Karl M. Heuer / David E. Roberson

Autumn 2022
Exercise class:
in 302 on October 11th

Exercises for Computational Tools for Data Science (02807)

Week 7: Clustering

For this exercise sheet we recommend installing and using the following Python libraries: *matplotlib* and *scikit-learn*.

The library *matplotlib* can be used to visualise the data. For further information see: https://matplotlib.org.

From *scikit-learn* you should use the relevant clustering algorithms. For further information regarding this see: https://scikit-learn.org/stable/modules/clustering.html.

Furthermore, the following exercises make use of data stored in files with the extension '.dat'. The files contain n many d-dimensional points. The first line consists of "n;d;". The following lines contain the coordinates of one point each, separated by semicolon, for example "1.0;3.45;34.5657585959;". A .zip-file containing all relevant .dat-files can be found on the course homepage or directly via this link: https://courses.compute.dtu.dk/02807/2022/data/clustering.zip

Exercises 1: Visualize the data

Use *matplotlib* to implement a viewer for two-dimensional data which reads a .dat-file and displays it graphically (use the data files from the course page). Make sure you can specify a color for each point (for later use).

Exercise 2: Run clustering algorithms from *scikit-learn*

Use *scikit-learn* and run the (implemented versions of the) following algorithms on the test data:

- Hierarchical clustering
- k-Means
- DBSCAN

You should try out all the test files, to see how well the different algorithms work on different kinds of data. Furthermore:

- Visualise the results for the 2-dimensional data.
- Experiment with the parameter settings, and explain the results.

Exercise 3: Evaluate the results

Compute the Davies–Bouldin index for your tests and compare it to your empirical impression.

Exercise 4: Implementation

Implement the k-Means algorithm by yourself, and ensure it works properly by comparing its results to the results from the *scikit-learn* version of the algorithm.