Interdisciplinary Master's Program Data Science

Prof. Dr. Stephan Doerfel — Summer Term 2022

Part II
«Portfolio-Exam»

MADS-MMS

This is the second part of the portfolio-exam for the Data Science course MADS-MMS (Mathematics and Multivariate Statistics), worth 51% of the final grade (102 points). This part of the exam is a term paper. 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 second part of the exam,

- complete the task below,
- create a Jupyter Notebook for the (commented) code, results, and diagrams as well as (textual) explanations, interpretations, and conclusions (English or German), and
- upload this file to Moodle **before 23:59 o'clock (German time) May 12th, 2022**.

Submissions after the deadline or via other means than Moodle will not be considered!
Carefully read these instruction and make sure nothing is forgotten in the final submission!

1 The Task

1. In the term paper, **demonstrate, describe, and interpret** all necessary steps for conducting a data analysis on a particular dataset. Draw conclusions explain your interpretation for a practical use case.
2. Choose a dataset, that you like to investigate and that is suitable for your experiments.
3. Conduct reproducible experiments in which you design experiments to analyze the dataset and discover and explain valuable insights.

2 The Notebook

1. Think of the Jupyter notebook you submit as a report of a (small) project which you – as a data scientist in a data science company – conduct for a customer. The report should demonstrate your ability to work with the customer's data and to discover valuable insights in the data. It should **tell a story and show and explain the efforts you made, and interpret the results for the customer and emphasize the customer's gain through your work.**
2. Your notebook should therefore contain MUCH MORE TEXT, introduction, comments, etc. than the notebooks we use to demonstrate methods in the lecture or the exercises!

3. Hint: Keep in mind that this is not your Master Thesis. The experiments should be comprehensive and created and conducted solely by you. However, limit your work towards solving one or two tasks – even though along the way you might recognize other interesting angles to follow up on.

4. End your report with concrete results (summarizing the report – what does the customer gain from your work) and with perspectives on future work (what is relevant as next steps that the customer should commission you to conduct).

3 The Rules

1. Choose a suitable dataset for your experiments. The data have to be real-life data. Do NOT use artificial datasets!

2. There are two possible availability scenarios for the data: A) It is available online with proper licence. In that case indicate in your notebook where the data can be downloaded. B) You (legally!) obtain a dataset and share it with me via Moodle. Such data should not come with any form of NDA or other obligations. If you are not sure about these criteria regarding the dataset you consider, please contact me before you invest too much time in the experiments.

3. Use Python for the experiments.

4. Conduct a full analysis of the data. That includes description, loading and initial analysis of the original data, preprocessing, and of cause the exploratory steps.

5. Explain your steps and choices.

6. Choose **at least three different methods from the realm of clustering and dimensionality reduction** to analyze (aspects of) the data. Select reasonable parametrizations and use additional steps to find good choices.[1]

4 Criteria for Grading

1. Your code is executable and yields reasonable and reliably replicable results

2. Your notebook tells a meaningful and interesting story, is understandable and the code is well documented.

3. Beyond that, the choices of methodology (different approaches, dataset specific choices, settings of hyperparameters, etc.) are reasonably explained.

4. Your results (numbers and diagrams) are referenced in the text, explained and interpreted.

5. All above aspects of the task, the notebook and the rules are met.

[1] E.g., computing silhouette coefficients and diagrams does not count as one of the three methods but is part of the expected additional steps to find a good parameter for k-Means and to interprete the result.