Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 4

18.05.2017
Deadline: 25.05.2017 - 23:55

Task 1: ROC Curve *(3 Points)*

Given the following predictions of a classifier and the true class, give the ROC curve for the classifier. Is the performance of the classifier good? Explain using the curve.

Class	Prediction
P	0.95
N	0.85
P	0.78
P	0.66
N	0.60
P	0.55
N	0.53
N	0.52
N	0.51
P	0.40

Table 1: Prediction of a classifier on new instances.

Task 2: Cross-Validation *(7 Points)*

The goal of this exercise is to implement stratified cross-validation (`Exercise_04_02.java`).

(1) Implement the stratification, which splits the entire dataset into a list of datasets, based on the value of the class attribute.

(2) Implement the `shuffle` method, which shuffles a dataset.

(3) Implement `trainCV` and `testCV` methods, which extract training and test subsets of the dataset for the specified fold index.

(4) Implement `stratifiedCrossValidation`, return the average accuracy.

Task 3: McNemar's test *(7 Points)*

In this task, you are supposed to compare random forests with decision trees again. But this time, the evaluation is more extensive and leaves some choices. You may choose any implementation of random forests and decision trees to perform the following tasks:

(1) Choose three datasets (e.g., `http://archive.ics.uci.edu/ml/` or `http://www.cs.waikato.ac.nz/ml/weka/datasets.html`).

(2) For each dataset, compare random forests with decision trees using McNemar's Test.

(3) Are random forests performing significantly better than decision trees?

(4) Briefly discuss your results.

Note: You may use the Weka framework for each step.

Task 4: Accuracy *(3 Points)*

Show that accuracy can be expressed through sensitivity, specificity, and prevalence $\left(\frac{P}{P+N}\right)$. Implement the formula in `Exercise_04_04.java`.