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

Machine Learning - Sheet 3

03.05.2018
Deadline: 11.05.2018 - 12:00

Task 1: Random Forests *(8 Points)*

In this task, you are supposed to compare the performance of decision trees with random forests on the car dataset (`http://archive.ics.uci.edu/ml/datasets/Car+Evaluation`). First, make yourself familiar with classification in WEKA (`http://www.cs.waikato.ac.nz/ml/weka/`) particularly random forests (use the API not the GUI!).

(a) What is the relation between bagging and random forests? Briefly describe both methods, point out the differences.

(b) Analyze how the performance of WEKA's random forest is affected when using different numbers of trees (e.g., 1, 5, 10, 20, 30, 40, 50). To do so, use the car dataset and a very simple procedure: Randomly split the dataset into a training set and a test set (two thirds and one third). Take the first to train the random forest. Afterward, use the test set to compute the percentage of correctly classified instances. Discuss your results.

(c) Repeat the previous step 10 times and compare the performance of WEKA's random forests with your decision tree results (part (f) from exercise sheet 2).

Task 2: Boosting *(10 Points)*

Read pages 358-362 of the book Data Mining [1] and make yourself familiar with Boosting and AdaBoost algorithms. Our goal is to implement the boosting pseudo code from page 359.

(a) Implement `sampling` method that samples from training data based on their input weights.

(b) Implement `modelGeneration` method that takes two arguments: a sampled input data, and a maximum number of iterations (use your decision tree implementation from the previous exercise as the base classifier).

(c) Implement `classification` method that returns the predicted class of a test instance.

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

(e) Evaluate your boosting algorithm using the car dataset and the same procedure as in part (b) of Task 1.

(f) Repeat the previous step 10 times and report the mean and standard deviation of the resulting accuracies. Compare your results with the results of random forest and your decision tree implementation.

(g) Change the maximum depth of base classifiers (e.g., $\{1, 2, 3, ..., 10\}$) and discuss your observations.

Task 3: Boosting *(1 Points)*
What happens if strong learners are used in AdaBoost?

Task 4: Bootstrap *(1 Points)*
What does 0.632 bootstrap mean? What happens if the number of drawn samples is doubled?

References

[1] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.