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

Machine Learning - Sheet 4

17.05.2018
Deadline: 24.05.2018 - 16:00

Task 1: Bias-Variance decomposition *(2 Points)*

Given in `curve_fitting.py` is a Python script for fitting polynomials (Python 3, numpy, and matplotlib required). Run this script using the values $1, 2, 3, 4, 5, 10$ for d. What does this little experiment tell you about the bias-variance decomposition?

Task 2: 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 3: Cross-Validation *(7 Points)*

The goal of this exercise is to implement stratified k-cross-validation.

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

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

(c) Implement `trainCV` and `testCV` methods, which extract training and test subsets of the dataset for the specified fold index and save them in a separate file.

(d) Implement `stratifiedCrossValidation`, which gets k as an input parameter and returns the mean and standard deviation of the accuracy.

(e) Run your `stratifiedCrossValidation` on the car dataset (use your decision tree implementation from Sheet 2 as a classifier) with $k = 10$.

Task 4: McNemar's test *(5 Points)*

In this task, you are supposed to compare random forests with decision trees again. This time use Weka API of random forests and J48 to perform the following tasks on the car and the diabetes datasets (`http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff`)

(a) Use your `stratifiedCrossValidation` with $k = 10$ to generate different tain and test subsets. Run random forests and J48 on them.

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

(c) Are random forests performing significantly better than decision trees? Discuss your results.

Task 5: Accuracy *(3 Points)*

Show that accuracy can be expressed through sensitivity, specificity, and prevalence $\left(\frac{P}{P+N}\right)$.