Institut for Matematik og Datalogi
Syddansk Universitet, Odense

6. november 2023
RF

Øvelsestimer DM573 Uge 45/46

Husk at læse de relevante sider i slides før du/I forsøger at løse en opgave.

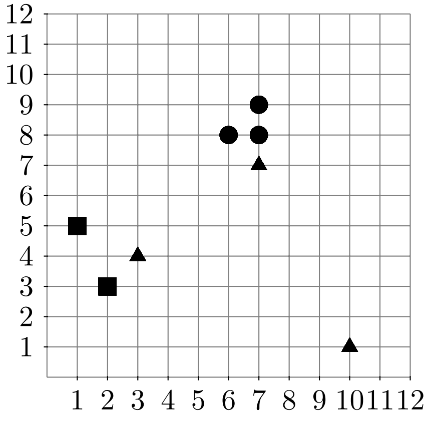
I: Løses i løbet af øvelsestimerne i uge 45

Indhentning af eventuelle hængende opgaver fra tidligere øvelsestimer. Dernæst nedenstående opgaver.

1. Repetér definitionen af en centroide for en cluster og beregn centroiden for en cluster C bestående af følgende tre punkter:

$$C = \{(2,3),(5,5),(4,1)\}.$$

2. Check beregningen af de to centroider i figuren på side 37 i Melih Kandemirs slides.
3. Consider the following data set (with 8 objects in $\mathbb{R}^2$) used in the lecture:



Compute a complete partitioning of the data set into $k = 3$ clusters using the basic k-means algorithm (the version by Forgy and Lloyd). The initial assignment of objects to clusters is given using the triangle, square, and circle markers.

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. As a help, you can use the data set figures last in this document.

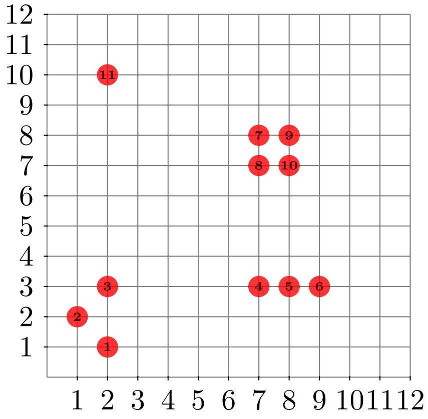
Calculate the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

4. Repetér forskellen på Forgy-Lloyd og MacQueen udgaverne af k-means algoritmen. Giver de to udgaver altid samme resultat?

5. I k-means algoritmen har initialiseringen (dvs. det første valg af centroider) ofte betydning for slutresultatet. Én metode til initialisering er et tilfældigt valg. En mere struktureret metode er *Furthest First*, som er beskrevet på side 68–78 i disse slides: `https://imada.sdu.dk/~rolf/Edu/DM534/E19/dm534_clustering.pdf`.

   Læs først om Furthest First metoden i ovenstående slides. Løs derefter nedenstående opgave.

   Consider the following data set with 11 objects (in $\mathbb{R}^2$):



   Perform a furthest-first initialization on this data set.

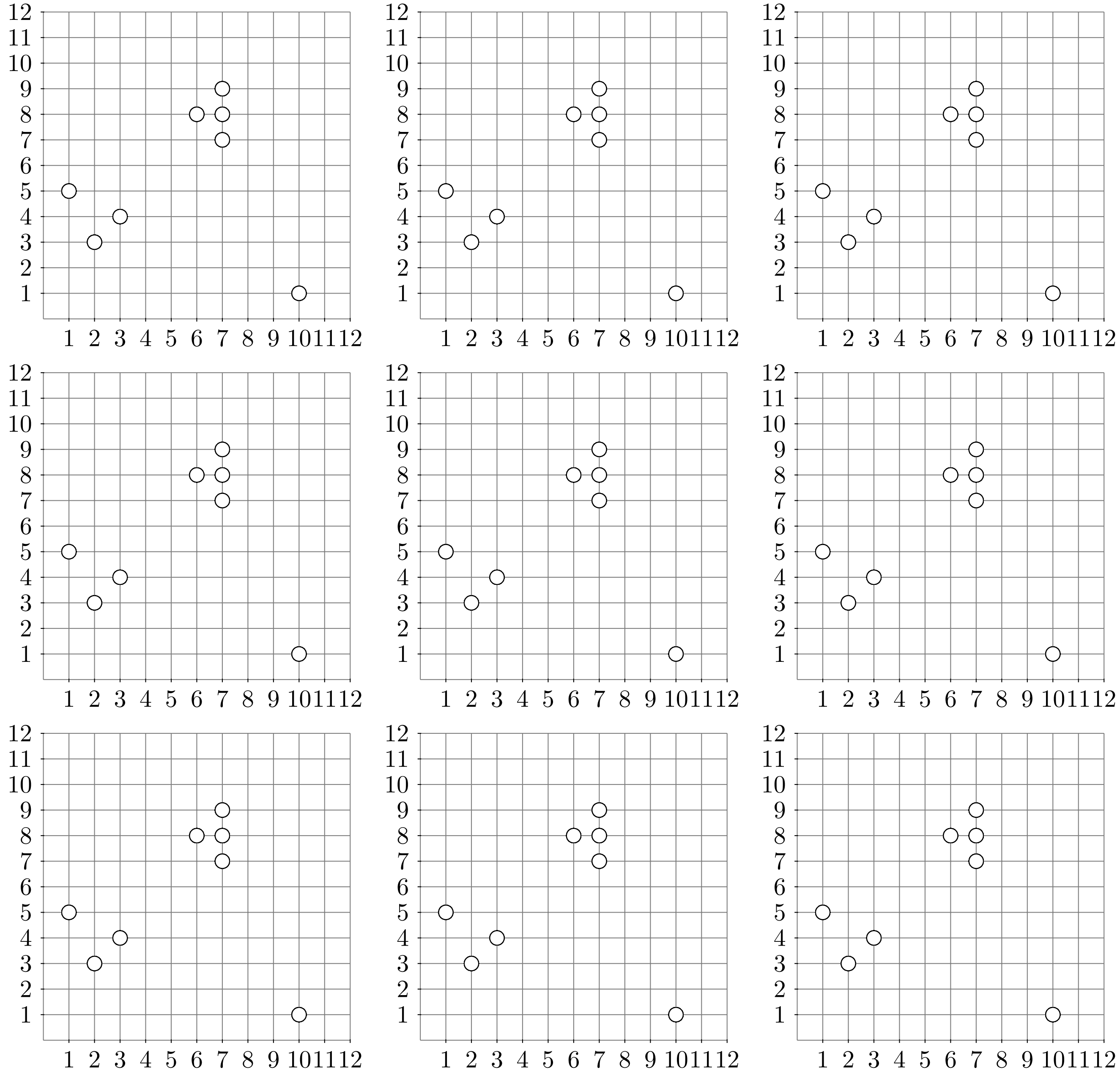
   In more detail: Let $k = 4$, i.e., in each initialisation choose four centers. Let the point with label 3 be the first center and calculate the next three centers according to the furthest-first method. In case two or more points have the same distance, choose the point with the smallest label.

The calculations needed will fill out parts of the matrix of all pair-wise distances between all points (note: only the upper triangle is required since the distance functions are symmetric). As a help, you can use the table last in this document to store your calculated distances.

II: Løses hjemme inden øvelsestimerne i uge 46

Ingen.

Datasæt figurer til brug for opgave I.3



Datasæt figurer til brug for opgave I.5

	1	2	3	4	5	6	7	8	9	10	11
1	▨										
2	▨	▨									
3	▨	▨	▨								
4	▨	▨	▨	▨							
5	▨	▨	▨	▨	▨						
6	▨	▨	▨	▨	▨	▨					
7	▨	▨	▨	▨	▨	▨	▨				
8	▨	▨	▨	▨	▨	▨	▨	▨			
9	▨	▨	▨	▨	▨	▨	▨	▨	▨		
10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	
11	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨