A simple and efficient kNN variant with embedded feature selection

Almudena Moreno-Ribera[1], Aida Calviño[1]
[1]Department of Statistics and Data Science, Complutense University of Madrid, Spain.

Abstract

Predictive modeling aims at providing estimates of an unknown variable, the target, from a set of known ones, the input. The k Nearest Neighbors (kNN) is one of the best-known predictive algorithms due to its simplicity and well behavior. However, this class of models has some drawbacks, such as the non-robustness to the existence of irrelevant input features or the need to transform qualitative variables into dummies, with the corresponding loss of information for ordinal ones. In this work, a kNN regression variant, easily adaptable for classification purposes, is suggested. The proposal allows dealing with all types of input variables while embedding feature selection in a simple and efficient manner, reducing the tuning phase. More precisely, making use of the weighted Gower distance, we develop a powerful tool to cope with these inconveniences by implementing different weighting schemes. The proposed method is applied to a collection of 20 data sets, different in size, data type and the distribution of the target variable. Moreover, the results are compared with previously proposed kNN variants, showing its supremacy, particularly when the weighting scheme is based on non-linear association measures and in datasets that contain at least one ordinal input variable.

Keywords: *Gower distance; weighting scheme; ordinal variables; Machine Learning; predictive modeling; regression.*

Editorial Universitat Politècnica de València