Robust risk estimation using exact resampling criteria for the kNN algorithm

Alain Célisse
Université Lille 1, Lille, France

Tristan Mary-Huard*
AgroParisTech - INRA, Paris, France. maryhuar@agroparistech.fr

We consider the problem of predicting the unknown label Y of an observation X based on a training sample, i.e. a set of n observations for which both X and Y are available. In the regression framework, the label to be predicted is a quantitative variable, whereas in binary classification the label is boolean. A classical strategy consists in predicting the label of a new observation according to the k most similar observations in the training sample. This strategy is known as the k-Nearest Neighbor algorithm (kNN), and has been successfully applied in a number of contexts (biomedecine, genomics, economics). However, the performance of this algorithm highly depends on the tuning of parameter k, that should be performed adaptively to the data at hand. Resampling strategies such as Bootstrap or Leave-p-out (LpO) cross-validation can be used to estimate the prediction performance obtained with different values of k, and select the optimal value k^* that minimizes the prediction error rate. While popular, theoretical guarantees for the use these resampling methods to select k were only obtained for the L1O. Moreover, due to computational cost considerations in practice one needs to limit the number of resamplings as soon as the training sample size is large, yielding a poor approximation of the actual risk. We will present computational shortcuts to obtain the exact LpO or Bootstrap risk estimators in a reasonable computational time (linear with respect to n in the case of LpO, whatever p). We will also investigate the theoretical properties of the LpO estimator, and illustrate on simulations how the choice of parameter p may be crucial to efficiently select the tuning parameter k of the kNN algorithm.

Keywords: Classification, Cross-validation, Model selection