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

Machine Learning - Sheet 2

26.04.2018
Deadline: 03.05.2018 - 12:00

Task 1: Decision Tree *(20 Points)*

We now finish the implementation of the basic decision tree algorithm, as described in Section 3.4 (page 55) in [1].

(a) Implement the data structure of the decision tree (inner nodes, leaves, the actual decision method).

(b) Implement a method or function that performs the attribute selection for a given node (reuse `informationGain` method already implemented in the previous exercise).

(c) Implement methods `trainModelOnSubset` and `trainModel`.

(d) Test your implementation on the Weather dataset (`weather.nominal.arff`).

(e) Evaluate your decision tree using the car dataset (`http://archive.ics.uci.edu/ml/datasets/Car+Evaluation`) and a very simple procedure: Randomly split the dataset into a training set (two thirds) and a test set (one third). Take the training set to train your decision tree. Afterward, compute the percentage of correctly classified instances from the test set.

(f) Repeat the previous step 10 times and report the mean and standard deviation of the resulting accuracies.

(g) Modify your implementation by taking the maximum depth of the tree as an input parameter.

(h) Repeat steps (d), (e), and (f) for depth of $\{3, 5, 10, 20\}$ and report your results. Discuss your findings.

References

[1] Tom M. Mitchell. *Machine learning.* McGraw Hill series in computer science. McGraw-Hill, 1997.