University of Southern Denmark
IMADA
Arthur Zimek

DM868/DM870/DS804: Data Mining and Machine Learning

Spring term 2023

Exercise 9: EM-Clustering, Density Estimation, DBSCAN, Comparison of Clusterings

Exercise 9-1 Assignments in the EM-Algorithm (1 point)

Given a data set with 100 points consisting of three Gaussian clusters A, B and C and the point p.

The cluster A contains 30% of all objects and is represented using the mean of all its points $\mu_A = (2, 2)$ and the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

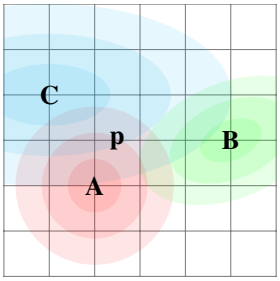
You will need the inverse: $\Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$.

The cluster B contains 20% of all objects and is represented using the mean of all its points $\mu_B = (5, 3)$ and the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$. $\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$.

The cluster C contains 50% of all objects and is represented using the mean of all its points $\mu_C = (1, 4)$ and the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$. $\Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$.

The point p is given by the coordinates $(2.5, 3.0)$.

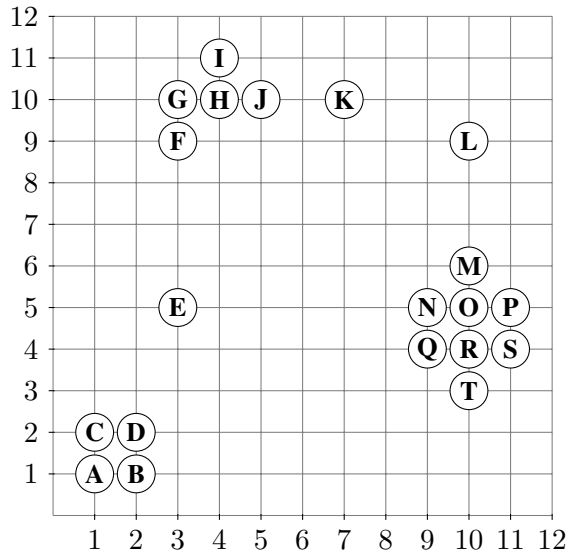
The following sketch is not exact, and only gives a rough idea of the cluster locations:



Compute the three PDF values and derived probability estimates of p belonging to the clusters A, B, and C.

Exercise 9-2 Density Estimation (1 point)

Given the following data set:



Estimate the density around each point in the dataset, using the discrete Kernel

$$\hat{f}(x) = \frac{k}{nV_k(x)}$$

based on Manhattan distance (L_1)

(a) with a fixed $k = 2$,

(b) with a fixed $k = 4$,

(c) with a fixed volume based on radius $\varepsilon = 1$,

(d) with a fixed volume based on radius $\varepsilon = 2$.

Explain what your choices are in computing the density estimate regarding

(a) including or excluding the point itself,

(b) ties in the neighborhood.

Note that using the Manhattan distance results in estimators that slightly differ from those discussed in the lecture.

What do you observe?

Exercise 9-3 Properties of DBSCAN (1 point)

Discuss the following questions or statements on DBSCAN:

- For *minPts* $= 2$, what about border points?
- The result of DBSCAN is deterministic for core and noise points, but not for border points.
- A cluster in DBSCAN can contain less than *minPts* objects.
- If the dataset has n objects, DBSCAN computes always exactly n neighborhood range queries.
- On uniformly distributed data, DBSCAN will typically put everything in one cluster or everything in noise. k-means will typically partition the uniformly distributed data in k approximately equal-size partitions.
- What is the relationship of DBSCAN with *minPts* $= 2$ to single-linkage clustering?

Exercise 9-4 Clustering Lab-Session

For each of the tasks you can use either ELKI, R, scikit-learn (Python), or another tool of your choice.

(a) Data Preprocessing:

- Go to `http://archive.ics.uci.edu/ml/datasets/seeds` and read about the seeds dataset. Download the dataset.
- The dataset is presented in a tab separated format. The format might not be immediately suitable to analyze the dataset with your preferred analysis tool.
  Try and reformat the dataset for use with the tool of your choice. The ELKI formatted version can be found on itslearning.

(b) k-means:

- Use a k-means implementation in your preferred tool to analyze the seeds dataset.
- What is a suitable choice for k?
- Try different k-means variants like MacQueen, Lloyd, Elkan, k-means++. Do you observe any differences or tendencies in the results?

(c) EM-clustering:

- Run EM clustering on the seeds dataset.
- What is a suitable choice for k?
- For different choices of k compare the result to a similar choice of k in the k-means algorithm.

(d) DBSCAN:

- Run DBSCAN on the seeds dataset.
- Find suitable parameter values for epsilon and minpts.

(e) You might also want to try SNN clustering or hierarchical clustering. Given your experience with this dataset by now – does using these algorithms make sense on this dataset?

(f) Comparison:

Which type of clustering algorithm do you consider the most suitable for this dataset?

(g) Try a different tool, and start from the preprocessing step again to get familiarized with that tool.