FACHHOCHSCHULE KIEL
University of Applied Sciences

Interdisciplinary Master's Program Data Science

Prof. Dr. Stephan Doerfel | Summer Term 2022

Part I
«Portfolio-Exam»

MADS-MMS

This is the first part of the portfolio-exam for the Data Science course MADS-MMS (Mathematics and Multivariate Statistics), 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 (three 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) April 29, 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!

For these tasks, you will use three datasets, named `portfolio_data_sose_2022_*.csv`. These are available on the Moodle course page and are henceforth referenced simply as Datasets A, B, and C respectively.

Exercise 1. (Initial Data Analysis on Dataset A. – 8 points)
Conduct a brief initial analysis of the raw Dataset A.

1. (1 point) How many different instances (rows) does the dataset contain?
2. (1 points) How many features (columns) are in the dataset?
3. (2 points) Determine each feature's range.
4. (2 points) Determine each feature's standard deviation.
5. (2 points) Plot the data in a scatter plot.

Exercise 2. (k-Means Clustering on Dataset A. – 38 points)

1. (14 points) On the raw Dataset A, compute k-Means clusterings of the dataset for different choices of $k : 2, 3, \ldots, 10$. For each k compute the silhouette coefficient and plot them against k in a diagram. Describe and interpret the diagram!
2. (8 points) Choose k according to your result in 2.1 and plot the data in a scatter plot! Visualize the clustering using colors! Additionally visualize the cluster centers in the diagram!
3. (4 points) Compare the resulting clustering to your intuition from looking at the data! Is the resulting clustering meaningful? Does the chosen k fit you intuition?
4. (7 points) Explain the issue with the dataset that let to this clustering! Did the clustering coefficient suggest a good or a bad clustering structure? Is the issue a problem with the silhouette coefficient, the clustering method, or the data modeling? Suggest and implement a fix for the issue!
5. (5 points) Conduct the same steps as in 2.1 and 2.2 for the data resulting from 2.4. Again, compare your intuition to the result.

Exercise 3. (k-Means Clustering on Dataset B. – 33 points)
In the following experiments, we process the Dataset B.

1. (6 points) On the raw Dataset B, compute k-means clusterings of the dataset for different choices of $k : 2, 3, \ldots, 10$. For each k compute the inertia of the resulting clustering and plot it against k in a diagram. Describe and interpret the diagram!
2. (2 points) In you own words, explain the idea of the elbow method!

3. (0 points) Familiarize yourself with the Python module `kneed`[2] and find a way to automatically determine the elbow of a curve.

4. (5 points) Compute the best k according to the elbow method using `kneed`!

5. (2 points) Determine the best choice of k by reading it from a silhouette plot like in 2.1 as well and compare it to your result of 3.4.

6. (2 points) Choose k according to your result in 3.4 and plot the data in a scatter plot! Visualize the clustering using colors! Additionally visualize the cluster centers in the diagram!

7. (8 points) Create and interpret the silhouette plot for the clustering from 3.6.

8. (8 points) For the clustering from 3.6, for each feature, plot the distribution of the feature in each cluster! Summarize the results about the k groups.

Exercise 4. (Geometry on Dataset C. 19 points)
In this task, sketch the mathematical steps you need to take for the representations of the lines.

1. (2 points) Compute and visualize a k-Means clustering for Dataset C with $k = 2$. Plot the data in a scatter plot! Visualize the clustering using colors! Additionally visualize the cluster centers in the diagram!

2. (7 points) Consider a line in $\mathbb{R}^2$ given through the vector equation $\langle \boldsymbol{x}, \boldsymbol{w} \rangle + b = 0$ with $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{w} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$, and $b = 11$. Add the resulting line into the diagram of 4.1. If you take intermediate steps to express the line in another form, include them in the notebook.

3. (7 points) Consider another line given through the following information:

   - the vector $\boldsymbol{u}_0 = \begin{pmatrix} -7 \\ 6 \end{pmatrix}$ is on the line and
   - the vector $\boldsymbol{w} = \begin{pmatrix} 10 \\ 12 \end{pmatrix}$ is orthogonal to the line.

   Add the resulting line into the diagram of 4.1. If you take intermediate steps to express the line in another form, include them in the notebook.

4. (3 points) Discuss the relation of the two lines of 4.2 and 4.3 to each other and to the clustered dataset!

[2] `https://pypi.org/project/kneed/`