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

Machine Learning - Sheet 2

07.05.2020
Deadline: 14.05.2020 - 18:00

Task 1: Decision Tree (2) *(10 Points)*

Continue with your code from the previous assignment.

1. (*2 points*) Implement a very simple evaluation procedure of your decision tree: Randomly split the dataset into a training set and a test set. Take the training set to train your decision tree. Afterward, compute the percentage of correctly classified instances from the test set. The seed for random selection and the ratio of the training data should be provided as input parameters.

2. (*1 point*) To reduce the effect of random seed for the training instances, implement a code that repeats the previous step n times with different random seeds and reports the mean and standard deviation of the resulting accuracy. n is an input parameter.

3. (*2 points*) Run task (b) on the car dataset (`http://archive.ics.uci.edu/ml/datasets/Car+Evaluation`) and $n = 10$. Change the portion of training data from $\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{9}{10}\}$ and compare the results.

4. (*3 points*) Modify your decision tree implementation by taking the maximum depth of the tree as a pre-pruning strategy and as an input parameter.

5. (*2 points*) Repeat task (c) for a depth of $\{1, 3, 5\}$ and report your results. Discuss your findings.

Please note that we will use this implementation in later exercises!