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

Machine Learning - Sheet 3

14.05.2020
Deadline: 21.05.2020 - 18:00

Task 1: Ensemble Learning *(20 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.

1. (*3 points*) Implement `sampling` method that samples from training data based on their input weights.
2. (*4 points*) Implement `modelGeneration` method that takes two arguments: an input data, and a maximum number of iterations (use your decision tree implementation from the previous exercise as the base classifier).
3. (*3 points*) Implement `classification` method that returns the predicted class of a test instance.
4. (*4 points*) Get yourself familiar with the `sklearn.ensemble` module in Python. Use the same procedure as in part (2) of the previous sheet to divide a dataset into train and test data, then run your decision tree implementation and three ensemble classifiers, i.e., your AdaBoost implementation, and Random Forest and Bagging from `sklearn` on the same training data, and report the mean and standard deviation of the resulting accuracies.
5. (*2 points*) Report the results of the last part for the car dataset with an ensemble size $k = 10$, $r = \frac{2}{3}$ training data, and $n = 10$ repeats. Compare the results and discuss your findings.
6. (*2 points*) Change the ensemble size parameter (e.g., 1, 5, 10, 20) in part (5) and analyze the performance change of the algorithms. Discuss your findings.
7. (*2 points*) Change the maximum depth of base classifiers (e.g., $\{1, 3, 5\}$) in part (5) and discuss your observations. What happens if strong learners are used in AdaBoost?

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.