Lecture “Advanced Data Analytics”

Problem Set 8

Simon Scheidegger

Exercise 1

k-Means

Given is the following 2-dimensional data set in R^2:

x_1	0.0	0.0
x_2	0.4	0.0
x_3	0.0	0.4
x_4	1.0	1.0
x_5	0.8	1.0
x_6	1.0	0.8
x_7	0.6	0.6

a) Apply a k-Means library code (with k = 2), and use the Euclidean Distance. Assume that the initial centroids were chosen as $\mu_1 = x_6$ und $\mu_2 = x_7$.

b) Describe what your results are after running k-means.

c) Visualize the different iteration steps.

Exercise 2

Gaussian Mixture Models

You are given the data set of the “Old faithful” geyser from the Yellowstone national park (supplementary material, in data/faithful.csv), and described here: https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat.

The dataset contains details of the eruption duration and the waiting time in between eruptions of the Old Faithful Geyser.

a) Apply Gaussian Mixture models library in combination with the expectation maximization algorithm to this data set. Use k = 1, 2, 3, 4, 5, 6, and plot the resulting hyper-parameters of the converged code.

b) How many clusters do you think are likely to be in the data? Justify your statement.

c) Visualize the individual steps of the EM algorithm by plotting the evolution of the normals.

Exercise 3

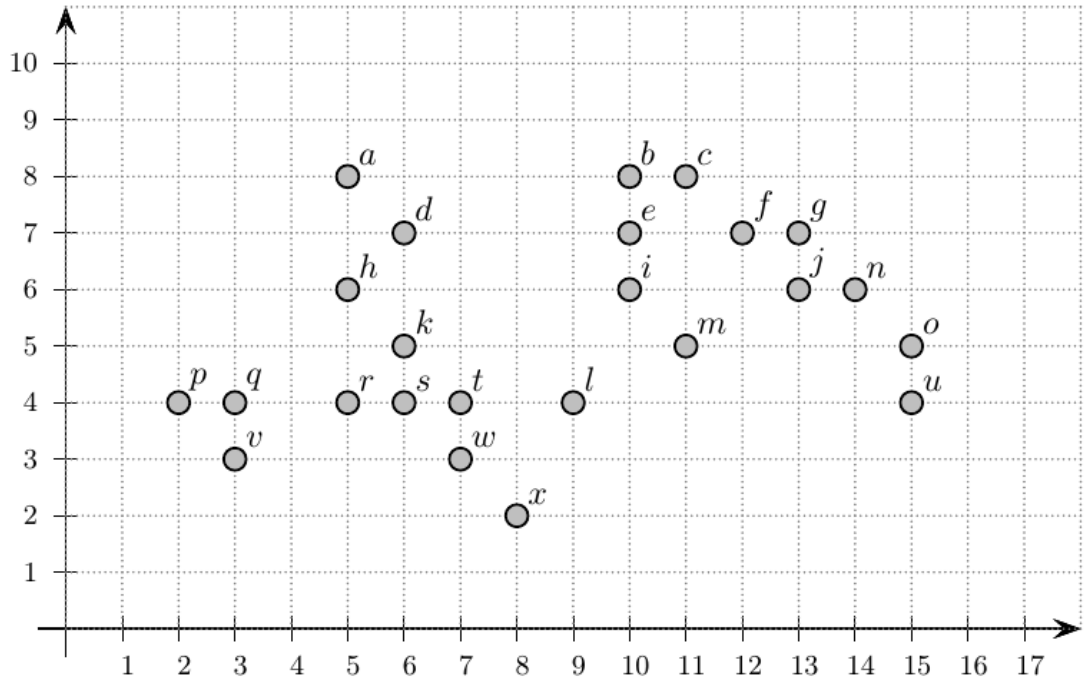
Hierarchical Clustering

a) Determine the distance matrix for the data points given in exercise 1.
b) Apply agglomerative clustering to the data set given in exercise 1, and use single linkage. Then, plot the resulting dendrogram.
c) Apply agglomerative clustering to the data set given in exercise 1, and use complete linkage. Then, plot the resulting dendrogram.

Exercise 4

DBSCAN (no computer to be used)

You are given the data set below. Reply the answers below under the assumption that *minpts = 3*, $\epsilon = 2$, and the Euclidean distance is used.



a) Which data points are cores?
b) Is the data point ***a*** directly reachable from point ***d***?
c) Is the data point ***o*** reachable from ***i***? Give the chain of intermediate points, or explain, where the chain breaks.
d) Are the points ***l*** and ***x*** connected?
e) Provide the resulting clusters. Moreover, mark the data points that can be considered as noise.