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-MMS

This is the first part of the portfolio-exam for the Data Science course MADS-MMS (Mathematics and Multivariate Statistics). 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 commented Jupyter Notebook for the code as well as one PDF file for your textual answers, and
- upload BOTH files to Moodle **before 23:59 o'clock (German time) December 12th, 2021**.

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 as ordered here.
- Some tasks are related to the ones above. E.g. they might require setting a parameter to some specific value, while in an exercise before, you were supposed to argue what a good choice for that same parameter might be. Please be aware that the choice in the later exercise might or might not be the one expected in that previous task.
- The main result that will be graded, is the PDF you hand in. The Jupyter Notebook will only be used in case your answer in the PDF is other than expected. Therefore, all required diagrams and numbers must be EXPORTED from the notebook and included in the PDF.
- Some of the experiments have to be repeated on different data. It is suggested to create dedicated parametrized functions for these purposes.
- The points per task are a result of the effort it takes and the complexity of the task. Thus, it is possible that a short answer of a rather complex task yields more points than a longer answer of a less complex task.
- For many tasks you will find suitable code samples in the notebooks of the lecture and you are free to copy and adapt them.
- 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.

For these tasks, we enter the domain of webometrics or altmetrics – the measurement and utilization of impact indicators on the Web. This field is relevant in many use cases, such as finding influential users and posts, social media marketing, information propagation, and many more.

Exercise 1. (Data Acquisition and Initial Data Analysis – 10 points)
Obtain the dataset [1] from the UCI Machine Learning Repository.
Conduct a brief initial analysis of the raw dataset (henceforth called Dataset A).

1. (2 points) What do the rows of the dataset represent?
2. (2 points) How many different instances does the dataset contain?
3. (2 points) How many attributes (columns) are in the dataset?
4. (4 points) What is the standard deviation of the feature `num_likes`?

Exercise 2. (k-Means Clustering on the Plain Data – 40 points)
Begin the analysis using the k-means approach.

1. (5 points) Which features of the dataset do *not* suggest themselves as features in a clustering analysis? For each of these features, briefly state why you exclude them.
2. (15 points) For the next tasks, restrict Dataset A to the following features: `num_reactions`, `num_comments`, `num_shares`, `num_likes`, `num_loves`, `num_wows`, `num_hahas`. We will call this Dataset B.

   On B, compute k-means clusterings of the dataset using different choices for $k : 2, 3, \ldots, 10$. Use a seed of 1 to make the experiments reproducible. For each k compute the silhouette coefficient and plot it against k in a diagram. Interpret the diagram!
3. (10 points) Create a silhouette plot for the k with the highest silhouette coefficient in the previous experiment. Interpret the diagram!
4. (10 points) For the same k, create a plot of the data where you use only the two features `num_reactions` and `num_likes` as the axes. Use color to distinguish instances from different clusters. Also highlight the cluster centroids of the k-means clustering. Interpret the diagram, considering only the above two features. Is there a clear clustering structure visible?

Exercise 3. (Scaling and Feature Selection – 50 points)
In these next experiments, we preprocess and restrict Dataset B further through scaling and using variance as a criterion for feature selection. Particularly, use the class `sklearn.feature_selection.VarianceThreshold`.

1. (10 points) Describe in your own words, what the class VarianceThreshold is used for and explain why looking at a feature's variance is meaningful.

2. (5 points) The features of the dataset are in different ranges. To be able to compare by variance we should scale the data first. Which is the better choice for the variance threshold method: Min-Max-Scaling or the Standard-Scaler (z-score transformation)? Explain your answer.

3. (10 points) Use Min-Max-Scaling on Dataset B to yield Dataset C and rerun the above experiments on C (k-means clusterings for $k = 2, 3, \ldots, 10$, silhouette plot for the best k, plot of clustered data and centroids). How does that compare to the previous experiments?

4. (10 points) Use the variance threshold method for feature selection on the scaled data to restrict the features further. Set the variance threshold to 0.005. Which features does the resulting dataset (Dataset D) include, what is their variance (computed on the scaled features)?

   Checkpoint to avoid subsequent errors: The dataset D should now contain the features `num_reactions` *and* `num_likes`

5. (5 points) Repeat the experiments from above on Dataset D (k-means clusterings for $k = 2, 3, \ldots, 10$, silhouette plot for the best k, plot of clustered data and centroids). Compare the outcome to the previous experiments.

6. (10 points) Compare the feature distributions over the clusters for features `num_reactions`, `num_likes`, and `num_shares`. Use violin plots for the original data (the unscaled values) and interpret your results.

References

[1] N. Dehouche and A. Wongkitrungrueng, "Facebook live as a direct selling channel, 2018," in *Proceedings of ANZMAC 2018: The 20th Conference of the Australian and New Zealand Marketing Academy. Adelaide (Australia)*, pp. 3–5 December, 2018. Available at `https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand`.