Ludwig-Maximilians-Universität München Institut für Informatik Prof. Dr. Thomas Seidl Julian Busch, Evgeniy Faerman, Florian Richter, Klaus Schmid

Knowledge Discovery in Databases SS 2016

Exercise 7: Clustering

Regarding tutorials on 08.06.-10.06.2016.

Exercise 7-1 Agglomerative Hierarchical Clustering

Given the following 2-dimensional data set:



Apply the Agglomerative Hierarchical Clustering to this data set using the Manhattan distance and:

- (a) Single-Link
- (b) Complete-Link
- (c) Average-Link

Note: It is sufficient to draw the resulting dendograms including the distances. You do not need to specify the computations for the distances.

Exercise 7-2 DBSCAN

Given the following data set:



As distance function, use Manhattan Distance:

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

Compute DBSCAN and indicate which points are core points, border points and noise points. Use the following parameter settings:

- Radius $\varepsilon = 1.1$ and minPts = 3
- Radius $\varepsilon = 2.1$ and minPts = 4
- Radius $\varepsilon = 4.1$ and minPts = 5

When minPts = 2, what happens to border points?

What is the relationship of DBSCAN with minPts = 2 to single-linkage clustering? Why does DBSCAN run in $\mathcal{O}(n^2)$ time while hierarchical clustering is usually denoted as $\mathcal{O}(n^3)$? Why is this not a contradiction?

Exercise 7-3 OPTICS

Considering the dataset in exercise 7-2 again, construct an OPTICS reachability plot for each of the following parameter settings:

- $\varepsilon = 5$ and minPts = 2
- $\varepsilon = 5$ and minPts = 4
- $\varepsilon = 2$ and minPts = 4