el grupo control, y debería tenerse en cuenta en la varianza de la estimación.

Podemos corregir el efecto de la correlación multiplicando el tamaño de la muestra en un factor $Deff = 1 + (m-1)\rho$, donde $\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}$ y σ_b^2, σ_w^2 son las desviaciones intra e inter-grupo respectivamente.

Otra forma de solucionar la correlación entre casos es tener en cuenta los grupos naturales que forman, diseñando Prueba controlada aleatoria por grupos (clustered RCT) (CRCT).

Prueba Controlada Aleatoria por grupos

Una CRCT tiene en cuenta covariables que agrupan a los casos en conjuntos naturales y resolviendo la comparación entre la intervención y el control teniendo en cuenta estos grupos naturales. En el siguiente ejemplo veremos como tratar este problema mediante el contraste de medias ponderadas vista en la sección 19.2.1.

Ejemplo 19.3 (CRCT de un CDSS para tumores cerebrales (cont. ejemplo 19.3)) El resultado del ejemplo 19.3 puede considerarse optimista debido a la correlación que puede haber entre los pacientes de un mismo radiólogo. Como disponemos de los resultados del estudio agrupados por radiólogo, podemos aplicar el contraste de medias ponderadas vista en la sección 19.2.1.

Con la muestra obtenida y teniendo en cuenta la agrupación por radiólogo de los pacientes, estudiamos la media del acierto de los radiólgogos, pudiendo decir que la diferencia de 0,07 entre la media de acierto de cada radiólogo al usar CDSS y no usarlo es significativa con un nivel $\alpha = 0{,}03$. Claramente este resultado corrige el optimismo del ejemplo 19.3 por no tener en cuenta el agrupamiento por radiólogo.

19.4. Evaluación de la aceptabilidad del CDSS

La utilidad y usabilidad percibidas por los usuarios son buenos indicadores de la actitud que luego tendrán en el momento de utilizar un sistema informático si se incorpora en sus procesos de trabajo.

La utilidad percibida se define como el grado que una persona cree que el uso de un sistema particular aumentará el rendimiento de su trabajo. La usabilidad percibida se define como el grado que una persona cree que podrá usar un software sin esfuerzo. TAM es una teoría de sistemas de información que modela la aceptabilidad y uso de una tecnología basada en la utilidad percibida y la usabilidad percibida.

Davis estableció las bases del TAM para medir la aceptabilidad mediante la usabilidad y la utilidad percibidas [157, 199]. Davis implementó TAM mediante un cuestionario de 12 preguntas con respuestas en escala Likert (escala del 1 al 7 donde 1 es muy en desacuerdo y 7 es totalmente de acuerdo). De las doce preguntas la mitad son sobre la utilidad percibida y la otra mitad sobre la usabilidad percibida:

- Q1. El uso del CDSS haría más sencilla mi labor de atención y seguimiento de pacientes
- Q2. El uso del CDSS mejoraría la productividad de mi tarea
- Q3. El CDSS mejoraría mi eficacia en la atención y seguimiento de pacientes
- Q4. El uso del CDSS en mi trabajo me permitiría cumplir mis tareas más rápidamente
- Q5. El uso del CDSS mejoraría la calidad de mi asistencia a pacientes
- Q6. Encuentro el CDSS útil para mi trabajo de atención y seguimiento de pacientes
- Q7. Consideró que me será fácil aprender a utilizar el CDSS
- Q8. Creo que sería sencillo para mí emplear el CDSS para realizar la tarea de atención y seguimiento de pacientes
- Q9. Mi interacción con el CDSS sería clara y entendible
- Q10. La interacción con el CDSS sería flexible
- Q11. Sería fácil para mi llegar a ser hábil usando el CDSS
- Q12. Considero que el CDSS será fácil de usar

De los resultados obtenidos midiendo la aceptabilidad con TAM mediante un RCT pueden extraerse estadísticas descriptivas y realizar contrastes de hipótesis, tanto de las preguntas por separadas como agrupadas por usabilidad y utilidad. Hay que tener en cuenta que la escala Likert obtiene poblaciones no Gaussianas, por lo que será mejor utilizar métodos para rangos.

19.5. Notas bibliográficas

El efecto de los sistemas de ayuda a la decisión clínica sobre la práctica clínca ha sido evaluado en [177, 200–202]

Pearson et al. [203] hacen una revisión del efecto de los sistemas de ayuda a la decisión basándose en 56 artículos publicados en Medline, Embase y PsychINFO desde 1990 hasta 2007.

Friedman y Wyatt desarrollan en [204] un análisis teórico/práctico sobre la evaluación de sistemas de salud.

Bibliografía

[1] Hood LE, Galas DJ (2008). P4 Medicine : Personalized , Predictive , Preventive , Participatory A Change of View that Changes Everything

[2] Tenenbaum JD, Avillach P, Benham-Hutchins M, Breitenstein MK, Crowgey EL, Hoffman MA, Jiang X, Madhavan S, Mattison JE, Nagarajan R, Ray B, Shin D, Visweswaran S, Zhao Z, Freimuth RR (2016) An informatics research agenda to support precision medicine: seven key areas. Journal of the American Medical Informatics Association 23(4): 791–795

[3] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. BMJ 312(7023): 71–72

[4] Eddy DM (2005) Evidence-based medicine: a unified approach. Health affairs 24(1): 9–17

[5] L Z, W Z (2018) Informatics, data science, and artificial intelligence. JAMA 320(11): 1103–1104

[6] Robledo JV (2012) Clinical Decision Support Systems for Brain Tumour Diagnosis: Classification and Evaluation Approaches. Ph.D. thesis, Editorial Universitat Politècnica de València

[7] Grain H (2007) Guide to the principles and desirable features of clinical decision support systems. Standards Australia, Sydney

[8] Berlin A, Sorani M, Sim I (2006) A taxonomic description of computer-based clinical decision support systems. Journal of Biomedical Informatics 39(6): 656–667

[9] Leaper DJ, De Dombal FT, Horrocks JC, Staniland JR (1972) Computer-assisted diagnosis of abdominal pain using estimates provided by clinicians. Br J Surg 59(11): 897–898

[10] Leaper DJ, Horrocks JC, Staniland JR, De Dombal FT (1972) Computer-assisted diagnosis of abdominal pain using "estimates" provided by clinicians. Br Med J 4(5836): 350–354

[11] Shortliffe EH, Scott AC, Bischoff MB, Campbell AB, va Melle W, Jacobs CD (1981) ONCOCIN: An expert system for oncology protocol management. In Seventh International Joint Conference on Artificial Intelligence. Vancouver

[12] Shortliffe EH (1981) ONCOCIN: An aid for the outpatient management of cancer patients. In Proceedings of the Society for Computer Medicine and the Society for Advanced Medical Systems. Washington, DC

[13] Shortliffe EH, Perreault LE, editors (2000) Medical Informatics, Computer Applications in Health Care and Biomedicine. Springer

[14] Shortliffe EH, Cimino JJ, editors (2006) Biomedical Informatics, Computer Applications in Health Care and Biomedicine. Springer, third edition

[15] Berner ES, editor (2007) Clinical decision support systems: theory and practice. Springer, health inf edition

[16] Lisboa PJG, Wong H, Harris P, Swindell R (2003) A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artif Intell Med 28(1): 1–25

[17] Tan KC, Yu Q, Heng CM, Lee TH (2003) Evolutionary computing for knowledge discovery in medical diagnosis. Artif Intell Med 27(2): 129–154

[18] Andrews RJ, Mah RW (2003) The NASA Smart Probe Project for real-time multiple microsensor tissue recognition. Stereotact Funct Neurosurg 80(1-4): 114–119

[19] Nattkemper TW, Arnrich B, Lichte O, Timm W, Degenhard A, Pointon L, Hayes C, Leach MO (2005) Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. Artif Intell Med 34(2): 129–139

[20] Dasmahapatra S, Dupplaw D, Hu B, Lewis PH, Shadbolt N (2005) Ontology-Mediated Distributed Decision Support for Breast Cancer. In S Miksch, J Hunter, ET Keravnou, editors, AIME 2005: 10th Conf on Artificial Intelligence in Medicine, volume 3581 of *Lect. Notes Comput. Sc.*, pp. 221–225. Springer-Verlag, Aberdeen, UK

[21] Markey MK, Tourassi GD, Margolis M, DeLong DM (2006) Impact of missing data in evaluating artificial neural networks trained on complete data. Comput Biol Med 36(5): 516–525

[22] Biganzoli EM, Boracchi P, Ambrogi F, Marubini E (2006) Artificial neural network for the joint modelling of discrete cause-specific hazards. Artif Intell Med 37(2): 119–130

[23] Leinsinger G, Schlossbauer T, Scherr M, Lange O, Reiser M, Wismuller A (2006) Cluster analysis of signal-intensity time course in dynamic breast MRI: does unsupervised vector quantization help to evaluate small mammographic lesions? Eur Radiol 16(5): 1138–1146

[24] Abidi SR, Abidi SSR, Hussain S, Shepherd M (2007) Ontology-based modeling of clinical practice guidelines: a clinical decision support system for breast cancer follow-up interventions at primary care settings. Medinfo 12(Pt 2): 845–849

[25] Hassanien AE (2007) Fuzzy rough sets hybrid scheme for breast cancer detection. Image Vision Comput 25(2): 172–183

[26] Iakovidis DK, Maroulis DE, Karkanis SA (2006) An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. Comput Biol Med 36(10): 1084–1103

[27] Zheng MM, Krishnan SM, Tjoa MP (2005) A fusion-based clinical decision support for disease diagnosis from endoscopic images. Comput Biol Med 35(3): 259–274

[28] Lucas PJ, Boot H, Taal BG (1998) Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. Methods Inf Med 37(3): 206–219

[29] Mofidi R, Deans C, Duff MD, de Beaux AC, Paterson Brown S (2006) Prediction of survival from carcinoma of oesophagus and oesophago-gastric junction following surgical resection using an artificial neural network. Eur J Surg Oncol 32(5): 533–539

[30] van Oosterhout EM, Talmon JL, De Clercq PA, Schouten HC, Jansen MP, Hasman A (2003) The PropeR way to support medical doctors in daily practice. Developing the protocol based DSS. Stud Health Technol Inform 95: 340–345

[31] Belacel N, Wang Q, Richard R (2005) Web-integration PROAFTN methodology for acute leukemia diagnosis. Telemed J E Health 11(6): 652–659

[32] Foran DJ, Comaniciu D, Meer P, Goodell LA (2000) Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy. IEEE Trans Inf Technol Biomed 4(4): 265–273

[33] Chen W, Meer P, Georgescu B, He W, Goodell LA, Foran DJ (2005) Image mining for investigative pathology using optimized feature extraction and data fusion. Comput Methods Programs Biomed 79(1): 59–72

[34] Covell DG, Wallqvist A, Rabow AA, Thanki N (2003) Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. Mol Cancer Ther 2(3): 317–332

[35] Futschik ME, Reeve A, Kasabov N (2003) Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. Artif Intell Med 28(2): 165–189

[36] Tung WL, Quek C (2005) GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data. Artif Intell Med 33(1): 61–88

[37] Yuan X, Yang Z, Zouridakis G, Mullani N (2006) SVM-based texture classification and application to early melanoma detection. Conf Proc IEEE Eng Med Biol Soc 1: 4775–4778

[38] Sboner A, Eccher C, Blanzieri E, Bauer P, Cristofolini M, Zumiani G, Forti S (2003) A multiple classifier system for early melanoma diagnosis. Artif Intell Med 27(1): 29–44

[39] Debeir O, Decaestecker C, Pasteels JL, Salmon I, Kiss R, Van Ham P (1999) Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions. Cytometry 37(4): 255–266

[40] Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H (2000) Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. Melanoma Res 10(6): 556–561

[41] Carrara M, Bono A, Bartoli C, Colombo A, Lualdi M, Moglia D, Santoro N, Tolomio E, Tomatis S, Tragni G, Santinami M, Marchesini R (2007) Multispectral imaging and artificial neural network: mimicking the management decision of the clinician facing pigmented skin lesions. Phys Med Biol 52(9): 2599–2613

[42] Tsai CA, Chen DT, Chen JJ, Balch CM, Thompson JF, Soong SJ (2007) An integrated tree-based classification approach to prognostic grouping with application to localized melanoma patients. J Biopharm Stat 17(3): 445–460

[43] Kawazu T, Araki K, Kanda S (2001) Application of neural networks to the prediction of lymph node metastasis in oral cancer. In CARS, pp. 1295–1296

[44] Nayak GS, Kamath S, Pai KM, Sarkar A, Ray S, Kurien J, D'Almeida L, Krishnanand BR, Santhosh C, Kartha VB, Mahato KK (2006) Principal component analysis and artificial neural network analysis of oral tissue fluorescence spectra: classification of normal premalignant and malignant pathological conditions. Biopolymers 82(2): 152–166

[45] Wigren T, Kolari P (1994) Evaluation of a decision-support system for inoperable non-small cell lung cancer. Methods Inf Med 33(4): 397–401

[46] Coppini G, Diciotti S, Falchini M, Villari N, Valli G (2003) Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. IEEE Trans Inf Technol Biomed 7(4): 344–357

[47] Campadelli P, Casiraghi E, Artioli D (2006) A fully automated method for lung nodule detection from postero-anterior chest radiographs. IEEE Trans Med Imaging 25(12): 1588–1603

[48] Matsopoulos GK, Mouravliansky NA, Asvestas PA, Delibasis KK, Kouloulias V (2005) Thoracic non-rigid registration combining self-organizing maps and radial basis functions. Med Image Anal 9(3): 237–254

[49] Bala M (2004) [Determining the possibility of collecting reliable data for use in decision making in health care on the example of cost-effectiveness analysis of methods used in smoking cessation]. Przegl Lek 61(10): 1180–1183

[50] Lenhard REJ, Waalkes TP, Herring D (1983) Evaluation of the clinical management of cancer patients. A pilot study. JAMA 250(24): 3310–3316

[51] Shi H, Lyons-Weiler J (2007) Clinical decision modeling system. BMC Med Inform Decis Mak 7: 23

[52] McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, Fry MJ (2007) Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 14(6): 736–745

[53] Spyridonos P, Cavouras D, Ravazoula P, Nikiforidis G (2002) A computer-based diagnostic and prognostic system for assessing urinary bladder tumour grade and predicting cancer recurrence. Med Inform Internet Med 27(2): 111–122

[54] Papageorgiou EI, Spyridonos PP, Stylios CD, Ravazoula P, Groumpos PP, Nikiforidis GN (2006) Advanced soft computing diagnosis method for tumour grading. Artif Intell Med 36(1): 59–70

[55] Sylvester RJ, van der Meijden APM, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, Newling DWW, Kurth K (2006) Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol 49(3): 466–465

[56] Tachibana M, Miyakawa A, Deguchi N, Baba S, Murai M, Tazaki H (1994) A new scoring system based on the histological behavior and proliferative activity of tumor cells for grading the malignant potential of bladder cancers. Int J Urol 1(1): 37–42

[57] Papageorgiou EI, Stylios CD, Groumpos PP (2003) An integrated two-level hierarchical system for decision making in radiation therapy based on fuzzy cognitive maps. IEEE Trans Biomed Eng 50(12): 1326–1339

[58] Kelm BM, Menze BH, Zechmann CM, Baudendistel KT, Hamprecht FA (2007) Automated estimation of tumor probability in prostate magnetic resonance spectroscopic imaging: pattern recognition vs quantification. Magn Reson Med 57(1): 150–159

[59] Zhu Y, Williams S, Zwiggelaar R (2006) Computer technology in detection and staging of prostate carcinoma: a review. Med Image Anal 10(2): 178–199

[60] Mattfeldt T, Trijic D, Gottfried HW, Kestler HA (2004) Classification of incidental carcinoma of the prostate using learning vector quantization and support vector machines. Cell Oncol 26(1-2): 45–55

[61] Ong K (2007) Medical Informatics, an executive primer. HIMSS, Ohio

[62] Chen, H And Fuller, S And Friedman, C And Hersh W, editor (2005) Medical Informatics, Knowledge Management and Data Mining in Biomedicine. Springer

[63] Escolar F (2003) Inferencia de un sistema de información sanitario basado en la historia de salud electrónica. SEIS, Sociedad Española de Informática de la Salud

[64] Zamorano, J L And Gil-Loyzaga, P And Miravet D, editor (2004) Telemedicina, análisis de la situación actual y perspectivas de futuro. Vodafone

[65] WW S (2018) Clinical implications and challenges of artificial intelligence and deep learning. JAMA 320(11): 1107–1108

[66] Accenture. Data Analysis Overview and Insights

[67] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE Big Data: Astronomical or Genomical? PLOS Biology 13(7): e1002195

[68] Fox G (2015). Big Data Applications and Analytics MOOC

[69] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: The next frontier for innovation, competition, and productivity | McKinsey

[70] Clinical Decision Support: The Road Ahead

[71] Clinical Decision Support: The Road to Broad Adoption

[72] Johnston D, Pan E, Walker J (2004) The value of CPOE in ambulatory settings. J Healthc Inf Manag 18(1): 5–8

[73] Consejería de Salud Junta de Andalucía (2001) Guía de diseño y mejora continua de procesos asistenciales

[74] Alfonsel M (2006) Las TIC en la sanidad del futuro. Ariel, colección edition

[75] Escolar, F And Martínez-Berganza MT (2007) Definición de necesidades por los diferentes actores en los distintos entornos, pp. 165–186. SEIS, Sociedad Española de Informática de la Salud

[76] Escolar F, Martínez-Berganza MT (2004) Asistencia clínica en la cabecera del paciente, pp. 95–122. SEIS, Sociedad Española de Informática de la Salud

[77] Carnicero J (2008) La gestión del medicamento en los servicios de salud. SEIS, Sociedad Española de Informática de la Salud

[78] (2007). INDICADORES CLAVE DEL SISTEMA NACIONAL DE SALUD

[79] Dotor Gracia M, Fernádez García E (2009). Guía de diseño y mejora continua de procesos asistenciales integrados

[80] de Salud C (2001). Guía de diseño y mejora continua de procesos asistenciales integrados

[81] EH S, MJ S (2018) Clinical decision support in the era of artificial intelligence. JAMA

[82] C ID, S PF, MJ VT, E MD, S P (2003) Determinantes de test de O'Sullivan positivo en gestantes. In IX Conferencia Española de Biometría

[83] Bernardo JM (1981) Bioestadistica. Una perspectiva bayesiana. Vicens-Vives

[84] Hoff PD (2009) A First Course in Bayesian Statistical Methods. Springer

[85] Bailer-Jones CA (2017) Practical Bayesian Inference. Cambridge University Press

[86] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. Journal of Machine Learning Research 3: 1157–1182

[87] Hunt EB, Martin J, Stone PJ (1966) Experiments in Induction. Academic Press, New York

[88] Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179–188

[89] Kira K, Rendell LA (1992) A practical approach to feature selection. In D Sleeman, P Edwards, editors, Machine Learning: Proceedings of International Conference (ICML'92), pp. 249–256. Morgan Kaufmann

[90] Kononenko I (1994) Estimating attributes: Analysis and extensions of Relief. In L De Raedt, F Bergadano, editors, Machine Learning: ECML-94, pp. 171–182. Springer Verlag

[91] Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97: 273–324

[92] Duda RO, Hart PE, Stork DG (2001) Pattern Classification. Wiley-Interscience, New York, NY

[93] Bishop CM (2006) Pattern Recognition and Machine Learning. Springer

[94] Cover TM, Thomas JA (2006) Elements of information theory 2nd edition. Wiley-interscience

[95] Jollife IT (2002) Principal Component Analysis. Springer

[96] Robnik-Sikonja M, Kononenko I (2003) Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning 53: 23–69

[97] Lee J, Verleysen M (2007) Nonlinear Dimensionality Reduction. Information Science and Statistics. Springer New York

[98] Maaten Lvd, Hinton G (2008) Visualizing data using t-SNE. Journal of machine learning research 9(Nov): 2579–2605

[99] Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press

[100] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215(3): 403–410

[101] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25(17): 3389–3402

[102] Aho AV, Corasick MJ (1975) Efficient string matching: an aid to bibliographic search. Communications of the ACM 18(6): 333–340

[103] Rangayyan RM (2002) Biomedical Signal Analysis: A Case-study Approach (IEEE Press Series on Biomedical Engineering). Wiley-Blackwell

[104] Dora L, Agrawal S, Panda R, Abraham A (2017) State-of-the-Art Methods for Brain Tissue Segmentation: A Review. IEEE Reviews in Biomedical Engineering 10: 235–49

[105] Pianykh OS (2012) Digital Imaging and Communications in Medicine (DICOM). Springer-Verlag Berlin Heidelberg

[106] Beutel J, Sonka M, Fitzpatrick JM (2000) Handbook of medical imaging. SPIE

[107] Semmlow JL (2004) Biosignal and Biomedical Image Processing. CRC Press

[108] Vapnik VN (1998) Statistical Learning Theory. John Wiley & Sons

[109] Juan A (2009). Reconocimiento de Formas

[110] Turck M. The 2018 Big Data & Artificial Intelligence Landscape

[111] Kuo E. AI in Healthcare: Industry Landscape

[112] for Health Technology Transformation (IHTT) I. Transforming Health Care Through Big Data

[113] Raghupathi W, Raghupathi V Big data analytics in healthcare: promise and potential. Health information science and systems 2

[114] The future awakes. Life sciences and health care predictions 2022.

[115] Bishop CM (1995) Neural Networks for Pattern Recognition. Springer

[116] Ripley BD, Hjort NL (1995) Pattern Recognition and Neural Networks. Cambridge University Press, New York, NY, USA, 1st edition

[117] Hornik K, Stinchcombe M, White H (1989) Multilayer Feedforward Networks Are Universal Approximators. Neural Netw 2(5): 359–366

[118] Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65(6): 386

[119] Minsky ML, Papert S (1988) Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge Mass., expanded ed. edition

[120] Glorot X, Bordes A, Bengio Y (2011) Deep Sparse Rectifier Neural Networks. In GJ Gordon, DB Dunson, M DudÌk, editors, AISTATS, volume 15 of *JMLR Proceedings*, pp. 315–323. JMLR.org

[121] Bengio Y, Simard P, Frasconi P (1994) Learning Long-term Dependencies with Gradient Descent is Difficult. Trans Neur Netw 5(2): 157–166

[122] Clevert DA, Unterthiner T, Hochreiter S (2015) Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). CoRR abs/1511.07289

[123] LeCun Y, Bengio Y, Hinton G Deep learning. Nature 521: 436

[124] Ronneberger O, Fischer P, Brox T U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiV

[125] Hosmer DW, Lemeshow S (2000) Applied Logistic Regression. John Wiley and Sons, Inc.

[126] McCullagh P, Nelder JA (1989) Generalized linear models (Second edition). London: Chapman & Hall

[127] Lauritzen SL, Spiegelhalter DJ (1988) Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society Series B (Methodological) 50(2): 157 – 224

[128] Jensen FV, Nielsen T (2001) Bayesian Networks and Decision Graphs (Information Science and Statistics). Springer-Verlag New York Inc.

[129] Ortony A, Clore G (1988) A. collins (1988). The cognitive structure of emotions

[130] Marsella S, Gratch J, Petta P, A EE (2010) Computational Models of Emotion. Oxford University Press, Oxford

[131] Ball G, Breese J (2000). Emotion and personality in a conversational agent

[132] Lang PJ (1995) The emotion probe: studies of motivation and attention. American Psychologist 50(5): 372–385

[133] Kenny P, Parsons T, Gratch J, Rizzo A (2008) Virtual humans for assisted health care. Proceedings of the 1st ACM international conference on PErvasive Technologies Related to Assistive Environments - PETRA '08 p. 1

[134] Larrañaga P (2009). Redes bayesianas, fundamentos

[135] Millán E (2000) Redes bayesianas, chapter Redes baye. UMA, Malaga

[136] Cover MT, Hart PE (1967) Nearest Neighbour Pattern Classification. IEEE Transactions on Information Theory 13: 21–27

[137] Holmes CC, Adams NM (2002) A probabilistic nearest neighbour method for statistica pattern recognition. Journal of the Royal Statistical Society Series B-Statistical Methodology 64(2): 295–306

[138] Manocha S, Girolami MA (2007) An empirical analysis of the probabilistic K-nearest neighbor classifier. Pattern Recognition Letters 28: 1818–1824

[139] Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA:

[140] Mangasarian OL, Wolberg WH (1990) Cancer diagnosis via linear programming. SIAM News 23(5): 1–18

[141] Kohavi R, et al. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, volume 14(2), pp. 1137–1145. Montreal, Canada

[142] Berrar D, Bradbury I, Dubitzky W (2006) Avoiding model selection bias in small-sample genomic datasets. Bioinformatics 22(10): 1245–1250

[143] Martin JK, Hirschberg DS (1996) Small Sample Statistics for Classification Error Rates I: Error Rate Measurements. Technical Report ICS-TR-96-22 `citeseer.ist.psu.edu/article/martin96small.html`

[144] Martin JK, Hirschberg DS (1996) Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests. Technical Report ICS-TR-96-22 `citeseer.ist.psu.edu/article/martin96small.html`

[145] Elter M, Schulz-Wendtland R, Wittenberg T (2007) The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical Physics 34(11): 4164–4172

[146] MacKay DJ (1992) Bayesian model comparison and backprop nets. In Advances in neural information processing systems, pp. 839–846

[147] Kim JH (2009) Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis 53(11): 3735–3745

[148] Paloplarrea V (2006) Documento de consenso sobre la utilización de antibióticos en atención primaria. Atención Primaria 38(3): 137–138

[149] Gérvas J (2000) La resistencia a los antibióticos , un problema de salud pública. Atencion Primaria 25

[150] Bretón J (2004) Vigilancia de la resistenica bacteriana en pediatría y su relación con el uso de antibióticos por medio del análisis de series temporales. Ph.D. thesis, Universidad de València

[151] Hernández-Orallo J, Ramírez-Quintana MJ, Ferri C (2004) Introducción a la Minería de Datos. Prentice Hall

[152] Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Mining Data Streams : A Review. ACM SIGMOD Record 34(2): 18–26

[153] Noy NF, Sintek M, Decker S, Crubezy M, Fergerson RW, Musen MA (2001) Creating Semantic Web contents with Protege-2000. IEEE Intelligent Systems 16(2): 60–71

[154] Winston (1992) Artificial Intelligence (A-W Series in Computerscience). Addison Wesley

[155] Laboratorio Clínico D (2004). Manual de toma de muestras para el estudio bacteriológico, parasitológico y micológico, selección, recolección, conservación y transporte

[156] Shearer C (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing 5(4): 13–21

[157] Davis FD (1989) Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13(3): 319

[158] Acosta D, Patkar V, Keshtgar M, Fox J (2009) Challenges in delivering decision support systems: the MATE experience. In International Workshop on Knowledge Representation for Health Care, pp. 124–140. Springer

[159] Groot P, Hommersom A, Lucas PJ, Merk RJ, ten Teije A, van Harmelen F, Serban R (2009) Using model checking for critiquing based on clinical guidelines. Artificial Intelligence in Medicine 46(1): 19–36

[160] Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann H (2006) Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians task force. Chest 129(1): 174–181

[161] Cruz-Correia RJ, Pereira Rodrigues P, Freitas A, Canario Almeida F, Chen R, Costa-Pereira A (2010) Data Quality and Integration Issues in Electronic Health Records. In Information Discovery On Electronic Health Records, pp. 55–96. V. Hristidis (ed.)

[162] Madnick SE, Wang RY (1992). Introduction to total data quality management (TDQM) research program

[163] Wang RY (1998) A Product Perspective on Total Data Quality Management. Commun ACM 41(2): 58–65

[164] Röthlin M (2010) Management of data quality in enterprise resource planning systems. Number Bd. 68 in Reihe: Wirtschaftsinformatik. Eul, 1. aufl edition

[165] Sebastian-Coleman L (2013) Measuring data quality for ongoing improvement: a data quality assessment framework. Morgan Kaufmann

[166] Karr AF, Sanil AP, Banks DL (2006) Data quality: A statistical perspective. Statistical Methodology 3(2): 137 – 173

[167] Madnick SE, Wang RY, Lee YW, Zhu H (2009) Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality 1(1): 1–22

[168] Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 12(4): 5–33

[169] Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys 41(3): 1–52

[170] Weiskopf NG, Hripcsak G, Swaminathan S, Weng C (2013) Defining and measuring completeness of electronic health records for secondary use. Journal of Biomedical Informatics 46(5): 830 – 836

[171] Liaw S, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo A, Talaei-Khoei A (2013) Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics 82(1): 10–24

[172] Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM (2012) Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform 180: 721–725

[173] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, et al. (2016) A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems 4(1)

[174] Sáez C, Rodrigues PP, Gama J, Robles M, García-Gómez JM (2015) Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery 29(4): 950–975

[175] Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM (2016) Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. Journal of the American Medical Informatics Association 23(6): 1085–1095

[176] Sáez C, Robles M, García-Gómez JM (2017) Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical methods in medical research 26(1): 312–336

[177] Kawamoto K, Houlihan Ca, Balas EA, Lobach DF (2005) Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ (Clinical research ed) 330(7494): 765

[178] García-Gómez JM, Esparza M, Robles M (2009) Herramientas de Bioinformática Clínica para la ayuda a la decisión médica asistida por Computador. Informática y Salud

[179] SáEz C, Bresó A, Vicente J, Robles M, GarcíA-GóMez JM (2013) An HL7-CDA wrapper for facilitating semantic interoperability to rule-based Clinical Decision Support Systems. Computer methods and programs in biomedicine 109(3): 239–249

[180] Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE (2007) A roadmap for national action on clinical decision support. Journal of the American medical informatics association 14(2): 141–145

[181] Sintchenko V, Coiera E, Iredell JR, Gilbert GL (2004) Comparative Impact of Guidelines, Clinical Data, and Decision Support on Prescribing Decisions: An Interactive Web Experiment with Simulated Cases. Journal of the American Medical Informatics Association 11(1): 71–77

[182] Gravel K, Légaré F, Graham ID (2006) Barriers and facilitators to implementing shared decision-making in clinical practice: a systematic review of health professionals' perceptions. Implementation Science 1(1): 16

[183] Légaré F, Ratté S, Gravel K, Graham ID (2008) Barriers and facilitators to implementing shared decision-making in clinical practice: update of a systematic review of health professionals' perceptions. Patient education and counseling 73(3): 526–535

[184] Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB (2005) Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 293(10): 1223–1238

[185] Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D (2009) Use of electronic health records in US hospitals. New England Journal of Medicine 360(16): 1628–1638

[186] Pearson SA, Moxey A, Robertson J, Hains I, Williamson M, Reeve J, Newby D (2009) Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007). BMC health services research 9(1): 154

[187] Demner-Fushman D, Chapman WW, McDonald CJ (2009) What can natural language processing do for clinical decision support? Journal of biomedical informatics 42(5): 760–772

[188] Friedman CP (1999) Information technology leadership in academic medical centers: a tale of four cultures. Academic medicine: journal of the Association of American Medical Colleges 74(7): 795–799

[189] Comission E (2011). "Quantifying the impact of PreCommercial Procurement (PCP) in Europe based on evidence from the ICT sector"

[190] Osheroff JA, Pifer EA, Teich JM, Sittig DF, Jenders RA (2005) Improving Outcomes with Clinical Decision Support: An Implementer's Guide. HIMSS, Ohio

[191] Bermejo J, editor (2008) Seguridad de la información en entornos sanitarios. SEIS, Sociedad Española de Informática de la Salud

[192] Food U, Administration D. Food and Drug Administration. FDA permits marketing of artificial intelligence–based device to detect certain diabetes-related eye problems.

[193] Miller RA (1996) Evaluating evaluations of medical diagnostic systems. Journal of the American Medical Informatics Association 3(6): 429

[194] Berner ES (2003) Diagnostic Decision Support Systems: How to Determine the Gold Standard? Journal of the American Medical Informatics Association 10(6): 608–610

[195] Liu JL, Wyatt JC (2011) The case for randomized controlled trials to assess the impact of clinical information systems. Journal of the American Medical Informatics Association 18(2): 173–180

[196] Ramnarayan P, Kapoor RR, Coren M, Nanduri V, Tomlinson AL, Taylor PM, Wyatt JC, Britto JF (2003) Measuring the Impact of Diagnostic Decision Support on the Quality of Clinical Decision Making: Development of a Reliable and Valid Composite Score. Journal of the American Medical Informatics Association 10(6): 563–572

[197] Holbrook A, Thabane L, Keshavjee K, Dolovich L, Bernstein B, Chan D, Troyan S, Foster G, Gerstein H, Investigators CI (2009) Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial. Canadian Medical Association Journal 181(1-2): 37–44

[198] Chuang JH, Hripcsak G, Heitjan DF (2002) Design and analysis of controlled trials in naturally clustered environments: implications for medical informatics. Journal of the American Medical Informatics Association 9(3): 230–238

[199] Venkatesh V (2000) Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Information systems research 11(4): 342–365

[200] Montgomery AA, Fahey T, Peters TJ, MacIntosh C, Sharp DJ (2000) Evaluation of computer based clinical decision support system and risk chart for management of hypertension in primary care: randomised controlled trial. BMJ 320: 686–690

[201] Hunt DL, Haynes RB, Hanna SE, Smith K (1998) Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review. JAMA 280: 1339–1346

[202] Hunt DL (2009) Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review. JAMA: The Journal of the American Medical Association 280(15): 1339–1346

[203] Pearson SA, Moxey A, Robertson J, Hains I, Williamson M, Reeve J, Newby D (2009) Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007). BMC health services research 9: 154

[204] Friedman CP, Wyatt JC, Owens DK (2006) Evaluation and technology assessment. In Biomedical Informatics, pp. 403–443. Springer

[205] Paridaens RJ, Gelber S, Cole BF, Gelber RD, Thürlimann B, Price KN, Holmberg SB, Crivellari D, Coates AS, Goldhirsch A (2010) Adjuvant!© Online estimation of chemotherapy effectiveness when added to ovarian function suppression plus tamoxifen for premenopausal women with estrogen-receptor-positive breast cancer. Breast cancer research and treatment 123(1): 303–310

[206] Engelhardt EG, van den Broek AJ, Linn SC, Wishart GC, Rutgers EJT, van de Velde AO, Smit VT, Voogd AC, Siesling S, Brinkhuis M, Seynaeve C, Westenend PJ, Stiggelbout AM, Tollenaar RA, van Leeuwen FE, van 't Veer LJ, Ravdin PM, Pharaoh PD, Schmidt MK (2017) Accuracy of the online prognostication tools PREDICT and Adjuvant! for early-stage breast cancer patients younger than 50 years. European Journal of Cancer 78: 37–44

[207] Sáez C, Martí-Bonmatí L, Alberich-Bayarri Á, Robles M, García-Gómez JM (2014) Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV 1H MRS: Evaluation as an additional information procedure for novice radiologists. Computers in biology and medicine 45: 26–33

[208] Juan-Albarracín J, Fuster-Garcia E, Pérez-Girbés A, Aparici-Robles F, Alberich-Bayarri n, Revert-Ventura A, Martí-Bonmatí L, García-Gómez JM (2018) Glioblastoma: Vascular Habitats Detected at Preoperative Dynamic Susceptibility-weighted Contrast-enhanced Perfusion MR Imaging Predict Survival. Radiology 287(3): 944–954. PMID: 29357274

[209] Kawamoto K (2011) OpenCDS: An open-source, standards-based, service-oriented framework for scalable CDS. In SOA in Healthcare 2011 Conference

[210] Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, McClay J, Parker C, Hrabak KM, Berg D, Weida T, et al. (2007) The SAGE Guideline Model: achievements and overview. Journal of the American Medical Informatics Association 14(5): 589–598

[211] de Chile CM (2000). EL SISTEMA DE SALUD CHILENO

[212] Bishop CM (2006) Pattern Recognition and Machine Learning. Springer

[213] Beal MJ (2003) Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London

[214] Minka T (2001) Expectation Propagation for approximate Bayesian inference. In Proceedings 17th Conference on Uncertainty in Artificial Intelligence, pp. 362–369. Morgan Kauffman

[215] Minka T (2001) A family of approximate algorithms for Bayesian inference. Ph.D. thesis, MIT

[216] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machine. J Chem Phys 21: 1087–1091

[217] Neal RM (1996) Bayesian Learning for Neural Networks. Springer. Lecture Notes in Statistics, 118

[218] Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian Data Analysis. Chapman & Hall, London

Apéndice A

Foros de CDSS

La tabla A.1 enumera los foros internacionales donde se promociona la investigación, desarrollo e implantación de CDSS:

La tabla A.3 enumera las revistas y congresos científicos con interés en las técnicas, desarrollo, integración y evaluación de CDSS.

Tabla A.1: Foros de CDSS, revisado en Febrero de 2010

Organismo	Grupo de Trabajo	Actividad	Referencia
AMIA: American Medical Informatics Association	Formal (Bio)Medical Knowledge Representation	Promover la representación formal del conocimiento médico	amia.org
AMIA: American Medical Informatics Association	Roadmap for National Action on Clinical Decision Support	Promover la implantación de CDSS (nivel 3) en ámbitos sanitarios	amia.org
CLINFOWIKI	CDS	Wiki sobre informática médica y CDS	clinfowiki.org
CONBIOMED	Grupo de trabajo en Bioinformática traslacional	Estudiar la posibilidad de usar datos extraídos de las historia clínicas para investigación	combiomed.isciii.es
COSSAC: IRC in Cognitive Science & Systems Engineering	-	IRC (Innovation Research Centre) formado por Oxford, UCL and UEDIN para la investigación en sistemas cognitivos en medicina	cossac.org
EFMI: European Federation of Medical Informatics	Working Group on Medical Image Processing	Integración y evaluación de CDSS basados en el procesamiento de imagen médica en la práctica clínica	efmi-wg-mip.net
EHTEL: European Health Telematics Association	Task Force Patient Safety and eMedication	Mejora de prescripción mediante ICT	ehtel.org
HL7 CDS Work Group	Decision Support Service (DSS) standards adopted by the Health Level 7 (HL7)	Estándares para DSS basados en HL7 (vMR)	www.hl7.org
HIMMS	Clinical decision support guidebook series	Grías sobre la implantación de ayuda a la decisión y su evaluación en instituciones sanitarias	himss.org
IMIA: International Medical Informatics Association	Biomedical Pattern Recognition WG	Promover el aprendizaje automático e interpretación de datos biomédicos en medicina y biología	imia.org
IMIA: International Medical Informatics Association	Intelligent Data Analysis and Data Mining WG	Dar a conocer y aceptar los métodos de minería de datos en medicina	imia.org

Tabla A.2: Foros de CDSS, revisado en Febrero de 2010 (2)

Organismo	Grupo de Trabajo	Actividad	Referencia
Open Clinical: Knowledge management for medical care	Decision Support Systems	Promover el uso de tecnologías para la gestión del conocimiento en salud	openclinical.org
SEIS: Sociedad española de informática de la salud	I+S	Publicación generalista de Informática Médica	seis.es
Standards Australia's	HB 307-2007	Guía de principios y características de los CDSS	e-health.standards.org.au
TICBiomed: cooperación en TIC salud	Innovación Alineada en Sanidad	Incorporación de innovación tecnológica entre las organizaciones sanitarias	ticbiomed.net
TICSalut:	-	Promoción de las tecnologías para la salud	ticsalut.cat

Tabla A.3: Revistas y congresos en CDSS, revisado en Febrero de 2010.

Revista	Temas
AMCIS	Ayuda a la decisión
Artificial Intelligence in Medicine	Inteligencia artificial en medicina
BioMed Central cancer (BMC cancer)	Biomedicina
BMJ: British Medical Journal	Decisión Médica
Breast cancer research	Biomedicina
Breast Cancer Research And Treatment	Biomedicina
Decision Support Systems	Ayuda a la decisión
Engineering Applications of Artificial Intelligence	Sistemas inteligentes y Tecnología Semántica
Evia (http://www.idi.aetic.es/evia)	Influencia en políticas de TIC Salud
IEEE Engineering in Medicine and Biology	Informática Biomédica
IEEE Intelligent Systems	Sistemas inteligentes y Tecnología Semántica
IEEE Transactions on Information Technology in Biomedicine	Informática Biomédica
INFORMED (Informática Médica)	Informática Biomédica
International Journal of Medical Informatics	Informática Biomédica
International Journal Of Technology Assessment (in Health Care)	Informática Biomédica
Journal of Biomedical Informatics	Informática Biomédica
Journal of Clinical Oncology	Oncología
Journal of Evaluation in Clinical Practice	Informática Biomédica
Journal of Information Technology	TIC
Journal of the American Medical Informatics Association	Informática Médica
Journal of Medical Internet Research	Informática Biomédica
Journal of Web Semantics	Sistemas inteligentes y Tecnología Semántica
Medical Decision Making	Ayuda a la decisión
Medinfo	Informática Biomédica
Medical Informatics and the Internet in Medicine	Informática Biomédica
Medical Informatics Europe	Informática Biomédica
MIE (Medical Informatics Europe)	Informática Biomédica
Methods of Information in Medicine	Informática Biomédica
New England Journal of Medicine	Evaluación de sistemas informáticos para la salud

Apéndice B

Listado de CDSS

B.1. Listado alfabético de CDSS con características

La tabla B.1 y sucesivas incluyen el listado alfabético de CDSS encontrados en Abril de 2011. Se ha analizado el tipo de CDSS según las características: nivel de complejidad, forma de interacción con el usuario o servicio, tecnología de inteligencia artificial utilizada, estado del desarrollo y referencias principales.

Tabla B.1: Listado de CDSS, revisado en Abril de 2011. La columna Nivel indica el tipo de CDSS según los niveles especificados en la sección 1.2: Nivel 1: acceso a conocimiento relacionado con el caso; Nivel 2: alertas basadas en reglas básicas compuestas por pocas variables; Nivel 3: Sistemas deductivos sobre reglas predefinidas y la situación del paciente; Nivel 4: Modelos predictivos aprendidos mediante conocimiento y casos del problema médico. La columna Interacción indica el tipo de interacción del CDSS según los tipos especificados en 1.2: A: CDSS Autónomo; W: Servicio Web; D: Interconectado con dispositivos específicos; I: Interoperable con EHR; S: Smart phone. La columna Tecnología indica la tecnología de inteligencia artificial utilizada en el CDSS: BI: Business Intelligent; RS: razonador semántico; AA: Aprendizaje Automático; MB: Meta-buscador. La columna Estado expresa el estado del desarrollo del CDSS: P: Producto comercial; C: Prototipo científico; I: Proyecto de investigación.

CDSS	Nivel	Interacción	Funcionalidades	Tecnología	Casos de Uso	Estado	Referencia
Accelerator Decision Manager	N3	D-I	Alertas, control de calidad	BI-RS	Laboratorios automatizados	P	abbott.com
Adjuvant! Online	N4	W	Pronóstico	AA	Cancer de mama	C	adjuvantonline .com [205, 206]
Alchemy	N2	A	Alertas	-	Prescripción, CPOE	I	alchemyrx.com
Anvita Insight analysis	N3	I	Gestión	BI	Gestión integral	P	anvitahealth.com
AREZZO	N3	I	Planificación (guías clínicas)	RS	Fallo cardíaco, HIV, farma, diabetes, etc	P	infermed.com
BestPractice (BMJ)	N3	W	Recomendador	MB	Segunda Opinión	P	bestpractice.bmj.com
Care Engine System	N3	I	Gestión, Alertas	BI	Clínica, laboratorio, farmacia	P	activehealth management.com
CHAMP	N2	A	Triaje	-	Primaria	I	ohsu.edu/champ
CINAHL	N1	A	Tratamiento, interpretación	-	Enfermería	P	ebscohost.com
Clinical Xpert Solution Suite	N1	A	Interpretación,	-	PoC	P	thomson reuters.com

Tabla B.2: Listado de CDSS, actualizado en Diciembre de 2018 (2). Ver leyenda la tabla B.1.

CDSS	Nivel	Interacción	Funcionalidades	Tecnología	Casos de Uso	Estado	Referencia
CPM Resource Center	N1	A	Calidad asistencial	-	Comunidades sanitarias	P	www.cpmrc.com
CURIAM	N4	I	Diagnóstico, pronóstico, tratamiento, riesgos, intepretación	AA	Tumores cerebrales, tumores de partes blandas, Depresión postparto	C	[207]
Doctor's toolbag	N3	S	Recomendador	MB	Triaje, planificación diagnóstica	P	group.bmj.com[a]
DynaMed	N1	A	Tratamiento, interpretación	-	PoC	P	ebscohost.com
e-lactancia	N1	W	Tratamiento	MB (bbdd)	Lactancia	I (web)	e-lactancia.org
Elsevier Clinical Decision Support	N1	W	Interpretación	-	Diagnóstico, farmacia	P	clinicaldecision support.com
emsor.es							
FirstConsult	N1	W	Interpretación	-	Estudiantes, residentes, medicos de familia y especialista	P	mdconsult.com
Flatiron Oncoanalytics	N2-3	I	Cuadros de mando	BI	Oncología	C	flatiron.com
help4mood	N[3..4]	A/S	Tratamiento	RS/AA	depresión	I	ibime.upv.es

[a] group.bmj.com/products/mobile-apps/doctors-toolbag-iphone-app

Tabla B.3: Listado de CDSS, actualizado en Diciembre de 2018 (3). Ver leyenda la tabla B.1.

CDSS	Nivel	Interacción	Funcionalidades	Tecnología	Casos de Uso	Estado	Referencia
i2b2	N[1..4]	W	Genérico	-	Farmacia, ensayos clínicos	P	i2b2.org
ict4depression	N[3..4]	A/S	Tratamiento	RS/AA	Stress, depresión	I	ict4depression.eu
IDx-DR	N4	D	Diagnóstico, triaje	AA	Oftalmología, retinopatía diabética	P	www.eyediagnosis.net
Infinicyt	N4	A	Tratamiento	AA	Citrometria, Leucemia	P	infinicyt.com
Interqual Decision Support	N3	I	Gestión	BI	Gestión por sectores	P	mckesson.com
Isabel	N4	I	Diagnóstico	AA-RS	CPOE, pediatría, emergencias	P	isabelhealthcare.com
Lexi-Comp	N1	W	Interpretación	-	Educación en PoC	P	lexi.com
Medicalis Decision Support	N3	I	Diagnóstico, Planificación (procedimientos diagnósticos)	-	Radiología	P	medicalis.com
Medicalis Analytics	N3	I	Gestión	BI	Radiología	P	medicalis.com
MedLEE							
Medworxx Clinical Utilization Management	N2-N3	I	Alertas, monitorización, control de calidad	BI-RS	Gestión hospitalaria integral	P	medworxx.com

Tabla B.4: Listado de CDSS, actualizado en Diciembre de 2018 (4). Ver leyenda la tabla B.1.

MICROMEDEX CLINICAL EVIDENCE SOLUTIONS	N1	W	Interpretación	-	Diagnóstico, farmacia, educación del paciente	P	thomson reuters.com
Odyssey	N3	A	Triaje	RS	Primaria, urgencias, personal sanitario	P	plain.co.uk
Oncohabitats	N4	W	Interpretación, pronóstico	AA	Glioblastoma	C	www.oncohabitats.upv.es [208]
Open CDS	N3	I	Alertas, planificación, genérico, open source	RS	NQF, diabetes	C	[209], HL7 CDS Work Group, opencds.org
Open EHR	N[1..4]	I	Guidlines, workflows	RS	Brasil (gubernamental)	P	openehr.org
CDSS	Nivel	Interacción	Funcionalidades	Tecnología	Casos de Uso	Estado	Referencia
PathIQ Inmunoquery	N1	A	Diagnóstico, interpretación	-	Patología	P	amirsys.com
Pediatric Knowledgebase (PKB)	-	W	Tratamiento	-	Farmacia, Pediatría	C	pkb.chop.edu
PeriOptimum	N3	I	Gestión	BI	Gestión integral	P	periop timum.com
Pinpoint Review	N4	W	Riesgo	AA	Riesgo de readmisión y otros	P	medai.com

Tabla B.5: Listado de CDSS, actualizado en Diciembre de 2018 (5). Ver leyenda la tabla B.1.

SAGE	N3	I	Planificación (guías clínicas)	RS	Neumonía	C	[210]
Sentri7	N3	I	Alertas	RS	Calidad Asistencial, Prioridades según riesgos	P	sentri7.com
Sistema Integrado de Apoio à Decisão	N3	A	Gestión	BI	Gestión integral	Daño cerebral	bi4all.pt
CDSS	Nivel	Interacción	Funcionalidades	Tecnología	Casos de Uso	Estado	Referencia
Soarian	N3	I	Planificación (procesos asistenciales)	RS	Primaria, Cardiología	P	siemens.com
StatDX	N1	A	Diagnóstico, interpretación	-	Radiología	P	amirsys.com
SuperDecisions	N4	A		superdecisions.com			
TheraDoc Expert System	N2-3	I	Alertas	RS	Infecciones, Antibióticos, prescripción	P	theradoc.com
Well Logic Consult	N2-3	I	Alertas, planificación (guías clínicas)	RS	CPOE, gestión hospitalaria	P	wellogic.com
Zebra	N4	I	Interpretación	AA	Radiología	P	www.zebra-med.com
Zynx solutions	N2-3	I	Alertas, Planificación (guías clínicas)	RS	CPOE, gestión hospitalaria	P	zynxhealth.com

Apéndice C

Los actores en salud

La tendencia actual de la medicina es situar a un paciente activo en el centro de los servicios sanitarios a través de grupos de profesionales multidisciplinares y recursos especializados.

Un equipo de atención primaria puede estar formado como mínimo por médicos de familia, pediatras, enfermeros y administrativos. Generalmente, un médico de familia estará encargado de atender entre 1.500 y 2.000 pacientes, realizando las tareas de diagnóstico, tratamiento, rehabilitación y prevención mediante consultas en un centro de salud o a domicilio, bajo una planificación programada o bajo demanda. Por su parte, la asignación en pediatría se ajusta a 1.000 niños. El servicio de enfermería se planifica para atender a 2.000 pacientes por profesional y centra su actividad en los cuidados, inyectables, extracciones y monitorizaciones a pacientes, el adiestramiento a los pacientes para autocuidados, consultas específicas programadas y atención a domicilio. Generalmente, un centro de salud dispondrá de un celador para la logística asociada a la información clínica y también necesitará de un conjunto de administrativos para gestionar las citas previas, la asignación de médicos y la tramitación de analíticas, recetas y adquisición de señales e imágenes biomédicas. Los problemas de salud más frecuentemente tratados en primaria son la hipertensión arterial, dislipemia, diabetes mellitus, EPOC, asma, artrosis, infecciones (respiratorias, urinarias, digestivas, piel o mucosas, ojos u oídos), traumatismos no laborales, depresiones, ansiedad, obesidad, insomnio, estreñimiento, y pacientes terminales.

La atención secundaria está formada principalmente por las especialidades médicas reguladas a través del concurso-oposición MIR (Médico Interno Residente). Las especialidades médicas reconocidas por los países de la Unión Europea son: alergología, anestesiología y Reanimación, aparato Digestivo, cardiología, endocrinología y nutrición, geriatría, hematología y hemoterapia, infectología, hidrología médica, medicina del deporte, medicina del trabajo (general y profesional), medicina familiar y comunitaria, medicina intensiva, medicina interna, medicina legal y forense, medicina preventiva y salud pública (comunitaria y hospitalaria), nefrología, neumología, neurología, oncología médica, oncología radioterápica, pediatría, psiquiatría, rehabilitación y reumatología.

Las especialidades quirúrgicas (cirujanos) se dividen por sistemas: cirugía cardiovascular, cirugía general y del aparato digestivo, cirugía oral y maxilofacial, cirugía pediátrica, cirugía plástica, estética y reparadora, cirugía torácica y neurocirugía.

Las especialidades médico-quirúrgicas incluyen un tratamiento integral al paciente, esto es, prevención, diagnóstico y tratamiento y se dividen en: angiología cirugía vascular, dermatología médico-quirúrgica y venereología, estomatología, obstetricia y ginecología, odontología, oftalmología, otorrinolaringología, ortopedia y traumatología y urología.

Algunas especialidades de laboratorio pueden ser cubiertas por profesionales diferentes a la medicina que también suelen tener su actividad regulada a través de exámenes, como por ejemplo RFIR (Radiofísica hospitalaria), EIR (Enfermero Interno Residente, para especialidades como

matrona y salud mental), BIR (Biólogo Interno Residente), el QIR (Químico Interno Residente), el FIR (Farmacéutico Interno Residente) y el PIR (Psicólogo Interno Residente). Las especialidades de laboratorio dan apoyo a las demás, realizan diagnósticos y sugieren tratamientos a los clínicos, siendo su relación con el paciente indirecta. Las principales especialidades son: análisis clínicos, anatomía patológica, bioquímica clínica, farmacología clínica, inmunología, medicina nuclear, microbiología y parasitología, neurofisiología clínica, y radiodiagnóstico.

C.1. Ejemplos de organizaciones sanitarias

Revisaremos algunos ejemplos de sistemas sanitarios para ilustrar diversas formas de organización de la atención sanitaria.

C.1.1. Sistema Nacional de Salud

El Sistema Nacional de Salud (SNS) aglutina las prestaciones sanitarias públicas de España desde 1986, en que fue creado mediante la Ley 14/1986, de 25 de abril, General de Sanidad (LGS). Dichas prestaciones son asumidas mayormente por las distintas Comunidades Autónomas (CC.AA.) o por el Instituto Nacional de Gestión Sanitaria, si bien la Sanidad Exterior, las Bases y coordinación general de la sanidad y la Legislación sobre productos farmacéuticos son competencias exclusivas del Estado. La actividad de la prestación sanitaria se armoniza mediante el Consejo Interterritorial del SNS (CISNS).

Las competencias del Estado en el ámbito sanitario son gestionadas por el Ministerio de Sanidad[a]. De la Secretaría General de Sanidad dependen y se organizan la Dirección General (D.G.) de Salud Pública y Sanidad Exterior, la D.G. Ordenación Profesional, Cohesión del SNS y Alta Inspección (que incluye el CISNS), la D.G. de Farmacia y productos Sanitarios, la Agencia de Calidad del SNS, la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), la Organización Nacional de Transplantes, el Instituto Nacional de Gestión Sanitaria y la Comisión Nacional de Reproducción Humana Asistida. Buena parte de esta organización viene derivada de la Ley 16/2003 de cohesión y calidad del Sistema Nacional de Salud, donde una vez derivadas las competencias a las CC.AA. se establecieron acciones de coordinación (a través del CISNS y se promovieron órganos especializados como la AEMPS y la Agencia de Calidad.

Cada CC.AA. tiene una figura administrativa y de gestión que integra los centros, servicios y establecimientos que componen sus servicio de salud y la prestación sanitaria de su población. Cada servicio de salud se constituye por un real decreto, por ejemplo el Servicio Andaluz de Salud (SAS) atiende a más de ocho millones de habitantes de Andalucía.

C.1.2. Agencia Valenciana de Salud

La Ley 3/2003 de Ordenación Sanitaria de la Comunidad Valenciana crea la Agencia Valenciana de Salud (AVS) para la organización de los servicios sanitarios públicos en la Comunidad Valenciana. La estructura del sistema sanitario valenciano se estructura en tres niveles:

1. La Conselleria de Sanidad es el máximo órgano encargado de la dirección y ejecución de la política del Consell de la Generalitat en materia de sanidad.

2. La AVS realiza la gestión y administración del sistema valenciano de salud y de la prestación sanitaria en la Comunidad Valenciana, incluyendo:

[a]En 2011, Miniterio de Sanidad, Política Social e Igualdad.

- Establecer y aplicar los criterios generales de coordinación de todos los recursos y medios sanitarios públicos de la Comunidad Valenciana.
- Establecer acuerdos, convenios y conciertos para la prestación de los servicios, en orden a la adecuada utilización de los recursos sanitarios públicos, y el establecimiento de fórmulas de gestión integrada o compartida con entidades públicas y privadas.
- Analizar y proponer a la Conselleria de Sanidad la constitución de fórmulas organizativas para la provisión y gestión de los servicios sanitarios.
- Autorizar y desarrollar en sus centros la creación de nuevas unidades/ fórmulas organizativas como son las áreas clínicas y las unidades de gestión clínica.

La AVS cuenta con una estructura central y una estructura periférica, basada en departamentos de salud. Los órganos centrales son:

- El Consejo de Administración, presidido por el Conseller de Sanidad y con representación de la administración y de los principales agentes sociales.
- El Consejo de Salud de la Comunidad Valenciana, como órgano superior colegiado de carácter consultivo.
- El Director Gerente, que ejerce la dirección y gestión de la misma, pudiendo a su vez ejercer la secretaría autonómica de Sanidad. De la dirección gerente de la Agencia dependen los cuatro directores de Asistencia Sanitaria de Zona, el Área de Coordinación y Planificación y el Área de Informática, Telecomunicaciones y Organización. Los Sistemas de Información Sanitaria están gestionados por esta última Área.
- Seis direcciones generales: Recursos Económicos, Recursos Humanos, Asistencia Sanitaria, Drogodependencias, Farmacia y Productos Sanitarios y Calidad y Atención al Paciente.

3. Los 22 Departamentos de Salud, equivalentes a las áreas de salud previstas en la LGS, que desarrollan las acciones de promoción, protección, prevención, curación y rehabilitación del estado de salud de sus poblaciones. En cada departamento de salud se garantizará una adecuada ordenación de la asistencia primaria y su coordinación con la atención especializada. La atención primaria es el primer nivel de acceso y se presta principalmente en los centros de atención primaria. La atención especializada es el segundo nivel de acceso a la sanidad pública e incluye la atención hospitalaria, la atención sociosanitaria, la atención psiquiátrica y salud mental, la atención a drogodependencias y otros recursos de atención especializada, por ejemplo los centros de especialidades y de rehabilitación, el tratamiento de la insuficiencia renal crónica (diálisis), los centros de prevención y control de enfermedades de transmisión sexual y los centros de vacunaciones y consejos a viajeros internacionales.

C.1.3. Departamento de Salud Valencia La Fe

El Departamento de Salud Valencia La Fe es el Departamento de referencia de la Comunidad Valenciana, por lo que atiende las necesidades sanitarias de su área de salud y está abierto, por su rol de referencia, a las demandas de otros departamentos y comunidades. Presta atención sanitaria universal, integral y personalizada en régimen comunitario, de urgencia, de ambulatorio, de hospitalización y domiciliario.

El Departamento de Salud Valencia La Fe engloba el Hospital Universitari La Fe, el Centro de Especialidades Ricardo Trénor, 7 centros de salud y 8 consultorios complementarios repartidos en 6 municipios, atendiendo a una población de más de 250.000 habitantes.

Un Centro de Especialidades es un centro asistencial de actividad programada de consultas externas que atiende la patología remitida por los centros de salud de los departamentos. Estos centros prestan atención personalizada en régimen ambulatorio en especialidades tales como medicina digestiva, urología, obstetricia, endocrinología, neurología, traumatología, rehabilitación, dermatología, odontología, oftalmología, ginecología, reumatología, cardiología, neumología, otorrinolaringología, cirugía, cirugía de mama, cirugía vascular y alergología. También en dichos centros desarrollan su actividad unidades específicas como la unidad de tabaquismo, la de pie diabético, la de coloproctología, etcétera.

C.1.4. Hospital Universitari La Fe

El Hospital Universitari La Fe se organiza en cuatro áreas y una macro-área, que conforman la estructura de dirección del hospital junto con la dirección gerencial: el Área Quirúrgica y de Críticos, el Área de Diagnóstico y Farmacoterapia, el Área Ambulatoria y de Alternativas a la Hospitalización, el Área de Hospitalización y las Áreas de Gestión Clínica, en las que se incluyen el Área de aparato locomotor, rehabilitacíon y neurociencias, el Área de Patología Médico-Quirúrgica del Adulto, el Área de Patología Infantil y Mujer y el Área Médica de Atención Integral.

Las áreas clínicas se estructuran a su vez en otras áreas. Por ejemplo, el Área de Patología Médico-Quirúrgica del Adulto contiene las áreas del tórax, del riñón, de tumores, del aparato digestivo u otras unidades quirúrgicas del área de cirugía y servicios de apoyo. Estas áreas clínicas agrupan los servicios y unidades por procesos clínicos comunes ofertados por el hospital. Por ejemplo, el área de tumores agrupa Oncología Radioterápica, Oncología, y Hematología/Hemoterapia).

Cada servicio tiene personal facultativo, residentes, personal investigador y personal de planta dirigido por un jefe de servicio. Algunos servicios tendrán una plantilla heterogénea de facultativos, además de relaciones con otros servicios. Así, el servicio de cirugía cardiovascular incluye cirujanos y cardiólogos y se relaciona con el servicio de anestesiología y reanimación.

Las unidades sirven para coordinar a los distintos especialistas en el tratamiento de ciertas patologías. Por ejemplo, la unidad de mama unifica toda la patología de mama. Además, los cirujanos que componen la unidad se encuentran en continuo contacto con otros especialistas relacionados con las patologías de la mama que forman parte funcional de la unidad. Por ejemplo, oncólogos médicos y radioterapeutas, los radiólogos de la Unidad de Prevención de Mama, cirujanos plásticos, psicólogos de apoyo y facultativos de medicina nuclear, anatomía patológica o biología molecular, entre otros.

La Fe es el Hospital Centro de Referencia del SNS para quemados críticos, trasplante renal infantil, trasplante pulmonar, trasplante cardio-pulmonar, tratamiento de las infecciones osteoarticulares resistentes, trasplante de progenitores hematopoyéticos alogénico infantil, epilepsia refractaria, ataxias y paraplejías hereditarias.

Diariamente se realizan 40 ingresos programados y 95 urgentes, 1.000 pacientes permanecen ingresados (estancia media de 7 días) y 135 reciben el alta médica. El Area de urgencias de La Fe atiende unos 650 pacientes diarios. Se realizan 150 tratamientos diarios en el Hospital de Día, 3.614 pacientes son atendidos en el Área de Consultas Externas y 89 pacientes son atendidos en sus casos por la Unidad de Atención Domiciliaria.

C.1.5. Sistema de Salud de Chile

El Sistema de Salud en Chile está compuesto por todas aquellas instituciones, públicas o privadas, y personas que ejecutan acciones de promoción, protección y recuperación de la salud y de rehabilitación de la personas enfermas [211].

El sector está integrado por instituciones, organismos y entidades pertenecientes al sector público y al sector privado, constituyendo un sistema de salud mixto coordinado por el Ministerio de Salud. El sector público está representado principalmente por el Fondo Nacional de Salud (FONASA) en su aspecto de seguro social y por el Sistema Nacional de Servicios de Salud (SNSS) en su componente de prestación de servicios. Por otra parte, el sector privado está representado principalmente por las Instituciones de Salud Previsional (ISAPRE) y mutuales en su aspecto financiero previsional y los profesionales y centros asistenciales privados en su componente de prestación de servicios.

El SNSS está compuesto por el Ministerio de Salud y sus organismos dependientes: los Servicios de Salud, FONASA, el Instituto de Salud Pública y la Central de Abastecimiento. Además, participan del Sistema todas aquellas instituciones que realizan convenios, destacando los municipios y servicios delegados.

Los Servicios de Salud son los responsables de ejecutar las acciones integradas de fomento, protección y recuperación de la salud y rehabilitación de los enfermos y de hacer cumplir las disposiciones del Código Sanitario en las materias que les compete. Son organismos estatales funcionalmente descentralizados, dotados de personalidad jurídica y patrimonio propio para la realización de las acciones referidas. Son 26 Servicios con asignación geográfica definida más el Servicio de Salud Metropolitano del Ambiente.

Para llevar a cabo sus funciones, el SNSS se ha estructurado en una red asistencial de establecimientos y niveles de atención. La red asistencial de los Servicios está constituida por Hospitales, Consultorios Generales Urbanos y Rurales, Postas Rurales de Salud y Estaciones Médico Rurales. El SNSS cuenta en todo el país con 197 establecimientos hospitalarios, incluyendo nueve hospitales delegados. Tiene además un total de 376 consultorios de los cuales 230 son generales urbanos, 146 son consultorios generales rurales y tiene además 1.102 postas rurales. La gestión de la mayoría de establecimientos asistencias de primaria se han traspasado a la administración municipal.

Hay tres niveles de atención sanitaria. El Nivel Primario tiene mínima complejidad y amplia cobertura. Realiza atenciones de carácter ambulatorio en las Postas Rurales de Salud y en los Consultorios Generales, Urbanos y Rurales. Allí se ejecutan principalmente los Programas Básicos de Salud de las Personas. El Nivel Secundario corresponde a una complejidad intermedia y de cobertura media que actúa por referencia de forma ambulatoria y hospitalaria. El Nivel Terciario se caracteriza por su alta complejidad y cobertura reducida.

Otros sistemas públicos de salud incluyen los establecimientos de las Fuerzas Armadas y de Orden, Penitenciaría, Empresa Nacional del Petróleo y Universidad de Chile.

Los sistemas previsionales de salud se apoyan en el Régimen de Seguro de Salud, y en el Régimen de Accidentes de Trabajo y Enfermedades Profesionales. El seguro de salud es financiado por el trabajador con el aporte de un 7 % de sus remuneraciones y, opcionalmente, por el empleador con un aporte adicional del 2 % para sus trabajadores afiliados a las ISAPRE. Estos aportes del trabajador son ingresados en su mayor parte al Fondo Nacional de Salud o a las Instituciones de Salud Previsional.

En el sector privado, ISAPRE tienen por objetivo otorgar prestaciones y beneficios de salud a sus afiliados (23-27 % de la población), ya sea mediante la entrega de ellas en sus propias unidades de atención o a través del financiamiento de las mismas por pago a personas, clínicas, hospitales u otras instituciones pertenecientes a terceros.

C.2. Otros actores en salud

El suministro de servicios sanitarios requiere de infraestructuras, procesos y administraciones de alta tecnología y complejidad. Así pues, alrededor de la atención sanitaria estudiada en el

apartado 2.1, existe toda una comunidad de instituciones, proveedores industriales, proveedores de servicios auxiliares, farmacéuticas, centros de investigación y sistemas financieros que hace de la sanidad uno de los sistemas más complejos desde el punto de vista estructural.

Apéndice D

Implementación de un sistema de vigilancia geográfica de la resistencia bacteriana y el uso de antibióticos basado en Pentaho CE

Este anexo implementa en Pentaho CE el sistema de vigilancia geográfica de la resistencia bacteriana y el uso de antibióticos, diseñado a lo largo del capítulo 15 como ejemplo práctico de almacenes de datos y procesamiento analítico en línea.

En primer lugar pondremos en marcha una estación con las herramientas de diseño y ejecución de *Pentaho BI Suite Community Edition (CE)*. Esta plataforma incluye herramientas ETL, un servidor de procesamiento analítico en línea OLAP, herramientas de *reporting*, diseño de cuadros de mandos y una suite gráfica de minería de datos.

En este anexo utilizaremos *Pentaho BI Suite Community Edition (CE)* para Windows, donde también se ha instalado un servidor de base de datos MySQL 5.1.

D.1. Obteniendo Pentaho BI Suite Community Edition (CE)

La versión Community Edition de Pentaho es de código libre[a], a diferencia de la versión *Enterprise Edition (EE)*.

De las múltiples formas de conseguir Pentaho CE, hemos optado por descargar los paquetes ya compilados desde `sourceforge.net/projects/pentaho` de las diferentes herramientas de la plataforma. Específicamente, las herramientas utilizadas han sido:

- Kettle (Pentaho Data Integration CE). Herramienta ETL en la que encontraremos principalmente *spoon*, el entorno gráfico de diseño que permite diseñar transformaciones y trabajos para integrar diversas fuentes de datos, realizar transformaciones de registros y campos, y preparar bases de datos para alimentar cubos multidimensionales. Kettle también incorpora el lanzador de transformadores *pan* y el planificador de trabajos *kitchen*. El paquete utilizado en nuestras pruebas ha sido pdi-ce-4.0.0.
- Mondrian (Pentaho BI server CE). Servidor OLAP de Pentaho que permite navegar por los almacenes de datos mediante los operadores de análisis. Además, también permite realizar

[a] `sourceforge.net/projects/pentaho`, community.pentaho.com

informes, gráficos y cuadros de mando. Dispone también de una consola de administración desde donde planificar ejecuciones y gestionar las conexiones con las bases de datos que implementan las arquitecturas ROLAP de los almacenes de datos El paquete utilizado en las pruebas es biserver-ce-3.6.0.

- Mondrian Schema Workbench. Herramienta que permite diseñar los esquemas con los cubos multidimensionales y enlazarlos con las tablas de la arquitectura ROLAP de la base de datos de respaldo. En las pruebas se utilizó el paquete psw-ce-3.2.0.13661.

D.2. Base de datos con las fuentes de datos

Tal como definimos en el capítulo 15, hemos simulado un conjunto de EHR que contiene los resultados de los estudios de laboratorio de resistencia de muestras de bacterias a antibióticos. Además se ha registrado el uso de antibióticos en los diferentes hospitales involucrados.

Como implementación del modelo-entidad relación de la figura 15.1 hemos utilizado tres tablas (*estudiomicrobiologia*, *cmiestudiomicrobiologiaantibiotico* y *usoantibiotico*) de la base de datos *resistenciaantibioticos* en una base de datos MySQL gestionada desde el cliente Navicat Lite. Tras simular un conjunto de datos mediante el software estadístico R, generando ficheros csv, realizamos la carga de las tablas mediante el *wrapper* de ficheros csv de Navicat.

Habiendo comprobado la disponibilidad de acceso a nuestra fuente principal de datos, podemos abordar el diseño conceptual del modelo multidimensional de nuestro problema. Para ello seguiremos el apartado 15.2, con lo que obtendremos el modelo de datawarehouse con dos estrellas y cuatro dimensiones de la figura 15.2.

D.3. Diseño de la arquitectura ROLAP y carga del almacén de datos

Lo primero que nos damos cuenta al comparar nuestro modelo multidimensional con la fuente principal de datos es que no disponemos de la información de correspondencia entre los hospitales y provincias que nos permita implementar la jerarquía de niveles de la dimensión Localización. Necesitaremos una fuente de datos que nos ofrezca dicha correspondencia. Optamos por ampliar nuestra base de datos "resistenciaantibioticos" con una tabla "provinciahospital" que relacione los niveles de la dimensión Localización.

Suponemos que los servicios centrales de la Agencia Valenciana de Salud nos proporcionan una hoja de cálculo excel con una tabla de dos columnas donde encontramos la correspondencia hospital-provincia. Comienza entonces el uso de Spoon, que es la herramienta ETL para el diseñador que ofrece Pentaho. Desde la transformación de la figura D.1, diseñada como flujo de trabajo compuesta de componentes de Spoon, podremos crear la tabla "provinciahospital" D.2 y realizar la carga de datos.

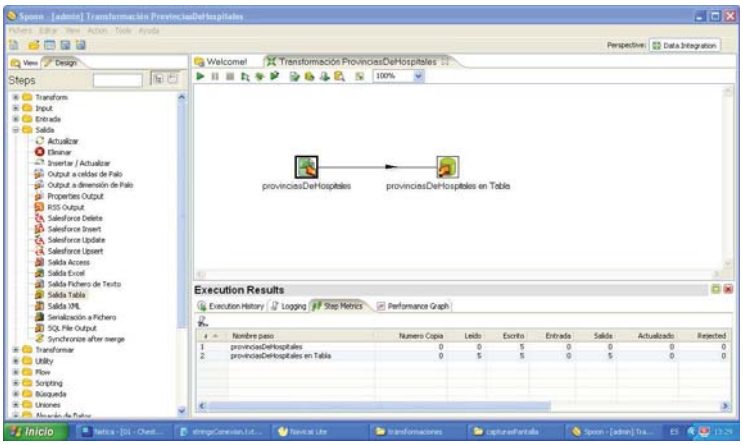


Figura D.1: Transformación para cargar la tabla "provinciahospital" desde una hoja de cálculo excel.

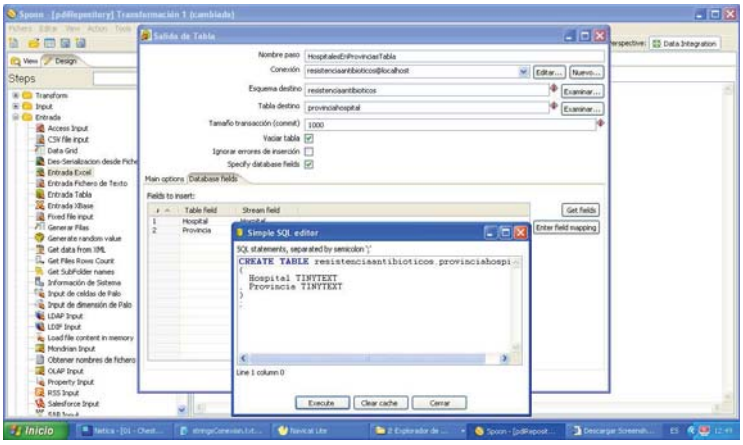


Figura D.2: Creación de la tabla"provinciahospital" desde *spoon*.

Teniendo la fuente de datos ampliada con la correspondencia de hospitales a provincias, podemos preparar la base de datos que servirá de arquitectura ROLAP de nuestro almacén de datos. Crearemos para ello una nueva base de datos en MySQL que llamaremos "resistenciaantibioticosdw". Una vez más, utilizaremos Spoon para diseñar la transformación que nos permita crear las tablas de hechos y dimensiones y realizar su carga desde nuestra fuente de datos "resistenciaantibioticos". Con la transformación de la figura D.3 crearemos la tabla de hechos "hechosresistencia" y las tablas de dimensiones "antibioticodimension", "bacteriadimension" y "localizaciondimension". Será necesario realizar las uniones de registros desde las tablas originales de "resistenciaantibioticos" para conseguir la forma de estrella del *datamart*. El componente "Búsqueda/Actualización en Dimensión" nos permite crear y actualizar las tablas de dimensiones con campos de *time stamping* para el mantenimiento del almacén de datos. Además, añadiremos

una clave primaria técnica a cada tabla de dimensión, cuyo campo identificaremos con el prefijo "idt".

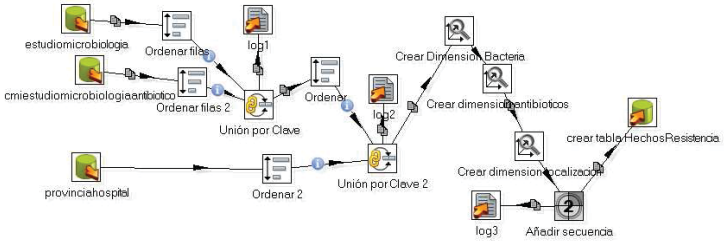


Figura D.3: Transformación para crear y cargar el *datamart* de resistencia bacteriana a antibióticos.

Como ya disponemos de las tablas de dimensiones, únicamente deberemos crear la tabla de hechos "hechosusoantibioticos" para crear la estrella de "uso de antibióticos". La transformación de la figura D.4 realiza la carga y mantenimiento de este *datamart*.

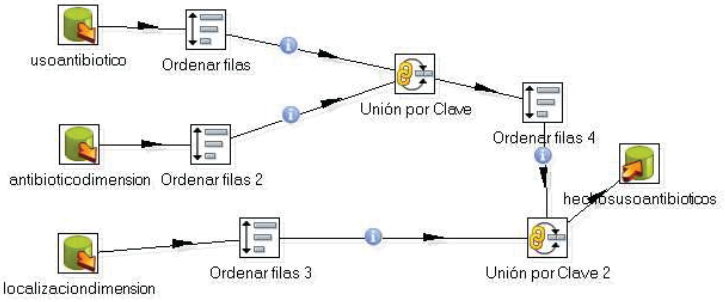


Figura D.4: Transformación para crear y cargar el datamart de uso de antibióticos.

D.4. Especificación y publicación del esquema del almacén de datos

Preparado el modelo conceptual y la arquitectura ROLAP con los datos cargados en nuestra base de datos relacional, es hora de diseñar el esquema de los cubos OLAP para Mondrian que es, en definitiva, la definición de nuestro almacén de datos. En Pentaho esta tarea podemos realizarla fácilmente mediante la herramienta Mondrian Schema Workbench que generará el esquema de nuestro datawarehouse como fichero XML y podremos publicarlo en el servidor Mondrian.

Lo primero que necesitaremos hacer antes de crear el cubo es establecer la conexión con la base de datos relacional "resistenciaantibioticosdw" que contiene nuestra arquitectura ROLAP. En nuestro caso, al tener una instalación estándar de MySQL en el servidor local, la URL de conexión será jdbc:mysql://localhost:3306/resistenciaantibioticosdw. Podremos utilizar el driver com.mysql.jdbc.Driver como clase del driver JDBC y tendremos que asegurarnos que en el

directorio "schema-workbench/drivers" tenemos el paquete jar del driver (e.g. mysql-connector-java-5.0.7.jar en nuestro caso).

Comenzaremos creando el cubo "Resistencia" que implementará la estrella "resistencia de bacterias a antibióticos". Añadiremos la tabla "hechoresistencia" de "resistenciaantibioticosdw" al cubo. A continuación añadiremos la medida "PorcentajeResistenciaAntibiotico", con la media (avg) como agregador (ver figura D.5). Enlazaremos la medida con la columna CMI de la tabla de hechos del cubo. Workbench irá validando que el esquema sigue una estructura bien formada y que las tablas y campos utilizados están disponibles en la conexión de la base de datos.

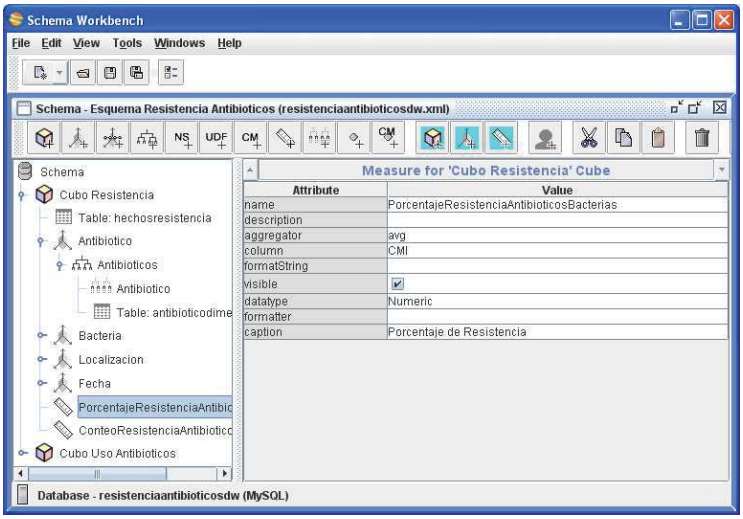


Figura D.5: Definición de la medida del cubo "Resistencia".

Podremos añadir ahora las cuatro dimensiones al cubo: "Antibiótico", "Bacteria", "Localizacion", "Fecha". Comenzaremos creando la dimensión "Antibiótico" (ver figura D.7), a la que añadiremos la tabla "antibioticodimension". Estableceremos la clave primaria de la dimensión al campo "idtAntibiotico" y la declararemos de tipo "StandardDimension". Añadiremos una jerarquía "Antibioticos" con la misma clave primaria (ver figura D.6), que contendrá un nivel de agregación que llamaremos "Antibiótico". Este nivel se enlazará con la columna "antibiótico" de la tabla "antibioticodimension" y que será de tipo cadena (ver figura D.8).

Podremos crear las dimensiones "Bacteria" y "Localización" de forma similar. Particularmente la dimensión "Localización" tendrá dos niveles "Provincia" y "Hospital" en la jerarquía, que deberán estar en orden de agregación descendente en el esquema y enlazados con sus campos de la tabla "localizaciondimension".

La dimensión fecha será de tipo "TimeDimension" y puede definirse sin tabla de dimensiones asociada, por lo que su clave primaria será el campo "fecha" de la tabla "hechosresistencia" a la dimensión. La jerarquía contendrá dos niveles, "Año" y "Mes", a los cuales añadiremos un campo "KeyExpression" con el dialecto "mysql" que ejecutará el código year(fecha) y month(fecha) respectivamente (ver figura D.9).

El cubo "Uso Antibióticos" se podrá crear siguiendo los mismos pasos que los descritos en el cubo "Resistencia", lo que completará el esquema de nuestro datawarehouse.

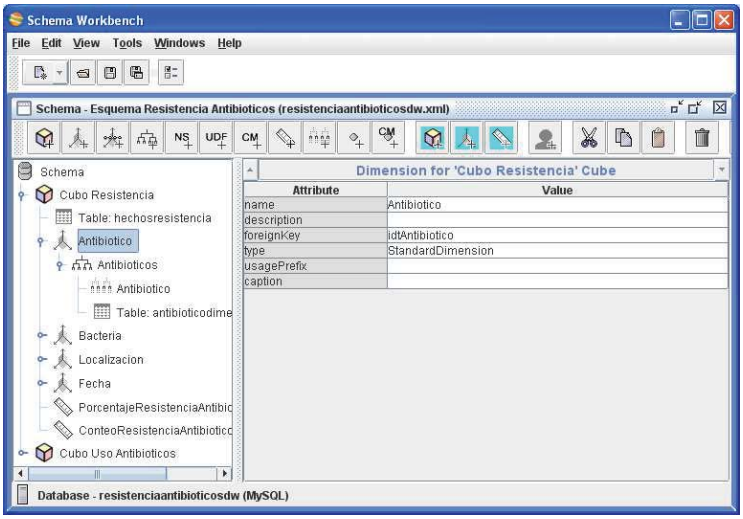


Figura D.6: Definición de la dimension “Antibiótico” del cubo “Resistencia”.

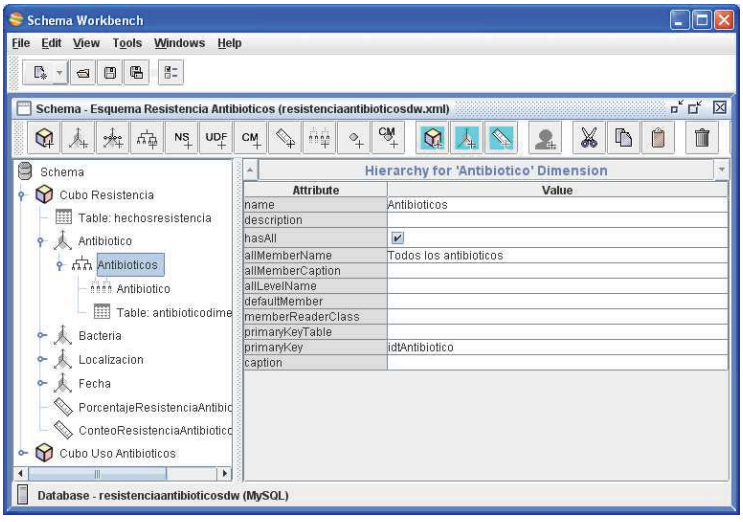


Figura D.7: Definición de la jerarquía de la dimension “Antibiótico” del cubo “Resistencia”.

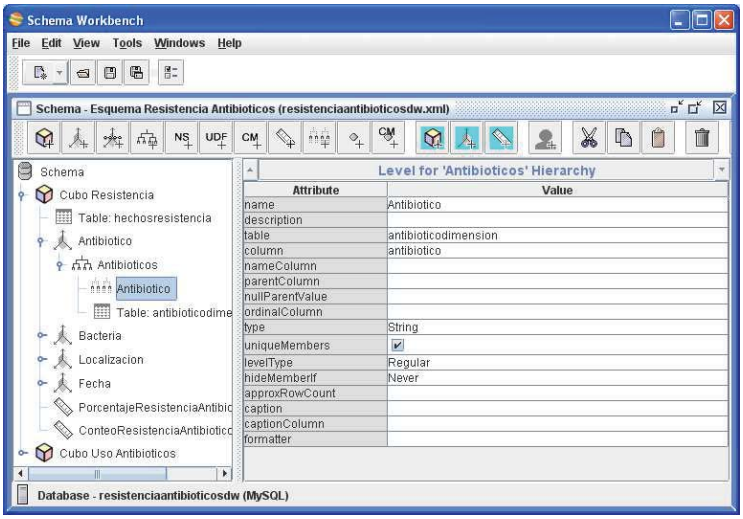


Figura D.8: Definición del único nivel de la dimensión Antibiótico del cubo "Resistencia".

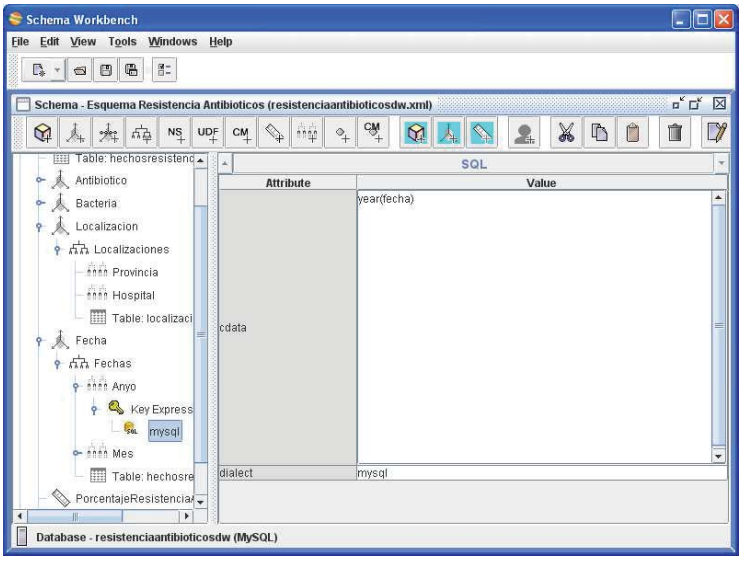


Figura D.9: Definición de la dimensión Fecha del cubo "Resistencia".

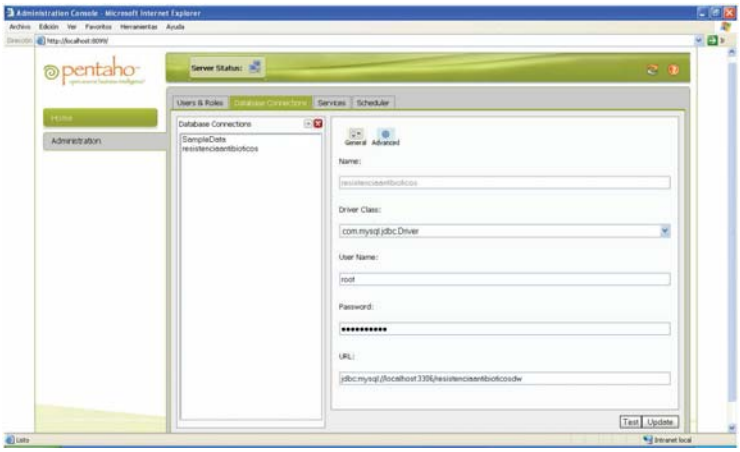


Figura D.10: Creación de la conexión desde Mondrian a la base de datos ROLAP del almacén de datos de resistencia bacteriana y uso de antibióticos.

En este momento podremos preparar el entorno de ejecución de nuestro cubo en Mondrian. Para ello, añadiremos en Mondrian la conexión a la base de datos "resistenciaantibioticosdw". Podremos realizar esta tarea cómodamente desde la consola de administración de nuestro servidor local Mondrian. Podremos acceder vía web, después de activar el servidor Mondrian y el servidor de administración de Mondrian en la dirección local http://localhost:8099, que despliega un menú de administración que incluyen las conexión de base de datos. Incluiremos nuestra conexión según la figura D.10 llamándola "resistenciaantibioticos".

Podremos finalmente publicar en el servidor Mondrian el esquema de nuestro datawarehouse recién creado desde el propio Workbench. Lo primero es asegurarnos que el servidor Mondrian está activo. Deberemos validarnos como usuarios del servidor Mondrian y añadir la contraseña de publicación para tener permiso en el servidor[b]. Tendremos la precaución de crear una carpeta "resistenciaantibioticos" en el servidor donde colocar nuestro esquema, no pudiendo publicarlo en la carpeta raíz. Por último, tendremos que especificar que nuestros cubos utilizan la conexión "resistenciaantibioticos" como fuentes de datos ROLAP. La figura D.11

Desde ahora, tenemos disponible nuestro datawarehouse en el servidor Mondrian, pudiendo analizar los cubos de resistencia bacteriana y uso de antibióticos mediante operadores OLAP.

D.5. Procesamiento analítico en línea con Mondrian

Accederemos al servidor Mondrian a través del navegador, en nuestra implementación por defecto a través de http://localhost:8080/pentaho. Tras la validación con un usuario registrado[c] tendremos un panel de control como el mostrado en la figura D.12 desde el que podremos crear nuevas vistas de análisis de nuestros cubos y también acceder a las vistas ya creadas anteriormente. Además, desde el panel de inicio podremos crear o acceder a nuevos informes y cuadros de mando.

[b]En la versión utilizada en nuestras pruebas, el usuario admin/password está creado por defecto, y la contraseña de publicación es password.

[c]En la implementación por defecto puede utilizarse el usuario joe/password. La gestión de usuarios se puede realizar desde la consola de administración (http://localhost:8099).

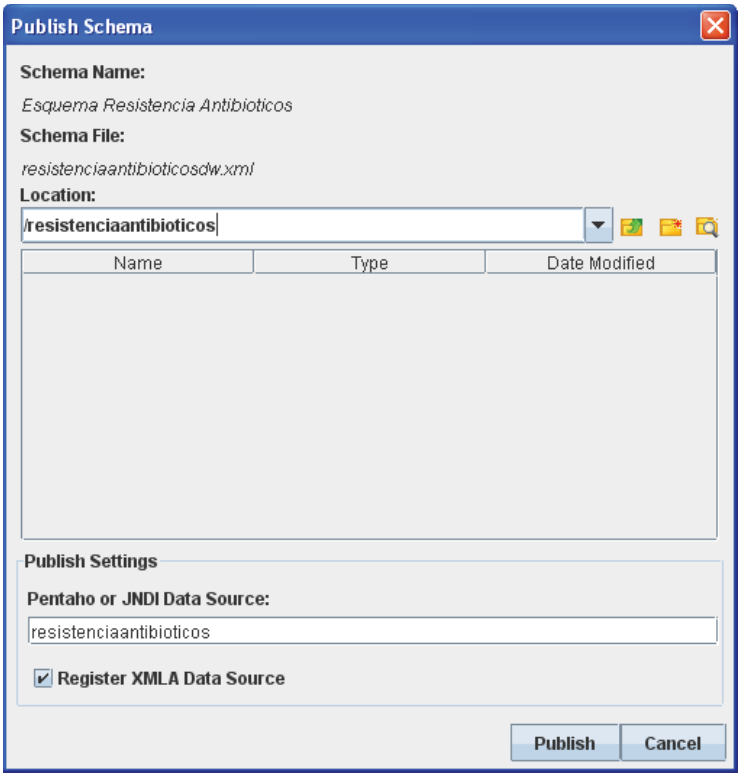


Figura D.11: Publicación del esquema del datawarehouse de resistencia bacteriana y uso de antibióticos en el servidor Mondrian.

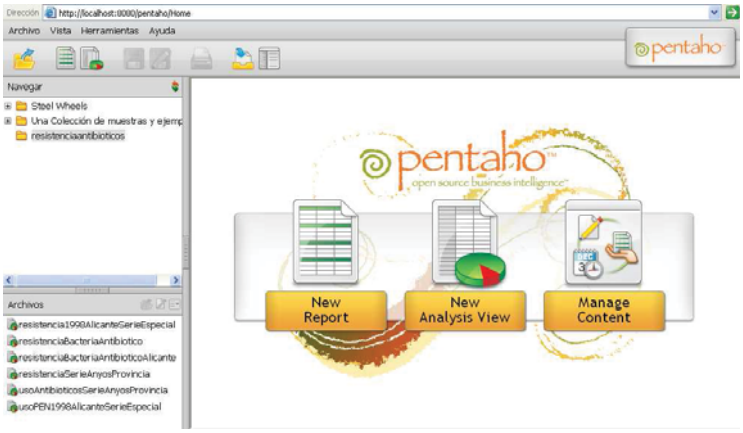


Figura D.12: Panel de control mostrado por Mondrian al inicio de sesiónl.

Figura D.13: Selección del datawarehouse y el cubo desde Mondrian.

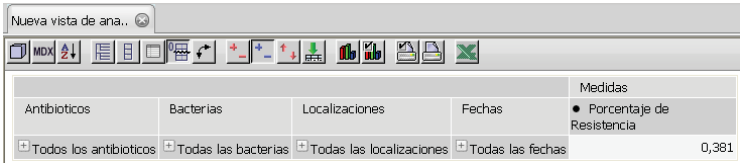


Figura D.14: Vista de análisis del cubo Resistencia bacteriana a antibióticos.

Crearemos una nueva vista de análisis basada en el cubo "Resistencia" del esquema "Resistencia Antibióticos" que es nuestro datawarehouse recién creado (ver figura D.13). Mondrian nos enseñará una primera vista de análisis del cubo "Resistencia" con todas las dimensiones agregadas, por lo que tendremos la medida resumen de todo nuestro almacén de datos, tal como vemos en la figura D.14.

Observamos que en la parte superior de la vista de análisis tenemos una barra de botones con las operaciones que podemos realizar sobre el cubo y las opciones de visualización y generación de gráficos y documentos de exportación.

El navegador OLAP nos dará acceso a las operaciones OLAP que podremos aplicar sobre nuestra vista del almacén de datos para detallar la tabla de contingencia conforme nos interese: cambiar de nivel de agregación en las dimensiones, seleccionar valores de las dimensiones, trasladar dimensiones desde filas a columnas y viceversa, etcétera. Cuando hayamos encontrado la tabla de contingencia que mejor resuma la vista que buscamos, podremos generar un gráfico que permita una inspección visual de los resultados. La figura D.15 muestra la tabla de contingencia con el porcentaje de resistencia de las bacterias frente a los antibióticos incluidos en el estudio. Las dimensiones de localización y fecha se han mantenido agregadas al máximo, lo que nos da una visión global estática de nuestro problema. Podremos guardar las vistas de análisis como ficheros "xaction" y acceder a ellos directamente o a través de cuadros de control. La sección 15.4 presenta otras posibles vistas de análisis con el fin de encontrar relaciones de causalidad en series temporales de nuestro problema.

	Medidas		
	Porcentaje de Resistencia		
	Antibioticos		
Bacterias	AMX	CIP	PEN
sau	0,742	0,037	0,915
sp	0,119	0,128	0,608
svi	0,279	0,221	0,304

Slicer:

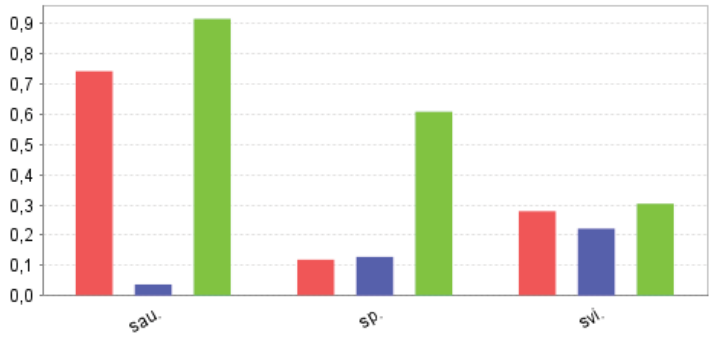


Figura D.15: Tabla de contingencia de porcentaje de resistencia de las bacterias frente a los antibióticos generada aplicando operadores OLAP al cubo Resistencia.

Apéndice E

Métodos matemáticos

E.1. Optimización Newton-Raphson

Si una función es cóncava, Newton-Raphson encuentra el óptimo global [212].

Al obtener una expresión no lineal, no podemos calcular una solución cerrada para $\hat{\mathbf{w}}$, por lo que se debe aplicar un esquema iterativo de optimización. La función de error es cóncava, por lo que es posible miminizar al función de error mediante Newton-Raphson, de la forma

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \mathbf{H}^{-1}\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}},$$

donde $\mathbf{H}^{-1}$ es la matriz hessiana de $L(\mathbf{w})$.

E.2. Aproximación de Laplace

Cuando el cálculo de la integral de una verosimilitud marginal o factor normalizador es analíticamente no resoluble, podemos hacer uso de una aproximación analítica determinista como alternativa. Esta aproximación es la *aproximación de Laplace*. Cuando la densidad de probabilidad tiene una forma unimodal convexa, se puede hacer uso de esta aproximación que es ampliamente empleada como aproximación determinista local.

El método de la aproximación de Laplace trata de aproximar, mediante una función de densidad de probabilidad gaussiana $q(z)$, una densidad de probabilidad no gaussiana $p(z)$ definida sobre un conjunto de variables continuas. Sea la función de distribución $p(z) = Z^{-1}f(z)$, siendo Z un coeficiente de normalización. La aproximación de Laplace busca una distribución gaussiana aproximada $q(z)$ centrada en la moda z_{max} de la distribución $p(z)$. Para ello, se aplica la expansión de la serie de Taylor para $\log\{f(z)\}$,

$$\log\{f(z)\} = \log f(z_0) + \left.\frac{\partial \log f(z)}{\partial z}\right|_{z=z_0}(z - z_0) + \frac{1}{2}\left.\frac{\partial^2 \log f(z)}{\partial z^2}\right|_{z=z_0}(z - z_0)^2 + \mathcal{O}(z^3) \quad \text{(E.1)}$$

donde se asume que los términos de orden mayor, representados por $\mathcal{O}(z^3)$, son despreciables. Supongamos que $z_0 = z_{max}$ es un máximo local en $f(z)$. En este caso, el término de primer orden es 0 puesto que z_{max} es un punto estacionario donde la derivada será nula. La expansión de la serie de Taylor será entonces

$$\log\{f(z)\} \approx \log f(z_{max}) + \frac{1}{2}\left.\frac{\partial^2 \log f(z)}{\partial z^2}\right|_{z=z_{max}}(z - z_{max})^2 \quad \text{(E.2)}$$

Si tomamos el exponencial y usamos

$$\beta = -\frac{\partial^2 \log f(z)}{\partial z^2}\bigg|_{z=z_{max}}$$

se obtiene

$$f(z) \approx f(z_{max}) \exp\left\{ -\frac{\beta}{2}(z - z_{max})^2 \right\} \tag{E.3}$$

lo que recuerda la forma de una distribución gaussiana. Cabe mencionar que la aproximación gaussiana solo estará bien definida si el punto estacionario z_{max} es un máximo local, ya que es imprescindible que la segunda derivada de $f(z)$ en z_{max} sea negativa. Por último, tendremos una distribución normalizada $q(z)$ si usamos el resultado de normalización estándar de una gaussiana,

$$q(z) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{ -\frac{\beta}{2}(z - z_{max})^2 \right\} \tag{E.4}$$

La aproximación de Laplace se puede extender para distribuciones multivariantes, donde una función de densidad de probabilidad $p(\mathbf{z}) = Z^{-1} f(\mathbf{z})$ está definida en un espacio multidimensional $\mathbb{R}^D$. Asumiendo que existe un punto estacionario $\mathbf{z}_{max}$ donde el gradiente $\nabla f(\mathbf{z})$ desaparece, la expansión de Taylor sobre este punto $\mathbf{z}_{max}$ es

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_{max}) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \tag{E.5}$$

donde $\mathbf{H}$ es la matriz hessiana con dimendiones $D \times D$, definida como

$$\mathbf{H} = -\nabla\nabla \log f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_{max}} \tag{E.6}$$

donde ∇ es el operador gradiente. Tomando el exponencial en ambos lados de la ecuación obtenemos

$$f(\mathbf{z}) \approx f(\mathbf{z}_{max}) \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \right\} \tag{E.7}$$

Usando el resultado estándar con el coeficiente adecuado para normalizar una densidad de probabilidad gaussiana multivariante la función $q(\mathbf{z})$ es

$$\begin{aligned} q(\mathbf{z}) &= (2\pi)^{-D/2} |\mathbf{H}|^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \right\} \\ &= \mathcal{N}(\mathbf{z}_{max}, \mathbf{H}^{-1}) \end{aligned} \tag{E.8}$$

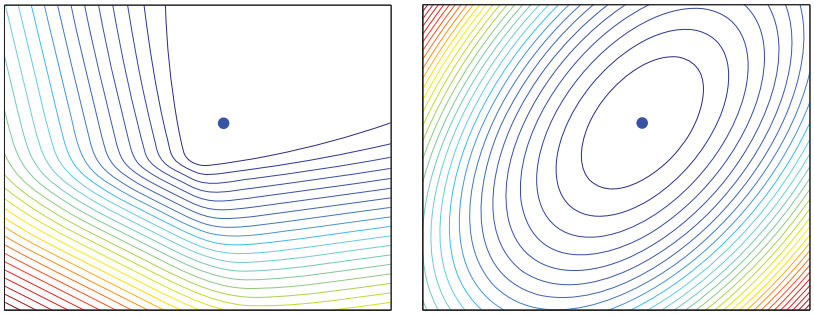
Igual que en la versión univariante, esta distribución gaussiana estará bien definida siempre que su matriz de precisión $\mathbf{H}$ sea definida positiva, lo que implica que el punto estacionario $\mathbf{z}_{max}$ sea un máximo local.

Veamos el proceso paso a paso adecuado para usar la aproximación de Laplace a una función de densidad de probabilidad unimodal $p(\mathbf{z})$ con una gaussiana $q(\mathbf{z})$. El primer paso es encontrar un máximo local $\mathbf{z}_{max}$ de la función de densidad de probabilidad $p(\mathbf{z})$ empleando algún tipo de algoritmo de optimizción numérica. Cabe decir que si la función $p(\mathbf{z})$ es multimodal, entonces pueden hallarse diferentes aproximaciones, al menos una por cada máximo local. El siguiente

paso es calcular la inversa de la matriz hessiana para el punto estacionario $\mathbf{z}_{max}$ empleando la siguiente expresión

$$\mathbf{H}^{-1} = -\left(\frac{\partial^2}{\partial \mathbf{z} \partial \mathbf{z}^{\mathrm{T}}} \log p(\mathbf{z}) \right)^{-1} \tag{E.9}$$

por último, podemos aproximar la función de densidad de probabilidad $p(\mathbf{z})$ usando $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}_{max}, \mathbf{H}^{-1})$. La figura E.1 muestra el resultado de aproximar una función de densidad de probabilidad bidimensional con esta aproximación.



(a) FDP no gaussiana (b) FDP gaussiana aproximada

Figura E.1: Se ha aproximado una función de densidad de probabilidad (FDP) no gaussiana (a) a otra FDP gaussiana usando la aproximación de Laplace (b). El valor máximo, que es el mismo para ambas densidades, se muestra con el punto azul.

E.3. Métodos de muestreo basados en cadenas de Markov Monte Carlo

Cuando los modelos probabilísticos no pueden inferirse de forma exacta se requiere algún método general de aproximación. Estas aproximaciones pueden ser deterministas, como la aproximación de Laplace, la aproximación Variacional [212, 213] o el método *Expectation-Propagation* [214, 215], o pueden ser no deterministas al depender de un muestreo aleatorio que siga una distribución concreta. En los modelos predictivos la distribución a posteriori de las variables no observadas (p.e. parámetros de un modelo lineal de clasificación por regresión logística) es necesaria para evaluar los valores esperados en muestras futuras.

Esta tarea requiere una evaluación de la esperanza de una función con respecto a la distribución a posteriori de los parámetros. ¿Qué significa esto? En primer lugar recordemos que la **esperanza matemática** de una función con variable aleatoria X es una media de la función con los posibles valores de X ponderada por la probabilidad de que tome esos valores $p(X = x)$, es decir,

$$E[X] = \sum_i f(x_i) p(x_i),$$

que para variables continuas es:

$$E[X] = \int f(x)p(x)dx. \tag{E.10}$$

Ahora, si se comparan las ecuaciones (8.39) o (8.40) y (**??**) se observarán las similitudes entre las expresiones. Por ejemplo, si hacemos $Q(\theta) = p(\theta|S)$ y $f(\theta) = p(s_{N+1}|\theta)$ en la expresión (8.39) o $f(\theta) = p(y_{N+1}|x_{N+1}, \theta)$ en la expresión (8.40), entonces la esperanza de $f(\theta)$ es:

$$E[f] = \int f(\theta)Q(\theta)d\theta.$$

Esta esperanza no es sencilla de calcular de forma analítica. Sin embargo, mediante métodos de *Monte Carlo* se pueden obtener muestras de parámetros empleando la distribución Q de forma que se generen un conjunto de extracciones de parámetros $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ a partir de los cuales se puede obtener una **estimación de la esperanza** de $f(\theta)$ haciendo:

$$E[f] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t). \tag{E.11}$$

El problema ahora es cómo generar estos valores Θ.

Existen varios métodos de *Monte Carlo* para llevar a cabo estas extracciones. Hay métodos que extraen θ_t tal que entre sí son independientes. Pero si Q es una distribución compleja, generar estos valores puede ser complicado. Sin embargo, es posible generar una serie de valores **dependientes** empleando *cadenas de Markov* que seguirán dando un valor no sesgado en la estimación de la ecuación (E.11), siempre que la dependencia entre cada extracción θ_t no sea demasiado grande[a]. Es decir, la estimación seguirá convergiendo al valor real cuando $T \to \infty$.

El marco de simulación MCMC permite la aproximación de gran número de modelos probabilísticos mediante muestreo. Además, las soluciones por MCMC escalan bien con la dimensionalidad de la muestra. A continuación, se ofrece una definición de lo que es una cadena de Markov y cómo se obtiene por muestreo de Monte Carlo. Después, veremos el algoritmo Metropolis, el algoritmo de Gibbs y la generalización de ambos, el algoritmo de Metropolis-Hastings, que permiten extraer muestras de distribuciones arbitrarias.

E.3.1. Markov Chain Monte Carlo (MCMC)

La serie de parámetros dependientes $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ se pueden generar mediante una **cadena de Markov**. La cadena de Markov se define mediante una *distribución inicial* para el primer estado de la cadena, θ_1, y un conjunto de *probabilidades (o densidades) de transición* de un estado θ_t a un nuevo estado θ_{t+1}, que expresaremos como $T(\theta_{t+1}|\theta_t)$.

Así pues, se puede estimar la esperanza de la ecuación (E.11) con respecto a Q con Θ, extrayendo cada parámetro $\theta_{t+1} \sim T(\theta_{t+1}|\theta_t)$. De tal modo que el conjunto de parámetros extraídos convergerá a la distribución que se desea calcular $\Theta \sim Q(\theta)$. Este objetivo se alcanzará si se cumple la condición de *ergodicidad*[b]. Para que la cadena de Markov sea ergódica, su distribución de equilibrio Q deberá ser *invariante (o estacionaria)*, es decir, debe cumplirse que si $\theta_t \sim Q(\theta)$, entonces $\theta_{t'}$ deberá seguir la misma distribución para todo $t' > t$. Es decir, $\theta_{t'} \sim Q(\theta)$.

[a] Esta afirmación deberá entenderse de manera general, ya que es vaga y no establece qué es una dependencia entre variables grande.

[b] En términos sencillos, un sistema es *ergódico* si a lo largo del tiempo converge a una distribución de *equilibrio* invariante desde cualquier estado inicial posible.

En muchas ocasiones, la convergencia a la distribución de equilibrio Q puede requerir algún tiempo. Por esta razón, se suelen descartar los primeros estados ya que pueden no ser representativos de la distribución de equilibrio Q que es, al fin y al cabo, la distribución que se desea muestrear.

En resumen, al usar un método *Markov chain Monte Carlo* para estimar la esperanza respecto a la distribución Q necesitamos:

1. Construir una cadena de Markov ergódica,
2. que tenga como distribución de equilibrio Q,
3. que converja tan rápido como sea posible y
4. que las distintas extracciones de θ_t cuando se alcance la distribución de equilibrio no sean altamente dependientes.

A continuación se presentan los algoritmos que pueden ofrecernos los métodos para obtener las cadenas de Markov apropiadas para extraer los parámetros Θ que sigan $Q(\theta)$ cumpliendo las condiciones propuestas.

E.3.2. Algoritmo de Metropolis

El *algoritmo de Metropolis* [216] define una cadena de Markov donde cada nuevo estado, θ_{t+1}, se genera a partir del estado anterior, θ_t, pero extrayendo en primer lugar un **estado candidato** que sigue una **distribución propuesta** y, después, decidiendo si se escoge o no dicho candidato, en base a una densidad de probabilidad relativa al estado anterior θ_t y con respecto a la distribución invariante $Q = p(\theta|S)$, que es la que buscamos simular y donde S son las observaciones disponibles.

En concreto, cada nueva propuesta θ^* se genera siguiendo la densidad de probabilidad de transición $\theta^* \sim T(\theta|\theta_t)$. En el algoritmo de Metropolis se asume que la probabilidad de transitar de un estado a otro es **simétrica**, es decir, que $T(\theta^*|\theta_t) = T(\theta_t|\theta^*)$[c]. Para decidir si este candidato se escoge, se calcula la siguiente expresión

$$\begin{aligned}\alpha(\theta^*|\theta_t) &= \text{mín}\left\{1, \frac{p(\theta^*|S)T(\theta_t|\theta^*)}{p(\theta_t|S)T(\theta^*|\theta_t)}\right\} \\ &= \text{mín}\left\{1, \frac{p(\theta^*|S)}{p(\theta_t|S)}\right\}.\end{aligned} \tag{E.12}$$

El candidato θ^* será finalmente aceptado si $\alpha(\theta^*|\theta_t) > \varphi$, donde $\varphi \sim U(0,1)$ se escoge aleatoriamente. Si la condición se cumple, entonces $\theta_{t+1} = \theta^*$, de lo contrario $\theta_{t+1} = \theta_t$. El pseudocódigo [E.1] describe el algoritmo de Metropolis.

E.3.3. Algoritmo de Gibbs

El *algoritmo de Gibbs* es un caso particular del algoritmo Metropolis-Hastings donde siempre se aceptan las extracciones muestreadas, es decir, $\alpha(\theta^*|\theta_t) = 1$. Obviamente, el objetivo es también construir una cadena de Markov cuyos valores converjan a la distribución de equilibrio. Este algoritmo se suele emplear cuando se tienen múltiples parámetros a muestrear

[c]Para comprender la importancia de esta información conviene conocer el algoritmo de Metropolis-Hastings.

$\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$. El problema es que muestrear directamente a partir de la distribución conjunta $Q(\Theta)$ puede ser inviable. La clave del *algoritmo de Gibbs* es que las muestras se obtienen a partir de distribuciones condicionales *univariantes*, esto es, distribuciones condicionales de cada parámetro individual θ_j dados los demás parámetros $\theta_{-j} = \{\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_m\}$[d].

Esto nos permite simular una cadena de Markov en la que Θ_{t+1} se genera a partir de Θ_t. En el proceso se generan conjuntos de parámetros donde $\{\theta_j^{(t+1)} | \theta_{-j}^{(t)}, S\} \sim Q(\theta_j | \theta_{-j}, S)$. El pseudocódigo [E.2] describe los pasos a seguir al emplear el algoritmo de Gibbs.

Algoritmo E.1 Algoritmo Metropolis

$\theta_0 \sim p(\theta_0), p(\theta_0|S) > 0$
for $t = 1, 2, \ldots$ **do**
  $\theta^* \sim T(\theta|\theta_t)$
  $\alpha(\theta^*|\theta_t) = \text{mín}\left\{1, \frac{p(\theta^*|S)}{p(\theta_t|S)}\right\}$
  $\varphi \sim U(0, 1)$
  if $\alpha(\theta^*|\theta_t) > \varphi$ **then**
    $\theta_{t+1} = \theta^*$
  else
    $\theta_{t+1} = \theta_t$
  end if
end for

Algoritmo E.2 Algoritmo Gibbs Sampling

$\theta_j^{(1)} \sim Q(\theta_j|S)$
for $t = 1, \ldots, T$ **do**
  $\theta_1^{(t+1)} \sim Q(\theta_1|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_m^{(t)}, S)$
  $\theta_2^{(t+1)} \sim Q(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_m^{(t)}, S)$
  $\vdots$
  $\theta_m^{(t+1)} \sim Q(\theta_m|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{m-1}^{(t+1)}, S)$
end for

E.3.4. Algoritmo Metropolis-Hastings

El algoritmo Metropolis-Hastings (MH) es una generalización del algoritmo de Metropolis. Este algoritmo genera una cadena de Markov $\theta^{(1)}, \theta^{(2)}, \ldots$ de muestras de una distribución arbitraria $p(\theta|y)$. Como en el algoritmo de Metropolis, la cadena se va generando por la aceptación o rechazo de una propuesta θ^* extraída de una distribución de transición $T(\theta|\theta_t)$. En este caso no se asume simetría en las probabilidades de transición entre estados, así pues $T(\theta^*|\theta_t) \neq T(\theta_t|\theta^*)$. Esto hace algo más complejo el cálculo de $\alpha(\theta^*|\theta_t)$. Por lo demás, el algoritmo de Metropolis-Hastings es igual al de Metropolis. El pseudocódigo E.3 describe el algoritmo MH.

[d] A este tipo de distribuciones se les llama distribuciones condicionales univariantes en Neal [217] y *full conditional distributions* en Hoff [84]

Algoritmo E.3 Algoritmo Metropolis-Hastings

$\theta_0 \sim p(\theta_0), p(\theta_0|S) > 0$
for $t = 1, 2, \ldots$ **do**
  $\theta^* \sim T(\theta|\theta_t)$
  $\alpha(\theta^*|\theta_t) = \text{mín}\left\{1, \frac{p(\theta^*|S)T(\theta_t|\theta^*)}{p(\theta_t|S)T(\theta^*|\theta_t)}\right\}$
  $\varphi \sim U(0, 1)$
  if $\alpha(\theta^*|\theta_t) > \varphi$ **then**
    $\theta_{t+1} = \theta^*$
  else
    $\theta_{t+1} = \theta_t$
  end if
end for

Como vemos, el algoritmo requiere calcular la $p(\theta|S)$ para todo θ, además de extraer una muestra propuesta θ^* de la distribución de transición $T(\theta^*|\theta_t)$ en todo t, y extraer una muestra aleatoria φ de una distribución uniforme. La propuesta θ^* se acepta con probabilidad α si es superior a φ. Típicamente, el algoritmo se puede aplicar a distribuciones $p(\theta|S) = \tilde{p}(\theta|S)/Z_p$, donde es únicamente necesario calcular $\tilde{p}(\theta|S)$, ya que los Z_p del numerador y denominador de la fracción para calcular α se cancelan.

Una vez obtenida la cadena de Markov $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ se puede calcular una aproximación de cualquier función $f(\theta)$ mediante E.11. Tanto en este algoritmo como en el de Metropolis se suelen descartar los b primeros casos obtenidos por muestreo como margen de convergencia del algoritmo a la distribución de equilibrio $p(\theta)$.

El algoritmo puede interpretarse como un algoritmo paso a paso estocástico para la búsqueda de $p(\theta)$, aceptando muestras que incrementan la densidad de probabilidad, pero aceptando también los pasos donde no se aumenta. La velocidad de convergencia del algoritmo MH depende de $T(\theta|\theta_t)$, por lo que las versiones del algoritmo suelen centrarse en el diseño de esta distribución.

Podemos encontrar la demostración de la convergencia de la distribución estacionaria de la cadena de Markov Θ obtenida por MH a la distribución objetivo $p(\theta)$ en los textos [93, 218].

Glosario

Notación matemática

$\mathbf{x}$	Vector columna $\mathbf{x}$
D	Dimensión de un vector D-dimensional $\mathbf{x} = (x_1, \ldots, x_D)$
$\mathcal{S}$	Conjunto de casos independientes y idénticamente distribuidos, e.g. $\mathcal{S}_T = (x_i, y_i), i = 1, \ldots, N; \mathbf{x_i} \in \mathbb{R}^d, y_i \in \{y_1, \ldots, y_C\}$ es una muestra de entrenamiento supervisado con N casos, donde el caso i tiene un vector de entrada $\mathbf{x}$ en un espacio $\mathbb{R}^d$ y clase de salida y_i en un conjunto de C categorías.
N	Dimensión de un conjunto de casos $\mathcal{S}$.
C	Número de clases que puede tomar una variable categórica $y \in \{y_1, \ldots, y_C\}$, típicamente utilizada como salida en problemas de clasificación.
$\alpha, \beta, \gamma, \ldots$	Parámetros.
$A, B, \Gamma, \ldots$	Espacios parámetricos.
$F(x)$	Función de distribución de probabilidad de x.
$p(x)$	Función de densidad de probabilidad de x.
$p(x\|y)$	Función de densidad de probabilidad de x condicionada a y.
$E_x[f]$	Valor esperado de f sobre x.
$E_{x\|y}[f]$	Valor esperado de f sobre x condicionado al valor de y.
$\log(x), \exp(x)$	Logaritmo de x y exponencial de x.
$\hat{y}$	Valor estimado de y
$\|\mathbf{x}\|$	Norma del vector $\mathbf{x}$.
$\mathbf{M}$	Matriz $\mathbf{M}$.
$\mathbf{M}^T$	transpuesta de la matriz $\mathbf{M}$
$\mathbf{M}^{-1}$	Inversa de la matriz $\mathbf{M}$.

Acrónimos y abreviaturas

2D	Bi-dimensional
A1	Astrocytoma grade I

A2 Astrocytoma grade II

AA Aprendizaje Automático

acc accuracy (acierto)

ADA American Diabetes Association

ADEs Efectos adversos de fármacos (Adverse drug event)

AEMPS Agencia Española de Medicamentos y Productos Sanitarios

AGG Aggressive tumor: GBM and MET

Ala Alanine

AMIA American Medical informatics Association

AMX Amoxicilina

ANN Artificial Neural Networks

API Application Programming Interface

AS Ability to Separate

a.u. arbitrary units

AUC Area Under the Curve

AVS Agencia Valenciana de Salud

BER Balanced Error Rate

BAR Balanced Accuracy Rate

BDK Bi-directional Kohonen Networks

BI Business Intelligence

BT Brain Tumor

CADS Computer Aided Diagnosis System

CC.AA. Comunidades Autónomas

cDNA complementary DNA

CDSS Clinical Decision Support Systems

CDVC Clinical Data Validation Committee

Cho Choline

CISNS Consejo Interterritorial del SNS

CIP Ciprofloxacino

CNS Central Nervous System

CPDF Conditional Probability Density Function

CPOE Computerized Physician Order Entry

CQCD Committee for Quality Control of Data

Cr Creatine

CRCT Prueba controlada aleatoria por grupos (clustered RCT)

CS Chemical Shift

CT Computed Tomography

CV Cross Validation

D.G. Dirección General

GAD Grafo Acíclico Dirigido

dDSS distributed Decision-Support System

DICOM Digital Imaging and Communication in Medicine

dLDA Linear Discriminant Analysis with diagonal covariance matrix

DM *Data Mining* (minería de datos)

DMG Data Mining Group

DNA Deoxyribonucleic acid

dQDA Quadratic Discriminant Analysis with diagonal covariance matrix

DSS Decision-Support System

DSSs Decision-Support Systems

DT Decision Trees

EbSS Evidence-based Search Service

ECC Eddy Current Correction

EFQM European Foundation for Quality Management

e.g. exempli gratia (for example)

EHR Historia Clinica Electrónica

EM Expectation-Maximization

EMBTD EM for Binned and Truncated Data

EMBTDr EM for Binned and Truncated Data with related means

EN-13606 Health informatics - Electronic Health Record Communication

err error

eTDB eTUMOUR database

ETL *Extract*, *transform* and *load* (extracción, transformación y carga de datos)

eTUMOUR the eTUMOUR EC project

etxxxx Acronym for a generic patient in the eTUMOUR database

EUROREC European Institute for Health Records

F Friedman's nonparametric two-way analysis of variance test

FDR False Discovery Rate

FDP Función de Distribución de Probabilidad

FE Feature Extraction

FID Free Induction Decay

FFT Fast Fourier Transform

FIPA Foundation of Intelligent Physical Agents

FLDA Fisher's LDA

FN Falso Negativo

FP Falso Positivo

FWHM Full Width at Half Maximum

GBM Glioblastoma

GE General Electric

GlioMET Glial (LGG + GBM) + MET

Gly Glycine

Glx Glutamate/Glutamine

GMOR Geometric Mean of Recalls

GMRP Geometric Mean of Recall and Precision

GUI Graphical User Interface

HEALTHAGENTS the HEALTHAGENTS EC project

HAL HEALTHAGENTS Language

HGG High grade glial

HIS Health Information System

HIV Human Immunodeficiency Virus

HLSVD Hankel-Lanczos Singular Value Decomposition

HR-MAS High-Resolution Magic Angle Spinning

HSVD Hankel Singular Value Decomposition

IA Inteligencia Artificial

IBIME Biomedical Informatics group

ICA Independent Component Analysis

ICT Information and Communication Technology

IDEF Integration Definition for Function Modelling

i.e id est (that is)

IHTSDO International Health Terminology Standards Development Organisation

i.i.d. independent and identically distributed

INTERPRET the INTERPRET project

IOTA International Ovarian Tumor Analysis

IT Independent Test

ITACA Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas

Ixxxx Acronym for a generic patient in the INTERPRET database

IRLS Iterative Reweighted Least Squares

JADE Cardoso's implementation for ICA

jMRUI Java Magnetic Resonance User Interface

kRSTT k-Random Sampling Train-Test

KNN k-Nearest Neighbors

KW Kruskal-Wallis nonparametric one-way analysis of variance

L2-norm Euclidean distance

L1 Lipid resonance at 1.29 *ppm*

L2 Lipid resonance at 0.92 *ppm*

Lac Lactate

LCC Lightweight Coordination Calculus

LDA Linear Discriminant Analysis

LGG Low-Grade Glial

LGS Ley 14/1986, de 25 de abril, General de Sanidad

LOO Leave-One-Out Cross-Validation

LSSVM Least-Squares Support Vector Machines

MAP Estimador Máximo a Posteriori, Maximum a Posteriori Estimation

MBE Medicina Basada en la Evidencia

MCMC Markov Chain Monte Carlo

MCT Tukey's honestly significant difference criterion for multiple comparison

MD Data Mining (Minería de datos)

MDM Cancer Multidisciplinary Meeting

MEN Low-grade meningiomas

MGP Modelo Gráfico Probabilístico

mMEN Meningothelial Meningioma

MET Metastases

MH Metropolis-Hastings

mI myo-Inositol

ML Mobile lipids

MLE Estimador Máximo Verosímil, Maximum Likelihood Estimation

MLP Multilayer Perceptron

MM Macromolecules

MiM Mismatch

MN Multinomial parametric classifier

MR (Nuclear) Magnetic Resonance

MRI Magnetic Resonance Imaging

mRNA messenger Ribonucleic acid

MRS Magnetic Resonance Spectroscopy

MRSI Magnetic Resonance Spectroscopic Imaging

MSE Mean Squared Error

MV Multi-voxel

NAA N-Acetyl Aspartate

NAc N-Acetyl groups

NHS National Health Service

NMR Nuclear Magnetic Resonance

NPV Negative predictive value

OA Oligoastrocytomas

OD Oligodendrogliomas

OLAP *On-line Analytical Processing* (procesamiento analítico en línea)

ONC Office of the National Coordinator for Health Information Technology

OWL Web Ontology Language

PR Pattern Recognition

PC principal component or loading

PCs principal components or loadings

PCA Principal Component Analysis

PCA-KNN K-nearest neighbours and local feature reduced by PCA

PEN Penicilina

PHS Personal Health System

PI Peak integration

PIB Producto Interior Bruto

pKNN K-vecinos probabilístico

PM Perfect Match

PMML Predictive Model Markup Language

PPV Positive predictive value

P Precision

PRESS Point-Resolved Spectroscopic Sequence

PPM Peak height of typical resonances

Q1 Low Grade Meningioma vs. Glioblastoma+Metastasis vs. Low Grade Glial

QC Quality Control

QDA Quadratic Discriminant Analysis

QH1 Hierarchical question 1: Meningioma vs. (Glioma and Metastasis)

QH2 Hierarchical question 2: Low grade glial vs. (Glioblastoma and Metastasis)

R Recall

RB red bayesiana

RBF Radial Basis Function

RCT Prueba controlada aleatoria

RDO Radiological diagnostic orientation

RDF Resource Description Framework

ReliefF ReliefF algorithm for Recursive Elimination of Features

ROC Receiver operating characteristic

RoI Retorno de la inversión

RF Radio frequency

RuleML Rule Markup Language

SAS Servicio Andaluz de Salud

SDM Sistemas de decisión compartida (Shared Decision-Making)

SNOMED CT Systematized Nomenclature of Medicine–Clinical Terms

SNR Signal-to-Noise Ratio

SNS Sistema Nacional de Salud

SNV Standard Normal Variate

STEAM Stimulated Echo Acquisition Mode sequence

SOC Sistema de Orientación Clínica

STT Soft Tissue Tumor

SV Single voxel

SVM Support Vector Machines

SW Stepwise algorithm for feature selection in classification

SWRL A Semantic Web Rule Language

Ta2 T2∗constant

TAM Technology Acceptance Model

TIC Tecnologías de la Información y Comunicaciones

Tau Taurine

TE Echo Time

TR Recycling Time

ULN Unit Length Normalization

VC Validation Committee

VN Verdadero Negativo

VP Verdadero Positivo

WAV Wavelet transform

WHO World Health Organization

XML eXtensible Markup Language

YP Yellow Pages