T-61.3050 Classification challenge 2014

Andrea Nodari 466165
andrea.nodari@aalto.fi

October 17, 2014

This summary explains the followed approach used to assign each element of the challenge dataset a *type* (i.e. "Red" or "White") and a *quality* (i.e. an integer between 1 and 7).

For the first task a logistic regression has been used. The dataset has been split in three parts: training, validation and testing set. The validation set is used to pick the best value for lambda (the regularisation parameter) and the testing set is used to analyse how the model generalise. Afterwards, all the training set is used to predict the labels for the challenge dataset using the lambda previously found. The accuracy on the testing set is 99.16%.

For the second task, after trying different methods with poor results, a k-nearest-neighbors classifiers has been used. In order to decide the best value for k,a lot of them have been tried using cross-validation to pick the best one. The experiment used two implementations of the model: one implemented by me, and the other using the matlab library. The first uses euclidean distance, the latter minkowski distance. If $k > 1$ in order to break a tie the class of the closest point is chosen. The best value for k is 1 and the accuracy using cross validation is 63%.

In order to increase the performance both exhaustive feature selection and dimensionality reduction with PCA have been used. Both of them do not significantely change the performance even if PCA reduces the computation time and allows to visualize the data in a 3-dimensional space.

All the code, charts and further details will be released open source at this URL: `https://github.com/nodo/machine-learning-challenge`.