Practice 7/8A: Naive Bayes and k-Nearest Neighbor

Supervised Learning - 10 December 2021

User Knowledge Modeling Data Set - classification problem

1. Load, explore and preprocess (if needed) the dataset. Consider we are going to train a Naive Bayes classifier, and a k-Nearest Neighbor classifier: which kind of preprocessing is required for the two models?

2. Train a Naive Bayes classifier (identify the correct classifier to train, according to the data types). Access the generalization performances of the model by means of a cross fold validation methodology, with $k = 4$ folds.

3. Split the dataset in train and test sets and train a single Naive Bayes classifier on the train set. Evaluate it on the test set (accuracy, precision, recall, f1-score) and calculate a confusion matrix.

4. Train a k-Nearest Neighbour classifier (identifying the best value of k) on the previously selected training set. Evaluate it on the test set (accuracy, precision, recall, f1-score) and calculate a confusion matrix.

5. Compare the results obtained by the two models.