FACHHOCHSCHULE KIEL
University of Applied Sciences

Interdisciplinary Master's Program Data Science

Prof. Dr. Stephan Doerfel WS 2021/2022

1. Exercises «Portfolio-Exam Part I»

MADS-ML

This is the first part of the portfolio-exam for the Data Science course MADS-ML (Machine Learning). This part of the exam is homework. Student's are allowed to exchange ideas. However, this is NOT a teamwork exercise. Every student must derive and write up their own solutions in their own words and programming style.
To complete this first part of the exam,

- solve ALL the following tasks (two pages!),
- create a commented Jupyter Notebook for the code as well as for your textual answers, and
- upload the file to Moodle **before 23:59 o'clock (German time) December 12th, 2021**.

Note: In the Notebook, please state clearly which task a piece of code or text belongs to. Some of the tasks are not strictly separable, e.g. setting up the cross validation and using it on the algorithms. In these cases, just indicate in the notebook, which tasks the respective piece of code addresses.

Exercise 1. (Random Forest, 20 points)
Research: Read up on the algorithm Random Forest [1]. You may select a reliable source of your choice for that purpose. You should be able to explain the basic idea of the algorithm and understand the application of the implementation in `sklearn`.

1. (10 points) Describe the relation between Random Forests and Decision Trees (for classification).
2. (6 points) Compare the Random Forest and the Decision Tree classifier in `sklearn` by discussing the parameters `n_estimators`, `criterion`, and `max_depth`. Explain what the parameters control and why they are applicable to both algorithms or just the one.
3. (4 points) Compare the two algorithms with respect to their application: Which are immediate advantages and disadvantages of Random Forest over Decision Trees?

Exercise 2. (Data Acquisition, 10 Points)
Find and download the dataset "Online Shoppers Purchasing Intention" [2]. Use the dataset to tackle to following question: *Given a user's browsing behavior during a session in a web system as well as some other features of that session, predict whether the user will buy something (indicated in the column revenue by true or false).* Load the dataset in python and answer the following questions:

1. (5 points) How many numerical features can we use for predicting whether revenue is true or false.

2. (5 points) Describe and comment on the class distribution in the dataset.

Exercise 3. (Machine Learning Setup – 15 points)
Setup a machine learning experiment by

- splitting the target attribute from the data, converting it into a form suitable for `sklearn` classifiers and preparing the numerical attributes as features,
- selecting 30% of the data as test data (choose random seed 42),
- scaling the data such that the features have similar average and standard distribution.

Exercise 4. (Cross Validation – 17 points)
Use a combination of the classes `GridSearchCV` and `RepeatedStratifiedKFold` to setup a cross validation procedure for hyper parameter optimization.

1. (5 points) Create a cross validation setting in which the data is split into 10 folds, where all experiments are repeated 10 times, and where algorithms are evaluated using balanced accuracy.
2. (2 points) Which dataset is used in the grid search cross validation (training data, test data, or full dataset)?
3. (3 points) Explain, what happens to that dataset during the grid search procedure!
4. (5 points) What is the difference between using `RepeatedStratifiedKFold` and the default cross validation in `GridSearchCV`?
5. (2 points) Explain the purpose that justifies repeating experiments on the same dataset and on different folds.

Exercise 5. (Evaluation of Classifiers – 10 points)
Use the above cross validation setup to optimize and compare tree based learners. Use

1. (5 points) Decision Trees with the Gini criterion and test parameters 2 through 14 for `max_depth`.
2. (5 points) Random Forests with 1, 10, or 100 trees and 2,3,5, or 10 for `max_depth`.

Exercise 6. (Oversampling – 8 points)
Use oversampling to create a balanced training dataset. Look at the class `imblearn.over_sampling.RandomOverSampler` for that purpose.

1. (2 points) What does the above class do?

2. (3 points) Why is oversampling only applied to the training dataset (not to the test data)?

3. (3 points) Optimize Random Forest with the same search grid as before, but trained on a balanced training dataset.

Exercise 7. (Interpretation – 20 points)
Evaluate the resulting three algorithms of the three above cross validation experiments (Decision Tree and two versions of Random Forest).

1. (2 points) Prepare a data frame in which the evaluation results of algorithms can be stored with columns for the algorithm, accuracy, balanced accuracy, confusion matrix and the best hyperparameters of the algorithm.

2. (2 points) On which dataset should the performance of algorithms (with already optimized hyper parameters) be compared (training data, test data, full data)?

3. (6 points) For each algorithm report the best choice of hyperparameters found using the above cross validations.

4. (6 points) Compare three classifiers regarding both accuracy and balanced accuracy. Recommend a setting for use in production.

5. (4 points) Explain differences in the values of the two quality measures using confusion matrixes.

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] C. Sakar and Y. Kastro, "Online Shoppers Purchasing Intention Dataset." UCI Machine Learning Repository, 2018. Available at `https://archive-beta.ics.uci.edu/ml/datasets/online+shoppers+purchasing+intention+dataset`.