RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. T. Seidl

November 30th, 2015

Exercise 6 for the lecture
Data Mining Algorithms
WS 2015/2016

Hand in your solutions on December 7th before the lecture. The tutorial for this exercise will be held on December 11st. Solutions of groups with less than 3 or more than 4 students will not be graded.

Note: All commands for the R-exercises are required to be provided with comments, indicating which task the commands belong to. All R script files should contain a comment-line with the names and matriculation numbers of all group-members. Send all R-files to siccha@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]".

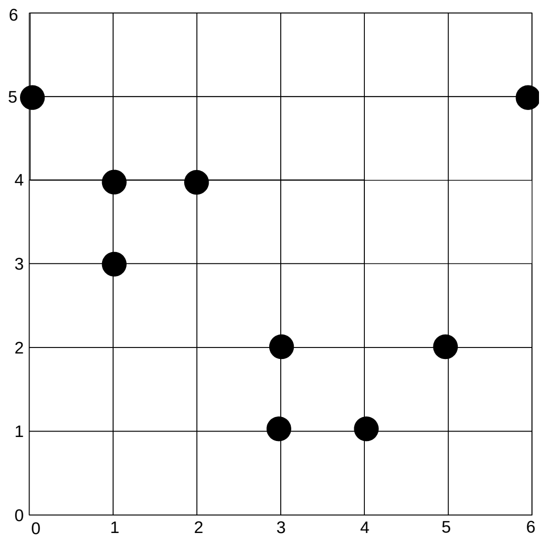
Exercise 6.1) Agglomerative Hierarchical Clustering **8 points**

Given the following 2-dimensional data set:
P_1=(0,5), P2=(2,4), P3=(1,4), P4=(1,3), P5=(5,2), P6=(3,2), P7=(3,1), P8=(4,1), P9=(6,5)

Apply the Agglomerative Hierarchical Clustering [AGglomerative NESting (AGNES), Slide 57, Chapter 4] to this data set using the *Manhatten 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.



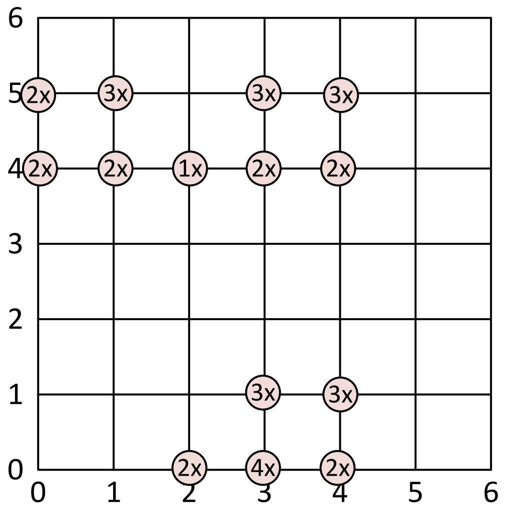
Exercise 6.2) OPTICS **8 points**

Draw the reachability plot and the core-distance plot for the following 2-d data set using the Manhattan-distance and MinPts = 6, ε = 2.
Start with o = (0,4). Then, once the ControlList is empty, restart with p = (2,0).

Data: {(2,0);(2,0);(3,0); (3,0); (3,0); (3,0); (4,0); (4,0); (3,1); (3,1); (3,1); (4,1); (4,1); (4,1); (0,4); (0,4); (0,5); (0,5); (1,4); (1,4); (1,5); (1,5); (1,5); (2,4); (3,4); (3,4); (3,5); (3,5); (3,5); (4,4); (4,4); (4,5); (4,5); (4,5)}

Given the resulting reachability and core-distance plot.

Based on the resulting plot, which two settings ε_1 and ε_2 correspond to a DBSCAN that yields two and three clusters as output, respectively?



Note: you do not need to do the actual computation, but you may refer to the figure for reading off the reachability and core distances, respectively.

Note: To make the corrections easier use the following heuristic. When multiple points have the same distance to the already processed points,

1. **the point with the smallest x-coordinate value is preferred,**
2. **if there exist several points with the same x-coordinate value, the one with the smallest y-coordinate value is preferred.**

Exercise 6.3) Ensemble Clustering **4 points**

Given the dataset depicted on the right and the three clusterings:

$\mathcal{C}_1 = \{\{P1, P2, P3, P4, P5, P6, P7\}, \{P8, P9, P10, P11, P12, P13\}\}$

$\mathcal{C}_2 = \{\{P1, P2, P3, P4, P8, P9, P10\}, \{P5, P6, P7, P11, P12, P13\}\}$

$\mathcal{C}_1 = \{\{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13\}\}$

a) Determine the co-association matrix $\mathbf{S}^{(\mathfrak{C})}$ based on $\mathfrak{C} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$.

b) Determine the ensemble clustering based on the co-association matrix from a) with the help of DBSCAN (MinPts=3, ϵ=3).

Note that the co-association matrix defines a pairwise similarity of the points. Deviating from the normal situation, the ϵ-Neighborhood of a point is thus defined as
$\mathbf{N_\epsilon(o) = \{p \in DB | similarity(o, p) \geq \epsilon\}}$.