Week 9: Clustering

02807 Computational Tools for Data Science

On the course page you can find a number of data files with the extension ".dat". The files contain n 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;".

In Todays exercise, we recommend you use matplotlib for visualizing your data, and scikit-learn for the clustering algorithms:

`pip3 install matplotlib` (`https://matplotlib.org/`)
`pip3 install scikit-learn` (`http://scikit-learn.org/stable/modules/clustering.html`)

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).

2 Run clustering algorithms Run the Hierarchical clustering, K-Means, and DBSCAN on the test data (use the algorithms from scikit-learn). You should try out all the test files, to see how well the different algorithms work on different kinds of data.

- For the 2-dimensional data files, visualize the results.
- Experiment with the parameter settings, and explain the results.

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

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 algorithm.