University of Southern Denmark
IMADA
Arthur Zimek

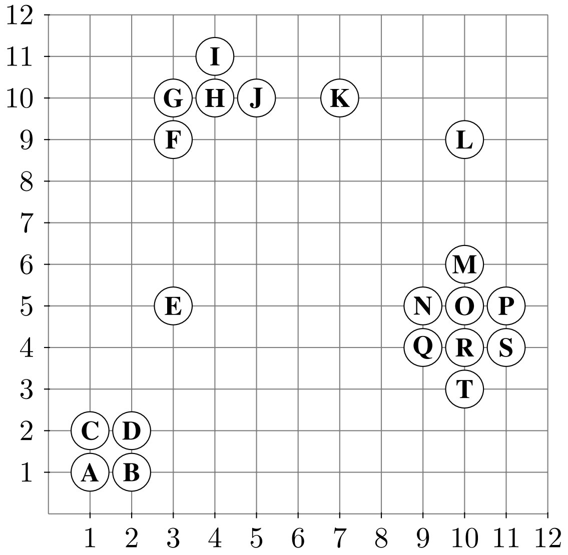
DM868/DM870/DS804: Data Mining and Machine Learning

Spring term 2023

Exercise 10: Shared Nearest Neighbors, Hierarchical Clustering, Outlier Detection

Exercise 10-1 Shared Nearest Neighbors (1 point)

Given the following data set:



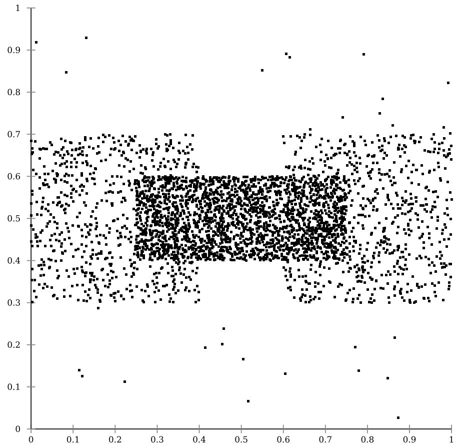
(a) Compute the pairwise shared-nearest-neighbor-similarities SNN_5 of the objects M, N, O, P, Q, R, S, and T.
Use Manhattan-distance L_1 to obtain the neighbors and neighborhoodsize 5.
The query point is a member of its neighborhood.

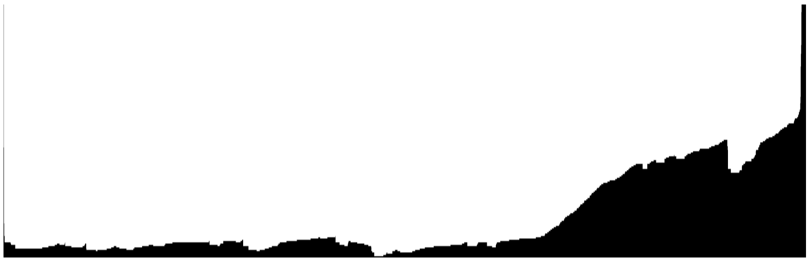
$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

(b) Give parameters ε and minpts such that the SNN variant of DBSCAN (Ertöz et al., 2003) identifies the 8 points as "dense" and connects them into a single cluster.

Exercise 10-2 OPTICS Plot (1 point)

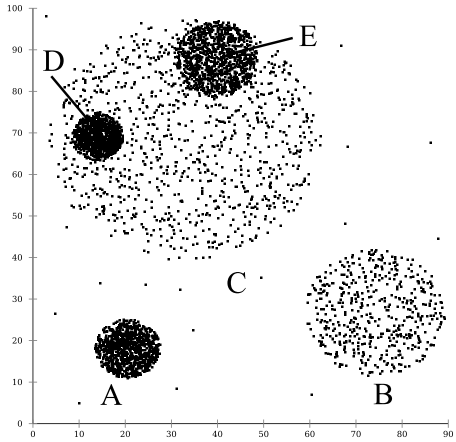
(a) For the data below we got computed the reachability diagram to the right.

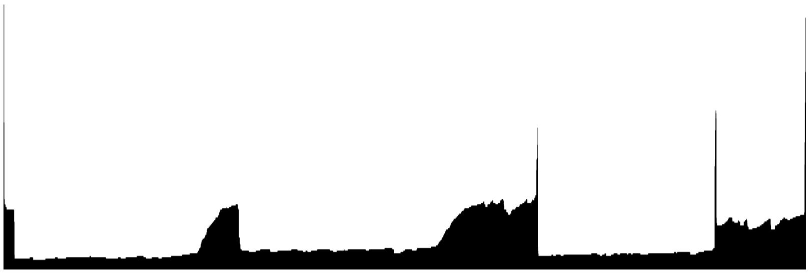




With a naïve understanding of hierachical clustering, wouldn't we have expected three valeys in the plot? Explain, why this is not the case and why the plot, instead, looks as it does and accurately describes the density structure of the data.

(b) For this dataset (left) we have the reachability plot (right).

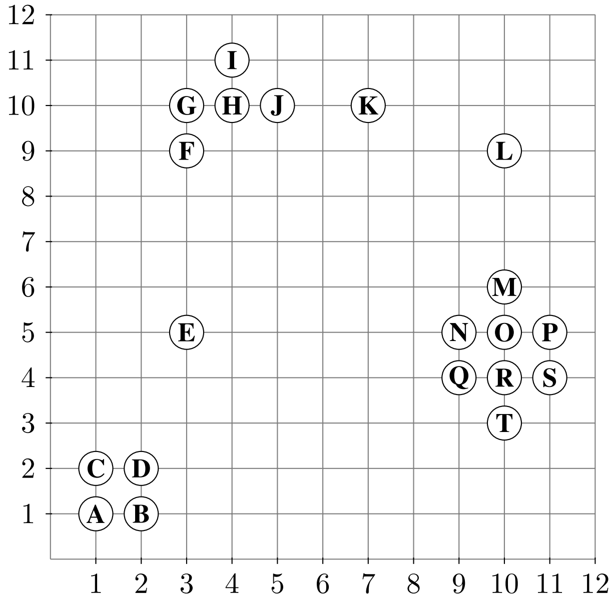




Mark in the reachability plot which areas relate to the clusters A, B, C, D, and E.

Exercise 10-3 Outlier Scores (1 point)

Given the following 2 dimensional data set:



As distance function, use Manhattan distance $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$.

Compute the following (without including the query point when determining the kNN):

- LOF using $k = 2$ for the points E, K and O.
- LOF using $k = 4$ for the points E, K and O.
- kNN distance using $k = 2$ for all points.
- kNN distance using $k = 4$ for all points.
- aggregated kNN distances for $k = 2$ and $k = 4$ for all points
  (aggregated kNN distance = averaged sum of the distances to all the kNN!)

Exercise 10-4 Evaluation of Outlier Scores (1 point)

A data set with known outliers $+$ was evaluated using two outlier detection methods S_1 and S_2. The results of the methods are given in the table below:

Object	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
Label	$-$	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$+$	$-$
S_1	1.0	1.1	1.1	1.3	3.0	2.0	1.5	0.9	1.4	1.2
S_2	.80	.80	.10	.81	.89	.50	.50	.91	.90	.20

Evaluate both outlier detection methods S_1 and S_2 using the following metrics:

- Precision, Recall and F-Measure,
  assuming that the top $k = 2$ ranked outliers were classified as outliers.
- Average Precision for $k = 1 \ldots 4$,
  assuming that the top k ranked outliers were classified as outliers.
- Draw the ROC curve, and compute the area under curve (AUC) measure.

Exercise 10-5 Outlier Detection – Practical (1 point)

(a) Work with some toolbox for data exploration (e.g., R, Python, ELKI) to try different outlier detection algorithms (e.g., knn, LOF) on some dataset (e.g., “3 clusters and noise 2d” from itslearning).

(b) How does the behavior change with the choice of the neighborhood size?

(c) Run OPTICS on the same dataset. Imagine, you would not know how the dataset looks like. What could you learn about the clusters and outliers in the dataset?