TU Dortmund 5.5.2022
Artificial Intelligence / Data Science and Data Engineering   **Machine Learning for Complex Data**
Prof. Dr. Erich Schubert
Prof. Dr. Emmanuel Müller   **Summer 2022**
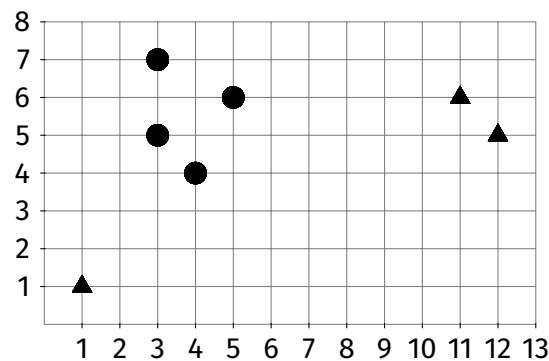Daniel Wilmes

Assignment 4: k-Means Clustering und PROCLUS
Due: Thursday, 12.5.2022

Problem 4-1   *k*-Means Clustering **7**

Given the following data set with 7 objects (in $\mathbb{R}^2$):

The data set are all points (both triangles and circles) in the above plot. All coordinates are integer.

The three triangles are the data points chosen as *initial cluster centers*.

a) Perform *k*-means clustering using the standard algorithm with $k = 3$ on this data set.
   At each iteration, give the new cluster assignment, then the new cluster centers.

b) What problem does the algorithm encounter in performing *k*-means on this data set?

c) Propose at least two strategies to handle this situation. Justify your answers.

d) Do the same with Hamerlys algorithm. Give the upper and lower bounds of all points after
   the first iteration, afterwards only the values recomputed by Hamerly's algorithm.
   Hint: the final result obviously must be the same; avoid unnecessary computations!

Calculate the following steps of a PROCLUS clustering using $k = 3$ and $l = 3$.
Use the entire dataset in the algorithm (no sample) and objects A, E and G as initial medoids.

	d1	d2	d3	d4
A	15	12	16	9
B	14	13	18	3
C	12	14	14	15
D	16	13	19	19
E	5	6	9	4
F	4	11	10	18
G	6	17	8	13
H	6	9	14	16
I	14	19	13	15
J	19	3	15	14

a) Compute the locality L_i and Z_{ij} values for each medoid

b) Determine the optimal dimension set D_i for each medoid m_i

c) Determine the three clusters (C_i, D_i)

d) Determine the clustering quality $Eval(C)$

e) Determine the bad medoids (for minDeviation $= 0.9$)