

DM566/DM868/DM870/DS804: Data Mining and Machine Learning
Spring term 2021

Exercise 11: Group Project 1: Clustering

Exercise 11-1 Exploration of Clustering Algorithms (5 points)

You should work on this project in a group of approximately four people.

Start early with your exploration and plan enough time. Exercises 8-10 should give you additional insights and guidance along the way.

Choose a data set from the “Shape sets” at <https://cs.joensuu.fi/sipu/datasets/> and study the behavior of the clustering algorithms discussed in the lecture. (You can choose a data mining tool of your preference or your own implementation.)

Document your experiments. Guiding questions for your documentation could be:

- Which data set did you choose and why?
- Which algorithms and which implementation did you use? (Why?)
- Showcase the achieved clustering results (e.g., scatter plots).
- Annotate the parameters used to achieve a solution.
- Annotate some quality measures to each solution (e.g., the silhouette coefficient).

Discuss your experimental results. Guiding questions for the discussion could be:

- Which algorithm is most suitable for the dataset?
- Which algorithm is most unstable in terms of quality achieved when changing parameters?
- Do the quality measures fit to the structure in the dataset?
- Do some quality measures have a systematic preference for some of the algorithms?
- Explain your findings!

Having studied the behavior of clustering algorithms and evaluation measures on synthetic 2 dimensional data, you can scale up and explore some more challenging data from any domain you find interesting. What can you learn about that data by using clustering?

Present your choices, experiments, results, and insights in a presentation in the exercise class in weeks 16+17 (prepare some slides for up to 15 minutes presentation per group and agree with your tutor on a time slot when exactly to present).