FACHHOCHSCHULE KIEL
University of Applied Sciences

Interdisciplinary Master's Program Data Science

Prof. Dr. Stephan Doerfel | Summer Term 2022

Part I
«Portfolio-Exam»

MADS-ML

This is the first part of the portfolio-exam for the Data Science course MADS-ML (Machine Learning), worth 49% of the final grade (98 points). This part of the exam is homework. Students 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 (four pages!),
- create a Jupyter Notebook for the (commented) code as well as your textual answers (English or German), and
- upload this as a single file to Moodle **before 23:59 o'clock (German time) June 13, 2022**.

Rules and Hints:

- The following exercises guide you through a set of experiments. Tasks build upon the results of previous tasks. Therefore, they must be completed in sequence.
- I will rerun the notebook. Make sure, everything is in the correct order and self-contained.
- Some of the experiments have to be repeated on different data. Follow the DRY principle[1] and create dedicated parametrized functions for these purposes.
- You are of course free to add additional helpful steps even if they are not required by the task (e.g., check the shapes of data structures and compare to your expectation).
- It is possible that a short answer of a rather complex task yields more points than a longer answer of a less complex task. Explain your results and reasoning! You can score points for partial answers.
- Whenever random functions are used, set their seed of 1 to make the experiments reproducible.
- You are free to reuse code from the lectures and exercises.
- It is well possible, that you will have to look up certain notions before you can answer a specific question. This is intended! Try to find reliable, valid sources.
- Do NOT FORGET to add textual answers to EACH task. Code alone is not sufficient!

[1] Don't repeat yourself!

Exercise 1. (k-Nearest Neighbors, 8 points)
The k-nearest neighbors algorithm takes the number of neighbors k, the distance function and the voting strategy as hyperparamters.

1. (4 points) Consider the following two parametrizations:
   a) $k = 1$, Euclidean distance, distance-based voting
   b) $k = 1$, Euclidean distance, uniform voting

   What can be said about the relation between the accuracies of the two resulting classifiers on the test data? Explain your answer.

2. (4 points) Consider the two parameter combinations:
   a) $k = 1$, Manhattan distance, distance-based voting
   b) $k = 2$, Euclidean distance, uniform voting

   What can be said about the relation between the accuracies of the two classifiers on the test data? Explain your answer.

Exercise 2. (Random Forest Algorithm, 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). Among others, address the goal of the algorithms, learning process, feature selection during learning, test and training data, the split criterion.

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 3. (Data Acquisition, 10 points)
Using the dataset "Online Shoppers Purchasing Intention" [2], we want to tackle to following research 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 and how many categorical features does the dataset offer with respect to the above research question? Note: Categorical features can be represented numerically, but still are categorical features!

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

Exercise 4. (Machine Learning Setup, 15 points)
Setup a machine learning experiment by

- restricting the data to the features:
  `'Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'OperatingSystems', 'Browser', 'Region', 'TrafficType'`
- splitting the target attribute from the data,
- selecting 30% of the data as test data,
- creating a function for adding configurations and results of an experiment to a data frame.

Exercise 5. (Training and Testing, 15 points)
Run experiments for learning and testing different parametrizations of Random Forest. Try all possible combinations with the following parameter sets:

1. number of trees: 1, 10, or 100
2. maximum tree depth: 2,3,5, or 10.

In your experiments

1. use the train-test-split from above,
2. use default parameters for all other than the above,
3. repeat each experiment 10 times using the seeds 1 through 10 to initialize Random Forest,
4. add the results and configurations to the data frame using the method created above.

Exercise 6. (Evaluation and Stability Analysis, 20 points)
For the stability analysis:

1. (2 points) Which parameter combination of the experiments yields the highest balanced accuracy (averaged over the 10 runs)?
2. (10 points) Create the following diagram:
   a) For each chosen number of trees, plot the balanced classification accuracy against the max-depth.
   b) Use the different runs with different seeds to plot the respective average scores together with their standard deviation.

3. (5 points) Interpret the diagram. In particular, comment on and explain the stability of the results for different parameter choices.

4. (3 points) Which further factor influences the evaluation, that is not addressed in the current setup? How would the setup have to be modified to include that as well? (Hint: Just explain, no further experiments are required.)

Exercise 7. (Feature Importance, 10 points)

1. (5 points) For a Random Forest with 100 trees and a maximum depth of 14, create a bar chart showing the importance of each individual feature with regard to its contribution in random forest.

2. (2 points) Which is the most important feature?

3. (3 points) Comment on the reliability of feature importances as implemented in `sklearn`'s Random Forest and propose an alternative.

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.