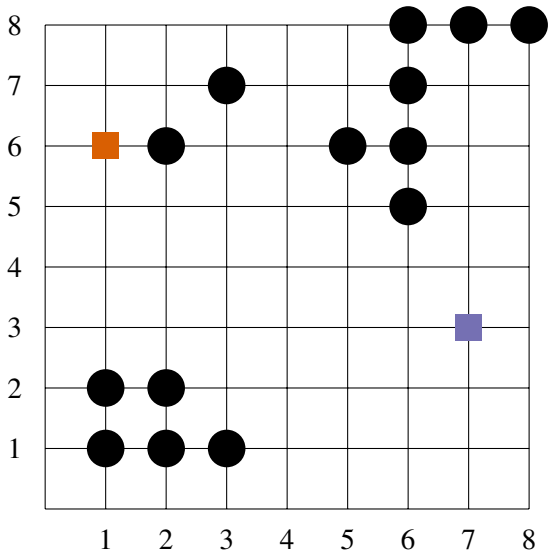
Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Janina Sontheim, Maximilian Hünemörder

Knowledge Discovery and Data Mining I
WS 2019/20

Exercise 6: k-Means, k-Modes, k-Medoids (PAM), EM

Exercise 6-1 k-Means

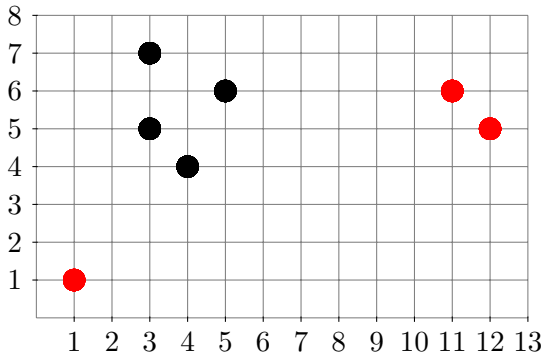
Given the following data set with 14 objects in $\mathbb{R}^2$ (the black dots):



Compute a partitioning into $k = 2$ clusters using the k-means algorithm. As initial representatives use the red and violet square. Start with computing the initial assignment. Explain and draw the assignments as well as the updated centroids after each step.

Exercise 6-2 k**-Means**

Given the following data set with 7 objects in $\mathbb{R}^2$ represented by the black and red dots:



In the following, we would like to compute complete partitionings of the dataset into $k = 3$ clusters using the k-means algorithm.

Let the initial cluster centers be given by the points marked in red. Carry out the k-Means algorithm as presented in the lecture. Which problem arises?

Exercise 6-3 k**-Mode**

Given the following Dataset of 15 Persons with their Jobs and Pets:

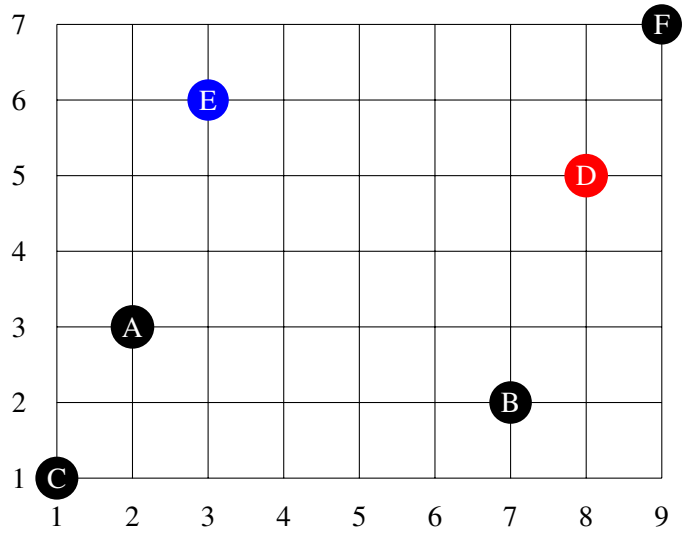
Name	Job	Pet
James	Programmer	Cat
Hans	Manager	None
Marcel	Programmer	Snake
Sebastian	Cook	None
Max	Technician	Cat
Michael	Cook	Cat
Anna	Manager	Dog
Friederike	Manager	None
Sarah	Programmer	Snake
Florian	Advisor	None
Theresa	Programmer	Cat
Jonas	Manager	None
Julian	Programmer	Cat
Nadine	Programmer	Dog
Thomas	Manager	None

Compute a partitioning using the k-Modes algorithm by Huang, Z. (Link). For initial modes choose a technician, who owns a snake and an advisor, who owns a dog.

Exercise 6-4 K-Medoid (PAM)

Consider the following 2-dimensional data set:

	A	B	C	D	E	F
x_1	2	7	1	8	3	9
x_2	3	2	1	5	6	7



(a) Perform the first loop of the PAM algorithm ($k = 2$) using the Manhattan distance. Select D and E (highlighted in the plot) as initial medoids and compute the resulting medoids and clusters.
Hint: When $C(m)$ denotes the cluster of medoid m, and M denotes the set of medoids, then the total distance TD may be computed as

$$TD = \sum_{m \in M} \sum_{o \in C(m)} d(m, o)$$

(b) How can the clustering result $C_1 = \{A, B, C\}, C_2 = \{D, E, F\}$ be obtained with the PAM algorithm ($k = 2$) using the weighted Manhattan distance

$$d(x, y) = w_1 \cdot |x_1 - y_1| + w_2 \cdot |x_2 - y_2|?$$

Assume that B and E are the initial medoids and give values for the weights w_1 and w_2 for the first and second dimension respectively.

Exercise 6-5 Assignments in EM-Algorithm

Given a data set with 100 points consisting of three Gaussian clusters A, B and C and the point p.

The cluster A contains 30% of all objects and is represented using the mean of all his points $\mu_A = (2, 2)$ and the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

The cluster B contains 20% of all objects and is represented using the mean of all his points $\mu_B = (5, 3)$ and the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

The cluster C contains 50% of all objects and is represented using the mean of all his points $\mu_C = (1, 4)$ and the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$.

The point p is given by the coordinates $(2.5, 3.0)$.
Compute the three probabilities of p belonging to the clusters A, B and C.

The following sketch is not exact, and only gives a rough idea of the cluster locations: