Universidad Politécnica de Valencia
Departamento de Física Aplicada
Programa de Promoción del Conocimiento

Incremental Learning approaches to Biomedical decision problems

DOCTORAL THESIS

Presented by **Salvador Tortajada Velert**

Supervised by Dr. Juan Miguel García-Gómez
and Dr. Montserrat Robles Viejo

Valencia - Spain
September 6, 2012

Agradecimientos

La finalización de este trabajo me brinda la oportunidad de reconocer a todas aquellas personas que, directa o indirectamente, han evitado que perdiese la cordura durante la realización del mismo, especialmente en su última fase. Me gustaría empezar por agradecer a mis directores de tesis, Montserrat Robles y Juan Miguel García, por la oportunidad que me han brindado y por la confianza que depositan en todo lo que hacemos. Agradecer también a los compañeros doctorandos Elies Fuster y Javier Vicente cuya ayuda mutua ha sido inestimable tanto profesional como anímicamente. Nada de esto sería posible sin el estupendo ambiente creado en el grupo Ibime, donde el trabajo duro, la colaboración, el respeto y la confianza, junto con algunas dosis de buen humor e inspiración han propiciado de un modo u otro el desarrollo de esta tesis. Agradezco por ello a Miguel Esparza, Carlos Sáez, Alfredo Navarro, César Vidal, Adrián Bresó, Javier Juan y Alfonso Pérez, del grupo de minería de datos biomédicos y al resto de compañeros y socios de Sistemas de Información e Imagen Médica.

Esta tesis ha sido desarrollada en el contexto de dos proyectos europeos: eTumour y HealthAgents. Participar en ambos ha sido una experiencia muy fructífera y una gran oportunidad para trabajar y aprender de grandes investigadores: Prof. Bernardo Celda, Prof. Sabine Van Huffel, Dr. Pieter Wesseling, Prof. Carles Arús, Prof. Lutgarde Buydens, Prof. Geert Postma, Prof. Paul Lewis, Prof. Srinandan Dasmahapatra, Dr. Daniel Monleón, Dr. Andrew Peet, Magí Lluch, Dra. Mariola Mier, Dra. Margarida Julià-Sapé, Dra. Ana Paula Candiota, Dr. Franklyn Howe, Dr. Iván Olier, Dr. Jan Luts, Dr. Jean-Baptiste Poullet, Dr. Bo Hu, Dra. Madalina Croitoru, and Dra. Anca Croitor.

Gran parte de esta tesis no habría sido posible sin la colaboración de aquellos pacientes que han contribuido aportando su información y datos biomédicos y cuya generosidad hace posible que podemos avanzar aunque sea con esta humilde tesis. También es necesario agradecer la colaboración de los expertos que han proporcionado los datos al consorcio, en concreto: Jaume Capellades (IDI-Badalona), Carles Majós (IDI-Bellvitge), Àngel Moreno (Centre Diagnòstic Pedralbes), Witold Gajewicz (MUL), Jorge Calvar (FLENI), Antoni Capdevila (H. Sant Joan de Déu), Arend Heerschap (RUNMC), and W. Semmier (DKFZ-Heidelberg).

También siento la necesidad de agradecer a los amigos y a la familia el ánimo y los momentos de ocio que, aunque han sido escasos, no por ello han sido menos necesarios en esta etapa. En especial, a mi hermana Marta y a mis padres, que también me han proporcionado todo lo necesario para poder empezar este trabajo: valores y educación; y a mi tía María del Mar cuyo ejemplo hace unos años sirvió de estímulo catalizador de esta aventura. Por último, quiero agradecer el cariño, la paciencia y el apoyo que me ha regalado María Elena en estos maravillosos años con ella.

Probablemente, todo habría sido más difícil sin vosotros.

Abstract

During the last decade, a new trend in medicine is transforming the nature of healthcare from reactive to proactive. This new paradigm aspires to detect diseases at an early stage and introduce diagnosis to stratify patients and diseases to select the optimal therapy based on individual observations. This paradigm transformation relies on the availability of complex multi-level biomedical data. In order to take advantage of this information, an important effort is being made to develop new mathematical and computational methods for extracting maximum knowledge from patient records. This requirement enables the use of computer-assisted Clinical Decision Support Systems for the management of individual patients.

The Clinical Decision Support System (CDSS) are computational systems that provide precise and specific knowledge for the medical decisions to be adopted for diagnosis, prognosis, treatment and management of patients. The framework and the origin of this Thesis is precisely the development of a CDSS based on Machine Learning algorithms to infer predictive models for non-invasive brain tumour diagnosis. This process began with the European project INTERPRET (2002) and went on with other two European projects eTUMOUR (2005) and HEALTHAGENTS (2008), which have endeavoured to develop an automatic diagnostic tool applied to Proton Magnetic Resonance Spectroscopy (^{1}H MRS) data from brain tumours. A major aim was to minimize the need for an invasive histological diagnosis of a brain tumour biopsy. Machine Learning has been successfully applied to this problem providing automated analysis of ^{1}H MRS. However, the development of brain tumour classifiers able to generalize requires a large number of cases to be acquired for each tumour type and at present the approach has only been used for a few common tumours. Cases are collected from a large number of hospitals over many years and data transferred to a centralised database. This approach has several disadvantages, ethical approval and patient consent needs to be obtained to send and store data. Distributed databases in which classifiers can be trained without moving the data from the hospital at which it was collected would provide a practical solution. The ability to retrain the classifiers as new data accumulates is also an important requirement and to meet these needs, incremental learning algorithms may give a practical optimal solution.

After the analysis of non-incremental ML approaches for automatic brain tumor diagnosis, this Thesis introduces new incremental learning algorithms of general purpose for stationary environments and particularly for adapting the predictive models to new centers in the framework of automatic brain tumour decision making using ^{1}H MRS.

Until now, the different CDSS developed for brain tumour diagnosis have only used non-incremental classification models. Non-incremental classifiers entail an implicit assumption that learning stops when the current training set has been processed. Hence, the performance of a non-incremental automatic classifier strongly depends on the avail-

ability of a representative training set for each class. However, the gathering of these data is often expensive and time-consuming. Under these circumstances the properties of incremental learning algorithms provide an effective solution.

An incremental learning algorithm sequentially produces a new predictive model when new observations are available. The new predictive model is determined by the knowledge held in the previous model and by the information provided by the current data. Therefore, an incremental learning algorithm should be able to learn additional information from new data without completely forgetting its previous knowledge, improving at the same time the performance of the models in the course of time.

The present Thesis introduces the design, development, and evaluation of two new incremental learning algorithms for general purpose dynamic CDSS with an application to brain tumour diagnosis. Unlike many state-of-the-art incremental learning algorithms, we assume that previous data are not accessible at all, which is a common constraint in medical decision problems with distributed databases.

The first incremental learning algorithm is based on a generative weighted combination of maximum-likelihood estimators where the data are assumed to follow a multivariate Gaussian distribution. The algorithm has the ability to learn in an incremental fashion, improving the performance of the models when new information is available, and converging in the course of time. Furthermore, it can incorporate new classes to its knowledge base if new diagnosis are available within the new data allowing the customization of the models at a particular clinical center. An evaluation using five benchmark databases has been used to characterize the behaviour of the algorithm and, finally, it has been applied to automatic brain tumour classification, comparing it with two state-of-the-art incremental algorithms.

The second algorithm is based on a discriminative logistic regression using a Bayesian inference paradigm where the posterior parameter distribution of one iteration is used as a prior parameter distribution for the training of the next model once the new data are available. This algorithm does not make any assumption on the underlying distribution of the data. The performance of the incremental algorithm is demonstrated by employing different benchmark datasets and comparing it to the previous incremental learning algorithm using also the brain tumour database.

Both algorithms show a good behaviour obeying the definition of incremental learning algorithm and achieving the desired properties they should have. The algorithms have the ability to customize an already trained predictive model to the specific distribution of a particular hospital assuming that new information is ready for supervised classification at different times, without the need to access to the previously seen data. This ability to customize a model to a specific clinical centre could be used to improve the behaviour of a state-of-the-art CDSS for aiding brain tumour diagnosis in the future.

Resumen

En los últimos diez años, una nueva tendencia en medicina está trasformando la práctica médica de reactiva a proactiva. Este nuevo paradigma aspira a detectar las enfermedades de forma precoz y usar el diagnóstico con el fin de seleccionar la terapia óptima en base a las observaciones individuales. Este cambio de paradigma depende en gran medida de la disponibilidad de datos biomédicos complejos. Para beneficiarse de esta información se está llevando a cabo un esfuerzo considerable por desarrollar nuevos métodos matemáticos y computacionales que sean capaces de extraer el máximo conocimiento posible de los registros médicos. Este requisito posibilita el uso de Sistemas de Ayuda a la Decisión Médica computerizados para la gestión individual de pacientes.

Los Sistemas de Ayuda a la Decisión Médica (SADM) son sistemas informáticos que proporcionan conocimiento preciso y específico para las decisiones médicas relacionadas con el diagnóstico, pronóstico, tratamiento y gestión de pacientes. El origen de esta Tesis es, precisamente, el desarrollo de un SADM basado en técnicas de Aprendizaje Automático para inferir modelos predictivos para el diagnóstico no invasivo de tumores cerebrales. La idea partió del proyecto europeo INTERPRET (2002) y continuó con otros dos proyectos europeos: eTumor (2005) y HEALTHAGENTS (2008), que llevaron a cabo un gran esfuerzo para desarrollar una herramienta de diagnóstico automático aplicada a datos de espectros de resonancia magnética nuclear (^{1}H MRS) de tumores cerebrales. Uno de los objetivos principales era reducir la necesidad de llevar a cabo un diagnóstico histopatológico invasivo a partir de la biopsia de un tumor cerebral. El Aprendizaje Automático se ha aplicado con éxito a dicho problema, proporcionando un análisis automático del ^{1}H MRS. Sin embargo, el desarrollo de clasificadores de tumores cerebrales capaces de generalizar requiere la adquisición de un gran número de casos para cada tipo de tumor y, hasta ahora, esta aproximación se ha empleado solo para un conjunto reducido de tumores comunes. Los casos se han recogido a lo largo de muchos años y a partir de un conjunto de hospitales y se han transferido a una base de datos centralizada. El problema de esta aproximación es que existen impedimentos éticos y legales para almacenar y enviar los datos. En cambio, las bases de datos distribuidas donde los modelos de clasificación pueden ser entrenados sin tener que mover los datos del hospital donde se adquirieron podrían proporcionar una solución práctica. Otro requisito interesante es la capacidad de reentrenamiento de los clasificadores a medida que se adquieren nuevos datos. una manera de proporcionar una solución práctica óptima que cumpla con ambos requisitos es aplicar algoritmos de aprendizaje incremental.

Tras analizar las aproximaciones de Aprendizaje Automático no incrementales para el diagnóstico automático de tumores cerebrales, esta Tesis presenta dos nuevos algoritmos de aprendizaje incremental de propósito general para entornos estacionarios y, en particular, para adaptar los modelos predictivos a nuevos centros sanitarios en el marco de la toma

de decisiones para tumores cerebrales empleando ^{1}H MRS.

Hasta ahora, los diferentes SADM desarrollados para el diagnóstico de tumores cerebrales habían empleado solo modelos no incrementales. Estos modelos asumen de forma implícita que el aprendizaje termina una vez que el conjunto de datos disponible ha sido procesado, por lo que las prestaciones de un clasificador automático no incremental depende de la disponibilidad de un conjunto de entrenamiento suficientemente representativo para cada clase. El problema reside en el considerable coste económico y temporal que la adquisición de estos datos suele suponer. Las propiedades de los algoritmos de aprendizaje incremental podrían proporcionar una solución efectiva ante esta situación.

Un algoritmo de aprendizaje incremental proporciona, de forma secuencial, un nuevo modelo predictivo siempre que se disponga de nuevas observaciones. Este nuevo modelo queda determinado por el conocimiento adquirido en el modelo anterior y por la información contenida en los nuevos datos. Por lo tanto, un algoritmo incremental debería ser capaz de incorporar información adicional a partir de los nuevos datos sin olvidar por completo su conocimiento previo. A la vez, las prestaciones de los modelos deberían mejorar con el paso del tiempo.

Esta Tesis presenta el diseño, desarrollo y evaluación de dos nuevos algoritmo de aprendizaje incremental de propósito general para SADM dinámicos, con una aplicación concreta al diagnóstico de tumores cerebrales. Al contrario que otros muchos algoritmos incrementales desarrollados, se asume que los datos anteriores no serán accesibles ya que esta es una restricción común en entornos de decisión médicos con bases de datos distribuidas.

El primer algoritmo se basa en una combinación ponderada de estimadores por máxima verosimilitud donde se asume que los datos siguen una distribución gaussiana multivariante. El algoritmo es capaz de aprender de manera incremental, mejorando las prestaciones de los modelos cuando se dispone de nueva información, convergiendo con el tiempo. Además, puede incorporar nuevas clases a su base de conocimiento si estos diagnósticos estuvieran disponibles. Estas capacidades le permiten adaptar los modelos a cada centro clínico particular. Para evaluar el comportamiento del algoritmo se han utilizado bases de datos de referencia y, finalmente, se ha aplicado al problema de clasificación de tumores cerebrales, comparando sus resultados con los de otros dos algoritmos incrementales de la literatura.

El segundo algoritmo se basa en una regresión logística que se ajusta mediante el paradigma de inferencia bayesiana donde la distribución a posteriori de los parámetros de una iteración se usa como distribución de los parámetros a priori para el entrenamiento del modelo siguiente. Este algoritmo no asume ningún tipo de distribución subyacente a los datos. Las prestaciones de este algoritmo se evalúan mediante diferentes bases de datos de referencia y comparándolo con el algoritmo anterior empleando también la base de datos de tumores cerebrales.

Ambos algoritmos muestran un buen comportamiento, cumplen con la definición de aprendizaje incremental y logran alcanzar las propiedades que deben tener. Ambos algoritmos pueden adaptar un modelo entrenado con datos de un hospital a la distribución específica de otro hospital siempre que se disponga de nueva información para poder llevar a cabo un entrenamiento supervisado. Además, pueden hacer esto sin tener que acceder a los datos anteriores. Esta capacidad de adaptación a un nuevo centro clínico podría emplearse en el futuro para mejorar el comportamiento de los SADM actuales en la ayuda al diagnóstico de tumores cerebrales.

Resum

En els últims deu anys, una nova tendència en medicina està transformant la pràctica mèdica de reactiva a proactiva. Aquest nou paradigma aspira a detectar les malalties de forma precoç i usar el diagnòstic amb la finalitat de seleccionar la teràpia òptima sobre la base de les observacions individuals. Aquest canvi de paradigma depèn en gran mesura de la disponibilitat de dades biomèdiques complexos. Per beneficiar-se d'aquesta informació s'està duent a terme un esforç considerable per desenvolupar nous mètodes matemàtics i computacionals que siguin capaços d'extreure el màxim coneixement possible dels registres mèdics. Aquest requisit possibilita l'ús de Sistemes d'Ajuda a la Decisió Mèdica computeritzats per a la gestió individual de pacients.

Els Sistemes d'Ajuda a la Decisió Mèdica (SADM) són sistemes informàtics que proporcionen coneixement precís i específic per a les decisions mèdiques relacionades amb el diagnòstic, pronòstic, tractament i gestió de pacients. L'origen d'aquesta Tesi és, precisament, el desenvolupament d'un SADM basat en tècniques d'Aprenentatge Automàtic per inferir models predictius per al diagnòstic no invasiu de tumors cerebrals. La idea va partir del projecte europeu INTERPRET (2002) i va continuar amb altres dos projectes europeus: eTumor (2005) i HEALTHAGENTS (2008), que van dur a terme un gran esforç per desenvolupar una eina de diagnòstic automàtic aplicada a dades d'espectres de ressonància magnètica nuclear (^{1}H MRS) de tumors cerebrals. Un dels objectius principals era reduir la necessitat de dur a terme un diagnòstic histopatológic invasiu a partir de la biòpsia d'un tumor cerebral. L'Aprenentatge Automàtic s'ha aplicat amb èxit a aquest problema, proporcionant una anàlisi automàtica del ^{1}H MRS. No obstant això, el desenvolupament de classificadors de tumors cerebrals capaços de generalitzar requereix l'adquisició d'un gran nombre de casos per a cada tipus de tumor i, fins ara, aquesta aproximació s'ha emprat solament per a un conjunt reduït de tumors comuns. Els casos s'han recollit al llarg de molts anys i a partir d'un conjunt d'hospitals i s'han transferit a una base de dades centralitzada. El problema d'aquesta aproximació és que existeixen impediments ètics i legals per emmagatzemar i enviar les dades. En canvi, les bases de dades distribuïdes on els models de classificació poden ser entrenats sense moure les dades de l'hospital on es van adquirir podrien proporcionar una solució pràctica. Un altre requisit interessant és la capacitat de reentrenament dels classificadors a mesura que s'adquireixen noves dades. Una manera de proporcionar una solució pràctica òptima que compleixi amb tots dos requisits és aplicar algorismes d'aprenentatge incremental.

Després d'analitzar les aproximacions d'Aprenentatge Automàtic no incrementals per al diagnòstic automàtic de tumors cerebrals, aquesta Tesi presenta dos nous algorismes d'aprenentatge incremental de propòsit general per a entorns estacionaris i, en particular, per adaptar els models predictius a nous centres sanitaris en el marc de la presa de decisions per a tumors cerebrals emprant ^{1}H MRS.

Fins ara, els diferents SADM desenvolupats per al diagnòstic de tumors cerebrals havien empleat sol models no incrementals. Aquests models assumeixen de forma implícita que l'aprenentatge acaba una vegada que el conjunt de dades disponible ha estat processat, per la qual cosa les prestacions d'un classificador automàtic no incremental depèn de la disponibilitat d'un conjunt d'entrenament suficientment representatiu per a cada classe. El problema resideix en el considerable cost econòmic i temporal que l'adquisició d'aquestes dades sol suposar. Les propietats dels algorismes d'aprenentatge incremental podrien proporcionar una solució efectiva davant aquesta situació.

Un algorisme d'aprenentatge incremental proporciona, de forma seqüencial, un nou model predictiu sempre que es disposi de noves observacions. Aquest nou model queda determinat pel coneixement adquirit en el model anterior i per la informació continguda en les noves dades. Per tant, un algorisme incremental hauria de ser capaç d'incorporar informació addicional a partir de les noves dades sense oblidar per complet el seu coneixement previ. Alhora, les prestacions dels models haurien de millorar amb el pas del temps.

Aquesta Tesi presenta el disseny, desenvolupament i avaluació de dos nous algorisme d'aprenentatge incremental de propòsit general per SADM dinàmics, amb una aplicació concreta al diagnòstic de tumors cerebrals. Al contrari que molts altres algorismes incrementals desenvolupats, s'assumeix que les dades anteriors no seran accessibles ja que aquesta és una restricció comuna en entorns de decisió mèdics amb bases de dades distribuïdes.

El primer algorisme es basa en una combinació ponderada d'estimadors per màxima versemblança on s'assumeix que les dades segueixen una distribució gaussiana multivariant. L'algorisme és capaç d'aprendre de manera incremental, millorant les prestacions dels models quan es disposa de nova informació, convergint amb el temps. A més, pot incorporar noves classes a la seva base de coneixement si aquests diagnòstics estiguessin disponibles. Aquestes capacitats li permeten adaptar els models a cada centre clínic particular. Per avaluar el comportament de l'algorisme s'han utilitzat bases de dades de referència i, finalment, s'ha aplicat al problema de classificació de tumors cerebrals, comparant els seus resultats amb els de altres dos algorismes incrementals de la literatura.

El segon algorisme es basa en una regressió logística que s'ajusta mitjançant el paradigma d'inferència bayesiana on la distribució a posteriori dels paràmetres d'una iteració s'usa com a distribució dels paràmetres a priori per a l'entrenament del model següent. Aquest algorisme no assumeix cap tipus de distribució subjacent a les dades. Les prestacions d'aquest algorisme s'avaluen mitjançant diferents bases de dades de referència i comparant-ho amb l'algorisme anterior emprant també la base de dades de tumors cerebrals.

Tots dos algorismes mostren un bon comportament, ja que compleixen amb la definició d'aprenentatge incremental i aconsegueixen aconseguir les propietats que han de tenir. Tots dos algorismes poden adaptar un model entrenat amb dades d'un hospital a la distribució específica d'un altre hospital sempre que es disposi de nova informació per poder dur a terme un entrenament supervisat. A més, poden fer això sense haver d'accedir a les dades anteriors. Aquesta capacitat d'adaptació a un nou centre clínic podria emprar-se en el futur per millorar el comportament dels SADM actuals en l'ajuda al diagnòstic de tumors cerebrals.

Contents

Chapter 1

Introduction

1.1 Motivation

During the last decade, a new trend in medicine is transforming the nature of healthcare from reactive to proactive. This new paradigm is changing into a personalized medicine where the prevention, diagnosis, and treatment of disease is focused on individual patients. Hence, its objective is to evolve from a classical reactive medicine, which waits until the patient is sick before reacting, to a personalized, predictive, preventive, and participatory medicine which aims to be cost effective and increasingly focused on wellness. This paradigm is known as P4 medicine [1, 2]. Among other key benefits, P4 medicine aspires to detect diseases at an early stage and introduce diagnosis to stratify patients and diseases to select the optimal therapy based on individual observations and taking into account the patient outcomes to empower the physician, the patient, and their communication.

This transformation relies on the availability of complex multi-level biomedical data that are increasingly accurate, since it is possible to find exactly the needed information, but also exponentially noisy, since the access to that information is more and more challenging. In order to take advantage of this information, an important effort is being made in the last decades to digitalize medical records and to develop new mathematical and computational methods for extracting maximum knowledge from patient records, building dynamic and disease-predictive models from massive amounts of integrated clinical and biomedical data. This requirement enables the use of computer-assisted Clinical Decision Support Systems for the management of individual patients.

The Clinical Decision Support System (CDSS) are computational systems that provide precise and specific knowledge for the medical decisions to be adopted for diagnosis, prognosis, treatment and management of patients. The CDSS are highly related to the concept of evidence-based medicine [3, 4] since they infer medical knowledge from the biomedical databases and the acquisition protocols that are used for the development of the systems, give computational support based on evidence for the clinical practice, and evaluate the performance and the added value of the solution for each specific medical problem. The CDSS have been cataloged into four categories, depending on the complexity of the operations performed to extract knowledge from the patient information [5]. The two most complex types of CDSS are related with the use of artificial intelligence in medicine, specifically with the use of deductive inference reasoning (type III) and inductive inference reasoning (type IV). The present Dissertation is in line with the development of CDSS of

type IV, which are based on Machine Learning (ML) algorithms to infer predictive models from real-world data.

The first CDSS used in clinical practice was developed by Leaper *et al.* [6] for the support of diagnosis and surgery of acute abdominal pain based on a naive Bayes approach. Shortliffe *et al.* designed and developed a ruled-based expert system for assisting physicians with the diagnosis and treatment for certain blood infections [7]. Since then, an increasing emergence of specific CDSS have been designed and developed with relative success. The framework and the origin of this Thesis is precisely the development of a CDSS for non-invasive brain tumour diagnosis that began with the European project INTERPRET (2002) [8] and went on with other two European projects eTUMOUR (2005) [9] and HEALTHAGENTS (2008) [10].

The three European projects (INTERPRET (2000-2002) [8, 11], eTUMOUR (2004-2009) [9], and HEALTHAGENTS (2005-2008) [10]) have endeavoured to develop a non-invasive automatic diagnostic tool using ML techniques applied to Proton Magnetic Resonance Spectroscopy (^{1}H MRS) data from brain tumours. A major aim was to minimize the need for an invasive histological diagnosis of a brain tumour biopsy as is currently required for the diagnosis and management of brain tumours. Non-invasive brain tumour diagnosis using ^{1}H MRS has shown considerable promise in aiding patient management but is not in widespread clinical use due mainly to the difficulties of data interpretation. ML has been successfully applied to this problem providing automated analysis of ^{1}H MRS [12, 13, 11, 14]. However, the development of robust brain tumour classifiers requires a large number of cases to be acquired for each tumour type and at present the approach has only been used for a few common tumours. Cases are accrued from a large number of hospitals over many years and data transferred to a centralised database. This approach has several disadvantages, ethical approval and patient consent needs to be obtained to send and store data. In order to expand the applicability of ML techniques to MRS of a wider range of tumours, more cases need to be collected over a more prolonged period of time and the logistics of using a centralised database to provide this have so far proved insurmountable. Distributed databases where the data is held at the data collecting hospitals have major advantages [10] and such a system in which classifiers can be trained without moving the data from the hospital at which it was collected would provide a practical solution. The ability to retrain the classifiers as new data accumulates is also an important requirement and to meet these needs, incremental learning algorithms may give a practical optimal solution.

Furthermore, a classification framework with a distributed architecture requires the classification models trained with data from one center to perform well when moved to another center, that is, to *generalize*. A model's performance can be assessed using new data from the same dataset following a hold-out evaluation strategy, but a further assessment of generalization requires evaluation on data from elsewhere [15]. Poor performance in new patients may arise because of deficiencies in the design of the model reaching overfitting, or because the setting of patients between the training and the new samples are different, considering factors such as healthcare systems, patient characteristics, and/or acquisition protocols. As an alternative to the re-calibration of the models [16, 17, 18], this Thesis proposes the use of incremental learning algorithms for the models to adapt to the center where they are going to be used.

After the analysis of non-incremental ML approaches, this Thesis introduces new in-

cremental learning algorithms of general purpose for stationary environments and –in particular– for adapting the predictive models to new centers in the framework of biomedical decision making, applying them to the automatic brain tumour classification using ^{1}H MRS.

Until now, the different CDSS developed for automatic brain tumour diagnosis have only used non-incremental classification models [11, 9, 10]. Non-incremental classifiers entail an implicit assumption that learning stops when the current training set has been processed. Hence, the performance of a non-incremental automatic classifier strongly depends on the availability of a representative training set for each class. However, the gathering of these data is often expensive and time-consuming, and a strategy to wait long enough as to gather enough data all in one set may be undesirable and/or impractical. Furthermore, there are situations where the access to previous data may be forbidden. There are also types of data sources where the underlying distributions may evolve over time rather than be stationary. In particular, this happens in the *concept shift* [19, 20, 21], where the target distribution $p(c|x)$ may change in the course of time, and in the *covariate shift* [22, 20, 21], where the data distribution $p(x)$ changes continously. Under these circumstances an incremental learning algorithm might be a practical and more effective solution.

Following previous state-of-the-art research on incremental learning, the proposed algorithms in this Thesis have to reach a trade-off between a stable classifier and a plastic classifier to face the stability/plasticity dilemma, which states that some information may be lost during incremental learning. Furthermore, the algorithms have to prove that they do not suffer from ordering effect bias, which means that presenting the observations in different order implies achieving different models. Finally, it is interesting if the incremental learning algorithms have also the ability to incorporate new classes or concepts as they appear. These remarkable properties are considered through the Thesis assuming the accessing previous data is forbidden.

1.2 Hypothesis

The present Thesis is based on three hypotheses:

I. The ML approach to automatic brain tumour classification yields predictive models with satisfactory performances.

II. The use of incremental learning techniques applied to biomedical data can help in the improvement of automatic classification models when they are moved to a new health organization and/or new cases are available for classification.

III. The Bayesian inference provides a straightforward framework to implement incremental learning algorithms to avoid assumptions over the data distribution.

1.3 Goals

In order to verify the first hypothesis, we need to guarantee that the ML paradigm is feasible for the automatic brain tumour classification task. To reach this goal, an inference

of predictive models to discriminate among different diagnoses, and the evaluation of the models using newly collected data are carried out.

The verification of the last two hypothesis defines the main goal of this Thesis: the design, development and validation of both a frequentist and a Bayesian incremental learning algorithms and their ability to take advantage of new observations. This goal is achieved by fulfilling the following specific goals:

- To design the mathematical framework of a maximum-likelihood approach for an incremental learning algorithm. Then, implement the algorithm and carry out several benchmark experiments, and evaluate the algorithm with a real brain tumour database. This evaluation required the design of an alternative evaluation procedure since the well-known hold-out and resampling methods were not indicated for incremental problems.

- To design the mathematical framework of a Bayesian incremental learning algorithm, to develop the algorithm and to evaluate it using benchmark databases as well as the brain tumour database.

- To compare the performance and the incremental ability of the algorithms with non-incremental algorithms and other state-of-the-art incremental learning algorithms.

1.4 Contributions

The scientific results of this Thesis concern the application of ML techniques and the development of new incremental learning algorithms for approximately stationary environments. These algorithms have a general ML purpose, but they are also applied for automatic brain tumour classification in this Thesis. The contributions of this Thesis have been published in scientific journals and proceedings of congresses in the fields of Applied Artificial Intelligence, Machine Learning, and Magnetic Resonance.

This Thesis continues the research initiated by Dr. Juan Miguel García-Gómez [23] in the framework of two European projects where the goal was to develop a ML-based CDSS for aiding brain tumour diagnosis. The contributions made included the development of ML-based models in a prospective multicenter evaluation [24] and the analysis of the effect of combining different MRS times of echo for the automatic classification of brain tumours [25]. Then, the development of a CDSS for automatic brain tumour diagnosis was carried out and published in [26, 27].

The main contribution of this Thesis is the development of two algorithms for incremental learning of ML-based models for stationary distributions or limited shifting distributions. Unlike many of the state-of-the-art incremental learning algorithms, we assume that, once the datasets are used for fitting the models, they will not be available again. The first incremental learning algorithm is based on maximum-likelihood parameter estimation and a weighted combination of the old model parameters with the new dataset to develop a new model [28]. The second incremental learning algorithm is based on the Bayesian inference paradigm. This paradigm assumes that the parameters are random variables. Taking into account this assumption, the Bayesian paradigm uses the information given by the observations together with prior beliefs about the parameters of the model to estimate the distribution of the parameters. This paradigm allows a new

model to use its parameters as the prior belief of a new model in light of new observed data. This Thesis shows an implementation of this paradigm in a discriminative logistic regression model to turn it into an incremental learning algorithm.

Additional contributions of this Thesis have been made to two other lines of research on brain tumour diagnosis leadered by Elies Fuster-Garcia, regarding the study of compatibility of PR models trained with 1.5T MRS with samples of 3T MRS [29], and Javier Vicente, regarding the design of paediatric brain tumour classifiers [30] and the design of a tool that allows the selection of the most suitable model in a CDSS [31], respectively. Collaboration on the mathematical framework, the writing and experimental design was carried out in these studies.

1.5 Projects and Partners

The development of this Thesis is part of a number of European projects in which the author has been actively involved:

eTUMOUR [9] *Web accesible (Nuclear) Magnetic Resonance (MR) decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data.* European Commission (VI Framework Program, LSHC-CT-2004-503094, 2004-2009).

Objectives: (1) Development of a web-accessible CDSS that has a Graphical User Interface (GUI) to display clinical, metabolomic and genetic brain tumor data. (2) To provide an evidence-based clinical decision-making computer-human interface by using statistical pattern recognition analysis of molecular images of brain tumours (using Magnetic Resonance Spectroscopy (MRS)) and incorporating new criteria such as genetic based tumour classifications and related clinical information.

Partners: University of Valencia (Valencia, Spain), Universitat Autònoma de Barcelona (Barcelona, Spain), St George's Hospital Medical School (London, UK), University Medical Center Nijmegen (Nijmegen, Netherlands), Stichting Katholieke Universiteit (Nijmegen, Netherlands), Université Joseph Fourier U594 (Grenoble, France), MicroArt S.L. (Barcelona, Spain), Hospital San Joan de Deu (Esplugues de Llobregat, Spain), Pharma Quality Europe, s.r.l. (Barcelona, Spain), Hyperphar Group SpA. (Milan, Italy), Katholieke Universiteit Leuven (Leuven, Belgium), Siemens AG, Medical Solutions (Erlangen, Germany), SCITO, S.A (Grenoble, France), Deutsche Krebsforschungs zentrum Heidelberg (Heidelberg, Germany), Bruker Biospin SA. (Wissembourg, France), Institute of Child Health - University of Birmingham (Birmingham, United Kingdom), INSERM U318 (Grenoble, France), Fundación para la Lucha contra Enfermedades Neurológicas de la Infancia (Buenos Aires, Argentina), Medical University Lodz (Lodz, Poland) and IBIME-ITACA group from Polytechnic University of Valencia (Valencia, Spain).

HEALTHAGENTS [10] *Agent-based distributed decision support system for brain tumour diagnosis and prognosis.* European Commission (VI Framework Program, IST-2004-27214, 2006-2009).

Objectives: To create a distributed datawarehouse with the world's largest network of interconnected databases of clinical, histological, and molecular phenotype data of brain tumour patients, providing evidence-based clinical decision-making by means of magnetic

resonance and genetic based tumour classifications, and to develop new methodologies to fulfill a dynamic clinical decision support system.

Partners: University of Valencia (Valencia, Spain), MicroArt S.L. (Barcelona, Spain), Universitat Autònoma de Barcelona (Barcelona, Spain), Pharma Quality Europe, s.r.l. (Barcelona, Spain), Katholieke Universiteit Leuven (Leuven, Belgium), University of Birmingham (Birmingham, UK), University of Edinburgh (Edinburg, UK), University of Southampton (Southampton, UK), and IBIME-ITACA group from Polytechnic University of Valencia (Valencia, Spain)

Furthermore, the author of this Thesis is a co-founding member of a spin-off company, *Veratech for Health*, that has emerged as a result of the knowledge and technologies developed over more than 12 years of experience in the research lines of the Biomedical Informatics Group (IBIME) from the ITACA Institute of the Polytechnic University of Valencia (UPV). Veratech for Health provides solutions for building large health information systems, integration and standardization of clinical information, analysis and knowledge extraction from biomedical data and development of clinical decision support systems. Some results of this Thesis have been already incorporated to develop technological products and some others will be useful to develop new ones in the near future.

1.6 Summary of the remaining chapters

Chapter 2 introduces a theoretical framework of the concepts that are used in this Thesis. It gives an introduction of the ML approach, and highlights the difference between incremental and non-incremental learning algorithms. In addition, different paradigms for fitting model parameters, such as maximum-likelihood and the Bayesian inference paradigm are explained.

Chapter 3 presents some initial results on classification of brain tumours using ^{1}H MRS from a multicenter European database of patients. The results obtained in Chapter 3 are the starting point to further investigate the development of incremental learning algorithms.

Chapter 4 introduces a new incremental algorithm for Gaussian Discriminant Analysis based on the weighted combination of the parameters of one model and the estimated parameters given the new observations. The weights to combine them are based on an unbiased estimator for combining different measures developed by Graybill and Deal [32]. The results show that the algorithm is able to learn from new data with a converging performance, it can include new classes if needed, and it has a negligible order effect. The algorithm is tested with some benchmark datasets and, finally, with the ^{1}H MRS brain tumour database.

Chapter 5 applies the Bayesian inference paradigm to develop an incremental learning algorithm to build a discriminative model. Since the analytical computation of the posterior probabilities are hard to obtain, a Laplace approximation is used. The algorithm follows two steps. First, the posterior probability of the parameters is approximated to a Gaussian. Then, the posterior probability is used as a prior probability that is combined with the likelihood when new observations are available to produce a new model. The results shows that the algorithm is able to learn

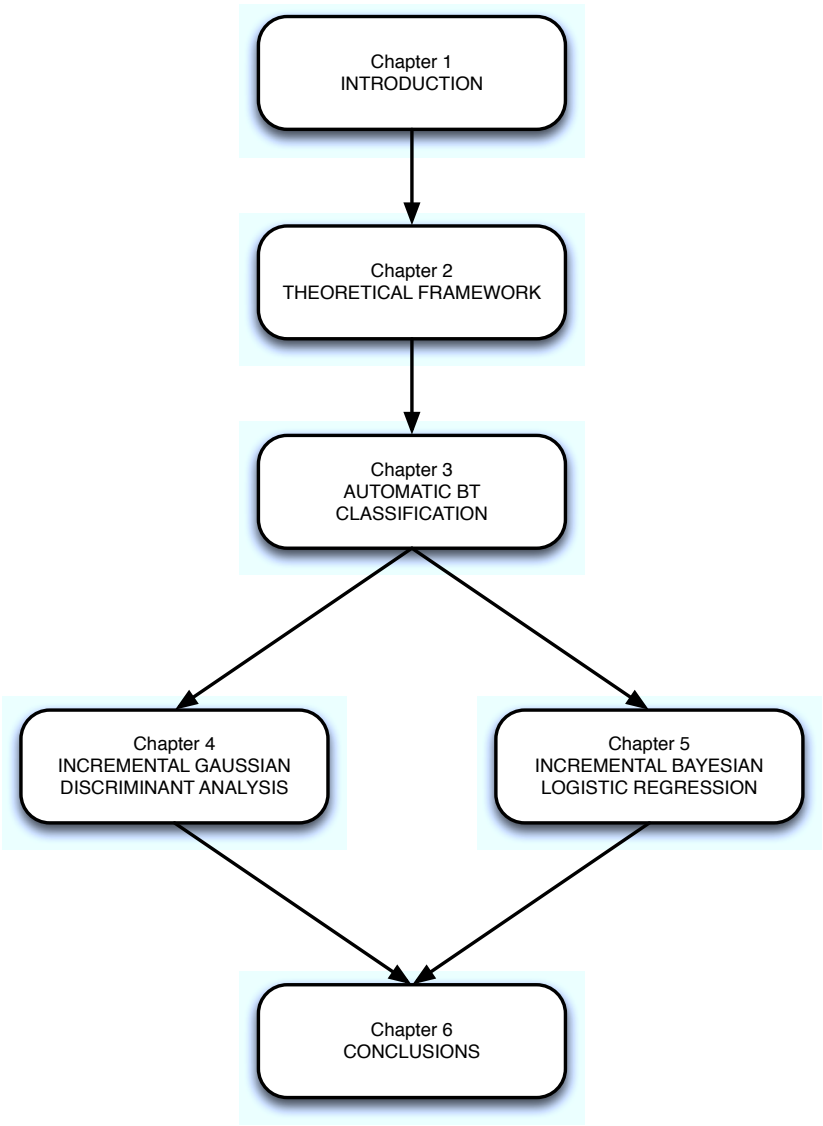


Figure 1.1: Scheme of the Chapters of the Thesis.

incrementally and has no order effects. As a result, an incremental learning algorithm that is independent of the data distribution is provided.

Chapter 6 summarizes the conclusions of the Dissertation and explains the future lines of research and development.

Chapter 2

Theoretical framework

This Chapter introduces the theoretical foundations that are the basis of this Thesis. We begin with a section about the machine learning and pattern recognition approaches to the automatic classification. Then, different modeling approaches for classification and the maximum-likelihood estimation methodology and the Bayesian inference paradigm are introduced as a way to estimate the parameters of a classification model. Then, the core concepts of incremental learning are presented. Finally, the evaluation methodologies followed in this Thesis to test the models are described.

2.1 Machine learning and Pattern recognition

A medical diagnosis is a cognitive process by which a clinician attempts to identify a health disorder or a disease in a patient. The diagnosis is based on a series of data sources that serve as the input information to yield the final result. Therefore, this process can be regarded as a classification problem. Since the field of Pattern Recognition (PR) concerns the automatic discovery of regularities in data to classify them into different categories, the problem of medical diagnosis can be automated by means of PR techniques. The PR models are often developed based on a Machine Learning (ML) approach, which provides the mathematical and computational mechanisms to infer knowledge from specific data of a given domain [33, 34].

The life cycle of a PR problem based on ML can be divided into two main phases: the training phase and the recognition phase (see Figure 2.1). During the training phase, a dataset is used to build the PR model. In this phase, a pre-processing and a feature selection or a feature extraction can be established. Then, an adaptive model is fitted, selected and evaluated in order to obtain the best generalization for solving new cases in the recognition phase. Once the model is ready, it can be incorporated into a CDSS to aid in future observations.

The general problem of ML is often described with a random observation $\mathbf{s}$ generating process, which are obtained following a two-stage process [35]. First, a generator produces random feature vectors $\mathbf{x} \in \mathcal{X}^D$ following a probability distribution function $p(\mathbf{x})$; then, a supervisor assigns the class $c \in \mathcal{C}$ given the feature vector and following the conditional probability distribution function $p(c|\mathbf{x})$, thus producing samples such as $\mathbf{s}_i = \{\mathbf{x}_i, c_i\}$ with probability $p(\mathbf{x}_i, c_i)$, where $p(\mathbf{x}_i, c_i) = p(\mathbf{x}_i)p(c_i|\mathbf{x}_i) = p(c_i)p(\mathbf{x}_i|c_i)$. The value $p(c)$ is known as the *prior* probability of the class and the value $p(\mathbf{x}|c)$ is called the *class-*

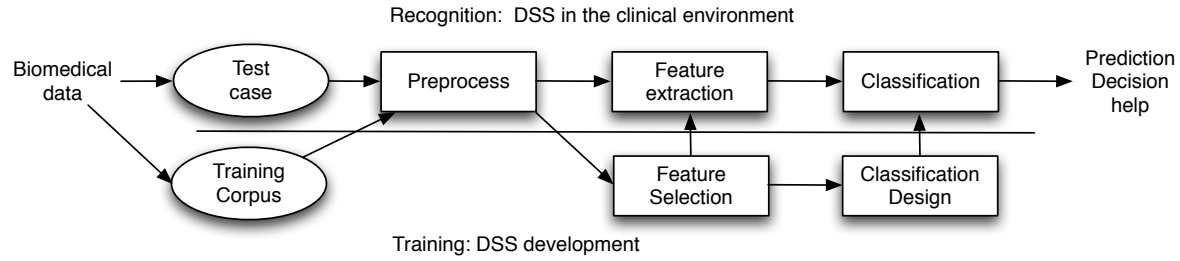


Figure 2.1: Methodology of the Machine Learning approach

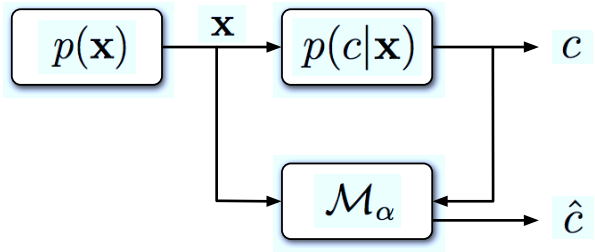


Figure 2.2: The two-stage generation of learning samples produces an observation $\mathbf{x}$ following the distribution $p(\mathbf{x})$, and then a class c is assigned to it. The supervised training of a model is carried out by observing the pairs $(\mathbf{x}, c)$ and fitting the parameters α of the model. Once the model is trained, it has to estimate the output $\hat{c}$, as close to the true value c as possible, given any new input observation $\mathbf{x}$.

conditional probability or simply the conditional probability.

The aim of ML is to develop a *decision rule*, also called a model, $\mathcal{M}$ that maps a random feature vector into a class. Hence, a model is a mapping

$$\mathcal{M} : \mathcal{X}^D \rightarrow \mathcal{C} \tag{2.1}$$

Generally, the model can be defined as a parameterized function $\hat{c} = f_{\mathcal{M}}(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$ that attempts to approximate the value c. It is thus possible to measure the consequences of approximating $\hat{c}$ given $\mathbf{x}$ by means of a *loss function* $L(c, \hat{c})$. In a classification problem, the loss function is often defined as

$$L(c, \hat{c}) = \begin{cases} 0 & \text{if } \hat{c} = c \\ 1 & \text{if } \hat{c} \neq c \end{cases} \tag{2.2}$$

The expression (2.2) is known as 0-1 loss function. When an object $\mathbf{x}$ is classified using a model $\mathcal{M}_\alpha$ and given a 0-1 loss function, it is possible to compute the *conditional risk* of such model, which is expressed as

$$R(\hat{c}|\mathbf{x}) = E_{\hat{c}|\mathbf{x}}[L(c, f_{\mathcal{M}}(\mathbf{x}; \alpha))] \tag{2.3}$$

since the set of classes is a finite set with discrete values, the equation (2.3) can be expressed as

$$R(\hat{c}|\mathbf{x}) = \sum_{c \in \mathcal{C}} L(c, f_{\mathcal{M}}(\mathbf{x}; \alpha)) p(c|\mathbf{x}) \tag{2.4}$$

In order to compare the performance of different classifiers independently of any specific observation $\mathbf{x}$ a *functional risk* can be defined and expressed as

$$\begin{aligned} R(\mathcal{M}_\alpha) &= E_{\mathbf{x}}[R(\hat{c}|\mathbf{x})] \\ &= \int R(\hat{c}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int \sum_{c\in\mathcal{C}} L(c,\hat{c})p(c|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int \sum_{c\in\mathcal{C}} L(c,\hat{c})p(\mathbf{x},c)d\mathbf{x} \\ &= E_{\mathbf{x},c}[L(c,\hat{c})] \end{aligned} \tag{2.5}$$

An optimal decision rule is defined as the one that achieves minimum probability of error. If the prior probabilities and the class-conditional probabilities are known, then the optimal decision rule is the *Bayesian decision rule*,

$$\hat{c}^* \leftarrow \arg\min_{\hat{c}\in\mathcal{C}} R(\hat{c}|\mathbf{x}) \tag{2.6}$$

When a 0-1 loss function is assumed, then the conditional risk is the average probability of error, which can be expressed as

$$R(\hat{c}|\mathbf{x}) = 1 - p(\hat{c}|\mathbf{x}) \tag{2.7}$$

However, the common situation for these distributions is to be unknown. Nevertheless, they can be approximated from a set of observations $\mathcal{S} = \{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_N, c_N)\} \in (\mathcal{X}^D \times \mathcal{C})$ that is supposedly drawn according to $p(\mathbf{x}, c)$. The basic assumption in ML is that both the observed and the unseen data are generated by the same process, which formally means that the data is sampled independently and from an identical probability distribution (*iid*). The main goal of ML is to find –based on $\mathcal{S}$– a function or a model $\mathcal{M}_\alpha$ whose risk is as close to $R(\hat{c}^*)$ as possible. Since the model $f_{\mathcal{M}}(\mathbf{x};\alpha)$ is approximated with a supervised training set $\mathcal{S}$, the estimation of the performance of the model is usually measured using an *empirical risk*:

$$R_{emp}(\mathcal{M}_\alpha) = \frac{1}{N}\sum_{n=1}^{N} L(c_n, f_{\mathcal{M}_\alpha}(\mathbf{x}_n, \alpha)) \tag{2.8}$$

When the 0-1 loss function is assumed, the empirical risk gives

$$R_{emp}(\mathcal{M}_\alpha) = \frac{1}{N}\sum_{n=1}^{N} \delta(c_n, \hat{c}_n) \tag{2.9}$$

where δ is the Kronecker delta function. Hence, the empirical risk using a 0-1 loss function simply counts the misclassifications on a set of observations. The inductive criterion to apply is naturally to select the model with the minimal empirical risk since it should arise as the one which minimizes the probability of error. This empirical risk minimization principle can be expressed as

$$\alpha^* \leftarrow \arg\min_{\alpha\in\Lambda} R_{emp}(\mathcal{M}_\alpha) \tag{2.10}$$

The Bayesian decision rule to minimize risk calls for selecting the class that minimizes the conditional risk. To minimize thus the conditional risk we should select the class that maximizes the posterior probability $p(\hat{c}|\mathbf{x})$. Therefore, the equation (2.6) is transformed into

$$\hat{c}^* \leftarrow \arg\max_{\hat{c}\in\mathcal{C}} p(\hat{c}|\mathbf{x}) \tag{2.11}$$

From a geometrical viewpoint, a decision rule labels an observation $\mathbf{x}$ in the sample space with a class c. As a consequence, the sample space $\mathcal{X}$ is divided into $|\mathcal{C}|$ disjoint *decision regions*, $\mathcal{R}_c$. Thus, a decision region can be defined as

$$\mathcal{R}_c = \{\mathbf{x} : p(c|\mathbf{x}) \geq p(c'(\mathbf{x})\}, \forall c' \neq c \tag{2.12}$$

The surface where the decision regions intersect is called a *decision boundary*. Finding the optimal decision rule using equation (2.11) gives the optimal decision boundary among classes, and it will be determined by the set of points where the class-posterior probabilities are equal,

$$\mathcal{F}^*_{c,c'} = \{\mathbf{x} : p(c|\mathbf{x}) = p(c'|\mathbf{x})\} \tag{2.13}$$

2.1.1 Generative and discriminative models

The equation (2.11) specifies that finding out the optimal class requires knowing the posterior probability of each possible class in order to choose the one that is maximum. If $p(c|\mathbf{x})$ were known for each class $c \in \mathcal{C}$ then the best possible classifier with minimum error rate –called the *Bayes error rate*– could be achieved. Since the posterior probabilities have to be approximated, it is possible to apply at least two different approaches for modeling them: a generative model, or a discriminative model.

Generative models

In the generative approach, the posterior probabilities $p(c|\mathbf{x})$ are computed applying the Bayes' theorem

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \tag{2.14}$$

where $p(\mathbf{x}|c)$ is the class-conditional probability density, $p(c)$ is the prior probability of each class, and $p(\mathbf{x})$ is the marginal distribution that serves as a normalization factor

$$p(\mathbf{x}) = \sum_{c'\in\mathcal{C}} p(\mathbf{x}|c')p(c') \tag{2.15}$$

In this approach, since the marginal distribution $p(\mathbf{x})$ is constant regardless of a particular class, the decision rule (2.11) can be expressed as

$$\begin{aligned} \hat{c}^* &\leftarrow \arg\max_{\hat{c}\in\mathcal{C}} \left\{ \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \right\} \\ &\leftarrow \arg\max_{\hat{c}\in\mathcal{C}} \Big\{ p(\mathbf{x}|c)p(c) \Big\} \end{aligned} \tag{2.16}$$

The final model depends on the assumptions about the class-conditional probability densities. A common assumption is that the variables follow a multivariate Gaussian distribution, that is, $p(\mathbf{x}|c) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$,

$$p(\mathbf{x}|c) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}_c|^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c) \right\} \tag{2.17}$$

with $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ being the mean and covariance matrix parameters for class c, respectively. This assumption is followed in the incremental learning algorithm introduced in Chapter 4.

Discriminative models

A discriminative approach attempts to model the posterior probabilities $p(c|\mathbf{x})$ directly by means of a parametric model that is optimized using a training set. This approach has typically fewer adaptive parameters to be determined. An interesting model for this Dissertation is the Logistic Regression model [36], where the posterior probability of class $c = 1$ is estimated with

$$p(c = 1|\mathbf{x}) = \frac{\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)}{1 + \exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)} \tag{2.18}$$

where $\mathbf{w}$ are the parameters of a discriminant function and $\phi(\cdot)$ is a vector of basis functions that results in a nonlinear transformation of the input data $\mathbf{x}$. This model is only useful for the two-class discrimination problem. The extension for discriminating more than two classes is accomplished using the multinomial logistic regression [36].

2.2 Maximum likelihood estimation

The frequentist approach to machine learning builds a model by finding the parameters that best fits the data. This is carried out by using the *maximum likelihood* estimation (MLE) method. This method has become widely used due to its interesting properties. The MLE selects a configuration of the parameters that maximizes the probability of the data given the model, which is equivalent to the likelihood of the parameters given the data. Let $\mathcal{S} = \{x_1, x_2, \ldots, x_N\}$ be the observed data, then the likelihood function is

$$\begin{aligned} \ell(\theta|\mathcal{S}) &= p(\mathcal{S}|\theta) \\ &= \prod_{n=1}^{N} p(x_n|\theta). \end{aligned} \tag{2.19}$$

The goal of the MLE method is to obtain the parameter values θ that maximize the likelihood function. It is often convenient to work with the logarithm of the likelihood function instead of the likelihood itself,

$$\mathcal{L}(\theta|\mathcal{S}) = \sum_{n=1}^{N} \log\Big\{p(x_n|\theta)\Big\} \tag{2.20}$$

Therefore, the MLE is based on solving a maximization problem

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{n=1}^{N} \log \Big\{ p(x_n|\theta) \Big\} \tag{2.21}$$

The most interesting property of the MLE method is that it is unbiased, i.e., $E[\hat{\theta}] = \theta$. The ML estimator is also consistent, that is, it is asymptotically unbiased when $N \to \infty$. The ML estimator is asymptotycally efficient, that is, it achieves the smallest mean squared error amongst all unbiased estimators. Although MLE has good properties, they are all based on asymptotic assumptions, which means that the MLE relies on the availability of a great amount of observations. In practical biomedical problems, the samples consist of a small number of observations and usually a high number of variables. Under these conditions the ML estimator can be biased. In this scenario, the MLE requires the use of regularization techniques to avoid overfitting. The overfitting problem appears when the models adjust their parameters to the noise of the sample rather than to the underlying distribution of the data avoiding the generalization to unseen data. An alternative is to use the Bayesian inference to estimate the parameters of the model. This is explained in the next section.

2.3 Bayesian inference

The *maximum-likelihood* method involves finding the parameters $\hat{\theta}$ of the model by maximizing (2.20) given the observations. Once the parameters are fitted, the model is ready for predicting new data using $p(x_{N+1}|\hat{\theta})$. However, the problem with this method is that we have solved the probability of the observed data given the estimated parameters. We may think that our problem to solve is really the probability of the parameters given the observed data. This problem can be solved using the Bayesian inference paradigm where, given that the parameters are unknown quantities, we can treat them as random variables and use the laws of probability to manipulate those uncertain quantities rather than maximize the likelihood.

Applying the Bayes theorem, the Bayesian approach to estimate the parameters of a model given the observed data is

$$\begin{aligned} p(\theta|\mathcal{S}, \mathcal{H}) &= \frac{p(\mathcal{S}|\theta, \mathcal{H}) p(\theta|\mathcal{H})}{p(\mathcal{S}|\mathcal{H})} \\ &\propto \mathcal{L}(\theta|\mathcal{S}) p(\theta|\mathcal{H}) \end{aligned} \tag{2.22}$$

where $\mathcal{H}$ denotes assumptions on which the probabilities are based. This expression shows that the parameters follow a probability distribution instead of being a unique value. This *posterior probability* of the parameters $p(\theta|\mathcal{S}, \mathcal{H})$ is proportional to the product of a *prior probability* of the parameters $p(\theta|\mathcal{H})$, which reflects the uncertainty about θ before the data is taken into account, and the *likelihood* $\mathcal{L}(\theta|\mathcal{S})$, which is equal to the probability $p(\mathcal{S}|\theta, \mathcal{H})$, as expressed in (2.20). This product is normalized using the *marginal likelihood* or *evidence* $p(\mathcal{S}|\mathcal{H})$, which is the integration of the numerator

$$p(\mathcal{S}|\mathcal{H}) = \int p(\mathcal{S}|\theta, \mathcal{H}) p(\theta|\mathcal{H}) d\theta \tag{2.23}$$

As we can see, the prior beliefs that the designer has about the parameter values are a key element of the Bayesian inference. These beliefs are expressed in mathematical form using a probability distribution. However, the assignment of a proper prior distribution depends on the kind of design assumptions that we are willing to apply. There are basically two different approaches to the definition of the prior probability distribution: the informative or subjective prior distributions and the non-informative or objective prior distributions[a]. In this Thesis, only the former approach is used in Chapter 5 to assign the parameter prior.

Informative priors

The informative prior approach tries to enclose the expert knowledge or the previous experience in the prior beliefs. Since it is difficult to express the knowledge in a mathematical form, a very convenient class of informative priors are the *conjugate* priors. A formal definition of *conjugate prior* is

Definition Let $\mathcal{P}$ be a family of prior parameter distributions $p(\theta)$ and $\mathcal{F}$ a family of likelihoods $p(x|\theta)$, then the family $\mathcal{P}$ is a **conjugate** of the family $\mathcal{F}$ if

$$\forall p(x|\theta) \in \mathcal{F} \wedge p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}.$$

That is, the product of the likelihood and the prior probability results in a posterior probability that belongs to the same family of distributions as the prior probability. This kind of conjugate priors often lead to analytically tractable integral expressions for estimating the parameters. Furthermore, if the posterior is tractable, then it can be used as a prior belief in subsequent inferences to obtain new posteriors that will be also tractable. This advantage is the core of the incremental algorithm that is proposed in Chapter 5; however, one disadvantage is that the only likelihood functions for which conjugate prior families exist are those belonging to general *exponential family* models. This poses a problem in the design of the incremental learning algorithm proposed as we will see later. For a deeper analysis of conjugate priors and exponential family likelihoods refer to Gelman *et al.* [37] or Bernardo and Smith [38].

2.3.1 Final predictive distribution

From a Bayesian approach, to predict the value of a new observation s_{new} we should integrate the posterior with respect to the parameters to obtain the *final predictive distribution.*

$$p(s_{new}|\mathcal{S}) = \int p(s_{new}|\theta)p(\theta|\mathcal{S})d\theta. \tag{2.24}$$

When we are dealing with a classification problem, the observations s_n are an ordered pair $\{\mathbf{x}_n, c_n\}$, where $\mathbf{x}_n$ are the independent variables and c_n is the output value. In order to predict a new output value c_{new} given $\mathbf{x}_{new}$, the final predictive distribution is

[a] Although the non-informative approach is sometimes called *objective* we should not forget that all Bayesian approaches are subjective since they require an expression of prior beliefs to be defined.

an integration of the predictions of the model with respect to the posterior distribution of the parameters,

$$p(c_{new}|\mathbf{x}_{new}, S) = \int p(c_{new}|\mathbf{x}_{new}, \theta)p(\theta|S)d\theta, \tag{2.25}$$

The expression of the final predictive distribution can be very difficult to obtain analytically. There are several ways to overcome this drawback. A widely used method is the Markov Chain Monte Carlo sampling methods can be used. However, they are often the source of considerable computational difficulties and cost. An alternative is to find the value with maximum posterior probability density, θ_{MAP}. This *Maximum A Posteriori* (MAP) estimate is only useful when it approximates the integral of equation (2.25).

2.3.2 Laplace Approximation

When calculating a marginal likelihood to obtain the posterior parameter probability is analytically intractable, we can make use of an analytical approximation –called the Laplace approximation– as an alternative. There are other types of approximations like the *Variational Inference* methods [39, 34] or the *Expectation-Propagation* method [40, 41]. However, since the posterior density of our method has a unimodal convex functional form we can restrict our research to the widely used Laplace approximation which is a straightforward deterministic local approximation.

A Laplace approximation is a method that aims to find a Gaussian approximation $q(z)$ to a non-Gaussian probability density $p(z)$ defined over a set of continuous variables. Let's consider the following distribution function $p(z) = Z^{-1}f(z)$, with Z being a normalization coefficient. The Laplace approximation searches for an approximated Gaussian distribution $q(z)$ centered on a mode z_{max} of the distribution $p(z)$. If we apply the Taylor series expansion for $\log\{f(z)\}$, then

$$\log\{f(z)\} = \log f(z_0) + \left.\frac{\partial \log f(z)}{\partial z}\right|_{z=z_0}(z-z_0) + \frac{1}{2}\left.\frac{\partial^2 \log f(z)}{\partial z^2}\right|_{z=z_0}(z-z_0)^2 + \mathcal{O}(z^3) \tag{2.26}$$

where it is assumed that the higher-order terms, represented by $\mathcal{O}(z^3)$, are negligible. Suppose that $z_0 = z_{max}$ is a local maximum in $f(z)$, then the first-order term is zero since it is a stationary point. Now, the Taylor expansion is

$$\log\{f(z)\} \approx \log f(z_{max}) + \frac{1}{2}\left.\frac{\partial^2 \log f(z)}{\partial z^2}\right|_{z=z_{max}}(z - z_{max})^2 \tag{2.27}$$

Taking the exponential and using

$$\beta = -\left.\frac{\partial^2 \log f(z)}{\partial z^2}\right|_{z=z_{max}}$$

we obtain

$$f(z) \approx f(z_{max}) \exp\left\{ -\frac{\beta}{2}(z - z_{max})^2 \right\} \tag{2.28}$$

which reminds the form of a Gaussian distribution. A noteworthy fact is that the Gaussian approximation will only be well defined if the stationary point z_{max} is a local maximum

because it is mandatory for the second derivative of $f(z)$ at the point z_{max} to be negative. Finally, we have a normalized distribution $q(z)$ by using the standard result for the normalization of a Gaussian,

$$q(z) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{ -\frac{\beta}{2}(z - z_{max})^2 \right\} \tag{2.29}$$

The Laplace approximation can be extended for multivariate distributions, where a distribution function $p(\mathbf{z}) = Z^{-1} f(\mathbf{z})$ is defined over a multidimensional space $\mathbb{R}^D$. Assuming that there is an stationary point $\mathbf{z}_{max}$ where the gradient $\nabla f(\mathbf{z})$ vanishes, the Taylor expansion around this point $\mathbf{z}_{max}$ is

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_{max}) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \tag{2.30}$$

where $\mathbf{H}$ is the Hessian matrix with dimension $D \times D$, which is defined by

$$\mathbf{H} = -\nabla\nabla \log f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_{max}} \tag{2.31}$$

where ∇ is the gradient operator. Taking the exponential of both sides of the equation we obtain

$$f(\mathbf{z}) \approx f(\mathbf{z}_{max}) \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \right\} \tag{2.32}$$

Using the standard result for a normalized multivariate Gaussian distribution with the appropriate normalization coefficient the distribution $q(\mathbf{z})$ is

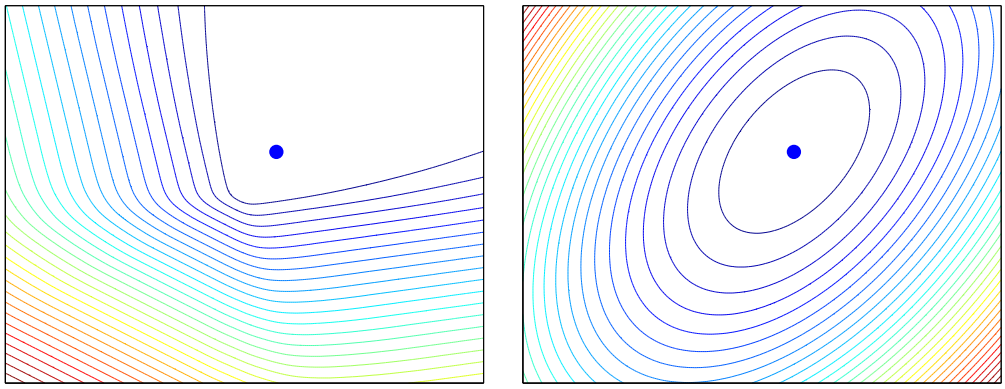
$$\begin{aligned} q(\mathbf{z}) &= (2\pi)^{-D/2} |\mathbf{H}|^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_{max})^{\mathrm{T}} \mathbf{H} (\mathbf{z} - \mathbf{z}_{max}) \right\} \\ &= \mathcal{N}(\mathbf{z}_{max}, \mathbf{H}^{-1}) \end{aligned} \tag{2.33}$$

As in the univariate version, this Gaussian distribution will be well defined provided its precision matrix, $\mathbf{H}$, is positive definite, which implies that the stationary point $\mathbf{z}_{max}$ must be a local maximum.

Now, there is a step-by-step process for using the Laplace approximation to properly approximate a probability density function $p(\mathbf{z})$ with a Gaussian $q(\mathbf{z})$ provided $p(\mathbf{z})$ has one single mode. The first step is to find a local maximum $\mathbf{z}_{max}$ of the given probability density function $p(\mathbf{z})$ by running some form of numerical optimization algorithm. It is worth to mention that if the distribution $p(\mathbf{z})$ is multimodal, then there can be different approximations. The next step is to calculate the inverse of the Hessian matrix for the stationary point $\mathbf{z}_{max}$ as

$$\mathbf{H}^{-1} = -\left(\frac{\partial^2}{\partial \mathbf{z} \partial \mathbf{z}^{\mathrm{T}}} \log p(\mathbf{z}) \right)^{-1} \tag{2.34}$$

finally, we can approximate the probability density function $p(\mathbf{z})$ using $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}_{max}, \mathbf{H}^{-1})$. Figure 2.3 shows the results of approximating a bidimensional probability density function with the Laplace approximation.



(a) Non-Gaussian pdf (b) Approximated Gaussian pdf

Figure 2.3: A probability density function (a) is approximated to a bidimensional Gaussian probability density function using the Laplace approximation (b). The maximum value, which is the same for both densities, is shown with a blue dot.

2.3.3 Monte Carlo sampling methods

The goal in Bayesian machine learning is to obtain a model to predict unseen cases. This may take the form of equation (2.24). If we look carefully to this equation, we can see that we are evaluating the expectation of a function with respect to the posterior distribution of the parameters of the model. If we write the posterior probability of the parameters as $Q(\theta) = p(\theta|\mathcal{S})$ and the predictions of the model is written as $f(\theta) = p(y_{new}|\mathbf{x}_{new}, \theta)$, then the expectation of $f(\theta)$ is

$$E[f] = \int f(\theta)Q(\theta)d\theta.$$

The analytical computation of this integration is often difficult. However, there are sampling *Monte Carlo* methods that can be applied to solve the problem. Markov chain Monte Carlo methods make no assumptions concerning the functional form of the distribution. The main disadvantage is however that they may in some circumstances require a very long time to converge to the desired distribution.

It is possible to obtain a sampling of the parameters by using the distribution $p(\theta|\mathcal{S})$ in order to obtain an estimate of the expectation. This is the core idea of the *Monte Carlo* sampling method, where a set of parameters $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ are generated by a process that results in each of them having the distribution defined by Q. From this set of generated parameters Θ the expectation can be approximated

$$E[f] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t). \tag{2.35}$$

The problem now is how to generate those values θ_t. There are several *Monte Carlo* methods to generate the sample. There exist basic simple methods that generate a set of independent parameters that follow $p(\theta|\mathcal{S})$ by means of a generator of pseudo-random numbers distributed uniformly over the interval $[0, 1]$. For multivariate distributions, we have to transform the uniformly distributed random numbers $\mathbf{z}$ using a function such that

$\mathbf{y} = f(\mathbf{z})$, which implies that $\mathbf{z} = f^{-1}(\mathbf{y})$. Hence, the multivariate distribution of $\mathbf{y}$ follows

$$p(\mathbf{y}) = p(\mathbf{z}) \left| \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right| \tag{2.36}$$

As explained in Chapter 5, our parameter posterior distribution is a multivariate Gaussian distribution and the sampling method uses the Box-Muller transformation to generate Gaussian-distributed samples from a uniform distribution inside a unit hypersphere [34].

When the distribution $p(\theta|\mathcal{S})$ is complicated however generating such independent values is often infeasible. Nevertheless, it may be possible to generate a series of dependent values using *Markov chains* that still can give an unbiased estimate of $E[f]$ as long as the dependence is not too great so the estimate will still converge to the distribution Q when $T \to \infty$. There are a number of Markov Chain Monte Carlo (MCMC) algorithms such as the Metropolis-Hastings algorithm [42], the Gibbs sampling [43], or the Hybrid Monte Carlo [44]. The interested reader can consult [45] for a deep understanding of these and other MCMC methods.

2.4 Incremental learning

During the last decades the development of the information and communication technologies has produced an explosion of data growth, which requires rates scalability of data analysis methods. As an alternative to the expensive computer cluster frameworks for large-scale analysis [46], incremental algorithms that build approximate models on continuous and possibly infinite data streams have been developed. At the same time, there are real-world scenarios, such as medical studies, financial data analysis or other data streaming applications, where obtaining a dense and representative dataset –essential for building a proper predictive model– is expensive and time consuming. Hence, data is often acquired in batches over time where an incremental algorithm may be also applied.

Historically, the development of the machine learning (ML) discipline has been focused on non-incremental methods where a large dataset is assumed to be available at design-time. Recently, the ML community has emphasized the research on incremental learning algorithms [47]. Hence, when a non-incremental ML model is trained, there is an implicit assumption that learning stops once the current gathered sample has been processed. Furthermore, it has been a common assumption to believe that the current observed data and the future data are sampled independently and from an identical probability distribution (*iid*). However, incremental ML models assume that learning is an adaptive process embedded within complex data processing systems where the data are available in small batches or where the underlying data distribution may change in the course of time (*dataset shift* [20, 21]).

The earliest formal description of an algorithm with the ability to learn from new data was the description of *learning in the limit* by Gold [48], which Sharma considers to be the ideal case of incremental learning [49]. In the initial description, a learning machine could use all the information seen so far to yield new hypotheses. This ideal scenario has been theoretically compared to more realistic scenarios where the past information is partly accessible or is completely unaccessible [50]. Following this research, Lange [51] concluded that, when noisy data is used, incremental learners with restricted access to data are as

powerful as unconstrained learning algorithms, which is an interesting conclusion for the incremental learning algorithms presented in this Thesis.

The easiest way to take advantage from new observations is to build a new model from scratch using a combination of the old and new data. But this solution may be more expensive than modifying an already trained system, or even impractical if older training set data is not readily accessible. Typically, an alternative has been to keep a relevant subset of the previous data available. This approach was used in the partial memory learning [52] and in the so-called boundary methods, or maximum margin methods [53, 54]. In our approaches, it is assumed that previous data are not accessible at all. In the last two decades, various approaches have been developed for providing learners with incremental learning ability. A number of incremental techniques were designed for decision trees [55, 56, 57], and then have been applied for data streaming [58]. Incremental learning has also been used for connectionist models based on structural adaptation [59, 60, 61, 62, 63] or on weight adaptation [64, 65]. There are some approaches to incremental principal component analysis [66, 67] that update the projection matrix incrementally. Moreover, incremental algorithms for Fisher's Linear Discriminant Analysis have also been developed in the last decade [68, 69, 70].

Up to the writing time, many state-of-the-art incremental algorithms assume that previous data are partially or totally accessible. Based on this assumption, these incremental models handle streaming datasets by time windows of fixed or adaptive size [71, 72, 73], by weighting the models in an ensemble [19, 74, 62], or by weighting the data [75]. An intermediate alternative approach has been to keep only a relevant subset of the previous data available [52, 54]. However, there are real scenarios with data protection policies or possible conflict of interests where the previous data are not available if a model trained with data from one organization is moved to another organization. For instance, when a model is moved from one hospital to another, patient consent and ethical approval needs to be obtained to send and store data [28]. In order to take into account this condition, we will assume that previous data are not accessible at all.

2.4.1 Definition of incremental learning algorithm

Following the definitions of Langley [76] and Giraud-Carrier [77], an **incremental learning algorithm** is a learning algorithm that produces a sequence of classifiers $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T$ for any given training set of samples $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_T$ available at different moments $t_1, t_2, \ldots, t_T$, such that $\mathcal{M}_{t+1}$ is determined by $\mathcal{M}_t$ and $\mathcal{S}_{t+1}$. The main characteristics of an incremental learning algorithm are: a) it should be able to learn additional information from new data without completely forgetting its previous knowledge; b) since each $\mathcal{M}_t$ can be viewed as the best approximation of the target application, the performance should improve over time.

This definition is related to a general problem for classification models called the *stability-plasticity dilemma* [78]. This dilemma reveals that some information may be lost when new information is learned (*gradual forgetting*) and highlights the difference between stable classifiers and plastic classifiers. On one hand, a completely stable classifier will preserve existing knowledge, but it will not incorporate any new information. On the other hand, a completely plastic classifier will learn any new information without preserving any previous knowledge. The latter case is also known as *catastrophic forgetting* [79, 80] and it happens when an already trained model learns a new set of patterns completely erasing

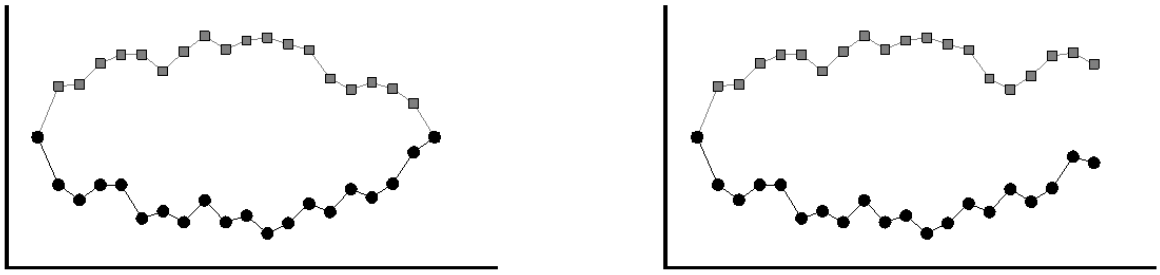


Figure 2.4: An incremental algorithm generates models in the hypothesis space for two different ordered sequences of training samples. On the left, there is an evolution of the models for a order independent algorithm. On the right, there is an evolution for an order sensitive incremental learning algorithm.

its previous knowledge. To summarize, the challenge is how to design a learning system that is sensitive to new input without being radically disrupted by such input.

A number of authors [77, 60, 81] include in the definition of incremental learning algorithms that it should not require access to previous data; however, this issue is not considered by the vast majority of incremental learning algorithms developed in the last years [71, 75, 58, 74, 72]. In this Thesis, we will consider that our incremental learning algorithms will not require access to previous data since this condition may be imposed by real-world health organizations.

In addition, Polikar et al. stated in [60] that an incremental learning algorithm should be able to admit new classes when they are introduced with the new data. This means that a new target concept appears over time while the rest of the target concepts remain stable. This issue entails a *prior probability shift* [21], which represents a significant change in the underlying joint distribution $p(\mathbf{x}, c)$. The admission of new classes is a solution that has been applied to identify additional arrhythmias with a system designed to detect other cardiovascular disorders by means of electrocardiograms [82].

Another issue to be considered is the problem of the *ordering effects* in incremental learning, which has been addressed by several authors [83, 76, 84]. An incremental learning algorithm suffers from an *order effect* when there exists two or more different ordered sequences of the same instances that lead to different models. This is illustrated in Figure 2.4, where an incremental algorithm searches through a model space based on different samples of the data. When two different orders of training samples are presented to the same algorithm it can evolve building different models but, when all the training samples are duly processed, it leads to the same one. This constitutes an example of an algorithm that is not affected by the order in which the samples are presented (left). However, it can happen that each different ordered samples lead to different final models. This embodies an example of an order effect (right). Therefore, the selection of the final models may be biased due to the ordering of the introduced inputs.

An incremental learning algorithm is said to be *order independent* if it never exhibits an order effect for any possible ordered sequence of samples. Otherwise, the algorithm is said to be *order sensitive*. When an incremental learning algorithm suffers from an order effect the performance of the models achieved usually reflect substantial differences for a given metrics. If the order effect produces models of nearly equal scores on a metric M,

then it is a *benign* order effect for metric M since this effect has relatively little impact on performance. In contrast, if the order effect produces models with very different results on the performance metric M, then it is a *malignant* order effect. The order effect can appear at three levels: attribute level, instance level, and concept or class level. The instance level order effect is the most studied order effect. It appears when different models are obtained when the order of the instances or observations are shuffled. The concept level order effect happens when the order of the different classes to discriminate entail the development of different models. Finally, the attribute order effect is related to a order sensitivity of the models to a different ordering sequence of the attributes or variables of the observations. This effect has been hardly studied in the state-of-the-art and thus is out of the scope of this Dissertation.

2.4.2 Incremental learning using Bayesian inference

An incremental learning task [77] consists of a set of observations that arrive over time in subsets of samples or batches $\mathcal{S}$. We consider that a sample $\mathcal{S}_t$ arrives at time t. Each sample has n_t observations which are ordered pairs $z_{t,i} = (\mathbf{x}_i, c_i)_t$ where $\mathbf{x} \in \mathcal{X}^D$ are the D-dimensional covariates and $c \in \mathcal{C}$ is the class label, and where the indices denote the i-th observation of sample t. We assume that only the sample $\mathcal{S}_t$ is available in time t for training an incremental model.

We will assume that a model $\mathcal{M}_t$ is determined by a set of parameters Θ_t. In the Bayesian paradigm the parameters of the model Θ are estimated given the data $\mathcal{S}$ and an overall hypothesis space $\mathcal{H}$ using the Bayes theorem,

$$p(\Theta|\mathcal{S}, \mathcal{H}) = \frac{p(\mathcal{S}|\Theta)p(\Theta|\mathcal{H})}{p(\mathcal{S}|\mathcal{H})} \tag{2.37}$$

where $p(\Theta|\mathcal{S}, \mathcal{H})$ is the *posterior probability* distribution of the parameters, $p(\mathcal{S}|\Theta)$ is the *likelihood* function, $p(\Theta|\mathcal{H})$ is the *prior probability* distribution of the parameters and $p(\mathcal{S}|\mathcal{H})$ is the *evidence* or the *marginal likelihood*, defined as

$$p(\mathcal{S}|\mathcal{H}) = \int p(\mathcal{S}|\Theta)p(\Theta|\mathcal{H})d\Theta \tag{2.38}$$

We use this approach to define our incremental algorithm by using the posterior probability estimated in time $t-1$ as the prior probability of the model estimated in time t. That is, $p(\Theta_t|\mathcal{H}) = p(\Theta_{t-1}|\mathcal{S}_t, \mathcal{H})$.

When the model parameters are estimated, a new observation $\mathbf{x}_{new}$ can be classified with the *final predictive distribution* expressed as

$$p(y_{new}|\mathbf{x}_{new}, \mathcal{S}, \mathcal{H}) = \int p(y_{new}|\mathbf{x}_{new}, \Theta)p(\Theta|\mathcal{S}, \mathcal{H})d\Theta \tag{2.39}$$

As previously mentioned, the expression (2.39) is often too complex to solve analytically and Monte Carlo (MC) sampling methods are required to obtain an approximation to the desired model. Alternatively, when the distribution is assumed to be sharply peaked, the *Maximum A Posteriori* (MAP) approach may be used instead to avoid the integral computation.

2.5 Evaluation

In this Thesis, different evaluation methodologies were followed depending on the type of algorithm we were evaluating. For a non-incremental learning algorithm (Chapter 3), the evaluation procedure for the experiments was a K-fold Cross Validation (CV) with stratified blocks. In the CV evaluation, the dataset is divided into K subsets. Then, one subset is used as a test set and the remaining $K-1$ subsets are used to train one model. This process is repeated K times taking a different subset as a test set and thus training K different models. Then, the average error of each model is computed. In this evaluation procedure, every data point appears in a test set exactly once, and comes out in a training set $K-1$ times. The bias of the resulting estimate is reduced as K is increased, while the variance increases. Thus, it is interesting to achieve a trade-off in the number of K subsets to be used. Kohavi in [85] proposed $K = 10$ as an optimum value. In the extreme case where $K = 1$ the evaluation method is often known as *Leave-one-out* evaluation. The disadvantage of the CV methodology is that the training algorithm has to be rerun from scratch K times.

The evaluation carried out for the incremental learning algorithms has been an adaptation of the k-Random Sampling Train-Test (kRSTT) with stratified test sets with K repetitions. In this methodology the dataset is randomly split into a set of instances to train the classification model and the remaining instances to test the model. The adaptation for the incremental learning evaluation takes the training set and splits it into several subsets $\mathcal{S}_t$ that are sequentially presented to the learning algorithm to re-train the model $\mathcal{M}_t$. Each incremental model $\mathcal{M}_t$ is then evaluated using the same test set. This procedure is repeated K times. These methods avoid underestimation of the true error when the evaluation is carried out in a nested-loop that covers the feature and model selection.

Evaluation metrics

The performances of the classifiers were measured in terms of accuracy (acc). Since we are using a 0-1 loss function and discrete values for each class, the acc metric is defined as the rate of well classified instances among all the classified instances,

$$\mathrm{a}cc = \frac{1}{N}\sum_{n=1}^{N}\delta(c_n, \hat{c}_n) \tag{2.40}$$

It can be seen that this expression is the same as the empirical risk (2.9).

The performances of the classifiers presented in this Dissertation were also measured with the Geometric mean of recalls or sensitivities for each class (G). The sensitivity of a class is defined as the ratio of the number of correct classifications of class c, T_c, divided by the total number of observations of class c, N_c, that is,

$$sen_c = \frac{T_c}{N_c} \tag{2.41}$$

This metric is a good non-linear measure for determining the average success even when working with highly imbalanced populations. The geometric mean of sensitivities is de-

fined as

$$G = \sqrt[|\mathcal{C}|]{\prod_{c=1}^{|\mathcal{C}|} sen_c} \tag{2.42}$$

the $|\mathcal{C}|$-th root of the product of the sensitivity of each class (sen_c), where $|\mathcal{C}|$ is the total number of classes. The measure G is only high when all the sensitivities are high and balanced.

Finally, the Balanced Error Rate (BER) was also used as another metric for imbalanced datasets. The BER is defined as

$$BER = \frac{1}{|\mathcal{C}|} \sum_{c=1}^{|\mathcal{C}|} (1 - sen_c) \tag{2.43}$$

that is, the arithmetic mean of the false positive rates for each class.

2.6 Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) (or MR) is the phenomenon where the nuclei of certain atoms absorb and emit energy because of the effect of an oscillating magnetic field when they are immersed in other static magnetic field [86]. Magnetic Resonance Spectroscopy (MRS) is the use of the NMR phenomenon to study the physical, chemical, and biological properties of organic and inorganic molecules in a non-destructive, non-invasive manner. Typically, for the study of brain tumours, the NMR phenomenon observed is performed with the ^{1}H sensitive nucleus.

For the protons ^{1}H, the spin quantum number s associated to the particle angular moment takes the half-integer value (1/2). For this particles, the *secondary spin quantum number* m_s takes the values $m_s = \{-1/2, 1/2\}$. That associates two possible potential energy levels to the ^{1}H particles (depending on m_s) in the presence of a magnetic field $\mathbf{B_0}$, being the energy difference between both states

$$\Delta E_{pot} = -\gamma \hbar |\mathbf{B_0}|,$$

where γ is the gyromagnetic constant and $\hbar$ is the reduced Planck's constant.

These particles do not align exactly with the axis of the external magnetic field $\mathbf{B_0}$ but precesses around it at a rate given by the Larmor frequency f_0,

$$f_0 = \gamma \mathbf{B_0}/(2\pi).$$

In the very first beginning time of the precessing motion around the field, the total magnetic moment $\mathbf{M}$ of material is still near $\mathbf{0}$. As the elements of the molecule have their magnetic momenta, they generate magnetic fields that change with the thermical motion of the environment, so each spin is precessing around a local and changing magnetic field instead of the applied $\mathbf{B_0}$, so the spins are slowly deviated. The probability of the low energy orientations are slightly higher than the probability of the high energy levels. Hence, when the thermical equilibrium is reached there will be more spins parallel to the $\mathbf{B_0}$ than anti-parallel. Consequently, a total magnetic moment $\mathbf{M} \neq \mathbf{0}$ is observed.

If a Radio frequency (RF) pulse is applied to the sample, the spins experiment the influence of two magnetic fields $\mathbf{B_0}$ and $\mathbf{B_1}$. The first is a static field, the second is an oscillating one of which the frequency is the resonant Larmor's frequency f_0. A pulse in the y-axis produces a progressive decay of the $\mathbf{M}$ vector to the XY-plane. When the pulse is over, the spins return to the precession around the static magnetic field, obtaining, as a result a macroscopic $\mathbf{M_{XY}}$ motion similar to the precession of the spins. Transversal to the axis coils can acquire the FID signal produced by the M_{XY} motion.

A noteworthy fact is that the magnetic field for each nucleus depends on the static magnetic field $\mathbf{B_0}$, but also on the local environment,

$$\mathbf{B_{eff}} = \mathbf{B_0}(1 - \sigma'),$$

where σ' is the shielding constant that depends on the electrical environment of the nucleus. This results in different frequencies of resonance (or Chemical Shift (CS)) of the same nuclei depending on the molecular environment and the main application of MRS in biochemistry and molecular biology.

As a consequence of these small differences in the resonance frequencies of each nuclei depending on their molecular environment, it is possible to analyze the composition of a sample by the study of the frequencies reemitted after a RF radiation stimulation. The analysis of this frequencies compound is done by the Fourier Analysis of the reemitted RF radiation. The frequencies where nuclei resonate use to be expressed in parts per million (ppm). This representation is due to the small distance between different resonance frequencies in contrast to the high frequencies on which they resound. Moreover the metabolite peaks positions expressed in ppm are invariant over NMR scanner magnetic field changes, and therefore allow direct comparisons between spectra obtained using different NMR scanners. The frequency shift δ in ppm is calculated as

$$\delta = \frac{f_0 - f_r}{f_r} \times 10^6 \tag{2.44}$$

where f_0 is the Larmor's frequency of resonance of the ^{1}H nuclei and f_r is the Larmor's frequency of resonance of a reference compound, which is usually tetrametilsilane.

Chapter 3

Automatic brain tumour classification by ^{1}H MRS

The ML-based approach to brain tumor classification by MRS has been under development for more than a decade. This Chapter summarizes some contributions to automatic brain tumour classification. First, we proposed a combination of Short Time of Echo (STE) and Long Time of Echo (LTE) single voxel ^{1}H MRS that profits from the advantages of each TE spectrum. The results showed that there was an improvement when using the combination of both TE with respect to using only one TE. Second, we contribute an independent external evaluation for the developed brain tumour predictive models. To our knowledge, there were no published evaluations of classification models with unseen cases subsequently acquired at different centers. A number of predictive models for automatic brain tumour diagnosis were fitted using the INTERPRET project (2000-2002) multicenter dataset and their performance was estimated using a cross-validation evaluation. Then, the eTUMOUR project (2004-2009) multicenter dataset was used as an independent test set to assess the performance of the models. Although the results were reasonably good, this evaluation showed that there was a bias trend between the performance estimations of the different datasets. This may indicate the need of incremental learning algorithms to take advantage of new available data as a way of re-adapting the models to the new centers.

The core of this Chapter were published paper journals in [25] and [24]. The author of this Thesis carried out contributions in both manuscripts regarding the design, development, and evaluation of the experiments. Specifically, Fisher's Linear Discriminant Analysis was used for modeling classifiers for automatic brain tumour classification using STE, LTE, and the combination of both times of echo. Also, Multilayer Perceptron (MLP) were used for pairwise brain tumour classification. In addition, the author of this Thesis contributed to the manual preprocessing of several ^{1}H MRS brain tumour cases. Finally, a critical analysis and discussion of the results obtained for different methodologies and machine learning techniques for automatic brain tumour diagnosis was carried out. This analysis led to conclude the necessity of incremental learning algorithms for the development of CDSS for automatic brain tumour diagnosis and for other general applications.

3.1 Introduction

Brain tumors are the second fastest growing cause of cancer deaths among people older than 65 years. Nowadays, the diagnosis and treatment of brain tumors is based on clinical observations, radiological appearance, and often a histopathological diagnosis of a biopsy, which is an invasive technique that removes tissue from the subject to determine the extent of the disease. During the last decade, three European projects (INTERPRET (2000-2002) [8, 11], eTUMOUR (2004-2009) [9], and HEALTHAGENTS (2005-2008) [10]) have endeavoured to develop a non-invasive diagnostic tool using machine learning (ML) techniques applied to ^{1}H MRS data from brain tumours. A major aim was to minimize the need for an invasive histological diagnosis of a brain tumour biopsy as is currently required for the diagnosis and management of brain tumours.

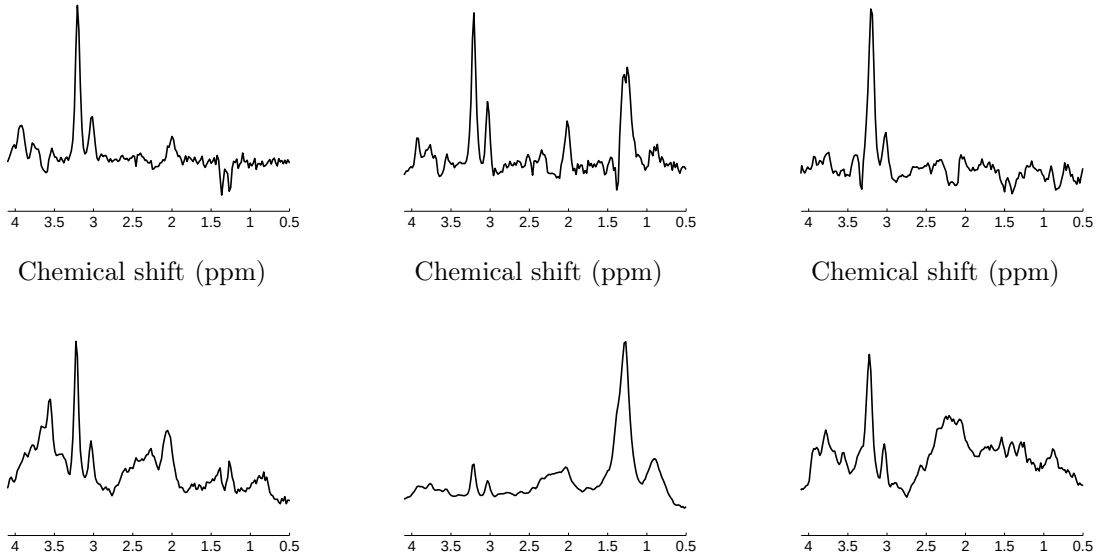
Non-invasive brain tumour diagnosis using ^{1}H MRS has shown considerable promise in aiding patient management, due to its ability to provide useful chemical information about different metabolites and other compounds for characterizing brain tumours [87, 88]. Currently, the acquisition of ^{1}H MRS use a time of echo (TE) that range between 18 and 288 ms in most studies. A spectrum acquired with a TE $<$ 45 ms is usually considered a STE spectrum, and a LTE spectrum otherwise. STE (20-35 ms) ^{1}H MRS allows to observe macromolecules (MM; 5.4ppm, 2.9ppm, 2.25ppm, 2.05ppm, 1.4ppm and 0.87ppm), Myo-Inositol (mI) and Mobile Lipids (ML) better than in LTE [89]. Single voxel (SV) Short TE ^{1}H MRS is fast (typically 5 min) and robust, so it is very useful for clinical studies [90, 91]. However, Short TE signals show a large number of overlapping peaks, a strong MM-/ML-originated baseline, and a certain sensitivity to artifacts [92]. LTE (about 135 ms) ^{1}H MRS, on the other hand, is less informative than STE because some resonances may be lost due to a short T_2. However, LTE signals are easier to analyze than STE signals [92]. Lipid resonances (1.3 and 0.9ppm) and MM will not be the dominating components at LTE, making possible the study of the contributions of lactate (Lac, doublet at 1.33ppm) and alanine (doublet at 1.47ppm) as inverted peaks [91]. Many successful applications of pattern recognition (PR) and machine learning (ML) for automated classification of brain tumours have been reported in brain tumor research [13, 93, 94, 14, 95, 96]. In [12], Hagberg summarizes classification of brain tumors with MRS based on pattern recognition and clustering methods. Eight of these studies were applied to brain tumor discrimination from normal tissue or other Central Nervous System (CNS) diseases. All of them were based on Linear Discriminant Analysis (LDA) or Artificial Neural Network (ANN) applied to relative metabolite levels or Principal Components Analysis (PCA) transformations, and they were all evaluated by leave-one-out cross-validation. More recent publications have also described results for classification of brain tumors based on the MR data available within the the INTERPRET project (INTERPRET) [13, 8], where linear and kernel-based methods on MRS features extracted by automatic procedures were applied [97, 93]. While [92] was focused on the classification of brain tumours using LTE ^{1}H MRS, other studies [13, 98, 93, 11] carried out experiments with STE ^{1}H MRS. Based on Least Squares Support Vector Machines (LS-SVMs) [99], Devos, Lukas et al. in [98, 92] developed different classifiers for in-vivo ^{1}H MRS and Magnetic Resonance Spectroscopic Imaging (MRSI) with good performance. Menze et al. [14] published an extensive benchmark study of quantitation and PR based feature extraction methods combined with learning strategies to discriminate between recurrent and non-recurrent brain tumors using LTE ^{1}H MRS. They reported that the PR methods perform at least as well as the ones based

on manual quantitation (5%-10% higher accuracy). However, the development of robust brain tumour classifiers requires a large number of cases to be acquired for each tumour type and at present the approach has only been used for a few common tumours. Cases are accrued from a large number of hospitals over many years and data transferred to a centralised database. This approach has several disadvantages, ethical approval and patient consent needs to be obtained to send and store data. In order to expand the applicability of ML techniques to MRS of a wider range of tumours, more cases need to be collected over a more prolonged period of time and the logistics of using a centralised database to provide this have so far proved insurmountable. Furthermore, despite its ability to provide useful information for characterizing brain tumours, the use of ^{1}H MRS is not in widespread clinical use due mainly to the difficulties of data acquisition and interpretation, which give rise to bias and variance from single-center or single-machine studies. Therefore, standardization of acquisition conditions and protocols should make data from different hospitals compatible and allow the development and evaluation of joint CDSSs. This standardization aims to reduce, or prevent, possible bias or variance and, additionally, increases the number of available cases for classifier development and test purposes.

In this Dissertation, several contributions to the ^{1}H MRS automatic brain tumour classification were carried out as the first steps of our research. Next sections summarizes these contributions.

3.2 Data acquisition and pre-processing

The datasets used for classifier development were acquired by six international centers in the framework of the INTERPRET project [8], eight in the eTUMOUR project [9], and four in the HEALTHAGENTS project [10]. The STE spectra acquired were single-voxel (SV) MRS signals at 1.5T using Point-Resolved Spectroscopic Sequence (PRESS), using a TE between 30-32 ms, or Stimulated Echo Acquisition Mode sequence (STEAM), using a TE of 20 ms. The acquisition was carried out avoiding areas of cysts or necrosis and with minimum contamination from the surrounding non-tumoral tissue. The volume of interest size ranged between $1.5 \times 1.5 \times 1.5$ cm3, (3.4 mL) and $2 \times 2 \times 2$ cm3, (8 mL), depending on tumor dimensions. The aim was to obtain an average spectroscopic representation of the largest possible part of the tumor. These signals were acquired with Siemens, General Electric (GE), and Philips instruments. The acquisition protocols included PRESS or STEAM sequences, with spectral parameters: Recycling Time (TR) between 1600 and 2020ms, TE of 20 or 30-32ms, spectral width of 1000-2500Hz, and 512, 1024, or 2048 data-points for STE, as described in previous studies [100]. In the acquisition of LTE spectra, the PRESS sequence was used, with a recycling time (TR) between 1500 and 2020 ms, TE of 135 or 136 ms, spectral width of 1000 or 2500 Hz and 512 or 2048 data points. Every training spectrum and diagnosis was validated by the INTERPRET Clinical Data Validation Committee (CDVC) and expert spectroscopists [11]. The classes considered for inclusion in this study were based on the histological classification of the CNS tumors set up by the WHO [101]: glioblastomas (GBM), meningiomas (MEN), metastasis (MET), and low grade gliomas (LGG), which consists of three types of brain tumours: Astrocytoma grade II, Oligoastrocytoma grade II, and Oligodendroglioma grade II.



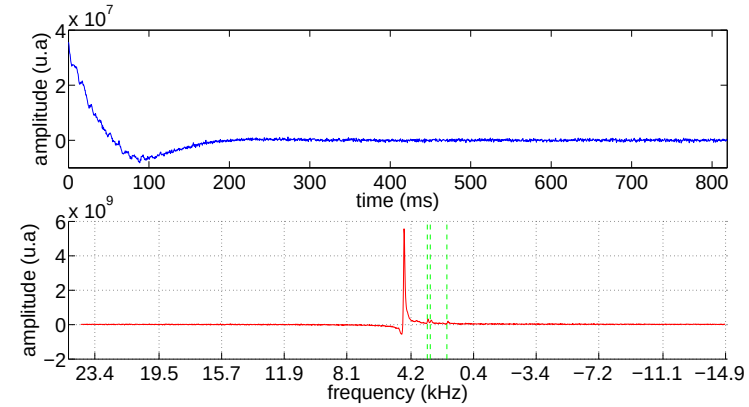
(a) Low grade glioma (b) Aggresive glioblastoma (c) Meningioma

Figure 3.1: Different spectra for LTE on the top row and for STE on the bottom row. The Y-axis displays arbitrary units and the X-axis shows the chemical shift in ppm. The spectra are examples of a low grade glioma (case et2354), a glioblastoma (case et2357), and a meningioma (case et3028).

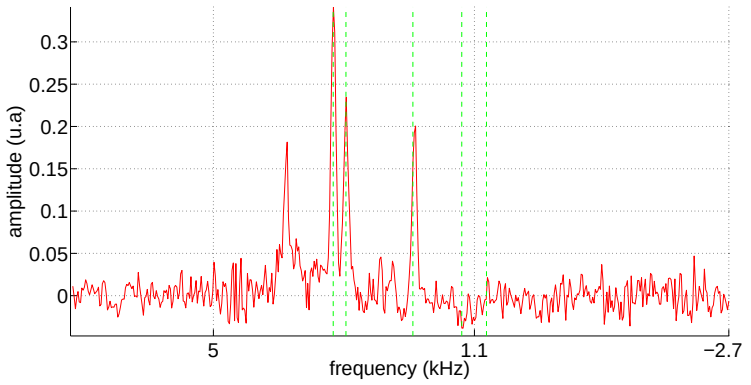
3.2.1 Automatic pipeline

During the INTERPRET project, a data acquisition protocol was defined to guarantee the compatibility of the signals coming from different hospitals [102, 98]. As a result, each signal was pre-processed according to the INTERPRET protocol. A fully automatic pre-processing pipeline was available for the training data. Besides, a semi-automatic pipeline was defined for some new file formats of the test cases from GE and Siemens manufacturers. The semi-automatic pipeline was designed to ensure compatibility of its output with the automatic one. This data acquisition protocol was then followed by the partners of the eTUMOUR project.

The steps of the automatic pre-processing pipeline were: (1) Eddy current correction was applied to the water-suppressed Free Induction Decay (FID) of each case using the Klose algorithm [103]. (2) The residual water resonance was removed using the Hankel-Lanczos Singular Value Decomposition (HLSVD) time-domain selective filtering using 10 singular values and a water region of [4.33, 5.07]ppm. (3) An apodization with a Lorentzian function of 1Hz of damping was applied. (4) Before transforming the signal to the frequency domain using the Fast Fourier Transform (FFT), an interpolation was needed in order to increase the frequency resolution of the low resolution spectra to the maximum frequency resolution used in the acquisition protocols (see [11] for details in the acquisition conditions and resolution). This was carried out with the zero-filling procedure. (5) Afterwards, the baseline offset, which was estimated as the mean value of the region $[11, 9] \bigcup [-2, -1]$ppm, was subtracted from the spectrum. (6) The normalization of the spectral data vector to the L2-norm was performed based on the data-points in the region



(a) Free Induction Decay (FID)



(b) Preprocessed spectrum

Figure 3.2: Illustration of a ^{1}H MRS brain tumour signal before the preprocessing pipeline is applied 3.2(a). It is possible to see that the contribution of water is dominant before any preprocess step is carried out. The FID in the time domain and in the frequency domain are shown. After preprocessing, the metabolite concentration arise providing useful chemical information. The region of interest of the spectrum that results from the complete preprocessing is shown in the bottom 3.2(b) (x-axis in the frequency domain, y-axis in arbitrary units).

$[-2.7, 4.33] \bigcup [5.07, 7.1]$ppm. (7) Depending on the Signal-to-Noise Ratio (SNR) and the tumor pattern, an additional frequency alignment check of the spectrum was performed by referencing –following the priority– the ppm-axis to the total Cr at 3.03ppm or to the Cho containing compounds at 3.21ppm or the ML at 1.29ppm. (8) Finally, the region of interest was restricted to $[0.5, 4.1]$ppm, obtaining a vector of 190 points for each spectrum where, after the preprocessing filters, the contribution of the residual water is expected to be minimal and the resonances of the main metabolites arise (see Figure 3.2).

3.2.2 Semi-automatic pipeline

Due to limitations of the automatic pre-processing software, a number of samples were preprocessed by a semi-automatic pipeline that was partially based on the Java Magnetic Resonance User Interface (jMRUI) [104]. Some modifications of the semi-automatic

pipeline with respect to the automatic pipeline were in the following steps: (1) The phase of the water-suppressed FID was mainly corrected with the reference water. Additional manual zero-order and first-order phase correction was performed when needed. (2) Residual water was removed by means of the jMRUI-implementation of the Hankel Singular Value Decomposition (HSVD) algorithm [105]. The filter was parametrized as in the automatic pipeline. From step 3 to step 8 the automatic pre-processing remained equivalent. As a result, a pre-processing pipeline based on different software implementations but compatible with the automatic one was set up, and comparable signals for testing the PR models were obtained.

3.3 Methods

A series of Pattern Recognition (PR) and ML techniques were applied in order to achieve empirical evidence supporting our experiments. These techniques included several feature extraction and selection methods that were applied to the real part of the spectra prior to any classification approach.

3.3.1 Feature selection and extraction methods

The feature extraction methods can be categorized into automatic methods and knowledge-based methods. The automatic feature extraction methods were Principal Component Analysis (PCA) [106], and Independent Component Analysis (ICA) [107, 108]. The feature selection methods included a Stepwise algorithm [109] and the ReliefF [110]. The knowledge-based methods included direct spectral peak integration on selected metabolite resonance region [111] and peak height of typical resonances [112].

PCA is a well-known projection method often used for feature extraction in PR [106]. PCA maps the original D-dimensional data into an orthogonal M-space, where the axes of this new coordinate system are a linear combination of the original variables. The new coordinate system lies along the direction of maximum variance of the original data. The more correlated the original variables are, the more the data variation is explained by the first principal components or loadings (PCs) of the analysis. Hence, feature reduction can be carried out discarding the remaining PCs.

Stepwise algorithm for feature selection in classification (SW) consists on a greedy hill climbing approach where the subset of features with the highest performance measure will be selected in each step and modified in the next step by the addition or deletion of one variable in the model. ReliefF algorithm for Recursive Elimination of Features (ReliefF) algorithm is a feature selection method based on how well features distinguish between instances that are near to each other [110]. In classification problems, the estimation of the quality of each variable is calculated by the accumulation of the distance between randomly selected instances and their K-nearest neighbors of a different class minus the distance to the K neighbors of the same class.

Coming from signal processing, the goal of ICA is to extract source signals when only a linear mixture of these source signals is available. The most commonly used assumption is that the sources are non-Gaussian and mutually statistically independent, as well as independent from the noise components [108].

Spectral Peak Integration (PI) is a knowledge-based feature extraction method that integrates the area under the peaks of the most relevant metabolites as a representation of the significant information contained in the spectra. To obtain the areas under the peaks we have considered an interval of 0.15 ppms from the assumed peak centre.

Peak height of typical resonances is another knowledge-based feature extraction method that uses the height of the peak of the ppm where the metabolites are known to appear.

3.3.2 Classification methods

The classification methods applied were parametric discriminant methods: linear and quadratic discriminant analysis; kernel-based methods: support vector machines and least-squares support vector machines; connectionist models: multilayer perceptron and bi-directional Kohonen networks; and a memory-based method, K-nearest neighbours. Some of these methods were applied to the full region of interest represented by a data vector of 190 points, and to the extracted features. The K-nearest neighbour was used with local feature reduced by PCA only.

The parametric Gaussian Discriminant Analysis techniques are designed to find a discriminant functions for each available class assuming that the classes follow a multivariate Gaussian distribution. The estimation of the parameters of the Gaussian distributions is based on the maximum-likelihood estimation. The most popular method is Linear Discriminant Analysis (LDA), which is based on the assumption of a common variance of the classes. In the Quadratic Discriminant Analysis (QDA) the covariances of the classes are independent, obtaining quadratic decision boundaries (see Appendix A). The Fisher's LDA (FLDA) is a reduced-rank version of LDA, which projects the variables into the lower-dimensional subspace that maximizes the rate of the between-variance and the within-variance on the training dataset.

The K-Nearest Neighbors (KNN) is a non-parametric classification method where the samples are assigned to the most frequent class among their neighbors based on the distances of the test cases to the training corpus in the feature space. KNN is a type of instance-based learning where the function is approximated locally and all computation is deferred until classification [33].

The Multilayer Perceptron (MLP) is a connectionist model consisting on a network of simple units or perceptrons [113], typically called neurons. One neuron computes an output as a non-linear function of the inner-product of the feature vector $\mathbf{x}$ and parameter vector $\mathbf{w}$, called the weight vector. In a MLP, the input signal forwards layer-by-layer obtaining an approximation of the probability distribution of each class. During the training phase, the weight parameters are updated by means of an error-backpropagation algorithm from the output to the previous layers.

The Bi-Directional Kohonen (BDK) network [114] is another type of connectionist model based on a supervised version of the Kohonen network [115]. Each neuron in the Kohonen map (a two-dimensional map of neurons) is associated with a weight vector which is adapted iteratively by some learning function, based on the properties of the objects. Individual objects are iteratively presented to the units in the network and the weight vector that is most similar to the particular object is assigned to be the winning unit. The winning unit and its neighborhood are then adjusted to become more similar to the particular object. The final Kohonen map then represents the structure of the data in an interpretable way. In the supervised Bi-directional Kohonen Networks (BDK) two separate

maps are updated, namely the input map (representing the features of the objects) and the output map (representing the class labels of the objects). The winning unit for an instance is mainly determined by the similarity of the target and the output unit. When convergence is achieved, after presenting each object to the network multiple times, the input and output map structures can be used to classify new unidentified objects.

Support Vector Machines (SVM) [116] are classification methodologies that define the optimal separating hyperplane between two classes with the maximal margin. This margin is the minimum distance of patterns of the training set to the hyperplane. SVM represent data in a higher dimensional space where the linear separating hyperplane is built. The explicit construction of a mapping to a higher dimensional space is avoided by using the kernel trick [116]. Least-Squares Support Vector Machines (LSSVM) is a reformulation of the SVM resulting in the solution of a linear system [99].

3.4 Contributions in automatic brain tumour classification

One contribution was to improve the performance of automatic brain tumour classifiers by combining the LTE and STE spectra of each patience in order to take advantage of the complementary views of the chemical composition of brain tumours that each TE offers. Despite the great effort and the contributions in automatic brain tumour classification, nobody had proposed a combination of both TE spectra for the development of a single brain tumour classifier before. A remarkable exception was the work carried out by Majos *et al.* [117], where they performed a clinical comparison between the STE and LTE discrimination capacity and pointed out the potential interest of combining both times of echo. A second contribution was related to the *external validation* of automatic brain tumour classification models to assess generalization [15]. The raw MR data acquired during INTERPRET were incorporated into the eTUMOUR dataset for classifier development. This provided a unique opportunity to evaluate INTERPRET-based models by means of cases of a later date from partly different hospitals with different instrumentation, but obtained using the same or compatible acquisition protocols. The multiproject-multicenter evaluation included in this Dissertation gives a close-up perspective of the conditions that predictive models may face under different real clinical environments, and points out the possibility of using incremental learning algorithms to adapt one model to the characteristics of different centers.

3.4.1 Combination of Long and Short Time of Echo

The combination of the information provided by the two different times of echo for automatic classification of brain tumours has been a new contribution in the PR approach for brain tumour CDSS. Specifically, the author of this Thesis contributed to the development of Fisher's Linear Discriminant Analysis (LDA) for the classification of three types of brain tumour. We were interested in obtaining a combination of the STE and the spectra without introducing any prior restriction or assumption of relationship between them. Therefore, it was considered that the a concatenation of the D_{ste} points from the STE spectrum vector followed by the D_{lte} points of the LTE spectrum vector was the

Table 3.1: Results for the Fisher's LDA for the Combined TE, the STE, and the LTE

Dataset	Accuracy (%)	Confidence Interval (%, $\alpha = 5\%$)
Short TE	88.8	[83.7, 92.8]
Long TE	82.5	[76.6, 87.5]
Combined TE	88.7	[83.6, 92.7]

most direct approach. This joint vector is treated as a $(D_{ste} + D_{lte})$-dimensional vector-valued observation of the distribution of the diagnosis. Then, the discrimination functions may choose simultaneously among the features from both spectra to solve the proposed prediction model. As a result, the Combined TE dataset of 185 samples with 380 data points was obtained. To compare the combined approach with single approaches, we also generated the STE dataset composed by 185 samples with the 190 values in the region of interest of the STE spectrum, and the LTE dataset of 185 samples with the 190 values in the region of interest of the Long TE spectrum.

The LDA technique has been successfully applied in many biomedical applications, including Decision Support Systems based on MRS for brain tumor diagnosis [11, 118]. This method was applied to the aforementioned datasets. One of the advantages of the method is the possibility of plotting the latent space where the variables are projected. When a multi-class task of three classes is solved, the latent space is bi-dimensional (2D), and it could be used to visualize the projection of the samples in a 2D plot. Before applying LDA, the input space should be reduced in a proper way; for this study we used two methods, the Stepwise algorithm for feature selection in classification (SW) algorithm and the PCA for feature selection and extraction.

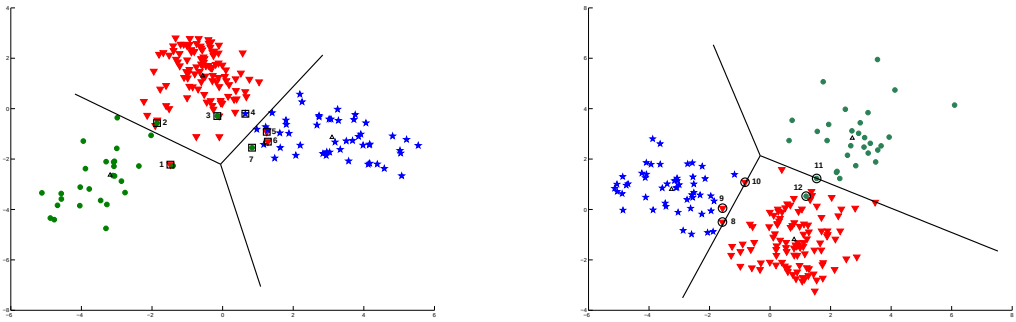
A kRSTT with stratified test sets with 150 repetitions was the evaluation procedure used for all the reported experiments. The partitions for repetitions were random and independent among the experiments with the training set composed by 70% of cases of each class. The evaluation was carried out in a nested-loop that covered the feature and model selection in order to avoid underestimation of the true error.

The best models were obtained using SW and Fisher's LDA. The developed multiclass classifiers discriminate among the three aforementioned superclasses (Aggressive tumor: GBM and MET (AGG), Low-grade meningiomas (MEN), and Low-Grade Glial (LGG)) simultaneously. A SW followed by LDA was applied to compare the Combined approach with STE or LTE based classifiers. Table 3.1 shows the kRSTT evaluation of the SW+LDA approach of the multiclass classifiers applied on the Combined TE, the STE, and the LTE datasets. Based on the predictions achieved by the different classifiers, some of the cases from the datasets were singled out for potential critical review by experts. In addition, Figure 3.3 shows the latent space of the Fisher's LDA.

Furthermore, three pairwise classification models were developed and evaluated using LS-SVM for automatic classification of AGG, MEN, and LGG brain tumours. Again, the evaluation was carried out with a kRSTT methodology. The results for these classifiers are shown in Table 3.2.

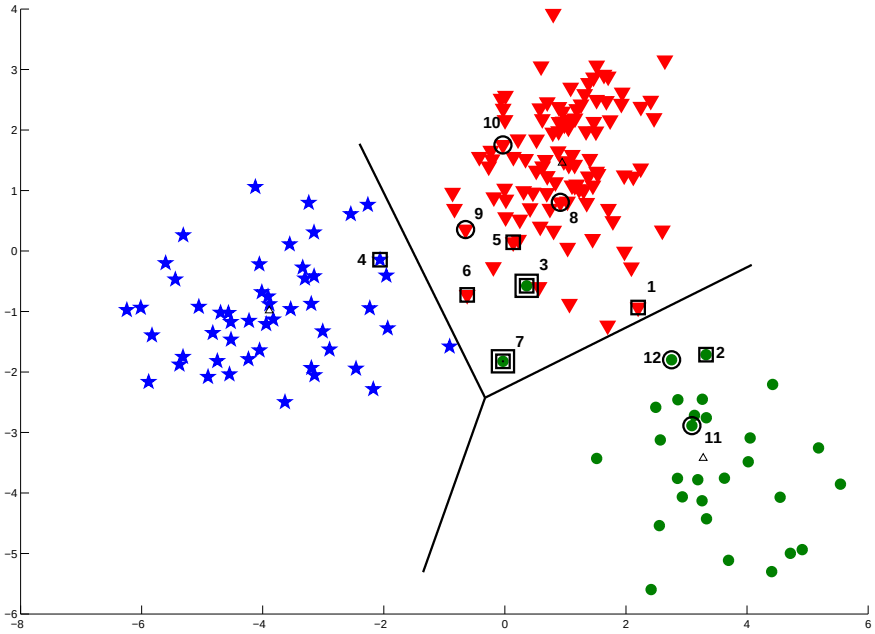
3.4.2 Multicenter evaluation of automatic BT classifiers

For the prospective multicenter evaluation of automatic brain tumour classifiers, six pairwise classifiers for Glioblastoma (GBM), MEN, Metastases (MET), and LGG diagnoses



(a) Short Time of Echo

(b) Long Time of Echo



(c) Combination of both Times of Echo

Figure 3.3: The latent space projection of the Fisher's LDA models for the (a) STE, (b) LTE, and (c) Combined TE. The Figure shows that the combination of both times of echo reaches a better latent space data projection.

were developed and tested on Single voxel (SV) STE MRS signals.

The aforementioned feature extraction methods were applied to the real part of the spectra prior to any classification approach. These methods also included direct spectral Peak integration (PI) on selected metabolite resonance regions, and Peak height of typical resonances (PPM), as well as the full region of interest represented by a data vector of 190 points (190). The selected features for the classifiers were derived from previous studies [98, 25] or from model validation based on the training dataset.

To develop the pairwise models, ten methods were applied: parametric discriminant analysis (LDA, FLDA, and QDA); Kernel-based models (SVM and LSSVM); Artificial

Table 3.2: kRSTT evaluation of the LS-SVM for pairwise classification of AGG, MEN and LGG classes. The percentage of LTE features selected by ReliefF with respect to the total number of features is shown in brackets in the features columns.

Task	Dataset	Features	Accuracy [CI] (%)	AUC
AGG vs. MEN	Combined TE	380 [LTE:50%]	95.3 [91.2,97.8]	0.992
	Short TE	100	92.6 [87.8,96.0]	0.982
	Long TE	190	92.2 [87.3,95.7]	0.975
AGG vs. LGG	Combined TE	10 [LTE:0%]	92.6 [87.3,96.1]	0.970
	Short TE	10	92.1 [86.7,95.7]	0.966
	Long TE	10	90.5 [84.8,94.6]	0.95
LGG vs. MEN	Combined TE	50 [LTE:42%]	97.5 [92.6,99.3]	0.996
	Short TE	50	96.0 [90.3,98.7]	0.993
	Long TE	100	94.5 [88.2,98.0]	0.993

Neural Networks (MLP and BDK); and single and ensemble classifiers using K-nearest neighbours and local feature reduced by PCA (PCA-KNN) were used. Specifically, the author of this Thesis contributed to the development, validation and evaluation of the MLP models and with the critical discussion of the results of the rest of the developed models.

For each task, different combinations of feature extraction and classification methods were applied in the study. An estimation of the error (ERR) and Balanced Error Rate (BER) for the INTERPRET dataset using a 10-fold CV was carried out for each model. Then, the models followed an independent prospective evaluation using the eTUMOUR database.

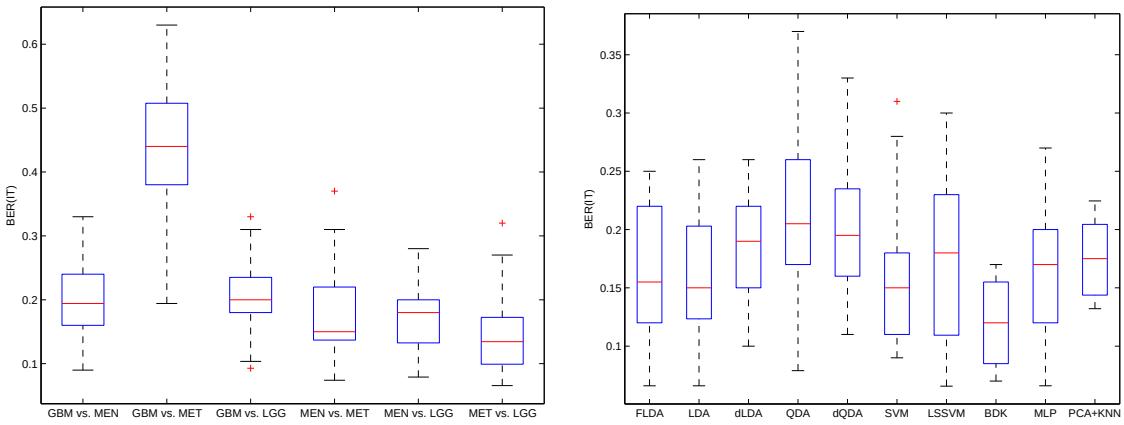
After estimating the models using a 10-fold cross-validation evaluation with the INTERPRET database, the estimation of the BER were obtained on the independent test dataset of eTUMOUR. Figure 3.5 shows the results with all the pairwise classifiers based on both the CV and the prospective independent test data.

For each binary classifier a multilayer perceptron (MLP) was trained. After the preprocessing of the data, a region of interest between 4.1 and 0.5 ppm's was selected. This means that 190 features were considered as the input of each MLP. After this region selection the data were normalized in order to scale the data between 0 and 1. The output consisted of two units, one for each class.

As a first step, the topology of each binary classifier and the parameters for training the nets were selected. In order to simplify the models without losing its predictive power only one hidden layer was used. For this single hidden layer 4, 8, 12 and 16 hidden units were tested. The algorithm for training the nets was the backpropagation with momentum. The learning rate, ρ, varied between 0.01 and 1 and the momentum factor, μ, varied between 0 and 0.5. Each feed-forward neural network was trained doing an exhaustive scan of these parameters and its error was estimated using a 10-fold cross validation evaluation.

In order to avoid overfitting, a stopping criterion had to be established. In this experiments, overtraining was avoided by stopping the learning procedure when the mean square error of the training set converged, i.e., when the learning process began to slow down rapidly [119]. In this study, learning began to slow down after 100-200 training cycles.

A summary of the results of all the models are shown in Figures 3.4 and 3.5.



(a) Pairwise classification problems

(b) Classification methods

Figure 3.4: A summary of the results obtained for each set of models regarding (a) each pairwise classification problem and (b) each type of classification method (dLDA and dQDA stands for a LDA with diagonal covariance matrix and a QDA with diagonal covariance matrices, respectively). The performance is measured with the BER.

3.5 Discussion and conclusions

For SW-based multiclass classifiers, the comparison between the use of the combined TE and the single STE show that both approaches are similar. However, the accuracy of the LTE approach is considerably lower with a significant p-value using a Friedman's non-parametric two-way analysis of variance test ($\alpha = 0.05$). Furthermore, in García-Gómez *et al.* [25], we showed that there are significant differences among the three approaches when pairwise classifications are delivered.

The classifiers developed from the INTERPRET dataset seem to be robust enough for predictive classification of prospective cases from eTUMOUR. We can conclude, from the multicenter evaluation, that accurate classification of new cases is feasible using data acquired in a mixed set of different hospitals, with different instrumentation, but similar acquisition protocols. The pairwise discrimination between Glioblastoma, Meningioma, Metastasis, and Low-grade Glial achieved accuracies of around 90%. However, the discrimination of Glioblastoma and Metastasis did not achieve a result better than 78% accuracy. Our results consolidate the conclusions of previous studies on automatic brain tumor classification using MRS but with multiproject-multicenter data for training and subsequent test.

An interesting conclusion from [24] is that the use of PI is comparable in terms of discriminative power to other feature extraction algorithms. Furthermore, in Menze [14] and Luts [120] it is shown that PI has also a comparable performance than using quantitation methods for estimating the concentration of the metabolites. Since PI is faster to calculate and easier to implement, these results justify the use of PI in the following models carried out using the incremental learning algorithms developed in this Thesis.

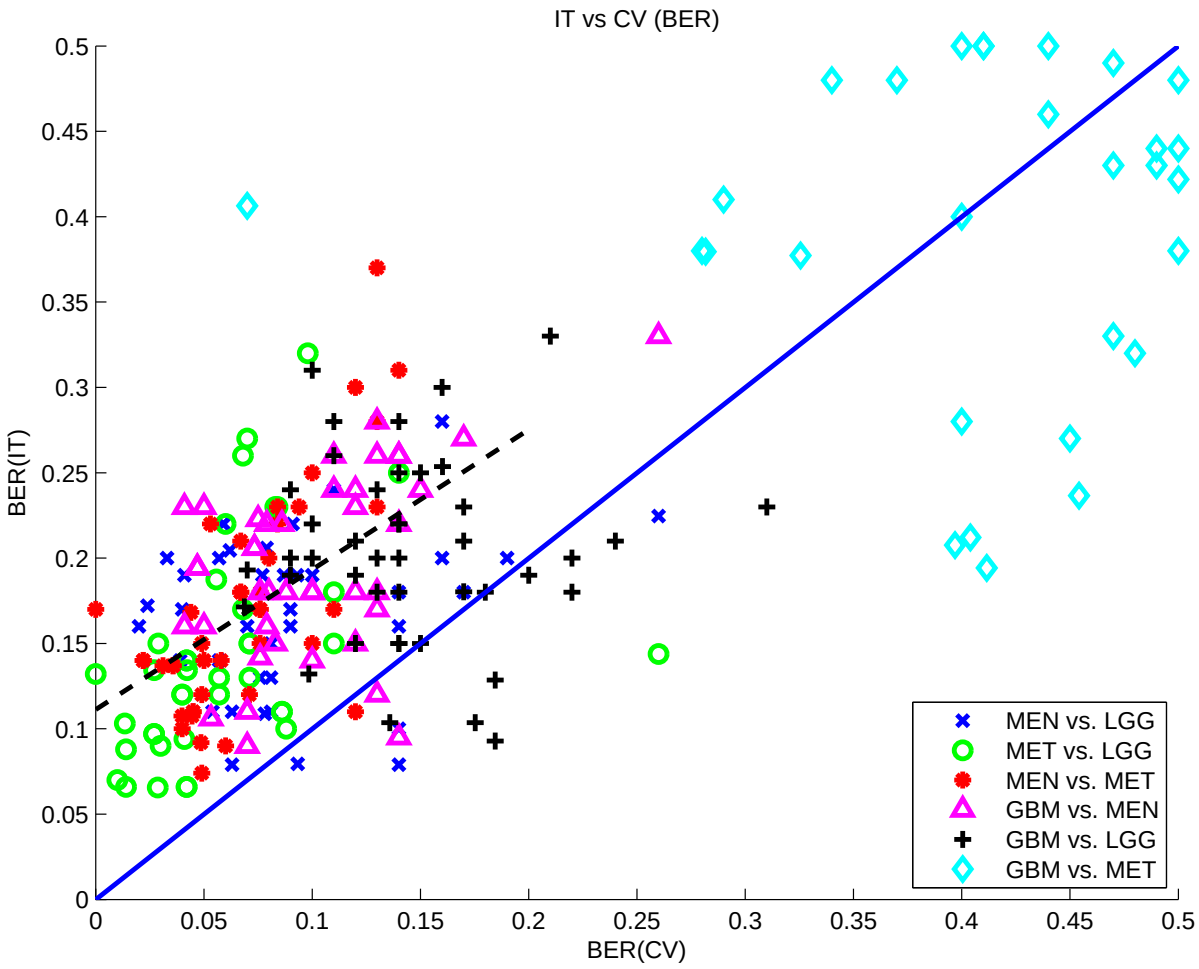


Figure 3.5: Scatter plot of the performance measured in BER estimated by the IT set consisting of new eTUMOUR cases and the BER estimated by the CV using the INTERPRET cases. The solid-blue line represents that BER(IT) and BER(CV) are equal. The black-dasehd line represents the trend of the cloud of points where BER(CV)$<$ 0.2 and BER(IT)$<$ 0.3. We can observe from the trend that there has been a subestimation of the error of the models since BER(IT) is almost always higher than BER(CV). This bias from the models trained with the INTERPRET dataset and evaluated with the eTUMOUR dataset points out that some differences may exist from one dataset to the other. One possible solution to avoid this bias could be to merge both datasets and develop new models from scratch. This solution may be useless unless we could evaluate the models with a different independent test set. A different solution is to apply an incremental learning algorithm using small subsets of new samples to develop incremental models provided they become available in the course of time.

There are still many remaining challenges in brain tumor classification by ^{1}H MRS. One of the most important is the limited number of available spectra per tumor type [94], which enables the use of incremental learning algorithms as a practical solution. The specific epidemiological distribution of tumors, with extremely low prevalent classes, and the increasing recognition of brain tumor molecular subtypes seems to be the reason for that limitation [121]. However, the development of robust brain tumour classifiers requires a large number of cases to be acquired for each tumour type. In standalone CDSS, the data is gathered from a large number of hospitals over many years and data is transferred to a centralized database. As explained in chapter 1, ethical approval and patient consent needs to be obtained to send and store data. Furthermore, to expand the applicability of ML techniques to MRS of a wider range of tumours, more cases need to be collected over a more prolonged period of time. This challenge has proven to be difficult to overcome using centralized databases. As an alternative, distributed databases allow the possibility to train new classification models without moving the data from the hospital at which it was collected. Hence, the ability to retrain the classifiers as new data accumulates is also an important requirement.

Furthermore, as shown in Figure 3.5, there may appear a bias between different health centers. The development of new models from scratch may be a useless solution unless we could evaluate the models with a different independent test set, which would entail the effort of obtaining new observations.

The aforementioned reasons justify the application of an incremental learning algorithm using small subsets of new samples to develop incremental models provided they become available in the course of time. For this purpose, two different incremental learning algorithms are proposed in the following chapters.

Chapter 4

Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis

In the last decade, ML techniques have been used for developing classifiers for automatic brain tumour diagnosis. However, the development of these ML models rely on a unique training set and learning stops once this set has been processed. Training these classifiers requires a representative amount of data. However, the gathering, preprocess, and validation of samples is expensive and time-consuming. Therefore, for a classical, non-incremental approach to ML, it is necessary to wait long enough to collect all the required data. In contrast, an incremental learning approach may allow us to build an initial classifier with a smaller number of samples and update it incrementally when new data are collected. In this Chapter, we introduce an Incremental Gaussian Discriminant Analysis (iGDA) learning algorithm based on the Graybill and Deal weighted combination of estimators. Each time a new set of data becomes available, a new estimation is carried out and a combination with a previous estimation is performed. iGDA does not require access to the previously used data and is able to include new classes that were not in the original analysis, thus allowing the customization of the models to the distribution of data at a particular clinical center. An evaluation using five benchmark databases has been used to characterize the behaviour of the iGDA algorithm in terms of stability-plasticity, class inclusion and order effect. Finally, the iGDA algorithm has been applied to automatic brain tumour classification, and compared with two state-of-the-art incremental algorithms. The empirical results obtained show the ability of the algorithm to learn in an incremental fashion, improving the performance of the models when new information is available, and converging in the course of time. Furthermore, the algorithm shows a negligible instance and concept order effect, avoiding the bias that such effects could introduce.

This Chapter has been published in the Journal of Biomedical Informatics in [28].

4.1 Introduction

In this work, an ML-based method is proposed to continously adapt an automatic brain tumour diagnosis model to reflect the most recent information included in newly acquired cases. An incremental learning algorithm based on a weighted combination of Gaussian parameter estimation is presented for automatic brain tumour diagnosis. Our method relies on the Graybill and Deal combination of unbiased estimators [32, 122] originally developed for the estimation of a common mean when several sets of data come from different measurement methods or different laboratories. The Graybill-Deal estimator is known to be unbiased for the mean [122, 123]. In this Thesis, it has been applied to discriminant analysis to develop a straightforward method for updating the parameters of each class when new observations arrive, adjusting the parameters of the model to incorporate new classes in the discriminant space when needed, and showing a benign order effect at instance and concept level. Some benchmark experiments have been carried out to show these issues and, finally, the incremental algorithm has been applied for brain tumour diagnosis.

4.2 Methods

The formal purpose of classification is to assign instances to one class among $|\mathcal{C}|$ possible classes based on a set of features obtained from each observation. A decision rule $\mathcal{M}$ is a function that maps an object $\mathbf{x} \in \mathbb{R}^d$ into a class $c \in \mathcal{C}$. An error is incurred if the decision rule assigns the instance to a wrong class. The final objective is to minimize the error for discriminating among different classes. In discriminant analysis, each class is represented by a function $g_i(\mathbf{x}), i = 1, \ldots, |\mathcal{C}|$. A classifier $\mathcal{M}(\mathbf{x})$ assigns the class c_j if $g_j(\mathbf{x}) > g_i(\mathbf{x}), \forall j \neq i$. When a 0-1 loss function is used, finding the class that maximizes the log-likelihood of the posterior probability $p(c|\mathbf{x})$ is equivalent. Using Bayes' rule, assuming that the density functions follow a multivariate normal, $p(\mathbf{x}|c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, and taking into account that the prior probabilities are parameters to be estimated, then the expression can be evaluated using

$$g_c(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{W}_c\mathbf{x} + \mathbf{w}_c^{\mathrm{T}}\mathbf{x} + w_{c0} \ , \tag{4.1}$$

where $\mathbf{W}_c = -\frac{1}{2}\boldsymbol{\Sigma}_c^{-1}$, $\mathbf{w}_c = \boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c$ and $w_{c0} = \log \pi_c - \frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c$. The mean, $\boldsymbol{\mu}_c$, the covariance matrix, $\boldsymbol{\Sigma}_c$, and the prior probabilities, π_c, of the class c are the parameters that can be estimated by the maximum-likelihood method using a set of labeled samples [33].

This decision rule divides the sample space into $|\mathcal{C}|$ decision regions. The points $\mathbf{x}$ of the sample space which satisfy that $g_j(\mathbf{x}) = g_i(\mathbf{x}), i \neq j$ make the decision boundary as explained in Chapter 2. These discriminant functions describe quadratic decision boundaries except when the covariance matrices of all the classes are identical. If a common covariance matrix is used the quadratic terms of Equation (4.1) cancel giving rise to a linear boundary. Linear and quadratic versions are both available in the proposed incremental algorithm.

4.2.1 Graybill-Deal combination of estimators

Given k sets with N_i instances $x_1, \ldots, x_{N_i}$ in each set, for $i = 1, \ldots, k$, it is possible to estimate the common mean of the population using a weighted mean, where the weights w_i depend on the number of instances and the population variance, provided that all the variances are known. When the true variance is not known, the sample variance is used instead. In this case, the weighted mean is computed by

$$\bar{X}_{GD} = \sum_{i=1}^{k} \hat{w}_i \bar{X}_i \ , \tag{4.2}$$

where $\bar{X}_i$ is the mean value of the i-th set. The weights are calculated using the sample variance as

$$\hat{w}_i = \frac{N_i / S_i^2}{\sum_{j=1}^{k} N_j / S_j^2} \ , \tag{4.3}$$

where S_i^2 is the sample estimate variance of the corresponding set. Notice that the estimation given by (4.3) gives higher weight to those sets with larger number of instances N_i and smaller variance S_i^2. Graybill and Deal [32] demonstrated that the estimation of the mean μ using $\bar{X}_{GD}$ is unbiased, that is $E[\bar{X}_{GD}] = \mu$.

It is trivial to extend this result to the combination of the mean for multivariate samples. However, the need of a combined estimator for the covariance matrix of the samples presents a harder challenge. In the next subsection, a solution is proposed as part of the developed incremental algorithm.

4.2.2 Incremental Gaussian Discriminant Analysis based on Graybill-Deal estimation of weights

Let the training dataset be obtained in different samples, $\mathcal{S}_i$, that are available in times $t_i, i = 1, \ldots, T$. The Graybill-Deal estimation for Gaussian discriminant analysis begins with the adjustment of the parameters for each class based on maximum log-likelihood using the first set of samples. Hence, the prior probabilities for each class, $\pi_c^{(1)}$; the mean vector, $\boldsymbol{\mu}_c^{(1)}$; and the covariance matrices, $\boldsymbol{\Sigma}_c^{(1)}$ are estimated following the usual maximum log-likelihood estimation [33].

A model $\mathcal{M}_i$ is composed of a mean $\boldsymbol{\mu}_c^{(i)}$, a covariance matrix $\boldsymbol{\Sigma}_c^{(i)}$, a prior probability $\pi_c^{(i)}$, and the number of instances $N_c^{(i)}$ of each class. The first iteration of the algorithm estimates the parameters of the first model. When a new dataset $\mathcal{S}_i$ is available in time t_i, $i > 1$, the first step is to carry out a new parameter estimation from $\mathcal{S}_i$, where $\hat{\boldsymbol{\mu}}_c$, $\hat{\boldsymbol{\Sigma}}_c$, π_c, and N_c are calculated for each class. Then, the prior probabilities of the new model $\mathcal{M}_i$ are updated based on the number of samples per class:

$$N_c^{(i)} = N_c^{(i-1)} + N_c \ , \tag{4.4}$$

$$N^{(i)} = N^{(i-1)} + \sum_c N_c^{(i)} \ , \tag{4.5}$$

$$\pi_c^{(i)} = \frac{N_c^{(i)}}{N^{(i)}} \ , \tag{4.6}$$

where $N_c^{(i)}$ is the number of instances of $\mathcal{S}_i$ and the class c. The new mean and a weighted covariance matrix are calculated as

$$\boldsymbol{\mu}_c^{(i)} = w_c^{(i)} \boldsymbol{\mu}_c^{(i-1)} + (1 - w_c^{(i)}) \hat{\boldsymbol{\mu}}_c \ , \tag{4.7}$$

$$\boldsymbol{\Sigma}_c^{(i)} = w_c^{(i)} \boldsymbol{\Sigma}_c^{(i-1)} + (1 - w_c^{(i)}) \hat{\boldsymbol{\Sigma}}_c \ , \tag{4.8}$$

where $w_c^{(i)}$ and $(1 - w_c^{(i)})$ are the weights for updating the parameters of class c. The weights are calculated using the Graybill-Deal combination of estimators (4.3) and they are subject to $0 \leq w_c^{(i)} \leq 1$ and $\sum_i w_c^{(i)} = 1$. We propose to adapt the variance to multivariate distributions by means of the total variation which is the sum of the variances $S_i^2 = tr(\boldsymbol{\Sigma}_i)$, which is equivalent to the sum of the eigenvalues of the covariance matrix.

After estimating and incrementally updating the parameters, the common covariance matrix can be used to obtain a linear discriminant instead of a quadratic discriminant as previously explained. The pseudocode for the iGDA algorithm is given at the end of the Chapter.

One interesting property is that the algorithm allows the possibility of introducing new classes if required. Therefore, if a new set of samples includes data from a new class, an estimation of the additional parameters is carried out. The prior probabilities are updated according to the new data set and the parameters of the new class are retained within the model, thus modifying the final decision boundaries and the regions described for each class. This is due to the generative model approach followed by the algorithm.

4.2.3 Comparison with other algorithms

Learn^{++} is a well-known incremental learning algorithm proposed by Polikar in [60]. Learn^{++} is inspired by the AdaBoost algorithm [124], which was developed to improve the classification performance of weak learners[a]. Schapire [125] showed that a *weak learner* can be transformed into a *strong learner* using a *boosting* procedure.

Learn^{++} uses the concept of boosting to incrementally improve the performance of the classification. In contrast with AdaBoost, Learn^{++} does not extract the subsets from the same training set but from the successive observations available throughout time. Learn^{++} uses a weak learner to generate multiple hypotheses from different subsets of data. Therefore, each hypothesis learns only a portion of the input space. The weak learner is based on a perceptron, thus each hypothesis defines a linear hyperplane as a decision boundary. When the algorithm learns with a new set of samples, it generates a new set of hypotheses. The outputs of all the hypotheses are combined using a weighted majority voting. Therefore, Learn^{++} does not require access to previously used data during the incremental learning and it does not forget previously acquired knowledge.

Another well-known incremental learning algorithm is the incremental Linear Discriminant Analysis (iLDA) proposed by Pang et al. [68]. iLDA uses a constructive method

[a] A *weak learner* is a learning algorithm that performs slightly better than random guessing.

for deriving an updated discriminant eigenspace for classification. A typical Linear Discriminant Analysis (LDA) seeks directions in the D-dimensional space that are efficient for discrimination, projecting the observations to a P-dimensional space where $P < D$. To obtain the projection matrix $\mathbf{W}$, the ratio between the *between-class* scatter matrix $\boldsymbol{S}_b$ and the *within-class* scatter matrix $\boldsymbol{S}_w$ must be maximized. Once the observations are projected, different ML techniques can be used for classification purposes [34].

The iLDA method aims to obtain a new discriminant eigenspace model Φ by combining two discriminant eigenspace models Ω_t and Ω_{t+1} from different samples $\mathcal{S}_t$ and $\mathcal{S}_{t+1}$ acquired at time t and $t+1$ respectively. This new model, Φ, updates the sample mean, the $\boldsymbol{S}_w$ matrix and the $\boldsymbol{S}_b$ matrix and results in a new projection matrix $\mathbf{W}$. Once the data are projected in the new discriminant eigenspace, a nearest neighbour algorithm is used for classification purposes. For technical details see [68]. iLDA does not require access to previously seen data and it can also include new classes if needed.

Finally, a naive incremental Gaussian model is used as a baseline for comparison with the above methods. This model updates its parameters from scratch. That is, the previous data and the current data are used to train a new model using quadratic discriminant analysis [33] (see Appendix A).

4.3 Benchmark experiments

The behaviour of the iGDA algorithm has been tested on several databases with a threefold purpose: 1) to show that the developed algorithm is able to incrementally learn and adapt the parameters of the classifier, improving its performance without incurring in catastrophic forgetting; 2) to show how the iGDA algorithm is able to introduce new concepts or classes into its knowledge representation; 3) to analyze whether the order in which the instances are introduced into the analysis have a crucial influence in the final hypothesis, that is, if the algorithm is order dependent or not. The selected datasets have only real attributes since the iGDA is restricted to that set of numbers. In order to avoid possible bias, every experiment was evaluated following a K random sampling train-test strategy, where $K = 100$.

4.3.1 Stability/Plasticity dilemma

Vehicle Silhouette Database

The vehicle silhouette database has been extracted from the UCI Machine Learning Repository [126]. The purpose of this database is to classify a given silhouette into one of four different types of vehicle using a set of 18 features. The database consisted of 846 instances. It was divided into a training partition (630 instances) and a test partition (216 instances). The training partition was split again into 7 training sets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ of 90 instances with a similar prevalence to the original database for each class. Table 4.1 shows that there is a gradual loss of information relating to the previous training datasets when new observations are introduced using the quadratic iGDA. However, the overall performance increases from 62% to 84%. The linear iGDA showed an increase from 73% to 78%, also with a gradual forgetting when new information was added (Table not shown). These results are comparable to the performance of a completely new quadratic classifier trained with the entire training dataset (85%) and to a linear classifier (80%).

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$
$\mathcal{S}_1$	99.93	97.42	95.48	94.22	93.22	92.57	92.11
$\mathcal{S}_2$	–	97.43	95.04	93.52	92.90	92.26	91.86
$\mathcal{S}_3$	–	–	95.31	94.16	93.18	92.53	91.90
$\mathcal{S}_4$	–	–	–	94.24	93.38	92.64	92.08
$\mathcal{S}_5$	–	–	–	–	92.68	92.06	91.54
$\mathcal{S}_6$	–	–	–	–	–	92.27	91.79
$\mathcal{S}_7$	–	–	–	–	–	–	91.73
TEST	62.00	79.08	81.53	82.82	83.62	84.09	84.54
CI ($\alpha = 1\%$)	±1.44	±0.71	±0.62	±0.65	±0.65	±0.63	±0.59

Table 4.1: Training and test accuracy for the Vehicle Silhouette Database using a quadratic iGDA. The rows indicate the different datasets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ and the columns show the hypothesis or models $\mathcal{M}_j$ built from a previous model $\mathcal{M}_{j-1}$ and the new dataset $\mathcal{S}_j$, except $\mathcal{M}_1$ which is built from $\mathcal{S}_1$ only. Each column shows the average performance (%) on the current and the previous training datasets for the current model. The last rows (TEST, CI) indicate the evolution of the average accuracy of the models in the course of time evaluated with an independent test set and the confidence interval ($\alpha = 1\%$).

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
$\mathcal{S}_1$	69.11	99.14	98.16	97.77	97.44
$\mathcal{S}_2$	–	99.16	98.32	97.70	97.31
$\mathcal{S}_3$	–	–	98.24	97.64	97.25
$\mathcal{S}_4$	–	–	–	97.55	97.17
$\mathcal{S}_5$	–	–	–	–	97.39
TEST	52.21	94.12	94.95	95.20	95.34
CI ($\alpha = 1\%$)	±3.56	±0.48	±0.46	±0.43	±0.42

Table 4.2: Training and test accuracy (%) for the Wisconsin Breast Cancer Database using a quadratic iGDA.

Wisconsin Breast Cancer Database

The Wisconsin Breast Cancer Database from the UCI Machine Learning Repository consists of 569 instances with 30 variables from a digitalized image of a fine needle aspirate (FNA) of a breast mass. The objective in this problem is to classify the instances into a malignant (37.3%) or a benign (62.7%) breast tumour. The database was divided into a test partition (169 instances) and a training partition (400 instances) that were also split into five different sets of 80 instances $\mathcal{S}_1, \ldots, \mathcal{S}_5$. Each partition had the same prevalence for each class as the whole database. The results of the quadratic classifier are shown in Table 4.2. The linear iGDA also showed an improvement on accuracy: from 91.14% to 94.38% for the independent test set. As shown in the previous experiment, there is generally an improvement in overall classification as the new data are used for incremental learning, but a gradual forgetting is observed with respect to the previous datasets. The poor performance of the first classifier in the quadratic iGDA may be due to the low number of instances in the first dataset $\mathcal{S}_1$ and it is known that quadratic discriminant classification rules generally require larger samples than those based on linear discriminant analysis [127].

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
$\mathcal{S}_1$	88.89	90.53	90.82	90.73	91.16	91.60
$\mathcal{S}_2$	–	89.97	90.70	90.45	90.64	91.11
$\mathcal{S}_3$	–	–	69.58	87.50	88.93	89.12
$\mathcal{S}_4$	–	–	–	87.44	88.97	89.15
$\mathcal{S}_5$	–	–	–	–	62.40	84.76
$\mathcal{S}_6$	–	–	–	–	–	84.99
TEST	50.78	52.93	60.52	68.28	71.73	83.50
CI ($\alpha = 1\%$)	$\pm$0.60	$\pm$0.48	$\pm$0.61	$\pm$0.78	$\pm$0.62	$\pm$1.01

Table 4.3: Training and test accuracy (%) for the Concentric Circle Database using a quadratic iGDA.

4.3.2 Introduction of new classes

Concentric Circle Database

The concentric circle database is a synthetic set of five classes each belonging to a concentric ring of data. This database is used to test the ability of the incremental algorithm to introduce new classes. The data is bidimensional with a uniform distribution inside each ring (see Figure 4.1, left). The database was split into 6 different sets: $\mathcal{S}_1$ and $\mathcal{S}_2$ included 50 instances from each of classes 1, 2, and 3; $\mathcal{S}_3$ and $\mathcal{S}_4$ included 50 instances from classes 1 to 3 and 100 instances from class 4; finally, $\mathcal{S}_5$ and $\mathcal{S}_6$ contained 100 instances from classes 1 to 4 and 200 instances from class 5. Therefore, equal prior probabilities were kept for the number of instances of each class. An independent test set was generated with 10 000 instances from each class. In order to simulate the general behaviour of the algorithm in a real scenario, the test set included all the five classes. Since the database describes quadratic boundaries, only quadratic iGDA was employed (see results in Table 4.3).

As demonstrated in Table 4.3, iGDA has the ability to include new classes with an increase in overall classification performance for the test set as soon as data from new classes appear in the new datasets.

Image Segmentation Database

The Image Segmentation database from the UCI Machine Learning Repository consists of 2 310 instances with 18 attributes for segmenting the images from 7 outdoor images. The seven classes are: brickface, sky, foliage, cement, window, path, and grass. The database was split into three training subsets $\mathcal{S}_1$ (including classes brickface, sky and foliage), $\mathcal{S}_2$ (including all the classes except path and grass), $\mathcal{S}_3$ (including all the classes), and one test partition (231 instances) were all the classes were represented. The prior probabilities of all classes were made equal as for the previous experiment. The results for the linear version of the iGDA algorithm are shown in Table 4.4 and are comparable to that in Muhlbaier et al. [62], where the best improvement went from a 42.2% to a 91.0% after the third dataset. Although there was an improvement for the quadratic version, the results obtained were poor: from 22.2% to 58.8%.

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$\mathcal{S}_1$	98.45	99.69	99.78
$\mathcal{S}_2$	–	88.25	87.45
$\mathcal{S}_3$	–	–	94.63
TEST	42.14	64.30	91.42
CI ($\alpha = 1\%$)	±0.15	±0.41	±0.43

Table 4.4: Training and test accuracy (%) for the Image Segmentation Database using a linear iGDA.

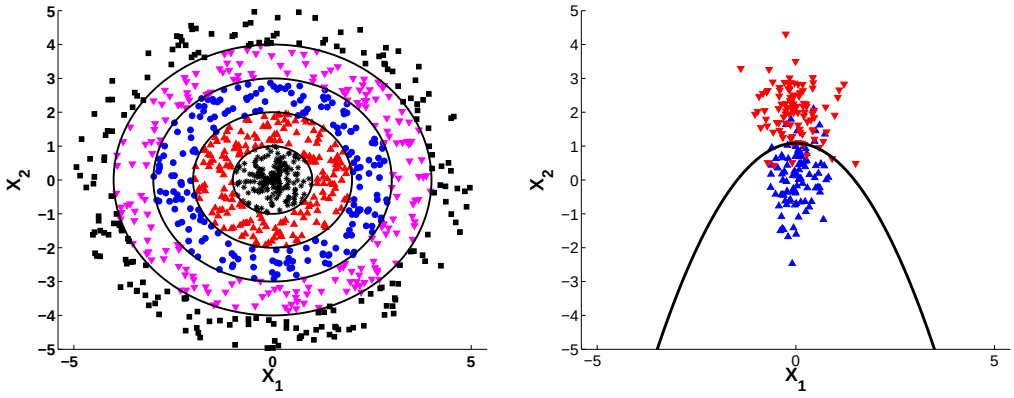


Figure 4.1: The Concentric Circle dataset is shown on the left. Five classes are drawn following a uniform distribution in their corresponding ring. Assuming Gaussian distributions the decision boundaries can be obtained. In addition, the two-dimensional synthetic dataset is shown on the right. The class c_1 follows $p(c_1|\boldsymbol{x}) \sim \mathcal{N}\left(\binom{0}{0}, \binom{1/8 \quad 0}{0 \quad 1/2}\right)$, and class c_2 follows a distribution $p(c_2|\boldsymbol{x}) \sim \mathcal{N}\left(\binom{0}{2}, \binom{1/4 \quad 0}{0 \quad 1/2}\right)$. The decision boundary is a parabolic curve.

4.3.3 Order effects

Instance level order effects

A synthetic dataset with two categories drawn from different multivariate normal distributions (shown right in Figure 4.1) has been used to analyze the instance level order effects. A training set of 400 instances and a test set of 4 000 instances were drawn from the distributions with equal prior probabilities for each category. The training set was split into 20 different training samples with 20 instances in each sample. The samples were used for incremental learning to build consecutive models as explained before. To evaluate the order effects, the instances were permuted in 100 experiments and the mean accuracies and the decision boundaries of the models of each iteration in the experiments were compared.

The Vehicle Silhouette database was also used to reinforce the analysis. The same configuration as in Section 4.3.1 was prepared, but the instances were permuted 100 times to test the effect of the instance order. Figure 4.2 shows the convergence in accuracy for these two experiments, whereas Figure 4.3 shows the iterative convergence of the decision boundary for the two-dimensional synthetic dataset.

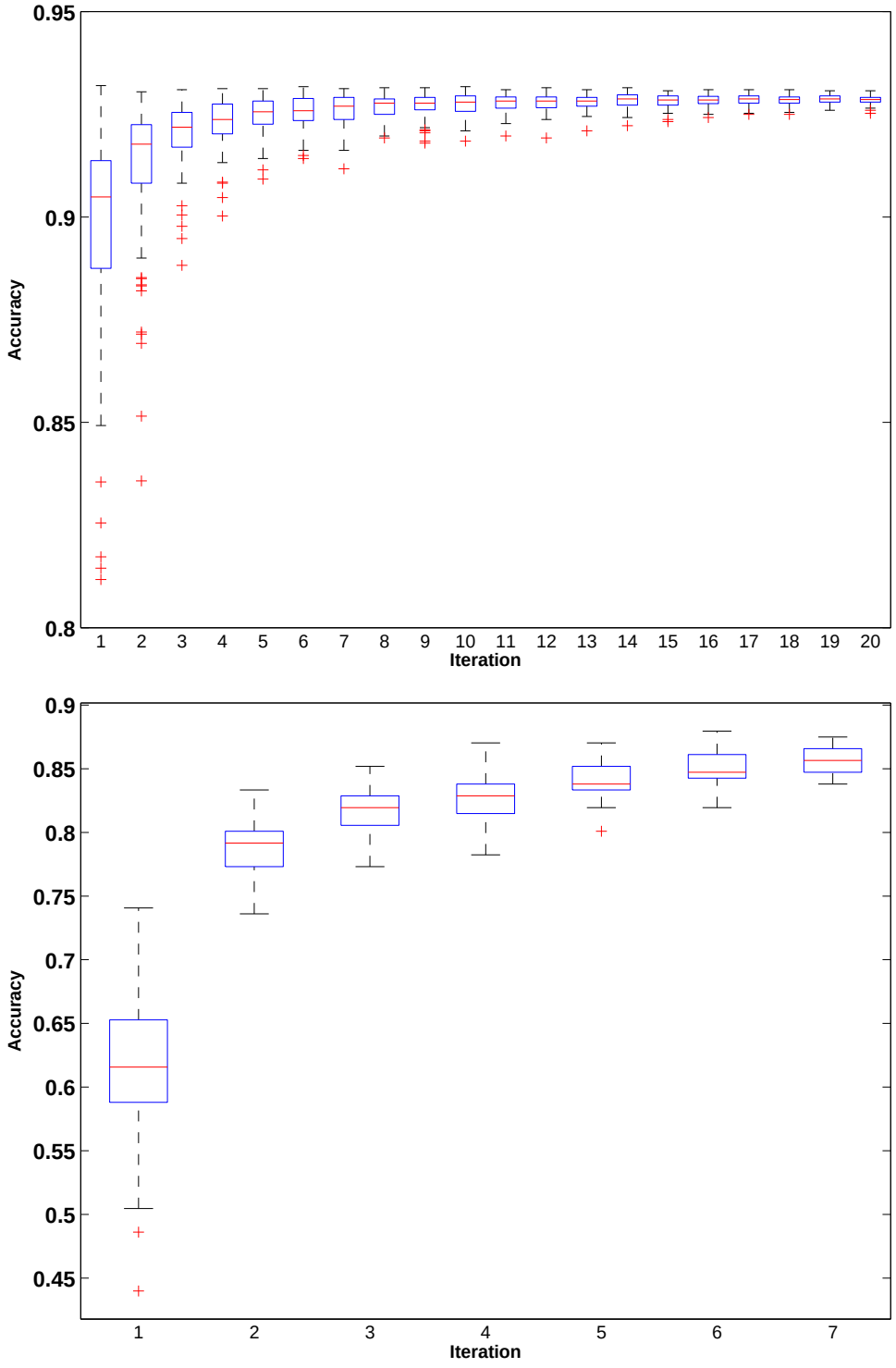


Figure 4.2: Boxplots of the accuracy of the models trained with different permutations of the instances. The X-axis shows the iterations of the incremental models. The top Figure shows the results for the two-dimensional synthetic database with 20 iterations. The bottom Figure shows the results for the Vehicle Silhouette database. The convergence of the accuracy proves that the instance order has a *benign* effect on the final models for both datasets.

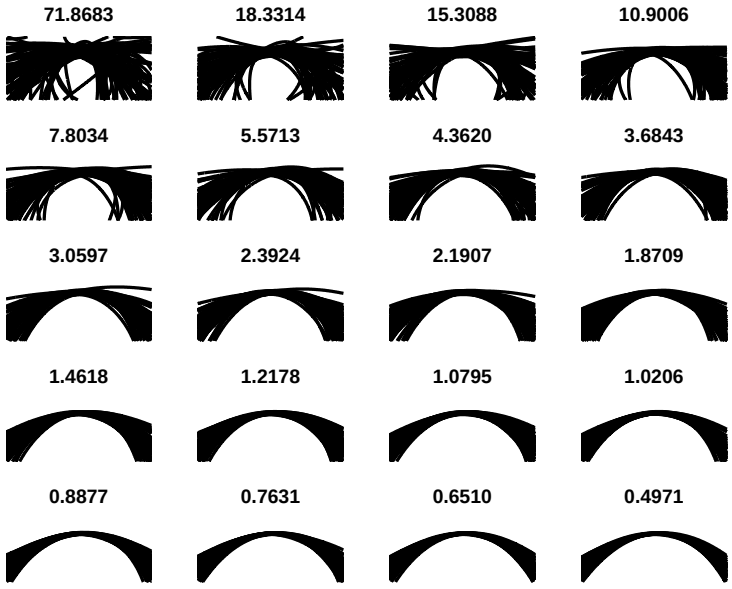


Figure 4.3: Convergence of the decision boundaries of each model in 20 iterations for the two-dimensional synthetic database. The variance of the different parameters of the decision boundaries is also shown at the top of each iteration. The iterations are shown left-to-right, top-to-bottom. It can be seen that the first models present arbitrary decision boundaries because their parameters are adjusted from the first sample only. When further samples are used for learning, the decision boundaries and their parameters begin to converge until the final iteration.

Concept level order effects

The Concentric Circle database was used to analyze the effect of the concept order on the iGDA. The database was divided into six different samples as in Section 4.3.2. To avoid the problem of imbalanced classes [128], the prior probabilities were forced to be equal. With this set-up of samples and classes and considering that there are five possible categories, the possible combinations for introducing different categories in each sample are 20. Therefore, 100 repetitions of 20 different combinations of samples were analysed. Figure 4.4 depicts the convergence of the incremental algorithm. The results show a *benign* concept lever order effect when the prior probabilities of the categories are equal.

4.4 Experimental design for brain tumour diagnosis

So far, the behaviour of the iGDA algorithm has been studied using different benchmark datasets with a focus on various properties. In this section, the iGDA algorithm is applied to a real biomedical problem of high medical relevance: automatic brain tumour classification with ^1H MRS. The current gold standard classification of a brain tumour is a histopathological analysis of biopsy; but this is an invasive surgical procedure with potential adverse consequences for the patient. An alternative is a diagnosis based on ^{1}H MRS, which is a non-invasive technique that provides biochemical information on tissue *in vivo*. The database used for our evaluation contains single voxel proton magnetic resonance spectra (SV ^{1}H MRS) acquired at 1.5T from brain tumours at nine European and

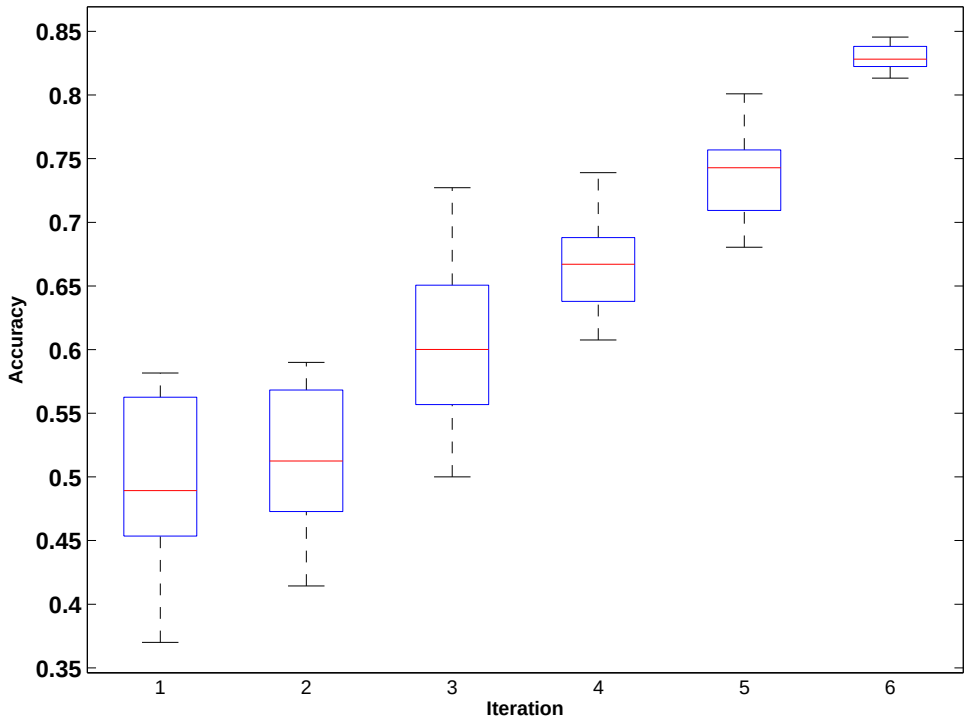


Figure 4.4: Convergence of the median accuracies of each combination of samples for the Concentric Circle database. The X-axis shows 6 iterations of the incremental models, each one corresponding to a sample $\mathcal{S}_i$. The convergence proves that the concept order has a slight effect on the accuracy of the models.

one Argentinian hospitals. Data used in this work was gathered during three European projects: INTERPRET, eTUMOUR, and HEALTHAGENTS. As explained in Chapter 3, an acquisition protocol was defined in INTERPRET to provide maximum compatibility of the spectra obtained using different MRS systems at the different participant hospitals [129, 130]. This acquisition protocol was extended to the data acquisition procedure in eTUMOUR and HEALTHAGENTS. The spectra were acquired with MR scanners of several manufacturers: Siemens, General Electric and Philips. The acquisition protocols included Point Resolved Spectroscopy (PRESS) and Stimulated Echo Acquisition Mode (STEAM) sequences [131] with a range in the Time of Repetition (TR, between 1600 and 2020 ms), the Time of Echo (TE, 20 or 30-32 ms), the spectral width (1000-200 Hz), and the number of data-points (512, 1024 or 2048) [11]. Each spectrum was semi-automatically pre-processed in order to suppress the water peak, perform a phase correction, suppress the base line, normalize the spectrum area and correct the frequency shift as described in [24].

Spectral patterns contain resonance peaks related to the concentration of different metabolites in the tissue analyzed which are useful for tumour classification purposes [25, 24]. Based on a biochemical prior knowledge, a total number of 15 features were obtained from the integration of the signal under a spectral region associated with each metabolite of interest (see Figure 4.5). Signal quality and the diagnosis associated with each spectrum was validated by the INTERPRET Clinical Data Validation Committee [11], the eTUMOUR Clinical Validation Committee, and expert spectroscopists. In INTERPRET and eTUMOUR the class of each case was determined by a panel of histopathologists, while in HEALTHAGENTS the class was established by the original histopathologist.

Three types of brain tumour classes were taken into account in the experiments: aggressive brain tumours (AGG), including Glioblastomas and Metastases; low-grade glial tumours (LGG), including grade II Astrocytoma, Oligodendroglioma and Oligoastrocytoma; and Meningioma (MEN). The prevalence of the brain tumour classes considered in this study is shown in Table 4.5.

A Gaussian assumption is made since all the variables are continuous. Furthermore, both quadratic and linear classifiers have previously been shown to be powerful enough to achieve good results in automatic brain tumour classification [11, 24]. Although there may be more sophisticated feature selection techniques for this problem [120, 24], the use of peak integration is a good trade-off between complexity and performance, and it is independent of the different incremental data subsets. Finally, the evaluation method is based on K-random sampling train-test where $K = 100$ because the iterative incremental procedure makes the use of cross-validation or bootstrapping difficult. From the K repetitions the mean accuracy is shown and the standard deviation is used to estimate the confidence interval.

In these experiments, three specific desired features of a clinical decision support system (CDSS) based on ML techniques were analyzed: 1) the convergence of the classifiers in terms of stability/plasticity; 2) the effect of including new classes; 3) the customization of the classifiers in relation to the distributions of data in different hospitals.

4.4.1 Convergence of the iGDA

Following the methodology applied in Section 4.3.1, we tried to show how the iGDA algorithm is able to learn brain tumour discrimination with MRS in an incremental fashion

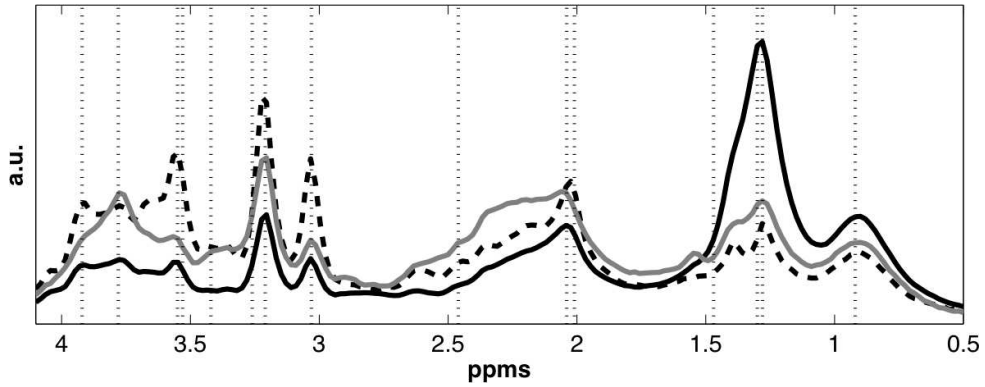


Figure 4.5: The features selected for classification are the peak integration of the metabolites observed in the brain (vertical dotted lines): Creatine (3.93 ppm and 3.02 ppm), Choline (3.21 ppm), N-Acetyl Aspartate (2.01 ppm), Myo-Inositol (3.26 ppm and 3.53 ppm), Glycine (3.55 ppm), Taurine (3.26 ppm), Glutamate/Glutamine (2.04 and 2.46 ppm), Alanine (1.47 and 3.78 ppm), Lactate (1.31 ppm), and Lipids (1.29 and 0.92 ppm). The peak integration computes the value of the area under the peaks considering an interval of 0.15 ppm from the assumed peak centre. The mean spectrum of each class of brain tumour is shown: aggressive (solid black line), low grade glial (dashed line), and meningioma (solid grey line).

from different subsets of training data. This was evaluated using the whole brain tumour database to show how the iGDA performance improved in the course of time when new observations were used to update the classifier. The whole database (see Table 4.5) was randomly split into a training partition (300 samples, 39.5%) and a test partition (460 samples, 60.5%). The decision of using only 39.5% of data for training is justified by the need of simulating a real scenario where the number of instances might be small. Although more incremental iterations could have been performed at the cost of having fewer instances for testing, the selected samples are enough to demonstrate the convergence of the algorithm and reduction in the standard errors of the results. The training partition was split into ten subsets of 30 samples. The whole test partition was used as an independent test set for each new updated classifier. The performance of the classifiers was measured in terms of the accuracy. The linear and quadratic versions of iGDA and the results were also compared to the performance of the other incremental algorithms.

4.4.2 Inclusion of new classes

In this second experiment, centers in Table 4.5 were used in order to address the inclussion of new classes. Each center initially contained only two classes (LGG, MEN). An initial classifier was trained from the first group of hospitals (CEN_0). Subsequently, using data from the rest of hospitals, the remaining class (AGG) was included in the following subsets and each generation of the classifier was evaluated with an independent test set. When introducing new classes, a problem of imbalanced classes may appear [128], resulting in a classifier with null sensitivity for the new class. In order to detect such a bias, a geometric mean of sensitivities was used to evaluate the classifiers in these experiments.

Center	Classes			Total
	AGG	LGG	MEN	
CEN_0	111	44	29	184
CEN_1	108	48	34	190
CEN_2	114	44	33	191
CEN_3	120	26	49	195
TOTAL	453	162	145	760

Table 4.5: The different centers and the number of instances per class. AGG: aggressive, LGG: low-grade glial, and MEN: Meningioma.

4.4.3 Customization to different centers

The third experiment simulates the customization of the classifier for a hospital by adapting a general model into the specific distribution of one hospital. Data from three hospitals (CEN_0) were used to train an initial classifier. Three other groups from two hospitals (CEN_1, CEN_2, and CEN_3) were made for testing the iGDA. These groups were chosen to balance the number of samples in each center. In addition, all the centers were grouped together in order to obtain a general behaviour of the convergence of the algorithm to compare with. This multicenter dataset is called CEN_{1-3} and is defined as $CEN_{1-3} = \bigcup_{i=1}^{3} CEN_i$. Table 4.5 shows the prevalence of each class in the dataset according to the four data groups used. Each center was divided into a test set and four subsets with 20 random samples in each one. Once the initial classifier was trained, it was used to automatically classify data from the test set of the other centers. Then, the first sample $\mathcal{S}_1$ of CEN_1 was used to update the classifier with the iGDA algorithm. The same process was performed with the first sample $\mathcal{S}_1$ of the other two centers, thus obtaining a total of three new incrementally updated classifiers. After incremental updating of the classifier of each center, a new evaluation was carried out using the independent test set of the corresponding center.

4.5 Results in brain tumour classification with MRS

4.5.1 Convergence of the iGDA

The comparison with the Learn^{++} and the iLDA algorithms shows that the accuracies of all these methods converge asymptotically (see Figure 4.6). This result suggest that the iGDA algorithm works properly as an incremental learning algorithm.

Generally speaking, the linear version of the iGDA algorithm performs better than the Learn^{++} and the iLDA algorithms. However, the quadratic version of iGDA needs three incremental updates to reach a comparable accuracy with the other algorithms. This behaviour may be explained by the low number of samples of the less prevalent classes in each subset. Nevertheless, there is asymptotic convergence of all methods: the data fits to the Gaussian model assumed by the iGDA, which describes linear or quadratic boundaries, as well as to the model assumed by the Learn^{++} algorithm, which divides the sample space using multiple hyperplanes.

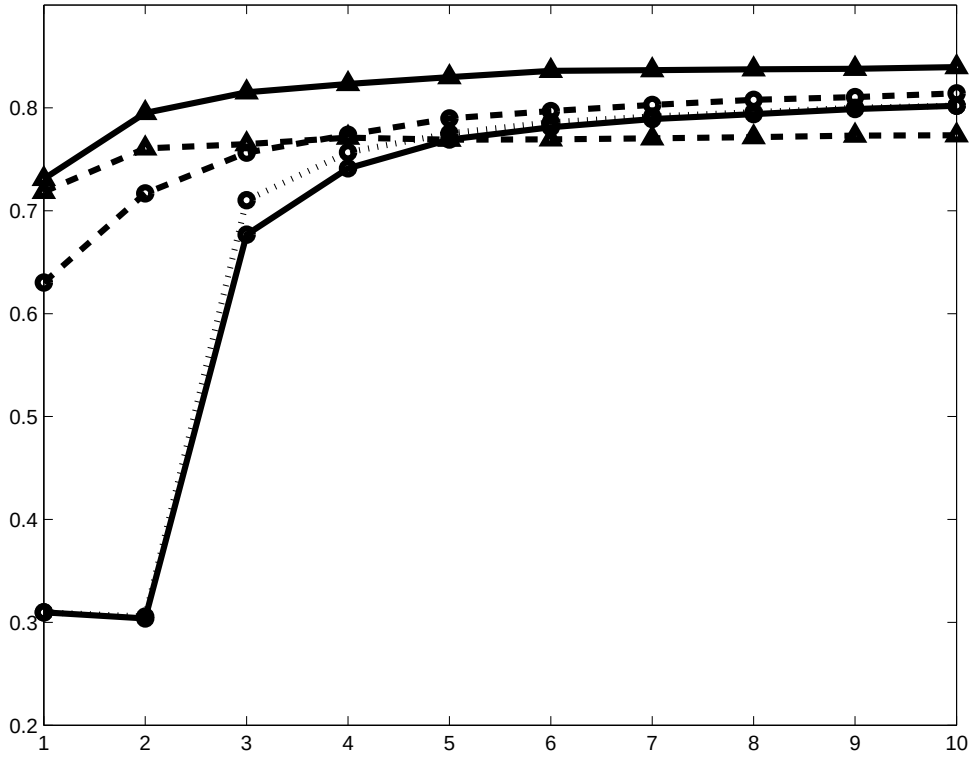


Figure 4.6: Comparison of the evolution of the accuracies of the linear iGDA (solid line and triangles), the quadratic iGDA (solid line and circles), iLDA (dashed line and triangles), and Learn^{++} (dashed line and circles) incremental learning algorithms. Also, a naive Gaussian classification model updated from scratch is compared (dotted line and circles). The first iteration is the performance of the initial classifier. From the second batch on, the incremental algorithm is executed. The experiment was repeated 100 times. The plots represent the mean value of all the experiments. The x-axis shows the different moments of time, t_i, of new observed data. The y-axis shows the accuracy. The iGDA using Graybill-Deal weight estimation shows a very good performance and it converges asymptotically.

The significance of differences ($\alpha = 5\%$) among algorithms was evaluated with a multiple comparison test using a Friedman's nonparametric two-way analysis of variance test with Tukey's honestly significant difference criterion from the first to the last iteration. The linear iGDA always displays a significant difference with respect to the other algorithms except with the iLDA in the first iteration. From iteration 8 to 10 the differences among the algorithms are all significant ($p < 0.01$), except between the quadratic iGDA and the naive incremental GDA retrained from scratch.

4.5.2 Inclusion of new classes

The mean accuracy of the results obtained when a new class appears inside the new observed samples improve from 0.29 to 0.78 in 10 incremental iterations. Since the convergence is asymptotic, the first two iterations show the biggest improvement: from 0.29 to 0.45 and to 0.57. Thereafter, the improvement is slower. The geometric mean of sensi-

tivities (G) improves from 0 to 0.76. Our results show that the first classifier is unable to correctly classify any sample belonging to the new class and thus $G = 0$. But, after further learning from two additional samples that include cases of the new class, the subsequent classifiers converge, obtaining not only a good accuracy but also a good G without forgetting to classify the initial classes. Our results show that the iGDA is able to introduce the new class into its knowledge base.

4.5.3 Customization to different centers

The third experiment tried to simulate a practical environment where a trained classifier is used for classification with data coming from different populations of patients and/or different acquisition machines. The results in Figure 4.7 show how the initial classifier exhibits a performance that clearly needs improvement. Therefore, when the classifier is updated with the new observations, the performance increases significantly with a small additional set of samples. In every new center, the accuracy of the incremental classifier improves in the course of new observations being used to incrementally train the classifier. Each observation included 20 new samples. The centers were joined in a unique set to compare the evolution of each center with the evolution of all the centers and show that the accuracy tends to converge asymptotically.

In general, the sensitivities for the first classification model in the centres CEN_1, CEN_2, and CEN_3 are between 0.71 to 0.76 for AGG, 0.85 to 0.86 for LGG, and 0.29 to 0.58 for MEN. After four incremental iterations the sensitivities vary from 0.79 to 0.83 for AGG, 0.74 to 0.84 for LGG, and 0.51 to 0.73 for MEN. Therefore, the incremental algorithm seems to be prone to balance the sensitivities of the different tumour types, increasing the sensitivities of the AGG and MEN tumour types while slightly decreasing the sensitivity of the LGG tumour types.

Again, a multiple comparison test ($\alpha = 5\%$) was carried out. Initially, only CEN_2 and CEN_3 showed significant differences but by iteration 5, only CEN_3 showed significant differences against the other centers ($p < 0.01$).

The same multiple comparison test ($\alpha = 5\%$) was used to analyze the statistical differences in the incremental models developed in the iterations of each center. These tests showed that the models of CEN_1 and CEN_3 had significant differences among iterations, except for the results of iteration 4 and 5, where the accuracies converge. With respect to the models of center CEN_2 there were significant differences between iteration 1 and 2 and between iterations 2 and 4, and iterations 3 and 5.

4.6 Discussion and conclusions

4.6.1 Technical aspects of the iGDA

The iGDA algorithm is presented as a new incremental algorithm for Gaussian discriminant analysis based on a weighted combination of different parameter estimations. It obeys the definition about the incremental learning algorithm given by several authors [76, 77, 60]. iGDA does not use any previous original datasets, but updates its knowledge by means of the information of the newly observed data and its already acquired knowledge. Therefore, it can be used when dealing with problems where past

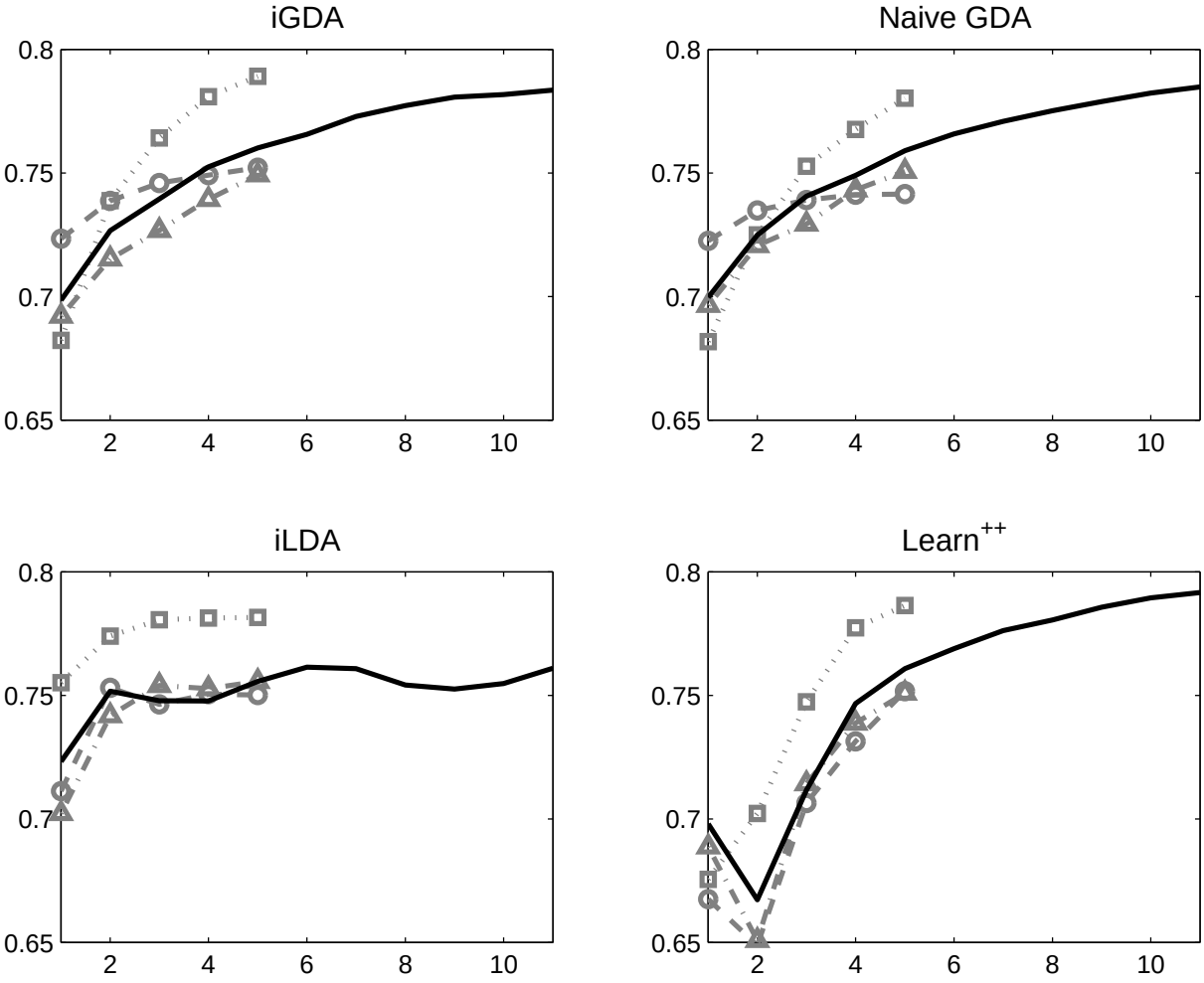


Figure 4.7: Comparison of the evolution of the mean accuracies of the different incremental learning algorithms trained with data from center CEN_0 and tested with data of new centers: CEN_1 (grey dash dotted line with triangles), CEN_2 (grey dashed line with circles), CEN_3 (grey dotted line with squares), and the evolution of the convergence for the union of centers CEN_1 to CEN_3, that is, CEN_{1-3} (black).

information is inaccessible or where there are problems gathering an appropriate dataset in a reasonable time. In such situations, this incremental learning algorithm can avoid the waiting time by using a small amount of information to build an initial simpler model and then update the model incrementally, and allow for additional classes, as new information arrives. Furthermore, the implementation of the algorithm is straightforward and the models can be estimated in polynomial time.

Figure 4.7 shows that the evolution of the updated classifiers in centers CEN_1 and CEN_2 is comparable to the evolution of the classifiers from all centers taken together. However, the evolution using the dataset from center CEN_3 shows the highest improvement. This is consistent with the Kullback-Leibler (KL) divergence between the joint distributions for all c, $p(\mathbf{x}, c)$, of CEN_0 and CEN_3, which are higher than the KL divergence of $p(\mathbf{x}, c)$ between CEN_0 and CEN_1, and between CEN_0 and CEN_2. This may be explained by the prevalence of the different brain tumour types in center CEN_3, which has

an influence on the prior probabilities of the models. Hence, while the updated classifiers of CEN_1 and CEN_2 are improving the knowledge concerning the conditional distributions $p(\mathbf{x}|c)$, the updated classifiers of CEN_3 are reinforcing the knowledge of the conditional distributions as well as the prior probabilities π_c. The final accuracy reached is similar to the median accuracy rate achieved in [24] for quadratic and linear discriminant analysis. In our results, the iGDA is comparable with the baseline model and the other incremental algorithms.

Since the experiments were repeated 100 times to avoid any possible bias, the results show a general behaviour of the iGDA algorithm. However, when the convergence to a minimum error has been achieved, there may be situations where addition of a new biased dataset results in a model with a slightly poorer performance than the previous one, but without statistical significance. Thus, when a convergence has been reached small oscillations in the accuracy of the models may be observed, similar to other iterative procedures.

An interesting feature of iGDA is that it does not have a *malignant* order effect [76], neither at instance level nor at concept level. This means that the order of the instances may give rise to slightly different models, but with similar discrimination accuracies. Our results show that the decision boundaries of the models are also similar regardless of the order in which the instances appear, or even the order in which the classes are introduced into the analysis.

One limitation of the iGDA algorithm is that it assumes that the data will follow a Gaussian distribution. This assumption may be useful for real number variables, even when they do not follow a Gaussian distribution, but this approach is useless for discrete distributions, such as Bernoulli or multinomial distributions. Nevertheless, the extension of these concepts may be of interest to other distributions, including discrete ones. The unimodal Gaussian assumption also restricts the type of decision boundaries to linear or quadratic boundaries.

Another feature of the iGDA is its ability to include new classes. However, this ability may lead to an imbalanced class problem [128] if the new class is underrepresented compared to the previous classes. This may be also related to the outvoting problem that occurs in incremental learners based on voting schemes such as the Learn^{++} [60, 62]. Furthermore, the behaviour of the weights in multivariate distributions and the combination of the covariance matrices using the Graybill-Deal estimation must still be theoretically studied and is the focus of future work.

4.6.2 Potential clinical interest of iGDA for brain tumour diagnosis

Primary brain tumours are proportionately less frequent than other cancers, but they are devastating diseases with high mortality. An accurate initial diagnosis of brain tumours has important consequences for therapeutic decisions and prognosis. Compared to most other tumours, obtaining brain tumour tissue for diagnostic purposes is relatively difficult even when using the advance technique of stereotactic biopsy [132]. The clinical DSS that are based on ML techniques and ^{1}H-MRS have shown a promising results for non-invasive brain tumour diagnosis. However, the development of robust classifiers requires acquisition of a large number of cases. Furthermore, in multicenter projects it is usually

assumed that the data have similar distributions, however in practice we may expect some differences in data distributions or class assignments. A straightforward application of the incremental method presented here is its ability to customize an already trained classifier to the specific distribution of a particular hospital. In other words, if a hospital has a limited number of samples for a particular class, a classifier trained with data from other hospitals can be used as an initial model and then adapted to the distribution of the patient population or the hospital scanner performance. Thus a classifier can be developed that has a customization to the hospital, but without the need for an unachievable acquisition of local data. The development of new models in the course of time as new data is acquired is related to the concepts of temporal and external validation reported by Altman et al. in [15]. Based on the results, our incremental algorithm could enhance the performance of such models when evaluated with subsequent patients coming from new hospitals.

In the framework of a clinical DSS the iGDA algorithm that has been developed may take advantage of the availability of new information to adapt the knowledge of the current system to the evolution of the data domain and also to extend the lifecycle of the system in a real clinical environment. Assuming that new information is ready for supervised classification at different times, the iGDA algorithm can learn from such new data without access to the previously seen data, even when a new class arises.

The ability to customize a model to a specific clinical centre could be used to improve the behaviour of a state-of-the-art CDSS for aiding brain tumour diagnosis. Further work will include the integration of the incremental algorithm developed in this work into a generic and dynamic DSS for clinical environments such as the aforementioned CDSSs and CURIAM [133]. The CURIAM Brain Tumour version [26] offers orientation on brain tumour diagnosis and is currently being tested in a clinical setting at several hospitals in Europe. The incremental learning method shown here may also complement to an audit model of brain tumour classifiers [31] and help provide dynamic optimisation of a CDSS.

Algorithm 1 Incremental Gaussian Discriminant Analysis

Input: $\mathcal{S}_{i+1} = \{(\mathbf{x}_n, c_n)_{n=1}^{N}\}; \mathcal{M}_i$
Output: $\mathcal{M}_{i+1}$
Require: $\forall c \in \mathcal{S}_i, N_c > 1$
  for all $c \in \mathcal{S}$ **do**
    $\pi_c \leftarrow N_c/N$
    $\boldsymbol{\mu}_c \leftarrow \frac{1}{N_c} \sum \boldsymbol{x}_n$
    $\boldsymbol{\Sigma}_c \leftarrow \frac{1}{N_c} \sum (\boldsymbol{x}_n - \boldsymbol{\mu}_c)^{\mathrm{T}} (\boldsymbol{x}_n - \boldsymbol{\mu}_c)$
  end for
  if $\mathcal{M}_{i-1} \neq \emptyset$ **then**
    for all $c \in \mathcal{S}$ **do**
      $\omega_{i+1}, \omega_i \leftarrow$ **Graybill-Deal**$(N_i, \Sigma_i, N_{i+1}, \Sigma_{i+1})$
      $\pi_c \leftarrow \frac{N_c^{i+1} + N_c^i}{N^{i+1} + N^i}$
      $\boldsymbol{\mu}_c \leftarrow \omega_{i+1} \boldsymbol{\mu}_c^{i+1} + \omega_i \boldsymbol{\mu}_c^i$
      $\boldsymbol{\Sigma}_c \leftarrow \omega_{i+1} \boldsymbol{\Sigma}_c^{i+1} + \omega_i \boldsymbol{\Sigma}_c^i$
    end for
  end if
  if linear **then**
    for all $c \in \mathcal{S}$ **do**
      $\boldsymbol{\Sigma}_c \leftarrow \boldsymbol{\Sigma}$
    end for
  end if
  for all $c \in \mathcal{S}$ **do**
    $\boldsymbol{W}_c \leftarrow -(1/2) \boldsymbol{\Sigma}_c^{-1}$
    $\boldsymbol{w}_c \leftarrow \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c$
    $w_{c0} \leftarrow \log \pi_c - (1/2) \log |\boldsymbol{\Sigma}| - (1/2) \boldsymbol{\mu}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c$
  end for
  return $\mathcal{M}_{i+1}$

Algorithm 2 Graybill-Deal computation of weights

Input: $N_1, N_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2$

Output: ω_1, ω_2

$S_1 = trace(\mathbf{\Sigma}_1)$

$S_2 = trace(\mathbf{\Sigma}_2)$

$\omega_1 = \frac{N_1/S_1}{\sum_{i=1}^{2} N_i/S_i}$

$\omega_2 = 1 - \omega_1$

return ω_1, ω_2

Chapter 5

Designing an incremental learning algorithm based on a Bayesian discriminative logistic regression

In this Chapter, we develop a new incremental learning algorithm for biomedical decision problems where no access to previous data is allowed. Unlike the previous iGDA algorithm, no assumptions over the underlying data distributions are made, and the estimation of the parameters are carried out using a Bayesian inference paradigm instead of the maximum-likelihood estimation. The Bayesian inference paradigm plainly fits with the design of an incremental learning algorithm by recursively using the posterior probability of the parameters of a model as the prior belief of a new model trained when new data is available. In this work, we introduce an incremental algorithm that uses a Bayesian paradigm to develop a discriminative logistic regression model based on an approximated posterior for two-class discrimination. Assuming that the parameters follow a multivariate Gaussian distribution, and taking into account that the posterior density of the parameters is unimodal, a Laplace approximation is carried out with a Newton-Raphson optimisation to find a local maximum. The performance of our algorithm is demonstrated by employing different benchmark datasets and is compared to a previous incremental algorithm and a non-incremental Bayesian model, showing that the algorithm is independent of the data model, iterative, and has a good convergence. Finally, we compare the Incremental Discriminative Bayesian Logistic Regression (iDBLR) and the iGDA algorithms using the brain tumour database.

This Chapter has been submitted as a journal paper with the title "Designing an incremental learning algorithm based on a Bayesian discriminative logistic regression" by Salvador Tortajada, Javier Vicente, Elies Fuster-Garcia, Montserrat Robles, and Juan Miguel García-Gómez.

5.1 Introduction

In this work, we introduce an application of the Bayesian inference paradigm for the design of an incremental learning algorithm for binary classification that works incrementally only with the new available data assuming that previous data are not available. From

now on, it will be referred to as incremental Discriminative Bayesian Logistic Regression (iDBLR). The Bayesian inference presents conceptual differences in the parameter estimation given the observations with respect to the traditional maximum-likelihood parameter estimation [38, 37]. Mainly, the Bayesian subjective interpretation of probability as a degree of belief assumes that the parameters are random variables and, consequently the inference results in a distribution of parameters. Precisely, one of its main advantages is the use of the prior beliefs about the parameters to estimate their posterior probability. This feature is the basis of our incremental approach since the posterior probability of the parameters of one iteration is used as the prior belief for the next iteration in a natural way as explained in the next section, where the full model is presented. Since the Logistic Regression is a two-class discriminative model it is thus unable to incorporate new classes. Therefore, this ability is not assessed for this algorithm. Section 5.3 introduces the benchmark dataset used for validating the algorithm and Section 5.4 shows the results compared to the iGDA algorithm, which also assumes that previous data are not available. Finally, some discussion and conclusions are drawn in Section 5.5.

5.2 Bayesian Discriminative Logistic Regression

The logistic regression is a generalized linear model that is used as a discriminative model for *binary* classification problems. When the outcome variable is binary some transformations have to be fulfilled which implies a different choice in the parametric model compared to a linear regression. Usually, the output variable is the membership of the observation into one of two possible classes $\mathcal{C} = \{0, 1\}$. Taking as a reference the class $c = 0$, the logarithm of the *odds ratio* of the probability of one class $p(c = 1|\mathbf{x})$ and the other $p(c = 0|\mathbf{x})$ can be used as a discriminative function

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \\ &= \log\left\{\frac{p(c=1|\mathbf{x})}{p(c=0|\mathbf{x})}\right\} \end{aligned} \tag{5.1}$$

where $p(c = 0|\mathbf{x}) = 1 - p(c = 1|\mathbf{x})$ and $\boldsymbol{\phi}(\mathbf{x})$ is an explicit and general basis expansion of the input such that $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x})]^{\mathrm{T}}$, and each $\phi_m(\mathbf{x})$ defines the m-th basis function applied to data vector $\mathbf{x}$. This expression for the logistic regression can be used to obtain a discriminative classifying method since we can classify an object $\mathbf{x}$ into class $c = 1$ if $p(c = 1|\mathbf{x}) > 0.5$. This is equivalent to decide that $\mathbf{x}$ belongs to class $c = 1$ if $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) > 0$.

Using the exponential of (5.1) it is possible to obtain the value of the probability of class 1 given an observation $\mathbf{x}$ (see Appendix B)

$$p(c = 1|\mathbf{x}) = \frac{\exp\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\}}{1 + \exp\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\}} \tag{5.2}$$

Now, our purpose is to estimate the parameters $\mathbf{w}$ of the discriminative model $g(\mathbf{x})$. With the Bayesian approach we can estimate the posterior probability of the parameters

$\mathbf{w}$ using

$$p(\mathbf{w}|\mathbf{c}, \mathbf{X}, \mathcal{H}) = \frac{p(\mathbf{w}, \mathbf{c}|\mathbf{X}, \mathcal{H})}{p(\mathbf{c}|\mathbf{X}, \mathcal{H})} \tag{5.3}$$

where $p(\mathbf{w}, \mathbf{c}|\mathbf{X}, \mathcal{H}) = p(\mathbf{c}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathcal{H})$. The overall hypothesis $\mathcal{H}$ is the first assumption we make about the parameters and it is established only in the first iteration as explained later. Therefore, we have to define the likelihood component, the prior probability, the evidence and the hypothesis $\mathcal{H}$. Let $\pi(\boldsymbol{\phi}(\mathbf{x}_n)) = p(c = 1|\boldsymbol{\phi}(\mathbf{x}_n))$, which is the model defined in (5.2) that is parameterized by the vector $\mathbf{w}$. Then, assuming that the instances $\mathbf{x}$ are independent and identically distributed, the likelihood function can be expressed as

$$\begin{aligned} p(\mathbf{c}|\mathbf{X}, \mathbf{w}) &= \prod_{n=1}^{N} \left[\pi(\boldsymbol{\phi}(\mathbf{x}_n))^{c_n} (1 - \pi(\boldsymbol{\phi}(\mathbf{x}_n)))^{1-c_n} \right] \\ &= \prod_{n=1}^{N} \frac{\left[\exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)) \right]^{c_n}}{1 + \exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n))} \end{aligned} \tag{5.4}$$

We assume that the prior probability of the parameters follows a multivariate Gaussian distribution with mean $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{C}$, that is, $\mathbf{w} \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{C})$. Furthermore, the first time a model is built an overall hypothesis $\mathcal{H}$ is required. Hence, we assume that the prior probability is $p(\mathbf{w}|\beta) = \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$, with β being an arbitrarily small precision. Now, a joint-likelihood consisting of the likelihood and the prior probability product is defined to obtain the posterior probability:

$$\begin{aligned} p(\mathbf{c}, \mathbf{w}|\mathbf{X}, \beta) &= p(\mathbf{c}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\beta) \\ &= \prod_{n=1}^{N} \frac{\left[\exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)) \right]^{c_n}}{1 + \exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n))} \mathcal{N}(\bar{\mathbf{w}}, \mathbf{C}). \end{aligned} \tag{5.5}$$

We still need to calculate the evidence or marginal likelihood $p(\mathbf{c}|\mathbf{X}, \beta)$ in the denominator of equation (5.3). The expression for this marginal likelihood is

$$\begin{aligned} p(\mathbf{c}|\mathbf{X}, \beta) &= \int p(\mathbf{c}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\beta)d\mathbf{w} \\ &= \int \prod_{n=1}^{N} \frac{\left[\exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)) \right]^{c_n}}{1 + \exp(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n))} \mathcal{N}(\bar{\mathbf{w}}, \mathbf{C}) d\mathbf{w}. \end{aligned} \tag{5.6}$$

Since there is no analytical solution to this integral, we can obtain a Laplace approximation to the posterior in order to obtain the analytical advantage and avoid the use of sampling techniques.

5.2.1 Laplace Approximation and the incremental DBLR

The Laplace approximation is a method that uses a Gaussian distribution to represent a given probability density function. This approximation for Bayesian logistic regression

has been used earlier [134]. Assuming that the posterior probability follows a multivariate Gaussian distribution it is possible to apply an analytical method to solve the estimation of $\mathbf{w}$. Furthermore, since the incremental learning algorithm uses the posterior probability estimated in time $t-1$ as the prior probability in the next estimation, it is required that both probabilities belong to the same class of density functions. As a consequence, it is also assumed that the Gaussian probability density function is a conjugate prior of the logistic regression likelihood function. Therefore, applying the Laplace approximation

$$p(\mathbf{w}|\mathbf{c},\mathbf{X},\beta) = \frac{p(\mathbf{c}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\beta)}{p(\mathbf{c}|\mathbf{X},\beta)} \approx \mathcal{N}(\mathbf{w}_{MAP},\mathbf{C}_t), \tag{5.7}$$

where $\mathbf{w}_{MAP}$ is the maximum of the posterior and the covariance $\mathbf{C}_t$ is defined as

$$\mathbf{C}_t = -\left(\frac{\partial^2}{\partial\mathbf{w}\partial\mathbf{w}^{\mathrm{T}}}\log p(\mathbf{c},\mathbf{w}|\mathbf{X},\beta)\right)^{-1}. \tag{5.8}$$

Therefore, we have to estimate the Maximum A Posteriori parameter and compute the curvature of the posterior at that point. Taking the logarithm of the joint likelihood and assuming that the prior probability is $p(\mathbf{w}|\beta) \sim \mathcal{N}(\mathbf{w}_{MAP},\mathbf{C}_t)$,

$$p(\mathbf{w}|\beta) = (2\pi)^{-\frac{D}{2}}|\mathbf{C}_t|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{MAP})^{\mathrm{T}}\mathbf{C}_t^{-1}(\mathbf{w}-\mathbf{w}_{MAP})\Big\}, \tag{5.9}$$

the logarithm of the joint log-likelihood is

$$\begin{aligned}\mathcal{L} &= \log p(\mathbf{c},\mathbf{w}|\mathbf{X},\mathbf{w}_{MAP},\mathbf{C}_t)\\ &= \sum_{n=1}^{N} c_n\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - \log\big(1+\exp(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))\big)\\ &- \frac{1}{2}(\mathbf{w}-\mathbf{w}_{MAP})^{\mathrm{T}}\mathbf{C}_t^{-1}(\mathbf{w}-\mathbf{w}_{MAP}) - \frac{D}{2}\log(2\pi) - \frac{1}{2}\ \log|\mathbf{C}_t|\end{aligned} \tag{5.10}$$

The first derivative is

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{w}} &= \sum_{n=1}^{N} c_n\boldsymbol{\phi}(\mathbf{x}_n) - p(c=1|\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n) - \mathbf{C}_t^{-1}(\mathbf{w}-\mathbf{w}_{MAP})\\ &= \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{c}-\mathbf{p}) - \mathbf{C}_t^{-1}(\mathbf{w}-\mathbf{w}_{MAP})\end{aligned} \tag{5.11}$$

where the $N\times 1$ vector of class-membership probabilities is defined as $\mathbf{p} = [p(c = 1|\mathbf{x}_1),\ldots,p(c=1|\mathbf{x}_N)]^{\mathrm{T}}$ and the $N\times M$ matrix $\boldsymbol{\Phi}$ is defined as

$$\boldsymbol{\Phi} = \begin{bmatrix}\phi_1(\mathbf{x}_1) & \ldots & \phi_M(\mathbf{x}_1)\\ \vdots & \phi_m(\mathbf{x}_n) & \vdots\\ \phi_1(\mathbf{x}_N) & \ldots & \phi_M(\mathbf{x}_N)\end{bmatrix}$$

The second derivative is

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^{\mathrm{T}}} &= -\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} p(c=1|\mathbf{x}_i)(1-p(c=1|\mathbf{x}_i)) - \mathbf{C}_t^{-1} \\ &= -\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} - \mathbf{C}_t^{-1}\end{aligned} \tag{5.12}$$

where $\mathbf{V}$ is a $N \times N$ dimensional diagonal matrix where each $v_{nn} = p(c=1|\mathbf{x}_n)(1-p(c=1|\mathbf{x}_n))$. Therefore, each v_{nn} can be interpreted as variances. As a result, the covariance matrix of the approximate posterior is

$$\mathbf{C}_t = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} + \mathbf{C}_{t-1}^{-1}\right)^{-1} \tag{5.13}$$

In order to find the parameter values $\mathbf{w}_{MAP}$ which yields the maximum of the parameter space, an iterative Newton-Raphson [135] optimization method can be used to find the roots of our function. This method is defined as

$$\mathbf{w}_{MAP}^{(t)} = \mathbf{w}_{MAP}^{(t-1)} - \left(\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^{\mathrm{T}}}\right)^{-1} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \tag{5.14}$$

Therefore, applying the first derivative (5.11) and the second derivative (5.12) into the definition of the Newton-Raphson step (5.14) we obtain

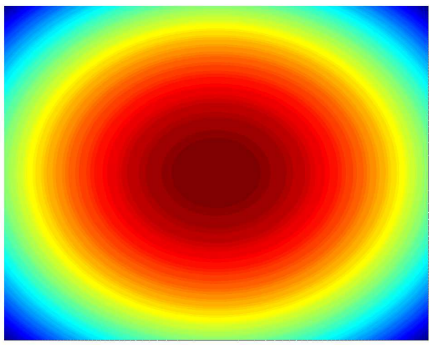
$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} + \mathbf{C}_{t-1}^{-1}\right)^{-1}\left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{c} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{p} - \mathbf{C}_{t-1}^{-1}(\mathbf{w}_{t-1} - \mathbf{w}_{MAP})\right) \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} + \mathbf{C}_{t-1}^{-1}\right)^{-1}\left(\left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} + \mathbf{C}_{t-1}^{-1}\right)\mathbf{w}_{t-1} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{c} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{p} - \mathbf{C}_{t-1}^{-1}(\mathbf{w}_{t-1} - \mathbf{w}_{MAP})\right) \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Phi} + \mathbf{C}_{t-1}^{-1}\right)^{-1}\left(\boldsymbol{\Phi}^{\mathrm{T}}\big(\mathbf{V}\boldsymbol{\Phi}\mathbf{w}_{t-1} + \mathbf{c} - \mathbf{p}\big) + \mathbf{C}_{t-1}^{-1}\mathbf{w}_{MAP}\right)\end{aligned} \tag{5.15}$$

An important issue to consider in the optimization algorithm is the stopping criterion for testing the convergence to a minimum. A widely used stopping criterion which does not require knowledge about the solution is to test that the norm of the gradient is less than a threshold, that is,

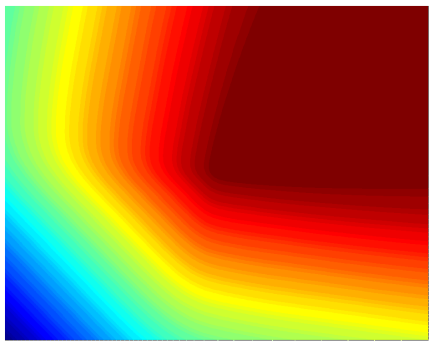
$$||\nabla \mathbf{w}|| = ||\mathbf{w}_t - \mathbf{w}_{t-1}|| \leq \epsilon \tag{5.16}$$

The disadvantage is that it can be difficult to choose the magnitude for ϵ. In our experiments however we checked that the algorithm had a similar behaviour irrespective of the order of magnitude for ϵ from 10^{-1} to 10^{-9}.

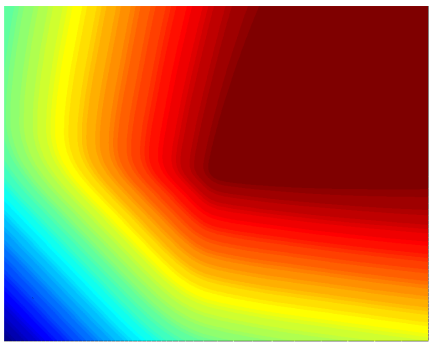
The approximation that we propose is illustrated in Figure 5.1, where the Laplace approximation to the product of the Gaussian prior distribution $p(\mathbf{w}|\beta) = \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$ and the likelihood of the logistic regression model $p(\mathbf{c}|\mathbf{X}, \mathbf{w})$ yields a posterior that follows



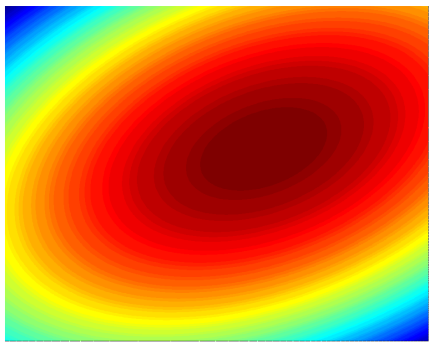
(a) Prior distribution $p(\mathbf{w}|\beta)$



(b) Likelihood distribution $p(\mathbf{c}|\mathbf{X},\beta)$



(c) Joint distribution $p(\mathbf{c},\mathbf{w}|\mathbf{X},\beta)$



(d) Laplace approximation distribution $p(\mathbf{w}|\mathbf{c},\mathbf{X},\beta)$

Figure 5.1: Illustration of the estimation of the posterior parameter distribution for a model with two parameters. The Laplace method approximates the joint distribution $\log p(\mathbf{c},\mathbf{w}|\mathbf{X},\mathbf{w}_{MAP},\mathbf{C}_t)$ to a multivariate Gaussian distribution.

a new multivariate Gaussian distribution $p(\mathbf{w}|\mathbf{c},\mathbf{X},\beta) \sim \mathcal{N}(\mathbf{w}_{MAP}^{(t)},\mathbf{C}_t)$. This solution enables the use of each posterior parameter distribution as an informative prior parameter distribution in the next iteration as explained in Chapter 2. The pseudocode of this incremental learning algorithm is shown in the Algorithm 3.

5.2.2 Bayesian Logistic Regression classification

Once the logistic regression parameters are estimated using the incremental Bayesian approach, it is possible to obtain a class prediction for a new observation $\mathbf{x}_{new}$ with the following expression

$$p(c=1|\mathbf{x}_{new},\mathbf{X},\mathbf{c},\mathcal{H}) = \int p(c=1|\mathbf{x}_{new},\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{c},\mathcal{H})d\mathbf{w} \tag{5.17}$$

An estimate of the above integral can be performed using samples simulated from our approximate posterior where each $\mathbf{w}_s$ is simulated or drawn from $p(\mathbf{w}|\mathbf{c},\mathbf{X},\mathcal{H})$ (5.7), that

is, $\mathbf{w}_s \sim \mathcal{N}(\mathbf{w}_{MAP}, \mathbf{C})$, such that

$$\begin{aligned} p(c=1|\mathbf{x}_{new}, \mathbf{X}, \mathbf{c}) &\approx \frac{1}{N}\sum_{n_s=1}^{N_s} p(c=1|\mathbf{x}_{new}, \mathbf{w}_s) \\ &= \frac{1}{N}\sum_{n_s=1}^{N_s} \frac{1}{1+\exp(-\mathbf{w}_s^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{new}))} \end{aligned} \tag{5.18}$$

However, although the sampling is easy to apply, it has a high computational cost because it needs a minimum number of necessary runs. The alternative to approximating by sampling averaging is to assume that the posterior is sharply peaked around the MAP value. Therefore, the object classification can be carried out using the MAP estimate to approximate the predictive posterior probability with

$$\begin{aligned} P(c=1|\mathbf{x}_{new}, \mathbf{X}, \mathbf{c}) &\approx P(c=1|\mathbf{x}_{new}, \mathbf{w}_{MAP}, \mathbf{X}, \mathbf{c}) \\ &= \frac{1}{1+\exp(-\mathbf{w}_{MAP}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{new}))} \end{aligned} \tag{5.19}$$

The assumption that the MAP value can approximate the whole posterior distribution seems reasonable since in our benchmark experiments the results of Monte Carlo sampling to obtain a set of generated parameters were comparable to the results obtained using MAP approximation, indeed the accuracy curves were perfectly overlapped. Therefore, we decided to use the MAP approximation for the rest of the experiments due to its simplicity and its lesser computational cost.

5.3 Materials

5.3.1 Stability/Plasticity dilemma

Synthetic datasets

The incremental learning algorithm proposed has been evaluated with different benchmark databases for dichotomous classification. The first three benchmarks are synthetic datasets with bidimensional data where each of the classes follow a unimodal bivariate Gaussian distribution with different mean vectors and covariance matrices obtaining different decision boundary configurations by selecting the appropriate class distributions. The values of the mean vectors and covariance matrices as well as the type of the decision boundary that achieves the theoretical Bayes error, $p(\text{error})$, are shown in Table 5.1. The fourth benchmark was a synthetic bidimensional dataset where each class belongs to a concentric ring of uniformly distributed data. This benchmark shows the property of having a 0% Bayes error. Finally, two dichotomous classification problems where each class follows a mixture of two bivariate distributions have been used to evaluate the incremental algorithm in front of multimodal Gaussian distributions. Figure 5.2 shows the six aforementioned synthetic benchmarks and their theoretical decision boundaries.

A total of $|\mathcal{B}| = 15$ incremental training subsets were drawn. Each training set or incremental sample had 20 instances with a different prevalence in each class. Initially,

Dataset	μ_1	μ_2	Σ_1	Σ_2	p(error)	Decision boundary
A	$(2\ 0)^{\mathrm{T}}$	$(0\ 2)^{\mathrm{T}}$	$\mathbf{I}$	$\mathbf{I}$	0.08	Line
B	$(0\ 0)^{\mathrm{T}}$	$(0\ 4)^{\mathrm{T}}$	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	0.02	Line
C	$(0\ 1)^{\mathrm{T}}$	$(0\ 3)^{\mathrm{T}}$	$\begin{pmatrix} 1/8 & 0 \\ 0 & 1/4 \end{pmatrix}$	$\begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}$	0.05	Ellipse

Table 5.1: True parameters of the distributions of the class-conditional probabilities.

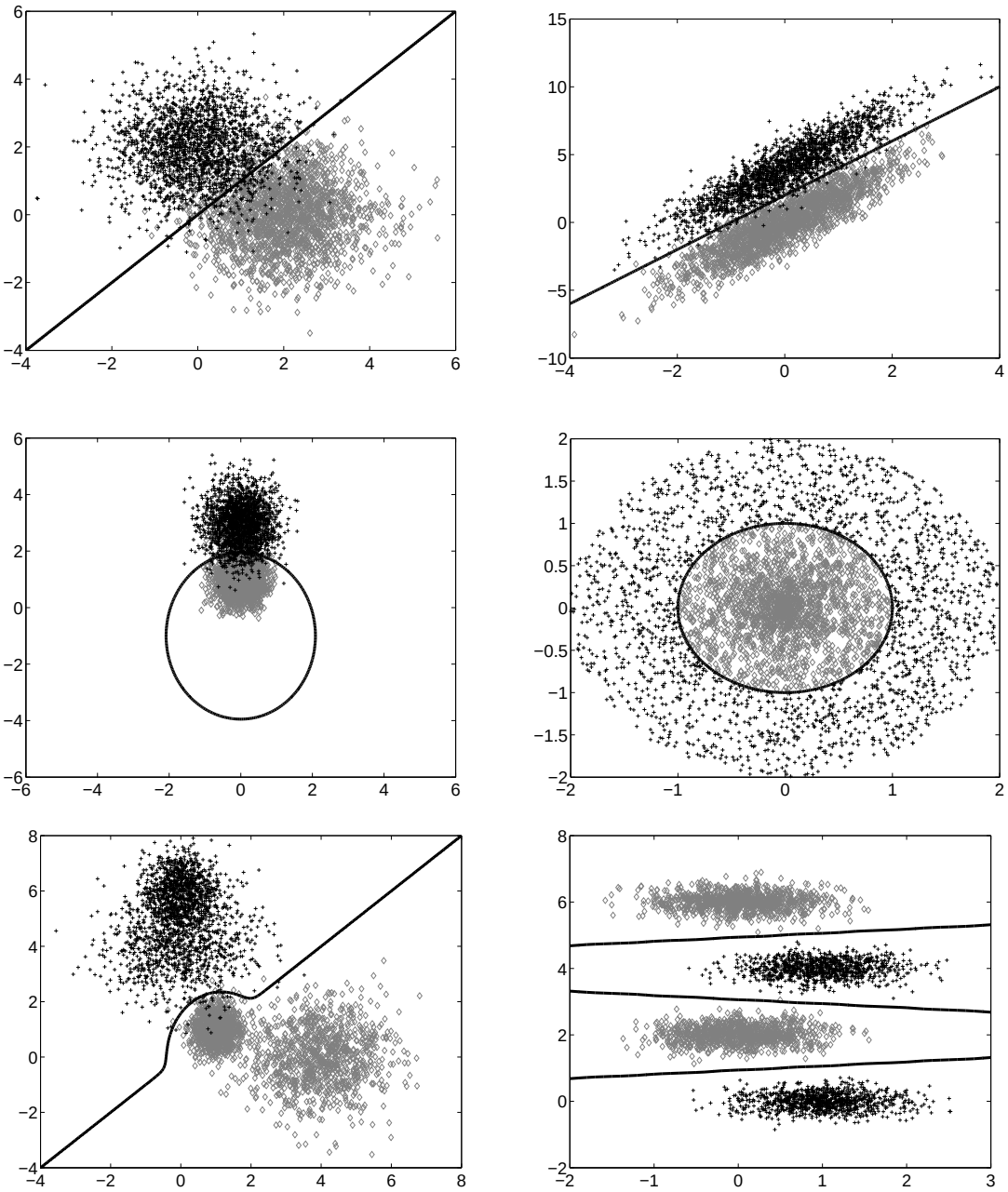


Figure 5.2: Distributions of the classes and theoretical decision boundary for the synthetic benchmark datasets.

a model is estimated using a prior distribution $p(\mathbf{w}_0|\beta) \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$. Then, an approximated posterior using Laplace approximation is calculated. Therefore, $p(\mathbf{w}_1|\mathbf{c}, \mathbf{X}, \mathcal{H}) \sim \mathcal{N}(\mathbf{w}_{MAP}^{(1)}, \mathbf{C})$ and successive models are estimated using the previous posterior as a prior, obtaining $|\mathcal{B}|$ incremental models where $p(\mathbf{w}_b|\mathbf{c}, \mathbf{X}, \mathcal{H}) \sim \mathcal{N}(\mathbf{w}_{MAP}^{(b)}, \mathbf{C})$. Furthermore, depending on the basis expansion function it is possible to obtain different decision boundaries. Finally, the predictions were obtained using a MAP approach to estimate the new test observations (equation 5.19). This process was repeated 100 times to avoid any bias.

The results were compared with another incremental algorithm: the iGDA algorithm [28] (see Chapter 4). This algorithm fits with most of these synthetic datasets since it assumes that the data follow a unimodal Gaussian distribution.

Two benchmark datasets from the UCI machine learning repository [126] were also used.

Vehicle Silhouette dataset

The purpose of the Vehicle Silhouette dataset is to classify a given silhouette into one of four different types of vehicle using a set of 18 features. Since we have a dichotomous classification algorithm the four original classes were merged into two classes (the first class included Opel and Saab original classes, while the second included Bus and Van original classes). The dataset consisted of 846 instances. It was divided into a training partition (630 instances) and a test partition (216 instances). The training partition was split again into 7 training sets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ of 90 instances with a similar prevalence to the original training dataset for each class. The sequential models obtained were tested with the previous training sets in order to observe if a gradual forgetting appeared, and with the independent test set to observe that the generalization performance increased asymptotically.

Wisconsin Breast Cancer dataset

The Wisconsin Breast Cancer dataset consists of 569 instances with 30 variables from a digitalized image of a fine needle aspirate (FNA) of a breast mass. The objective in this problem is to classify the instances into a malignant (37.3%) or a benign (62.7%) breast tumour. The database was divided into a test partition (169 instances) and a training partition (400 instances) that were also split into five different sets of 80 instances $\mathcal{S}_1, \ldots, \mathcal{S}_5$. Each partition had the same prevalence for each class as the whole dataset.

5.3.2 Order effects

Instance level order effects

The two-class multivariate Gaussian distribution synthetic dataset shown in Figure 4.1 from Chapter 4 has been used to assess the instance level order effects for the iDBLR algorithm. In order to compare the results between the iGDA and the iDBLR algorithms the configuration of the training set and its splitting in different incremental batches, and the configuration of the test set was the same as in Section 4.3.3: the training set consisted of 400 instances divided into 20 different batches with 20 instances in each one; whereas the test set consisted of 4 000 instances. The order effects were evaluated by

Center	Classes		Total
	AGG	NON	
CEN_0	111	73	184
CEN_1	108	82	190
CEN_2	114	77	191
CEN_3	120	75	195
TOTAL	453	307	760

Table 5.2: The different centers and the number of instances per class. AGG: aggressive, and NON: non-aggressive (a mixture of low grade gliomas and meningiomas).

permuting the training instances in 100 experiments in order to compare the distribution of the accuracies and the decision boundaries yielded by the models.

5.3.3 ^{1}H MRS Brain Tumour dataset

Finally, the iDBLR algorithm is evaluated with the ^{1}H MRS Brain Tumour database. As explained in Chapters 3 and 4, the dataset consists of ^{1}H MRS of brain tumour tissue that are labeled with one class among three different types of tumour classes; however, since the logistic regression is a two-class classification model, the three classes AGG, MEN, and LGG have been transformed into only two classes: AGG and a mixture of non-aggressive (NON) that includes the MEN and the LGG classes. The dataset has been divided into four different centers as explained in Chapter 4 (see Table 5.2). The center CEN_0 was used to train an initial model whose accuracy is estimated using a Leave-One-Out evaluation. Then, the model is evaluated with an independent test set from centers CEN_1, CEN_2, and CEN_3. Then, with subsequent subsets of data from these centers, different incremental models for each center are developed and their performances are evaluated with the same test set (see Section 2.5 for details).

A comparison with the iGDA algorithm has been carried out for the two-class classification problem in order to compare the ability of each algorithm to customize the parameters of the corresponding models to the new available data from each new center. A noteworthy fact is that, while the iGDA assumes an underlying Gaussian data distribution, the iDBLR does not assume any specific data distribution.

5.4 Results

The results of the simulated datasets, the different benchmark datasets, and the Brain Tumour dataset are explained in the subsections below.

5.4.1 Stability/Plasticity dilemma

Synthetic datasets

The results for the different synthetic bidimensional datasets are shown in Figure 5.3 and 5.4. Each subfigure shows the incremental results for the iDBLG, compared with the

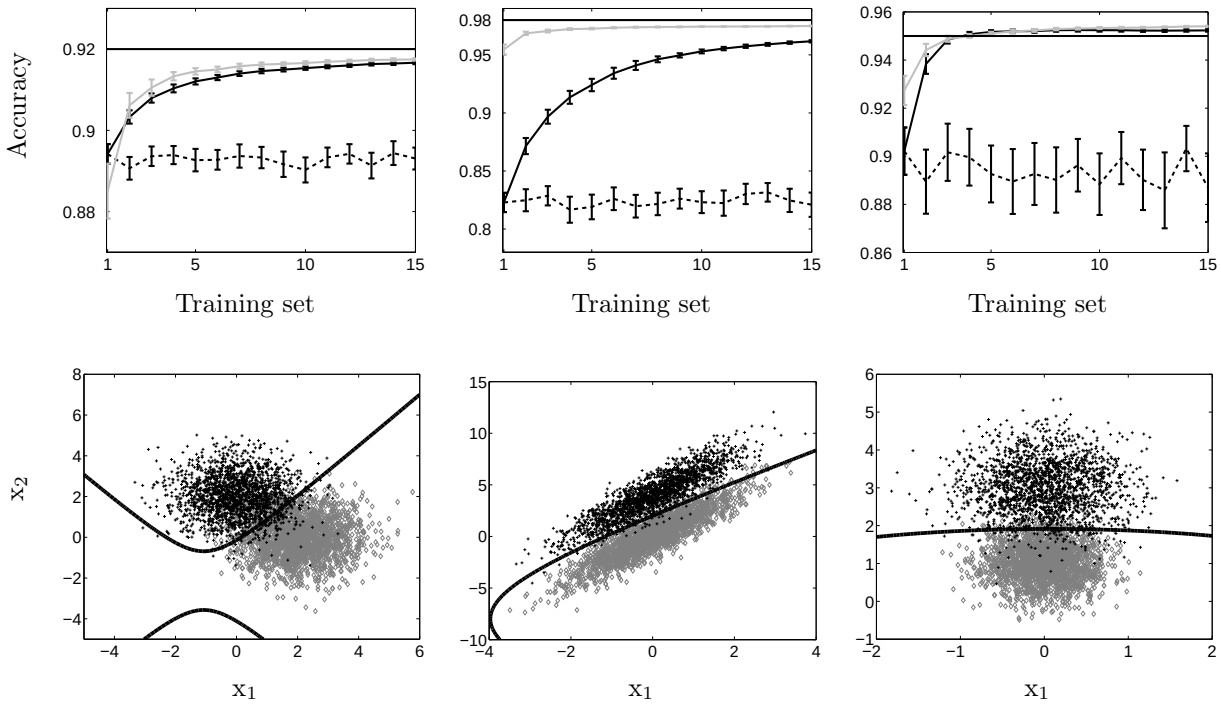


Figure 5.3: Results for the first three synthetic bidimensional Gaussian datasets. The Figure shows the theoretical Bayes error (solid black line), the incremental evolution of the iGDA models (solid gray line with whiskers) and the iDBLR models (solid black line with whiskers), and the accuracy of a non-incremental Bayesian logistic regression trained only with the current data (dashed black line with whiskers). The whiskers represent the confidence intervals ($\alpha = 0.05$). In the first dataset (left) the error converges to the theoretical Bayes error (8%). In the second synthetic bidimensional dataset (center) the iGDA error converges to the theoretical Bayes error (2%) while the iDBLG still shows room to improve. Finally, in the third synthetic bidimensional dataset (right), the error converges to the theoretical Bayes error (5%). The decision boundaries of the iDBLG models with the best accuracy of all the repetitions are shown on the bottom frames.

iGDA algorithm and the results obtained with a discriminative logistic regression algorithm that is trained only with the data of each iteration (top frame); also, an example of the decision boundary extracted by the best iteration of the iDBLG algorithm is shown (bottom frame). The results show that each new incremental model outperforms the previous ones. The incremental performance is always better than the performance of a model trained using only each new dataset. An iDBLG model $\mathcal{M}_i$ had the same performance than the one obtained by creating a new model from scratch with the subset $\bigcup_{j=1}^{i} \mathcal{S}_j$. This curve is not shown because it was overlapped with the incremental one.

The iGDA has a better performance than the iDBLG algorithm when the data distributions are Gaussian (see Figure 5.3). Nevertheless, the latter outperforms the iGDA when the data do not follow Gaussian distributions (see Figure 5.4). This is consistent with the design of both algorithms since the iGDA assumes that the underlying data distributions follow a multivariate Gaussian while the iDBLR does not consider any assumption about the data.

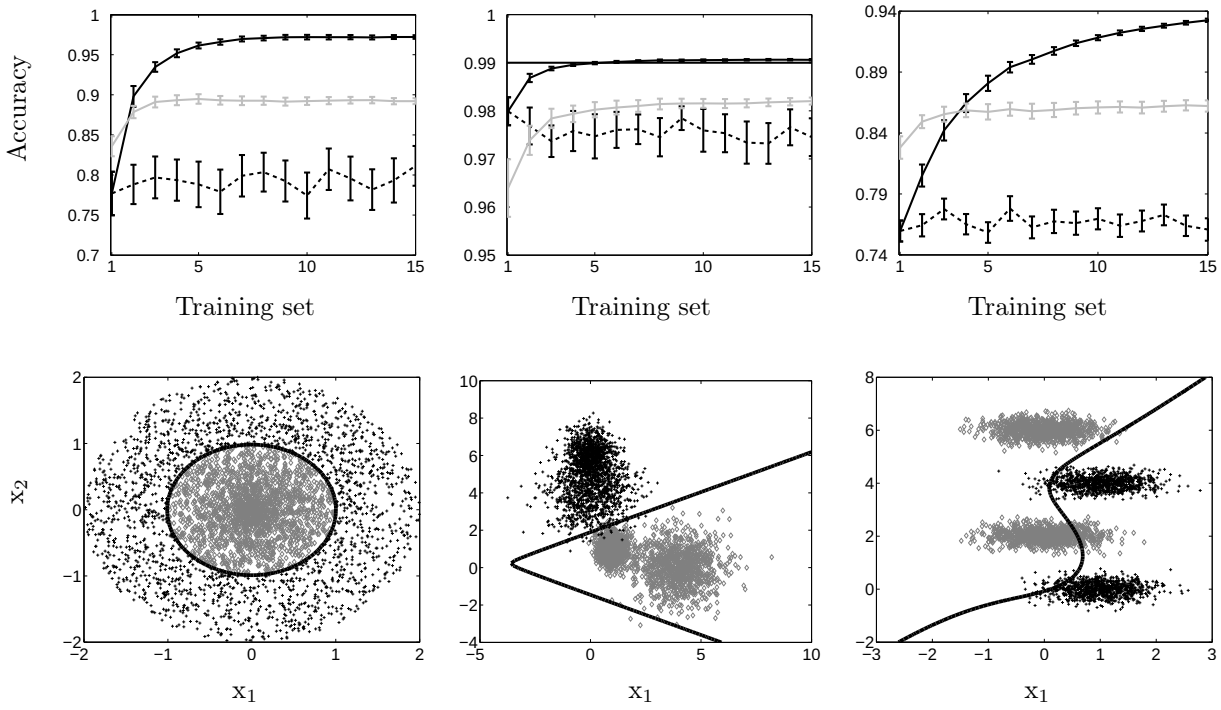


Figure 5.4: Results for the last three synthetic bidimensional non-Gaussian datasets. The Figure shows the theoretical Bayes error (solid black line), the incremental evolution of the iGDA models (solid gray line with whiskers) and the iDBLR models (solid black line with whiskers), and the accuracy of a non-incremental Bayesian logistic regression trained only with the current data (dashed black line with whiskers). The whiskers represent the confidence intervals ($\alpha = 0.05$). In the first one (left) the error converges but the theoretical Bayes error is lower (0%). In the second dataset (center) the iDBLR error converges to the theoretical Bayes error (1%). In the last dataset (right) the error converges, but the theoretical Bayes error is lower (0.05%). However, the iDBLR still has room to improve. On the contrary, the iGDA has a higher error rate because it is unable to describe cubic decision boundaries.

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$
$\mathcal{S}_1$	96.44	95.40	95.36	95.27	95.10	95.08	95.03
$\mathcal{S}_2$	–	94.75	94.52	94.33	94.33	94.47	94.25
$\mathcal{S}_3$	–	–	94.95	94.97	94.81	94.80	94.79
$\mathcal{S}_4$	–	–	–	94.78	94.76	94.66	94.67
$\mathcal{S}_5$	–	–	–	–	94.92	94.76	94.60
$\mathcal{S}_6$	–	–	–	–	–	94.34	94.00
$\mathcal{S}_7$	–	–	–	–	–	–	94.53
TEST	89.29	91.85	92.72	93.11	93.40	93.52	93.57
CI ($\alpha = 5\%$)	$\pm$0.54	$\pm$0.70	$\pm$0.48	$\pm$0.37	$\pm$0.33	$\pm$0.32	$\pm$0.32

Table 5.3: Training and test accuracy for the Vehicle Silhouette Database using a linear basis function $\boldsymbol{\phi}(\mathbf{x}) = [\mathbf{x}^0, \mathbf{x}^1]$ within the incremental algorithm. The rows indicate the different datasets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ and the columns show the models $\mathcal{M}_t$ built from a previous model $\mathcal{M}_{t-1}$ and the new dataset $\mathcal{S}_t$ using the posterior probability in time $t-1$ as the prior probability in time t; except $\mathcal{M}_1$ which is built from $\mathcal{S}_1$ and assuming that $p(\mathbf{w}|\beta) = \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$. Each column shows the average performance (%) on the current and the previous training datasets for the current model. The bottom rows (TEST, CI) indicate the evolution of the average accuracy of the models in the course of time evaluated with an independent test set and the confidence interval ($\alpha = 5\%$).

Vehicle Silhouette dataset

Table 5.3 shows that there is a gradual loss of accuracy relating to the previous training datasets when new observations are introduced using Bayesian incremental algorithm. We call this effect *gradual forgetting* and it is related to the *stability-plasticity dilemma* [78, 62]. However, the overall performance increases from 89% to 93%. This performance is lower than the one obtained by the iGDA algorithm, which increases from 93% to 97%.

Wisconsin Breast Cancer dataset

The results are shown in Table 5.4. There is an improvement in overall classification as the new data are used for incremental learning, but no gradual forgetting is observed with respect to the previous datasets. In this case, the overall performance –an increase from 95% to 97.5%– is better than the iGDA results, which increase from a 94% to a 97%.

Dataset	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
$\mathcal{S}_1$	97.44	97.84	97.87	98.18	98.15
$\mathcal{S}_2$	–	97.75	98.00	98.14	98.20
$\mathcal{S}_3$	–	–	97.69	97.81	97.85
$\mathcal{S}_4$	–	–	–	97.86	98.00
$\mathcal{S}_5$	–	–	–	–	98.06
TEST	95.69	96.56	97.03	97.33	97.50
CI ($\alpha = 5\%$)	$\pm$3.98	$\pm$3.57	$\pm$3.33	$\pm$3.16	$\pm$3.06

Table 5.4: Training and test accuracy (%) for the Wisconsin Breast Cancer Database using a linear basis function $\boldsymbol{\phi}(\mathbf{x}) = [\mathbf{x}^0, \mathbf{x}^1]$ within the incremental algorithm.

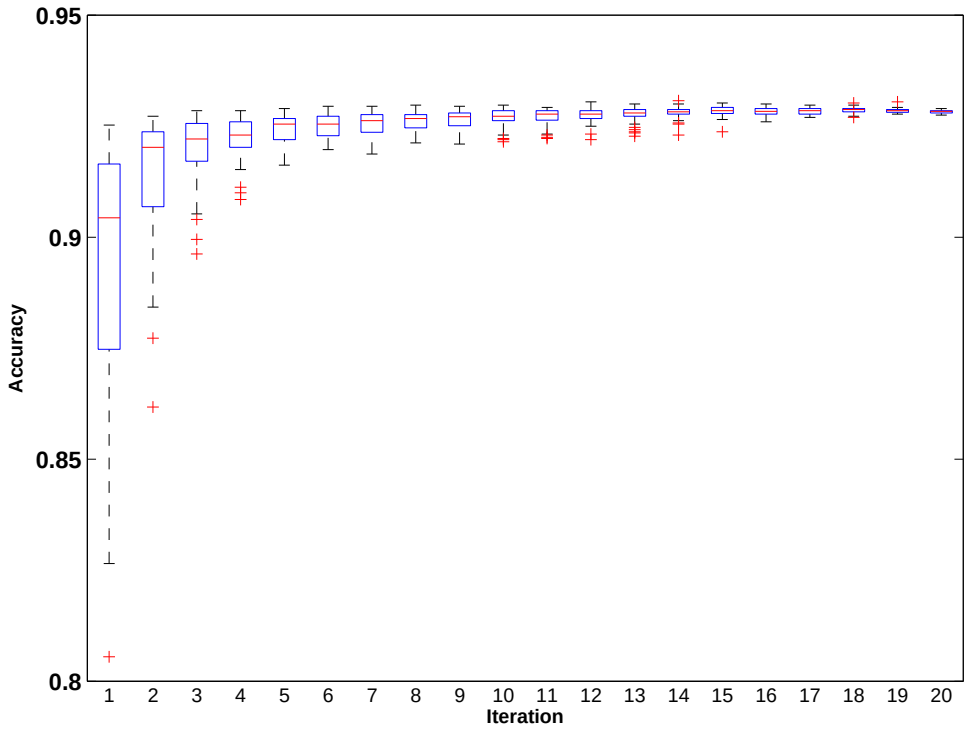


Figure 5.5: Boxplots of the accuracy of the models trained with different permutations of the instances. The X-axis shows the iterations of the incremental models. The Figure shows the results for the two-dimensional synthetic database with 20 iterations.

5.4.2 Order effects

Instance level order effects

The results for the evaluation of the ordering effects at instance level show that the iDBLR algorithm has also a negligible order effect as Figure 5.5 shows. The different permutations of the instances show that, after learning from all the available incremental batches, the obtained final models have a convergent accuracy. The convergence of the decision boundaries for the different models obtained for the bidimensional synthetic dataset is shown in Figure 5.6. These results prove that the algorithm has a *benign* order effect.

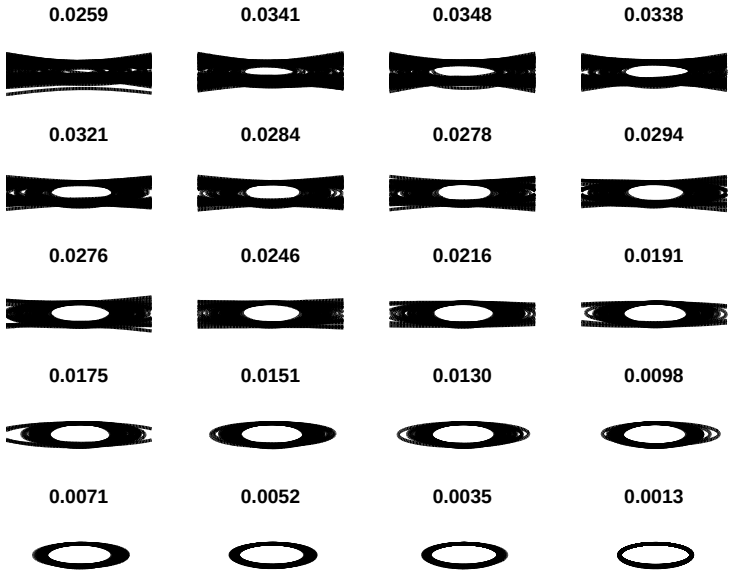


Figure 5.6: Convergence of the decision boundaries of each model in 20 iterations for the two-dimensional synthetic database. The variance of the Maximum A Posteriori parameter vector $\mathbf{w}_{MAP}$ is shown at the top of each iteration. The iterations are shown left-to-right, top-to-bottom. Again, it can be seen that the first models present arbitrary decision boundaries since their parameters are fitted from one single sample. When further samples are used for learning, the decision boundaries and their parameters begin to converge until the final iteration.

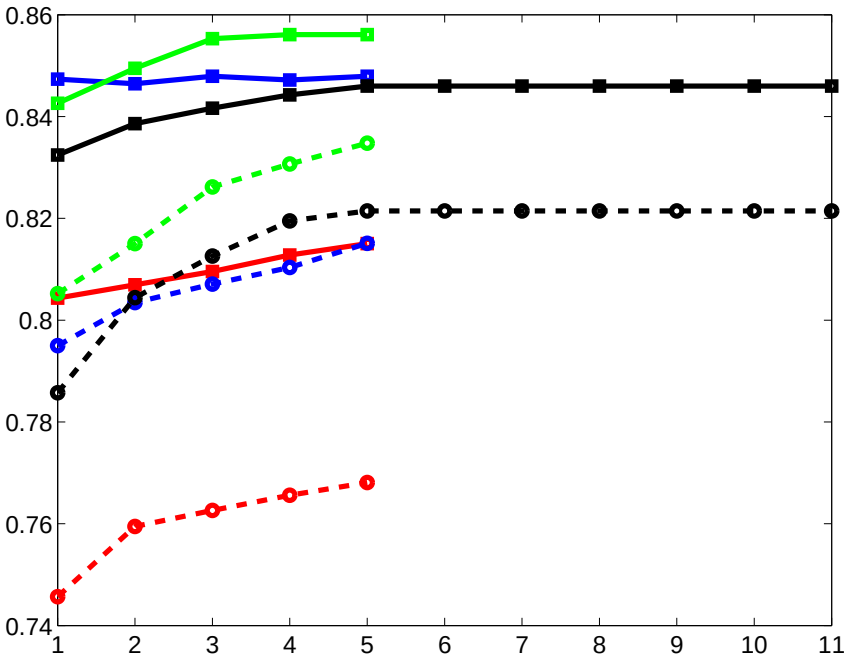


Figure 5.7: Comparison of the evolution and improvement of the mean accuracies (x-axis) of the iDBLR (solid lines) and the iGDA (dashed lines) and for the new centers: CEN_1 (blue), CEN_2 (red), CEN_3 (green), and the union of the three aforementioned centers (black).

5.4.3 ^{1}H MRS Brain Tumour dataset

The results for the iDBLR for automatic brain tumour diagnosis show the ability of this algorithm to improve the performance of the subsequent incremental models when new observations are available to re-adapt its parameters. Figure 5.7 shows that the performance of the iDBLR is better than the one obtained by the iGDA. This may be due to a non-Gaussian underlying distribution of the brain tumour dataset. The performance difference of the iDBLR models compared to the iGDA as we saw in the synthetic experiments supports this. Another reason is that the Bayesian approach regularises better than the traditional maximum-likelihood approach which explains the difference in the first iteration where no incremental learning has been performed yet. Again, the CEN_3 shows the better improvement among the other centers. This is consistent with the results obtained in Chapter 4 for the iGDA algorithm.

5.5 Discussion

In this Chapter we introduce an algorithm for updating the parameters of a discriminative logistic regression model by using the posterior probability approximation of iteration t as a prior probability on iteration $t+1$. Unlike many previous works [19, 75, 52, 74, 72], no access to previous data is allowed, which satisfies a common constraint found in many organizations where the data may be distributed [28]. Hence, we assume that only the data of the current batch t is available in time t. Thus, the classification model is able to classify data of a future batch $t+1$ only with the knowledge extracted from the previous model $\mathcal{M}_{t-1}$ and the current data $\mathcal{S}_t$. This is equivalent to having a sliding time window of size 0 [72].

Our solution to the incremental learning problem achieves the following desired properties defined by Street in [19]: it is iterative, processing batches of data at a time instance rather than requiring the whole set at the beginning; the algorithm requires only one pass through each data sample; since the parameters of the model are always the same and the incremental step only changes their values, the model structure requires a constant amount of memory and does not depend on the size of the data; finally, if the evolution of the algorithm is stopped at moment t, the model $\mathcal{M}_t$ provides the best answer at that moment. Furthermore, it has a negligible order effect which provides robustness to the algorithm.

Compared to the iGDA algorithm, the new Bayesian algorithm does not make any assumption over the data. Instead, it assumes that the parameters follow a multivariate Gaussian distribution allowing a good model fitting to the data independently of its distribution. Since the algorithm is based on basis expansion functions $\boldsymbol{\phi}(\mathbf{x})$ it can describe any polynomial decision boundary. One interesting research may be set out to automatically select the degree of the polynomial to best fit the observed data while keeping parsimony.

From a clinical viewpoint, a noteworthy fact is that the iDBLR algorithm achieves the same performance as its non-incremental version trained from scratch. This implies that waiting long enough to gather a representative amount of clinical data to produce a first classification model may be unnecessary since using the incremental version it is possible to obtain a similar model from a previous one that could have been trained using the available clinical observations at each moment. Therefore, a dynamic CDSS may be working from the first moment it has a sample available regardless of its representativeness and improve the models as soon as new data become available. This may be also a complement to an audit model that may choose the best classifier for any clinical question [31].

One obvious limitation of the model is that it is unable to discriminate more than two classes. A multinomial logistic regression model can be applied instead to extend its ability to discriminate more classes. In addition, since the Laplace approximation is based on a unimodal Gaussian distribution, the iDBLG is only applicable to such distributions, capturing only local properties of the true distribution. Hence, if the joint parameter distribution follows a multimodal distribution, which may happen when a concept drift occurs, it may be a wrong approximation. This can be overcame by using a deterministic global approximation, such as the Variational Bayes approach [136, 34] or the Expectation Propagation approach [137], on (5.7) instead of the Laplace approximation. Finally, the use of the Newton-Raphson method to optimize the MAP value of the distribution may be risky when a drifting concept happens, since the old $\mathbf{w}_{\text{MAP}}$ may be a bad starting point –or a saddle point– for the method which may imply a diverging behavior of the method.

Therefore, the iDBLG algorithm may not be suited when drifting concepts occurs [21]. Hence, although the iDBLG algorithm works well in stationary environments, it should be used with caution when applied to problems in non-stationary environments. In the latter case, a tracking of drifting concepts could be applied in order to detect changes [138] and, therefore, develop a new strategy to deal with these problems.

In summary, our contribution proposes an algorithm where the Bayesian approach can be used to develop an incremental learning algorithm for discriminative models by using the knowledge of one model, represented by the posterior probability, as a prior knowledge for the development of a new model given new observations and without making any assumption on the underlying distribution of the data.

Algorithm 3 Laplace Approximation

Input: $\mathbf{X}, \mathbf{t}, \mathbf{w}, \mathbf{C}, \text{degree}$
Output: $\mathbf{w}_{new}, \mathbf{C}_{new}$

$\mathbf{\Phi} \leftarrow \phi(\mathbf{X}, \text{degree})$
$\mathbf{w}_{old} \leftarrow \mathbf{w}$
$\mathbf{C}_{old} \leftarrow \mathbf{C}$
while $||\nabla \mathbf{w}|| > \epsilon$ **do**
  $\mathbf{P} \leftarrow 1/(1 + \exp(\mathbf{\Phi}\mathbf{w}_{old}))$
  $\mathbf{V} \leftarrow \text{diag}(\mathbf{P}(1 - \mathbf{P}))$
  $\mathbf{C}_{new} \leftarrow (\mathbf{\Phi}^{\mathrm{T}}\mathbf{V}\mathbf{\Phi} + \mathbf{C}_{old}^{-1})^{-1}$
  $\mathbf{w}_{new} \leftarrow \mathbf{C}_{new}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{V}\mathbf{\Phi}\mathbf{w}_{old} + \mathbf{t} - \mathbf{P}) + (\mathbf{C}_{new}\mathbf{C}_{old}^{-1}\mathbf{w}_{old})$
  $||\nabla \mathbf{w}|| \leftarrow ||\mathbf{w}_{new} - \mathbf{w}_{old}||$
  $\mathbf{w}_{old} \leftarrow \mathbf{w}_{new}$
end while
return $\mathbf{w}_{new}$, $\mathbf{C}_{new}$

Chapter 6

Concluding remarks and future work

6.1 Conclusions

The present Thesis deals with the use of ML discipline to solve biomedical decision problems. Specifically, we have applied them to develop computer-assisted CDSS for automatic brain tumour classification. The first contributions are focused on the development of ML-models for brain tumour diagnosis and their prospective evaluation. These models however lack the ability to adapt their parameters when new data is available or when the models are moved from one hospital to another. The remaining of this Thesis introduces two new incremental learning algorithms, which assume that previous data are not accessible, for the development of models having adaptability to new centers or when data are obtained is small batches.

The technical aspects covered in the Thesis include the mathematical design of two different incremental learning algorithms and the development of a series of benchmark experiments to carry out an evaluation of their performances and the final accuracies of the yielded models. The incremental learning algorithms assume that previous observations are not accessible once they have been used to train the model. This assumption is consistent with the problems found during the development of the HEALTHAGENTS project. The advantages and disadvantages of both, the iGDA, and the iDBLR algorithms are summarized in Table 6.1. This Table shows that since the iGDA is a generative model, the number of parameters is higher than the discriminative iDBLR, unless $M \approx c \cdot d$. The use of basis expansion functions allows the iDBLR to describe a polynomial decision boundary surface, while the Gaussian distribution assumption of the iGDA allows to describe at most a quadratic decision boundary surface. The use of Bayesian inference in the iDBLR entails an implicit regularization while the iGDA needs an explicit smoothing technique to regularize the models built. However, the use of logistic regression in the iDBLR restricts the classification models to a two-class discrimination problem, while the iGDA can deal with multiclass problems, and therefore include new classes if needed.

The conclusions extracted from this Thesis are:

- The result of a prospective multicenter-multiproject evaluation show that the prediction of in-vivo MRS is possible using models inferred by multicenter datasets, where the data comes from a mixed set of different hospitals using different instrumentation although they are obtained under the similar acquisition parameters. However,

Table 6.1: Comparison of the features of the incremental learning algorithms. In the table, M is the number of basis expansion functions, c is the number of classes and d is the number of variables or dimensions.

	iGDA	iDBLR
Type of model	Generative	Discriminative
Parameter estimation	Weighted maximum-likelihood	Bayesian inference
Data distributions	Assumes Gaussian distributions	No assumptions
Decision boundary	Quadratic	Polynomial
# of parameters	$\frac{c}{2}(d(d+3))$	$M(d+1)$
Discriminates	Multiple classes	Two classes
Regularization	Explicit	Implicit
Order effects	Benign	Benign
Incorporates new classes	Yes	No

prospective evaluation shows the need of new cases to improve the ML models. Also, the combination of Short and Long Times of Echo has been proposed for ^{1}H MRS-based Brain Tumor (BT) classification. These results entail the interest in taking advantage of the new data as they become available which justifies the research on incremental learning algorithms for these problems.

- A new incremental learning algorithm for Gaussian discriminant analysis based on a weighted combination of different parameter estimations has been designed, implemented, characterized, and validated. It obeys the definition about the incremental learning algorithm given by several authors [76, 77, 60]. The algorithm does not use any previous original datasets, but updates its knowledge by means of the information of the newly observed data and its already acquired knowledge. Therefore, it can be used when dealing with problems where past information is unavailable or where there are problems gathering an appropriate dataset in a reasonable time. In such situations, this incremental learning algorithm can avoid the waiting time by using a small amount of information to build an initial simpler model and then update the model incrementally, and allow for additional classes, as new information arrives and showing a negligible order effect.

- Another new incremental learning algorithm based on the Bayesian inference paradigm has been also designed, implemented, characterized, and evaluated. This algorithm updates the parameters of a discriminative logistic regression model by using the posterior probability approximation of one iteration t as a prior probability on the following iteration $t+1$ without making any assumption on the underlying distribution of the data. Unlike many previous works [19, 75, 52, 74, 72], no access to previous data is allowed, which satisfies a common constraint found in many organizations where the data may be distributed [28]. Hence, we assume that only the data of the current batch t is available in time t. Thus, the classification model is able to classify data of a future batch $t+1$ only with the knowledge extracted from the previous model $\mathcal{M}_{t-1}$ and the current data $\mathcal{S}_t$. This is equivalent to having a sliding time window of size 0 [72]. Furthermore, prior information can also be incorporated for the initialization of the learning process. Since the algorithm does

not depend on the data distribution, it can be applied to several practical problems in medicine.

- Both incremental learning algorithms are iterative, processing batches of data at a time instance rather than requiring the whole set at the beginning; both algorithms require only one pass through each data sample; since the parameters of the models are always the same and the incremental step only changes their values, the structure of the models require a constant amount of memory and do not depend on the size of the data; finally, if the evolution of the algorithms are stopped at moment t, the models $\mathcal{M}_t$ provide the best answer at that moment. Finally, the performance of the incremental learning algorithm is equivalent to the performance of the non-incremental learning algorithms, but the former has the advantage of being able to adapt themselves to new data learning in an incremental fashion.

6.2 Future work

Some of the future lines of research directly related to the results of this Thesis are:

- The incremental learning algorithms have been developed mainly for real number variables. The extension of the introduced incremental learning concepts may be of interest to discrete distributions.

- The iGDA assumes a unimodal Gaussian distribution of the data. The iDBLR assumes a unimodal distribution since the Laplace method is a local approximation. A future line of work is to develop incremental learning algorithms that are able to deal with multimodal distribution. In the case of the iGDA algorithm it may be possible to research for incremental EM algorithms. While in the case of the iDBLR it may be interesting to apply a deterministic global approximation, such as the Variational Bayes approach [39, 34] or the Expectation Propagation approach [137].

- The iDBLR algorithm is unable to discriminate more than two classes. Future work will include the design and development of a multinomial logistic regression model to extend its ability to discriminate more classes.

- In the iGDA, the behaviour of the weights in multivariate distributions and the combination of the covariance matrices using the Graybill-Deal estimation must be still theoretically studied.

- In the iDBLR, an interesting research may be set out to incorporate the degree of the linear generalized model. This may be accomplished by using a Bayesian model comparison framework [139, 134].

- Further work will include the integration of the incremental algorithm developed in this work into a generic and dynamic **DSS** for clinical environments such as the aforementioned CDSSs and CURIAM [133]. The incremental learning method shown here may also complement to an audit model of brain tumour classifiers [31] and help provide dynamic optimisation of a CDSS.

Appendix A

Gaussian Discriminant Analysis

A generative classification model estimates the class posterior probability $p(c|\mathbf{x})$ by means of the Bayes' Theorem

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})}$$

where $p(c)$ is the prior probability of class c, and $p(\mathbf{x}|c)$ is the likelihood or class-conditional probability density. A Gaussian Discriminant Analysis (GDA) is a generative model that assumes that the class-conditional densities follow a multivariate normal or Gaussian distribution, that is, $p(\mathbf{x}|c) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where $\boldsymbol{\mu}_c$ is the vector mean and $\boldsymbol{\Sigma}_c$ is the covariance matrix of the respective multivariate Gaussian distribution of class $c \in \mathcal{C}$. Hence,

$$p(\mathbf{x}|c) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\Big\} \tag{A.1}$$

where D is the dimension of the data.

As explained in Chapter 2, to minimize the probability of error of a classification model, the decision rule of a Bayes' classifier should be

$$c^* \leftarrow \arg\max_{c \in \mathcal{C}} g_c(\mathbf{x}) \tag{A.2}$$

where $g_c(\mathbf{x})$ is a discriminant function for class c expressed as

$$\begin{aligned} g_c(\mathbf{x}) &= p(c|\mathbf{x}) \\ &= \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \end{aligned} \tag{A.3}$$

The decision rule (A.2) is transformed by using logarithmic notation and taking into account that the denominator $p(\mathbf{x})$ is independent of the class,

$$
\begin{aligned}
c^* &= \arg\max_{c \in \mathcal{C}} g_c(\mathbf{x}) \\
&= \arg\max_{c \in \mathcal{C}} p(c|\mathbf{x}) \\
&= \arg\max_{c \in \mathcal{C}} \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \\
&= \arg\max_{c \in \mathcal{C}} p(\mathbf{x}|c)p(c) \\
&= \arg\max_{c \in \mathcal{C}} \log\{p(\mathbf{x}|c)\} + \log\{p(c)\}
\end{aligned}
$$

If the class-conditional probability of the expression of the decision rule is replaced with the multivariate Gaussian equation (A.1),

$$
c^* = \arg\max_{c \in \mathcal{C}} \log\{p(c)\} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c \tag{A.4}
$$

We see that the equation (A.4) has the form of a quadratic function since the discriminant function $g_c(\mathbf{x})$ can be written

$$
g_c(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}_c\mathbf{x} + \mathbf{L}_c^{\mathrm{T}}\mathbf{x} + K_c \tag{A.5}
$$

where

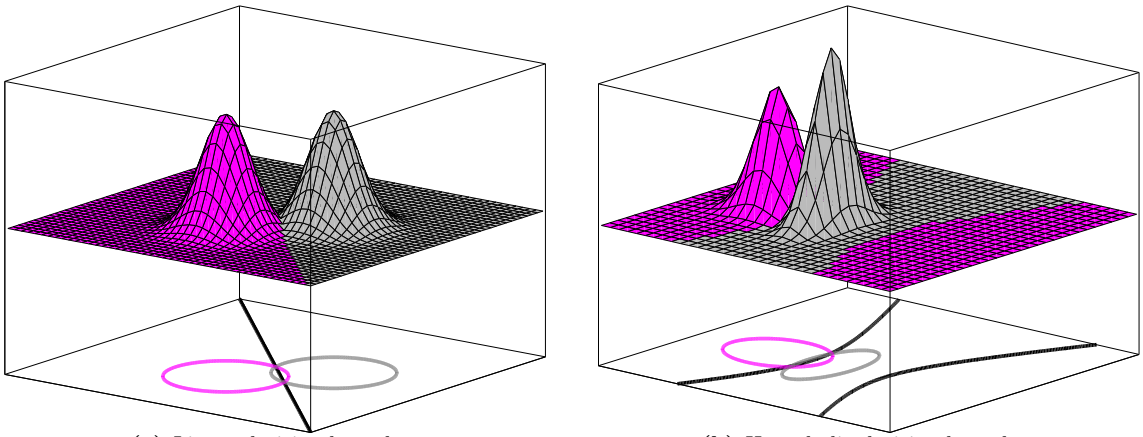
$$
\mathbf{Q}_c = -\frac{1}{2}\boldsymbol{\Sigma}_c^{-1} \tag{A.6}
$$

$$
\mathbf{L}_c = \boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c \tag{A.7}
$$

$$
K_c = \log\{p(c)\} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c \tag{A.8}
$$

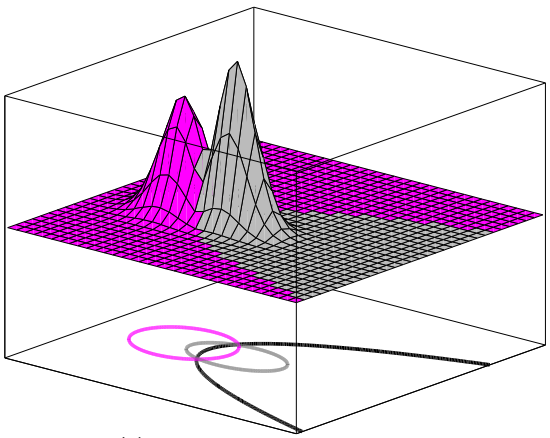
Geometrically, when the covariance matrices of the classes are arbitrary, the decision boundary surface is *hyperquadratic*, and thus they can take any of the general forms: hyperplanes, pairs of hyperplanes, hyperellipsoids, hyperspheres, hyperhyperboloids, or hyperparaboloids.

When the covariance matrices of each class are identical, say $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, then the decision boundary surface is an hyperplane since the quadratic terms of the discriminant function of each class cancel each other (see Figure A.1).

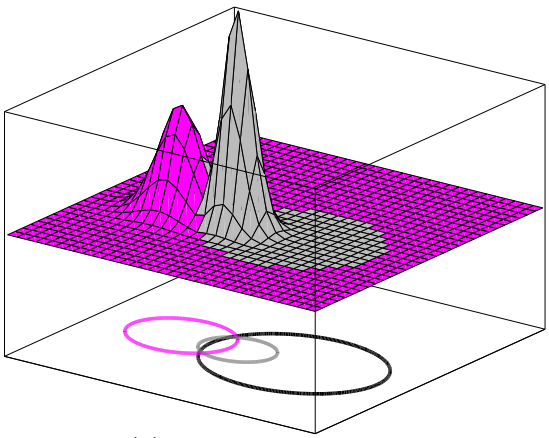


(a) Linear decision boundary

(b) Hyperbolic decision boundary



(c) Parabolic decision boundary



(d) Eliptic decision boundary

Figure A.1: Illustration of a linear decision boundary (a), when both distributions have the same covariance matrix. When the covariance matrices of each class are different, then the decision boundary can take any quadratic form (b,c,d). The decision boundaries are shown with a black solid line under the distributions, where the Mahalanobis ball containing the 95% of the data are also displayed with the color of its corresponding class.

Appendix B

Logistic regression

The logistic regression model is a well-known algorithm for solving a two-class classification problem, where $\mathcal{C} = \{0, 1\}$. We have seen that, taking one class as a reference, for instance $y = 0$, the logarithm of the *odds ratio* can be used as a discriminant function

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \\ &= \log\left\{\frac{p(c=1|\mathbf{x})}{p(c=0|\mathbf{x})}\right\} \end{aligned} \tag{B.1}$$

where $p(c = 0|\mathbf{x}) = 1 - p(c = 1|\mathbf{x})$, and $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_m(\mathbf{x})]$ is a basis expansion function. Let $p = p(c = 1|\mathbf{x})$, then from equation (B.1) we can find out the expression for p using the exponential function and solving,

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) &= \log\left\{\frac{p(c=1|\mathbf{x})}{p(c=0|\mathbf{x})}\right\} \\ \exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right\} &= \frac{p}{1-p} \\ p &= \frac{\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right\}}{1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right\}} \end{aligned}$$

and, consequently,

$$1 - p = \frac{1}{1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right\}}$$

This function of $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ is also called the *logistic sigmoid function*, which is often applied as an activation function for Multilayer Perceptrons (see [140]). To determine the parameters $\mathbf{w}$ of the logistic regression function it is possible to use a maximum-likelihood approach or a Bayesian approach (see Chapter 5). For this purpose, we shall make use of the derivative of the logistic sigmoid function. Considering that $\pi(\boldsymbol{\phi}(\mathbf{x}_n)) = p(c = 1|\mathbf{x}_n)$,

then

$$\begin{aligned}\frac{\partial \pi(\boldsymbol{\phi}(\mathbf{x}_n))}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \frac{\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}}{1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}} \\ &= \frac{\boldsymbol{\phi}(\mathbf{x}_n)\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}(1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}) - \exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}}{(1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\})^2} \\ &= \frac{\boldsymbol{\phi}(\mathbf{x}_n)\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}}{1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}} \frac{1}{1+\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}} \\ &= \boldsymbol{\phi}(\mathbf{x}_n)\pi(\boldsymbol{\phi}(\mathbf{x}_n))(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n)))\end{aligned} \tag{B.2}$$

This interesting final results expresses the derivative of the logistic sigmoid function in terms of itself.

Whether we apply a frequentist or maximum-likelihood approach or a Bayesian approach, the parameters $\mathbf{w}$ are fitted making use of the likelihood function. Given a dataset, $\mathcal{S} = \{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_N, c_N)\}$, where $c_n \in \mathcal{C}$, the likelihood function for the logistic regression can be written

$$\ell(\mathbf{w}) = \prod_{n=1}^{N} \left[\pi(\boldsymbol{\phi}(\mathbf{x}_n))^{c_n}(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n))^{1-c_n}\right] \tag{B.3}$$

As usual, to facilitate the calculation and to avoid underflow computation problems, we can take the logarithm of the likelihood (*log-likelihood function*), which gives

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \Big\{ c_n \log\left[\pi(\boldsymbol{\phi}(\mathbf{x}_n))\right] + (1-c_n)\log\left[(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n)))\right] \Big\} \tag{B.4}$$

If we wish to find the value of $\mathbf{w}$ that maximizes $\mathcal{L}(\mathbf{w})$ we have to differentiate the log-likelihood with respect to $\mathbf{w}$.

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \sum_{n=1}^{N} \Big\{ c_n \log\left[\pi(\boldsymbol{\phi}(\mathbf{x}_n))\right] + (1-c_n)\log\left[(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n)))\right] \Big\} \\ &= \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) c_n \frac{\pi(\boldsymbol{\phi}(\mathbf{x}_n))(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n)))}{\pi(\boldsymbol{\phi}(\mathbf{x}_n))} - \boldsymbol{\phi}(\mathbf{x}_n)(1-c_n)\frac{\pi(\boldsymbol{\phi}(\mathbf{x}_n))(1-\pi(\boldsymbol{\phi}(\mathbf{x}_n)))}{1-\pi(\boldsymbol{\phi}(\mathbf{x}_n))} \\ &= \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\left[c_n - \pi(\boldsymbol{\phi}(\mathbf{x}_n))\right]\end{aligned} \tag{B.5}$$

We see that the factor involving the derivative of the logistic sigmoid function has cancelled, leading to a simplified form for the gradient of the log-likelihood. Therefore, the contribution from a particular observation $\mathbf{x}_n$ is given by $c_n - \pi(\boldsymbol{\phi}(\mathbf{x}_n))$, which can be understood as the error between the true value and the prediction of the model, times the basis function vector $\boldsymbol{\phi}(\mathbf{x}_n)$. Once we have obtained the partial derivative we must set the

resulting expression of equation (B.5) equal to zero and solve. For logistic regression the expressions in these equations are nonlinear and there is no closed-form solution. However, since the error function is convex it has a unique maximum. It requires thus special methods for their solution like the *iterative reweighted least squares* based on the *Newton-Raphson* iterative optimization scheme (see McCullagh and Nelder [141] or Rubin [142]).

Glossary

Mathematical notation

$\mathbf{x}$	Column vector $\mathbf{x}$
D	Dimension of a D-dimensional array $\mathbf{x} = (x_1, \ldots, x_D)$
$\mathcal{S}$	A sample or a set of observations $\mathcal{S}_t = (x_i, c_i), i = 1, \ldots, N; \mathbf{x}_i \in \mathbb{R}^D, c_i \in \mathcal{C} = \{c_1, \ldots, c_{\lvert\mathcal{C}\rvert}\}$ is a supervised training sample with N cases, where the i-th case has an input vector $\mathbf{x}$ in a $\mathbb{R}^D$ space and the output target or class $y_i \in \mathcal{Y}$
N	Number of cases or observations in a sample $\mathcal{S}$.
$\lvert\mathcal{C}\rvert$	Number of concepts or classes $c \in \mathcal{C}$ expressed as the cardinal number of a set.
$\mathbf{w}, \theta, \beta$	Parameters.
$p(x)$	Probability density function of a random variable x.
$p(x\lvert y)$	Conditional probability density function of a random variable x given y.
$p(x, y)$	Joint probability density function of two random variables x and y.
$E_x[f]$	Expected value of f over x.
$E_{x\lvert y}[f]$	Expected value of f over x given y.
$\hat{y}$	Estimated value of y
$\lVert\mathbf{x}\rVert$	Norm of $\mathbf{x}$.
$\mathbf{M}$	Matrix $\mathbf{M}$.
$\mathbf{M}^{\mathrm{T}}$	Matrix transpose $\mathbf{M}$
$\mathbf{M}^{-1}$	Inverse of a matrix $\mathbf{M}$.
$\lvert\mathbf{M}\rvert$	Determinant of a matrix $\mathbf{M}$.

Acronyms

acc	accuracy
AGG	Aggressive tumor: GBM and MET
BER	Balanced Error Rate
BDK	Bi-directional Kohonen Networks
BT	Brain Tumor
CDSS	Clinical Decision Support System
CS	Chemical Shift

CV	Cross Validation
GBM	Glioblastoma
GDA	Gaussian Discriminant Analysis
GUI	Graphical User Interface
^{1}H MRS	Proton Magnetic Resonance Spectroscopy
ICA	Independent Component Analysis
iDBLR	Incremental Discriminative Bayesian Logistic Regression
iGDA	Incremental Gaussian Discriminant Analysis
kRSTT	k-Random Sampling Train-Test
LDA	Linear Discriminant Analysis
LGG	Low-Grade Glial
LSSVM	Least-Squares Support Vector Machines
LTE	Long Time of Echo
MEN	Low-grade meningiomas
MET	Metastases
ML	Machine Learning
MLP	Multilayer Perceptron
MR	(Nuclear) Magnetic Resonance
MRS	Magnetic Resonance Spectroscopy
NMR	Nuclear Magnetic Resonance
PR	Pattern Recognition
PCA	Principal Component Analysis
PI	Peak integration
PR	Pattern Recognition
ReliefF	ReliefF algorithm for Recursive Elimination of Features
RF	Radio frequency
STE	Short Time of Echo
SV	Single voxel
SVM	Support Vector Machines
SW	Stepwise algorithm for feature selection in classification

Bibliography

[1] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin, "Systems Biology and New Technologies Enable Predictive and Preventative Medicine," *Science*, vol. 306, no. 5696, pp. 640–643, 2004.

[2] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nat Rev Clin Oncol*, vol. 8, no. 3, pp. 184–187, 2011.

[3] D. L. Sackett, W. M. Rosenberg, J. Gray, R. B. Haynes, and W. S. Richardson, "Evidence Based Medicine: What it is and what it isn't," *BMJ*, vol. 312, no. 7023, pp. 71–72, 1996.

[4] D. M. Eddy, "Evidence-based medicine: a unified approach.," *Health Aff (Millwood)*, vol. 24, no. 1, pp. 9–17, 2005.

[5] H. Grain, *Guide to the principles and desirable features of clinical decision support systems*. Standards Australia, 2007.

[6] F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, and J. C. Horrocks, "Computer-aided Diagnosis of Acute Abdominal Pain," *BMJ*, vol. 2, no. 7023, pp. 9–13, 1972.

[7] B. G. Buchanan and E. H. Shortliffe, *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1984.

[8] INTERPRET Consortium, "INTERPRET." Web site, 1999-2001. IST-1999-10310, EC, http://gabrmn.uab.es/interpret/.

[9] eTUMOUR Consortium, "eTumour: Web accessible MR Decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data." Web site. FP6-2002-LIFESCIHEALTH 503094, VI framework programme, EC, http://www.etumour.net.

[10] H. González-Vélez, M. Mier, M. Julià-Sapé, T. N. Arvanitis, J. M. García-Gómez, M. Robles, P. H. Lewis, S. Dasmahapatra, D. Dupplaw, A. C. Peet, C. Arús, B. Celda, S. V. Huffel, and M. Lluch i Ariet, "HealthAgents: Distributed multiagent brain tumor diagnosis and prognosis," *Applied Intelligence*, vol. 30, pp. 191–202, June 2009.

[11] A. R. Tate, J. Underwood, D. M. Acosta, M. Julià-Sapé, C. Majós, A. Moreno-Torres, F. A. Howe, M. van der Graaf, V. Lefournier, M. M. Murphy, A. Loosemore, C. Ladroue, P. Wesseling, J. L. Bosson, M. E. Cabañas, A. W. Simonetti, W. Gajewicz, J. Calvar, A. Capdevila, P. R. Wilkins, B. A. Bell, C. Rémy, A. Heerschap, D. Watson, J. R. Griffiths, and C. Arús, "Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra," *NMR in Biomedicine*, vol. 19, no. 4, pp. 411–434, 2006.

[12] G. Hagberg, "From magnetic resonance spectroscopy to classification of tumors. a review of pattern recognition methods," *NMR in Biomedicine*, vol. 11, no. 4-5, pp. 148–156, 1998.

[13] A. R. Tate, C. Majós, Àngel Moreno, F. A. Howe, J. R. Griffiths, and C. Arús, "Automated classification of short echo time in in vivo 1H brain tumor spectra: a multicenter study," *Magnetic Resonance in Medicine*, vol. 49, no. 1, pp. 29–36, 2003.

[14] B. H. Menze, M. P. Lichy, P. Bachert, B. M. Kelm, H.-P. Schlemmer, and F. A. Hamprecht, "Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors," *NMR in Biomedicine*, vol. 19, pp. 599–609, Aug 2006.

[15] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, "Prognosis and prognostic research: validating a prognostic model," *BMJ*, vol. 338, no. b605, pp. 1432–1435, 2009.

[16] H. C. van Houwelingen, "Validation, calibration, revision, and combination of prognostic survival models," *Statistics in Medicine*, vol. 19, no. 23, pp. 3401–3415, 2000.

[17] E. W. Steyerberg, G. J. Borsboom, H. C. van Houwelingen, M. J. Eijkemans, and J. D. F. Habbema, "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage," *Statistics in Medicine*, vol. 23, no. 16, pp. 2567–2586, 2004.

[18] K. J. Janssen, K. G. Moons, C. J. Kalkman, D. E. Grobbee, and Y. Vergouwe, "Updating methods improved the performance of a clinical prediction model in new patients," *Journal of Clinical Epidemiology*, vol. 61, no. 1, pp. 76–86, 2008.

[19] N. W. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 377–382, ACM, 2001.

[20] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, eds., *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[21] J. G. Moreno-Torres, T. Raeder, R. Alaíz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[22] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, pp. 227–244, Oct. 2000.

[23] J. M. García-Gómez, *Pattern Recognition Approaches for Biomedical Data in Computer-Assisted Cancer Research*. PhD thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2009.

[24] J. M. García-Gómez, J. Luts, M. Julià-Sapé, P. Krooshof, S. Tortajada, J. Vicente, W. Melssen, E. Fuster-Garcia, I. Olier, G. Postma, D. Monleón, A. Moreno-Torres, J. Pujol, A.-P. Candiota, M. C. Martínez-Bisbal, J. Suykens, L. Buydens, B. Celda, S. V. Huffel, C. Arús, and M. Robles, "Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy.," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 22, no. 1, pp. 5–18, 2009.

[25] J. M. García-Gómez, S. Tortajada, C. Vidal, M. Julià-Sapé, J. Luts, A. Moreno-Torres, S. V. Huffel, C. Arús, and M. Robles, "The effect of combining two echo times in automatic brain tumor classification by MRS," *NMR in Biomedicine*, vol. 21, pp. 1112–1125, Sep 2008.

[26] C. Sáez, J. M. García-Gómez, J. Vicente, S. Tortajada, E. Fuster-Garcia, M. Esparza, A. T. Navarro, and M. Robles, "Curiam BT 1.0, decision support system for brain tumour diagnosis," in *ESMRMB 2009: 26th Annual Scientific Meeting*, Springer, Oct 2009.

[27] C. Sáez, J. M. García-Gómez, J. Vicente, S. Tortajada, J. Luts, D. Dupplaw, S. V. Huffel, and M. Robles, "A generic and extensible automatic classification framework applied to brain tumour diagnosis in HealthAgents," *The Knowledge Engineering Review*, vol. 26, pp. 283–301, 2011.

[28] S. Tortajada, E. Fuster-Garcia, J. Vicente, P. Wesseling, F. A. Howe, M. Julià-Sapé, A.-P. Candiota, D. Monleón, À. Moreno-Torres, J. Pujol, J. R. Griffiths, A. Wright, A. Peet, M. C. Martinez-Bisbal, B. Celda, C. Arús, M. Robles, and J. M. García-Gómez, "Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis," *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 677–687, 2011.

[29] E. Fuster-Garcia, C. Navarro, J. Vicente, S. Tortajada, J. M. García-Gómez, C. Sáez, J. Calvar, J. Griffiths, M. Julià-Sapé, F. a. Howe, J. Pujol, A. C. Peet, A. Heerschap, A. Moreno-Torres, M. C. Martínez-Bisbal, B. Martínez-Granados, P. Wesseling, W. Semmler, J. Capellades, C. Majós, A. Alberich-Bayarri, A. Capdevila, D. Monleón, L. Martí-Bonmatí, C. Arús, B. Celda, and M. Robles, "Compatibility between 3T 1H SV-MRS data and automatic brain tumour diagnosis support systems based on databases of 1.5T 1H SV-MRS spectra.," *Magma (New York, N.Y.)*, vol. 24, no. 1, pp. 35–42, 2011.

[30] J. Vicente, E. Fuster-Garcia, S. Tortajada, J. M. García-Gómez, N. Davies, K. Natarajan, M. Wilson, R. G. Grundy, P. Wesseling, D. Monleón, B. Celda, M. Robles, and A. C. Peet, "Accurate classification of childhood brain tumours by *in vivo* ^{1}H MRS – a multi-centre study," in *Proceedings of the 15th International Symposium on Pediatric Neuro-Oncology, Toronto, Canada*, 2012. In press.

[31] J. Vicente, J. M. García-Gómez, S. Tortajada, A. T. Navarro, F. A. Howe, A. C. Peet, M. Julià-Sapé, B. Celda, P. Wesseling, M. Lluch-Ariet, and M. Robles, "Ranking of Brain Tumour Classifiers Using a Bayesian Approach," in *Bio-inspired systems: Computational and Ambient Intelligence (Part I)*, vol. 5517 of *Lecture Notes in Computer Science*, pp. 33–50, Springer, 2009.

[32] F. A. Graybill and R. B. Deal, "Combining unbiased estimators," *Biometrics*, vol. 15, pp. 543–550, 1959.

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY: Wiley-Interscience, 2001.

[34] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[35] V. N. Vapnik, *Statistical learning theory*. Wiley, 1st ed., Sept. 1998.

[36] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, second ed., 2004.

[37] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed., July 2003.

[38] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Wiley, 1994.

[39] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[40] T. Minka, "Expectation Propagation for approximate Bayesian inference," in *Proceedings 17th Conference on Uncertainty in Artificial Intelligence*, pp. 362–369, Morgan Kauffman, 2001.

[41] T. Minka, *A family of approximate algorithms for Bayesian inference*. PhD thesis, MIT, 2001.

[42] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applica- tions," *Biometrika*, vol. 57, pp. 97–109, 1970.

[43] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[44] R. M. Neal, *Bayesian Learning for Neural Networks*. Springer. Lecture Notes in Statistics, 118, 1996.

[45] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo Methods for Bayesian computation*. Springer, 2001.

[46] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, pp. 107–113, 2008.

[47] A. Cornuéjols, *Ubiquitous Knowledge Discovery*, ch. On-line learning: Where are we so far?, pp. 129–147. Springer-Verlag, LNAI, 2010.

[48] E. M. Gold, "Language identification in the limit," *Information and Control*, vol. 10, no. 5, pp. 447–474, 1967.

[49] A. Sharma, "A note on batch and incremental learnability," *Journal of Computer and System Sciences*, vol. 56, no. 3, pp. 272–276, 1998.

[50] K. P. Jantke, "Types of Incremental Learning," in *Proceedings of the AAAI Spring Symposium: Training Issues on Incremental Learning*, vol. SS-93-06, pp. 26–32, 1993.

[51] S. Lange and G. Grieser, "On the power of incremental learning," *Theoretical Computer Science*, vol. 288, pp. 277–307, october 2002.

[52] M. A. Maloof and R. S. Michalski, "Incremental learning with partial instance memory," *Artificial Intelligence*, vol. 154, no. 1-2, pp. 95–126, 2004.

[53] J.-L. Sancho, W. E. Pierson, B. Ulug, A. R. Figueiras-Vidal, and S. C. Ahalt, "Class separability estimation and incremental learning using boundary methods," *Neurocomputing*, vol. 35, no. 1-4, pp. 3–26, 2000.

[54] A. Shilton, M. Palaniswami, D. Ralph, and A. C. Tsoi, "Incremental training of support vector machines," *IEEE Transactions on Neural Networks*, vol. 16, pp. 114–131, 2005.

[55] J. C. Schlimmer and D. H. Fisher, "A case study of incremental concept induction," in *5th National Conference on Artificial Intelligence*, pp. 496–501, 1986.

[56] P. E. Utgoff, "Incremental induction of decision trees," *Machine Learning*, vol. 4, pp. 161–186, 1989.

[57] Z.-H. Zhou and Z.-Q. Chen, "Hybrid decision tree," *Knowledge-Based Systems*, vol. 15, no. 8, pp. 515 – 528, 2002.

[58] J. Gama and P. Medas, "Learning decision trees from dynamic data streams," *Journal of Universal Conputer Science*, vol. 11, no. 8, pp. 1353–1366, 2005.

[59] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Transactions on Neural Networks*, vol. 3, pp. 698–713, Sep 1992.

[60] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: an incremental learning algorithm for supervised neural networks," *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, vol. 31, no. 4, pp. 497–508, 2001.

[61] K. Yamauchi, T. Oohira, and T. Omori, "Fast incremental learning methods inspired by biological learning behavior," *Artificial Life and Robotics*, vol. 9, no. 3, pp. 128–134, 2005.

[62] M. Muhlbaier, A. Topalis, and R. Polikar, "Learn++.NC: Combining Ensemble of Classifiers Combined with Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 152–168, 2009.

[63] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517–1531, 2011.

[64] S. Wan and L. E. Banta, "Parameter Incremental Learning Algorithm for Neural Networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1424–1438, 2006.

[65] L. Fu, "Incremental knowledge acquisition in supervised learning networks," *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, vol. 26, pp. 801–809, Nov. 1996.

[66] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, and H. Zhang, "An eigenspace update algorithm for image analysis," *Graphical Models Image Processing*, vol. 59, no. 5, pp. 321–332, 1997.

[67] P. Hall, D. Marshall, and R. Martin, "Merging and splitting eigenspace models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1042–1049, 2000.

[68] S. Pang, S. Ozawa, and N. Kasabov, "Incremental linear discriminant analysis for classification of data streams," *IEEE Transactions on System, Man and Cybernetics*, vol. 35, no. 5, pp. 905–914, 2005.

[69] T.-K. Kim, S.-F. Wong, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.

[70] Z. Huang, K. Ding, L. Jin, and X. Gao, "Writer Adaptive Online Handwriting Recognition Using Incremental Linear Discriminant Analysis," pp. 91–95, 2009.

[71] G. Widmer and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning*, vol. 23, pp. 69–101, 1996.

[72] M. Scholz and R. Klinkenberg, "Boosting classifiers for drifting concepts," *Intelligent Data Analysis (IDA), Special Issue on Knowledge Discovery from Data Streams*, vol. 11, pp. 3–28, 2007.

[73] P. P. Rodrigues, J. a. Gama, J. a. Araújo, and L. Lopes, "L2GClust: local-to-global clustering of stream sources," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, pp. 1006–1011, ACM, 2011.

[74] J. Z. Kolter and M. A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts," *J. Mach. Learn. Res.*, vol. 8, pp. 2755–2790, December 2007.

[75] R. Klinkenberg, "Learning drifting concepts: Example selection vs. example weighting," *Intelligent Data Analysis*, vol. 8, pp. 281–300, 2004.

[76] P. Langley, "Order effects in Incremental Learning," in *Learning in humans and machines: Towards an interdisciplinary learning science* (P. Reimann and H. Spada, eds.), pp. 1–17, Oxford: Elsevier, 1995.

[77] C. Giraud-Carrier, "A note on the utility of incremental learning," *AI Communications*, vol. 13, no. 4, pp. 215–223, 2000.

[78] S. Grossberg, "Nonlinear neural networks: principles, mechanisms and architectures," *Neural Networks*, vol. 1, no. 1, pp. 17–61, 1998.

[79] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *The psychology of learning and motivation*, vol. 24, pp. 109–164, 1989.

[80] R. Ratcliff, "Catastrophic models of recognition memory: Constraints imposed by learning and forgetting functions," *Psychological review*, vol. 97, pp. 285–308, 1990.

[81] S. Lange and S. Zilles, "Formal models of incremental learning and their analysis," in *International Joint Conference Neural Networks*, vol. 4, pp. 2691–2696, 2003.

[82] J. Macek, "Incremental learning of ensemble classifiers on ECG data," in *IEEE Symp. Comput. Based Med. Syst.*, pp. 315–320, 2005.

[83] A. Cornuéjols, "Getting order indepence in Incremental Learning," in *AAAI Spring Symposium on Training Issues in Incremental Learning*, pp. 43–54, 1993.

[84] N. Di Mauro, F. Esposito, S. Ferilli, and T. M. A. Basile, "Avoiding Order Effects in Incremental Learning," in *In S. Bandini and S. Manzoni (Eds.), Advances in Artificial Intelligence (AI*IA05) LNCS*, pp. 110–121, Springer-Verlag, 2005.

[85] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, (San Francisco, CA, USA), pp. 1137–1143, Morgan Kaufmann Publishers Inc., 1995.

[86] A. Dodd, "An introduction to the theory of nuclear magnetic resonance," tech. rep., University of Melbourne, 2003.

[87] M. Julià-Sapé, D. Acosta, C. Majós, Àngel Moreno-Torres, P. Wesseling, J. J. Acebes, J. R. Griffiths, and C. Arús, "Comparison between neuroimaging classifications and histopathological diagnoses using an international multicenter brain tumor magnetic resonance imaging database," *J Neurosurg*, vol. 105, pp. 6–14, Jul 2006.

[88] A. P. Lin, T. T. Tran, and B. D. Ross, "Impact of evidence-based medicine on magnetic resonance spectroscopy," *NMR Biomed*, vol. 19, pp. 476–483, Jun 2006.

[89] I. Barba, A. Moreno, I. Martinez-Perez, A. R. Tate, M. E. Cabanas, M. Baquero, A. Capdevila, and C. Arús, "Magnetic resonance spectroscopy of brain hemangiopericytomas: high myoinositol concentrations and discrimination from meningiomas," *J Neurosurg*, vol. 94, pp. 55–60, Jan 2001.

[90] M. Kaminogo, H. Ishimaru, M. Morikawa, M. Ochi, R. Ushijima, M. Tani, Y. Matsuo, J. Kawakubo, and S. Shibata, "Diagnostic potential of short echo time MR spectroscopy of gliomas with single-voxel and point-resolved spatially localised proton spectroscopy of brain," *Neuroradiology*, vol. 43, pp. 353–363, May 2001.

[91] F. A. Howe and K. S. Opstad, "1H MR spectroscopy of brain tumours and masses," *NMR Biomed*, vol. 16, pp. 123–131, May 2003.

[92] L. Lukas, A. Devos, J. A. K. Suykens, L. Vanhamme, F. A. Howe, C. Majós, A. Moreno-Torres, M. V. D. Graaf, A. R. Tate, C. Arús, and S. V. Huffel, "Brain tumor classification based on long echo proton MRS signals," *Artif. Intell. Med.*, vol. 31, pp. 73–89, 2004.

[93] A. W. Simonetti, W. J. Melssen, F. Szabo de Edelenyi, J. J. A. van Asten, A. Heerschap, and L. M. C. Buydens, "Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification," *NMR in Biomedicine*, vol. 18, pp. 34–43, Feb 2005.

[94] D. Galanaud, F. Nicoli, O. Chinot, S. Confort-Gouny, D. Figarella-Branger, P. Roche, S. Fuentes, Y. Le Fur, J.-P. Ranjeva, and P. J. Cozzone, "Noninvasive diagnostic assessment of brain tumors using combined in vivo MR imaging and spectroscopy," *Magnetic Resonance in Medicine*, vol. 55, pp. 1236–1245, Jun 2006.

[95] F. F. González-Navarro, L. A. Belanche-Muñoz, E. Romero, A. Vellido, M. Julií-Sapé, and C. Arús, "Feature and model selection with discriminatory visualization for diagnostic classification of brain tumors," *Neurocomput.*, vol. 73, pp. 622–632, Jan. 2010.

[96] A. Vellido, E. Romero, M. Julià?Sapé, C. Majós, A. Moreno?Torres, , and C. Arús, "Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single–voxel proton mrs.," *NMR in biomedicine*, vol. 25, no. 6, pp. 819–828, 2012.

[97] Y. Huang, P. J. G. Lisboa, and W. El-Deredy, "Tumour grading from magnetic resonance spectroscopy: a comparison of feature extraction with variable selection," *Stat Med*, vol. 22, pp. 147–164, Jan 2003.

[98] A. Devos, L. Lukas, J. A. K. Suykens, L. Vanhamme, A. R. Tate, F. A. Howe, C. Majós, À. Moreno-Torres, M. van der Graaf, C. Arús, and S. Van Huffel, "Classification of brain tumours using short echo time 1H MR spectra," *J Magn Reson*, vol. 170, pp. 164–175, Sep 2004.

[99] J. A. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300, 1999.

[100] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, D. Watson, and T. I. Consortium, "A Multi-Centre, Web-Accessible and Quality Control-Checked Database of in vivo MR Spectra of Brain Tumour Patients, journal = Magnetic Resonance Materials in Physics, Biology and Medicine, pages = 22–33, volume = 19, issue = 1, year = 2006,"

[101] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, "The 2007 who classification of tumours of the central nervous system.," *Acta Neuropathol*, vol. 114, 2007.

[102] M. van der Graaf, M. Julià-Sapé, F. A. Howe, A. Ziegler, C. Majós, À. Moreno-Torres, M. Rijpkema, D. Acosta, K. S. Opstad, Y. M. van der Meulen, C. Arús, and A. Heerschap, "Mrs quality assessment in a multicentre study on mrs-based classification of brain tumours.," *NMR in biomedicine*, vol. 21, pp. 148–158, 2008.

[103] U. Klose, "In vivo proton spectroscopy in presence of eddy currents," *Magnetic Resonance in Medicine*, vol. 14, pp. 26–30, April 1990.

[104] A. Naressi, C. Couturier, I. Castang, R. de Beer, and D. Graveron-Demilly, "Java-based graphical user interface for MRUI, a software package for quantitation of in vivo/medical magnetic resonance spectroscopy signals," *Comput Biol Med*, vol. 31, pp. 269–286, Jul 2001.

[105] E. Cabanes, S. Confort-Gouny, Y. Le Fur, G. Simond, and P. J. Cozzone, "Optimization of residual water signal removal by HLSVD on simulated short echo time proton MR spectra of the human brain," *J Magn Reson*, vol. 150, pp. 116–125, Jun 2001.

[106] I. T. Jolliffe, *Principal Component Analysis*. Springer, second ed., Oct. 2002.

[107] P. Comon, "Independent Component Analysis, a new concept?," *Signal Processing, Elsevier*, vol. 36, pp. 287–314, Apr. 1994.

[108] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[109] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[110] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, pp. 23–69, 2003.

[111] J. Hoch and A. Stern, *NMR Data Processing*. New York, NY: John Wiley & Sons, 1996.

[112] M. Preul, Z. Caramanos, D. Collins, J. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, and D. Arnold, "Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy," *Nature Medicine*, vol. 2, no. 3, pp. 323–325, 1996.

[113] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*, pp. 318–362. Cambridge, MA, USA: MIT Press, 1986.

[114] W. Melssen, R. Wehrens, and L. Buydens, "Supervised kohonen networks for classification problems," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 99–113, 2006.

[115] T. Kohonen, *Self-Organizing Maps*. Springer, 3rd ed., 2001.

[116] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 ed., 2000.

[117] C. Majós, M. Julià-Sapé, J. Alonso, M. Serrallonga, C. Aguilera, J. J. Acebes, C. Arú, and J. Gili, "Brain tumor classification by proton MR spectroscopy: comparison of diagnostic accuracy at short and long TE," *AJNR Am J Neuroradiol*, vol. 25, pp. 1696–1704, Nov 2004.

[118] K. S. Opstad, C. Ladroue, B. A. Bell, J. R. Griffiths, and F. A. Howe, "Linear discriminant analysis of brain tumour 1H MR spectra: a comparison of classification using whole spectra versus metabolite quantification," *NMR in Biomedicine*, vol. 20, no. 8, pp. 763–770, 2007.

[119] H. Poptani, J. Kaartinen, R. K. Gupta, M. Niemitz, Y. Hiltunen, and R. A. Kauppinen, "Diagnostic assessment of brain tumours and non-neoplastic brain disorders in vivo using proton nuclear magnetic resonance spectroscopy and artificial neural networks.," *J Cancer Res Clin Oncol*, vol. 125, no. 6, pp. 343–349, 1999.

[120] J. Luts, J.-B. Poullet, J. M. García-Gómez, A. Heerschap, M. Robles, J. A. Suykens, and S. V. Huffel, "The effect of feature extraction for brain tumour classification based on short echo time 1h mr spectra," *Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 288–298, 2008.

[121] P. S. Mischel, R. Shai, T. Shi, S. Horvath, K. V. Lu, G. Choe, D. Seligson, T. J. Kremen, A. Palotie, L. M. Liau, T. F. Cloughesy, and S. F. Nelson, "Identification of molecular subtypes of glioblastoma by gene expression profiling," *Oncogene*, vol. 22, pp. 2361–2373, Apr 2003.

[122] N. F. Zhang, "The uncertainty associated with the weighted mean of measurement data," *Metrologia*, vol. 43, pp. 195–204, 2006.

[123] N. Pal, J.-J. Lin, C.-H. Chang, and S. Kumar, "A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill-Deal estimator," *Computational Statistics & Data Analysis*, vol. 51, pp. 5673–5681, August 2007.

[124] Y. Freund and R. E. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," *Computer Systems Science*, vol. 57, no. 1, pp. 119–139, 1997.

[125] R. E. Schapire, "The Strength of Weak Learnability.," *Machine Learning*, vol. 5, pp. 197–227, 1990.

[126] A. Frank and A. Asuncion, "UCI machine learning repository." [http://archive.ics.uci.edu/ml], University of California, Irvine, School of Information and Computer Sciences, 2010.

[127] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165–175, Mar 1989.

[128] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis Journal*, vol. 6, pp. 429–449, Nov 2002.

[129] M. Julià-Sapé, D. M. Acosta, M. Mier, C. Arús, and D. Watson, "A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 19, pp. 22–33, Feb 2006.

[130] M. van der Graaf, M. Julià-Sapé, F. A. Howe, A. Ziegler, C. Majós, A. Moreno-Torres, M. Rijpkema, D. M. Acosta, K. S. Opstad, Y. M. van der Meulen, C. Arús, and A. Heerschap, "MRS quality assessment in a multicentre study on MRS-based classification of brain tumours," *NMR Biomed*, vol. 21, no. 2, pp. 148–158, 2008.

[131] L. Landini, V. Positano, and M. F. Santarelli, eds., *Advanced Image Processing in Magnetic Resonance Imaging*. CRC Press, 2005.

[132] R. Dammers, J. W. Schouten, I. K. Haitsma, A. J. Vincent, J. M. Kros, and C. M. Dirven, "Towards improving the safety and diagnostic yield of stereotactic biopsy in a single centre," *Acta Neurochir (Wien)*, vol. 152, no. 11, pp. 1915–21, 2010.

[133] C. Sáez, J. M. García-Gómez, J. Vicente, S. Tortajada, M. Esparza, A. T. Navarro, E. Fuster-Garcia, M. Robles, L. Martí-Bonmatí, and C. Arús, "A generic Decision Support System featuring an assembled view of predictive models for Magnetic Resonance and clinical data," in *ESMRMB 2008: 25th Annual Scientific Meeting*, Springer, Oct. 2008.

[134] D. J. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[135] R. Fletcher, *Practical Methods of Optimization*. New York: John Wiley & Sons, 2nd ed., 1987.

[136] M. A. Girolami and S. Rogers, "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors," *Neural Computation*, vol. 18, no. 8, pp. 1790 – 1817, 2006.

[137] T. P. Minka, "Expectation Propagation for approximate Bayesian inference," in *UAI*, pp. 362–369, 2001.

[138] J. Gama and P. Kosina, "Learning about the Learning Process," in *Advances in Intelligent Data Analysis X*, vol. 7014 of *Lecture Notes in Computer Science*, pp. 162–172, Springer Berlin / Heidelberg, 2011.

[139] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 415–447, 1991.

[140] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[141] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Chapman and Hall, 2nd ed., 1989.

[142] D. B. Rubin, *Encyclopedia of Statistical Sciences*, vol. 4, ch. Iteratively reweighted least squares, pp. 272–275. Wiley, 1983.

List of Figures

List of Tables

List of Algorithms