Indicative exercises for clustering algorithms

Solved exercises

Exercise 1 ***(proximity measures)*:** Prove that the **Euclidean distance** between two l-dimensional vectors $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_l]^T$, $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2}$, is a **metric dissimilarity measure**.

Hint: The *Minkowski inequality* states that for two l-dimensional vectors $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_l]^T$ and a positive integer p, it holds $\left(\sum_{i=1}^{l}|x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{l}|x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{l}|y_i|^p\right)^{1/p}$

Solution: In order to prove that $d(\boldsymbol{x}, \boldsymbol{y})$ is a metric dissimilarity measure, we need to prove that it possesses all the following properties:

(a) $\exists d_0 \in R$, so that $d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty$, for all $\boldsymbol{x}, \boldsymbol{y} \in R^l$

(b) $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$, for all $\boldsymbol{x} \in R^l$

(c) $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$, for all $\boldsymbol{x}, \boldsymbol{y} \in R^l$

(d) $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$

(e) $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$, for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in R^l$

For (a): It is straightforward to see that for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, it is $0 \leq d(\boldsymbol{x}, \boldsymbol{y})$. Therefore, in this case, it is $d_0 = 0$.

For (b): It is $d(\boldsymbol{x}, \boldsymbol{x}) = \sqrt{\sum_{i=1}^{l}(x_i - x_i)^2} = 0 \equiv d_0$

For (c): It is $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2} = \sqrt{\sum_{i=1}^{l}(y_i - x_i)^2} = d(\boldsymbol{y}, \boldsymbol{x})$

For (d): It is $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2} = 0 \Leftrightarrow x_i = y_i, for\ all\ i = 1, \ldots, l \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$

For (e): Here, the Minkowski inequality will be utilized. It is

$$d(\boldsymbol{x}, \boldsymbol{z}) = \left(\sum_{i=1}^{l}|x_i - z_i|^2\right)^{1/2} = \left(\sum_{i=1}^{l}|(x_i - y_i) + (y_i - z_i)|^2\right)^{1/2}$$

$$\leq \left(\sum_{i=1}^{l}|x_i - y_i|^2\right)^{1/2} + \left(\sum_{i=1}^{l}|y_i - z_i|^2\right)^{1/2} = d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$$

Therefore, since the Euclidean distance possesses all the properties from (a) to (e), it follows that it is a metric dissimilarity measure.   Q.E.D.

Exercise 2 *(proximity measures)*: Prove that the **distance** between for two l-dimensional vectors $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_l]^T$, defined as $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i{}^2 - y_i{}^2)^2}$ is a **dissimilarity measure** but **not** a **metric**.

Solution: In order to prove that $d(\boldsymbol{x}, \boldsymbol{y})$ is a dissimilarity measure, we need to prove that it possesses all the following three properties:

(a) $\exists d_0 \in R$, so that $d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty$, for all $\boldsymbol{x}, \boldsymbol{y} \in R^l$

(b) $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$, for all $\boldsymbol{x} \in R^l$

(c) $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$, for all $\boldsymbol{x}, \boldsymbol{y} \in R^l$

For (a): It is straightforward to see that for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, it is $0 \leq d(\boldsymbol{x}, \boldsymbol{y})$. Therefore, in this case, it is $d_0 = 0$.

For (b): It is $d(\boldsymbol{x}, \boldsymbol{x}) = \sqrt{\sum_{i=1}^{l}(x_i{}^2 - x_i{}^2)^2} = 0 \equiv d_0$

For (c): It is $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i{}^2 - y_i{}^2)^2} = \sqrt{\sum_{i=1}^{l}(y_i{}^2 - x_i{}^2)^2} = d(\boldsymbol{y}, \boldsymbol{x})$

Therefore, since the distance $d(\cdot,\cdot)$ possesses all the properties from (a) to (c), it follows that it is a dissimilarity measure.

In order to prove that it is not a metric, we need to prove that at least one of the following two properties are not possessed by the distance under study:

(d) $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$

(e) $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$, for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in R^l$

For (d): It is $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \sqrt{\sum_{i=1}^{l}(x_i{}^2 - y_i{}^2)^2} = 0 \Leftrightarrow x_i{}^2 = y_i{}^2, for\ all\ i = 1, \ldots, l \Leftrightarrow x_i = \pm y_i, for\ all\ i = 1, \ldots, l$

This implies that there exist vectors that although they are different, they have the minimum possible distance value (for example, for l=3, the vectors $\boldsymbol{x} = [2, 3, 4]^T$, $\boldsymbol{y} = [2, -3, -4]^T$, although they are different, their distance is equal to 0). Therefore the distance under study, is not a metric. Q.E.D.

Exercise 3 *(k-means cost function optimization algorithm)*:

Consider the data set $Y = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [0, 0]^T$, $\boldsymbol{x}_2 = [3, 0]^T$, $\boldsymbol{x}_3 = [0, 6]^T$, $\boldsymbol{x}_4 = [0, 7]^T$, $\boldsymbol{x}_5 = [-3, 7]^T$.

(a) Run the k-means clustering algorithm, for two representatives, $\boldsymbol{\theta}_1$and $\boldsymbol{\theta}_2$, whose initial positions are $\boldsymbol{\theta}_1(0) = [1, 6]^T$ and $\boldsymbol{\theta}_2(0) = [0, 8]^T$, respectively. Report the formed clusters, along with their respective representatives.

(b) What would be the clustering result for the case where $\boldsymbol{\theta}_2(0) = [0, 20]^T$?
(c) How many clusters will be obtained if three representatives where employed?

Solution: We remind that, at each iteration of the k-means algorithm, two processing steps are involved. At the first one the data vectors are assigned to the clusters (each data vector $\boldsymbol{x}$ is assigned to the cluster whose representative is closest to $\boldsymbol{x}$, in terms of the squared Euclidean distance) and at the second one each cluster representative $\boldsymbol{\theta}_j$ is re-estimated (as the mean of the data vectors that belong to the respective cluster). Based on this, we proceed as follows:

1st iteration

1st step: The ***squared Euclidean distances*** of each data point from the two representatives are given in the following table

	$\boldsymbol{\theta}_1(0) = [1, 6]^T$	$\boldsymbol{\theta}_2(0) = [0, 8]^T$
$\boldsymbol{x}_1 = [0, 0]^T$	$(0-1)^2 + (0-6)^2 = \mathbf{37}$	$(0-0)^2 + (0-8)^2 = \mathbf{64}$
$\boldsymbol{x}_2 = [3, 0]^T$	$(3-1)^2 + (0-6)^2 = \mathbf{40}$	$(3-0)^2 + (0-8)^2 = \mathbf{73}$
$\boldsymbol{x}_3 = [0, 6]^T$	$(0-1)^2 + (6-6)^2 = \mathbf{1}$	$(0-0)^2 + (6-8)^2 = \mathbf{4}$
$\boldsymbol{x}_4 = [0, 7]^T$	$(0-1)^2 + (7-6)^2 = \mathbf{2}$	$(0-0)^2 + (7-8)^2 = \mathbf{1}$
$\boldsymbol{x}_5 = [-3, 7]^T$	$(-3-1)^2 + (7-6)^2 = \mathbf{17}$	$(-3-0)^2 + (7-8)^2 = \mathbf{10}$

Since:

$||\boldsymbol{x}_1 - \boldsymbol{\theta}_1(0)||^2 = min_{j=1,2}||\boldsymbol{x}_1 - \boldsymbol{\theta}_j(0)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.[1]

$||\boldsymbol{x}_2 - \boldsymbol{\theta}_1(0)||^2 = min_{j=1,2}||\boldsymbol{x}_2 - \boldsymbol{\theta}_j(0)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_3 - \boldsymbol{\theta}_1(0)||^2 = min_{j=1,2}||\boldsymbol{x}_3 - \boldsymbol{\theta}_j(0)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_4 - \boldsymbol{\theta}_2(0)||^2 = min_{j=1,2}||\boldsymbol{x}_4 - \boldsymbol{\theta}_j(0)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

$||\boldsymbol{x}_5 - \boldsymbol{\theta}_2(0)||^2 = min_{j=1,2}||\boldsymbol{x}_5 - \boldsymbol{\theta}_j(0)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

Thus, $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ and $C_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$.

2nd step: Letting $n_1 = 3$ and $n_2 = 2$ be the cardinalities of C_1 and C_2, respectively, the cluster representatives are re-estimated as

$$\boldsymbol{\theta}_1 \equiv \boldsymbol{\theta}_1(1) = \frac{1}{n_1} \cdot (\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3) = \frac{1}{3} \cdot \left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}3\\0\end{bmatrix} + \begin{bmatrix}0\\6\end{bmatrix}\right) = \frac{1}{3} \cdot \begin{bmatrix}3\\6\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}$$

and

[1] Note that cluster C_1 (C_2) is associated with the representative $\boldsymbol{\theta}_1$ ($\boldsymbol{\theta}_2$).

$$\boldsymbol{\theta}_2 \equiv \boldsymbol{\theta}_2(1) = \frac{1}{n_2} \cdot (\boldsymbol{x}_4 + \boldsymbol{x}_5) = \frac{1}{2} \cdot \left(\begin{bmatrix} 0 \\ 7 \end{bmatrix} + \begin{bmatrix} -3 \\ 7 \end{bmatrix} \right) = \frac{1}{2} \cdot \begin{bmatrix} -3 \\ 14 \end{bmatrix} = \begin{bmatrix} -1.5 \\ 7 \end{bmatrix}$$

2nd iteration

<u>*1st step:*</u> The squared Euclidean distances of each data point from the two representatives are given in the following table

	$\boldsymbol{\theta}_1(1) = [1,2]^T$	$\boldsymbol{\theta}_2(1) = [-1.5,7]^T$
$\boldsymbol{x}_1 = [0,0]^T$	$(0-1)^2 + (0-2)^2 = \mathbf{5}$	$(0-(-1.5))^2 + (0-7)^2 = \mathbf{51.25}$
$\boldsymbol{x}_2 = [3,0]^T$	$(3-1)^2 + (0-2)^2 = \mathbf{8}$	$(3-(-1.5))^2 + (0-7)^2 = \mathbf{69.25}$
$\boldsymbol{x}_3 = [0,6]^T$	$(0-1)^2 + (6-2)^2 = \mathbf{17}$	$(0-(-1.5))^2 + (6-7)^2 = \mathbf{3.25}$
$\boldsymbol{x}_4 = [0,7]^T$	$(0-1)^2 + (7-2)^2 = \mathbf{26}$	$(0-(-1.5))^2 + (7-7)^2 = \mathbf{2.25}$
$\boldsymbol{x}_5 = [-3,7]^T$	$(-3-1)^2 + (7-2)^2 = \mathbf{41}$	$(-3-(-1.5))^2 + (7-7)^2 = \mathbf{2.25}$

Since:

$||\boldsymbol{x}_1 - \boldsymbol{\theta}_1(1)||^2 = min_{j=1,2}||\boldsymbol{x}_1 - \boldsymbol{\theta}_j(1)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_2 - \boldsymbol{\theta}_1(1)||^2 = min_{j=1,2}||\boldsymbol{x}_2 - \boldsymbol{\theta}_j(1)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_3 - \boldsymbol{\theta}_2(1)||^2 = min_{j=1,2}||\boldsymbol{x}_3 - \boldsymbol{\theta}_j(1)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

$||\boldsymbol{x}_4 - \boldsymbol{\theta}_2(1)||^2 = min_{j=1,2}||\boldsymbol{x}_4 - \boldsymbol{\theta}_j(1)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

$||\boldsymbol{x}_5 - \boldsymbol{\theta}_2(1)||^2 = min_{j=1,2}||\boldsymbol{x}_5 - \boldsymbol{\theta}_j(1)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

Thus, $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and $C_2 = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$.

<u>*2nd step:*</u> Noting that the cardinalities of C_1 and C_2 are $n_1 = 2$ and $n_2 = 3$, respectively, the cluster representatives are re-estimated as

$$\boldsymbol{\theta}_1 \equiv \boldsymbol{\theta}_1(2) = \frac{1}{n_1} \cdot (\boldsymbol{x}_1 + \boldsymbol{x}_2) = \frac{1}{2} \cdot \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \end{bmatrix} \right) = \frac{1}{2} \cdot \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$$

and

$$\boldsymbol{\theta}_2 \equiv \boldsymbol{\theta}_2(2) = \frac{1}{n_2} \cdot (\boldsymbol{x}_3 + \boldsymbol{x}_4 + \boldsymbol{x}_5) = \frac{1}{3} \cdot \left(\begin{bmatrix} 0 \\ 6 \end{bmatrix} + \begin{bmatrix} 0 \\ 7 \end{bmatrix} + \begin{bmatrix} -3 \\ 7 \end{bmatrix} \right) = \frac{1}{3} \cdot \begin{bmatrix} -3 \\ 20 \end{bmatrix} = \begin{bmatrix} -1 \\ 6.7 \end{bmatrix}$$

3rd iteration

<u>*1st step:*</u> The squared Euclidean distances of each data point from the two representatives are given in the following table

	$\boldsymbol{\theta}_1(2) = [1.5, 0]^T$	$\boldsymbol{\theta}_2(2) = [-1, 6.7]^T$
$\boldsymbol{x}_1 = [0, 0]^T$	$(0-1.5)^2 + (0-0)^2 = \mathbf{2.25}$	$(0-(-1))^2 + (0-6.7)^2 = \mathbf{45.89}$
$\boldsymbol{x}_2 = [3, 0]^T$	$(3-1.5)^2 + (0-0)^2 = \mathbf{2.25}$	$(3-(-1))^2 + (0-6.7)^2 = \mathbf{60.89}$
$\boldsymbol{x}_3 = [0, 6]^T$	$(0-1.5)^2 + (6-0)^2 = \mathbf{38.25}$	$(0-(-1))^2 + (6-6.7)^2 = \mathbf{1.49}$
$\boldsymbol{x}_4 = [0, 7]^T$	$(0-1.5)^2 + (7-0)^2 = \mathbf{51.25}$	$(0-(-1))^2 + (7-6.7)^2 = \mathbf{1.09}$
$\boldsymbol{x}_5 = [-3, 7]^T$	$(-3-1.5)^2 + (7-0)^2 = \mathbf{69.25}$	$(-3-(-1))^2 + (7-6.7)^2 = \mathbf{4.09}$

Since:

$||\boldsymbol{x}_1 - \boldsymbol{\theta}_1(2)||^2 = min_{j=1,2}||\boldsymbol{x}_1 - \boldsymbol{\theta}_j(2)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_2 - \boldsymbol{\theta}_1(2)||^2 = min_{j=1,2}||\boldsymbol{x}_2 - \boldsymbol{\theta}_j(2)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_1.

$||\boldsymbol{x}_3 - \boldsymbol{\theta}_2(2)||^2 = min_{j=1,2}||\boldsymbol{x}_3 - \boldsymbol{\theta}_j(2)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

$||\boldsymbol{x}_4 - \boldsymbol{\theta}_2(2)||^2 = min_{j=1,2}||\boldsymbol{x}_4 - \boldsymbol{\theta}_j(2)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

$||\boldsymbol{x}_5 - \boldsymbol{\theta}_2(2)||^2 = min_{j=1,2}||\boldsymbol{x}_5 - \boldsymbol{\theta}_j(2)||^2$, $\boldsymbol{x}_1$ is assigned to cluster C_2.

Thus, $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and $C_2 = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$.

2nd step: Noting that the cardinalities of C_1 and C_2 are $n_1 = 2$ and $n_2 = 3$, respectively, the cluster representatives are re-estimated as

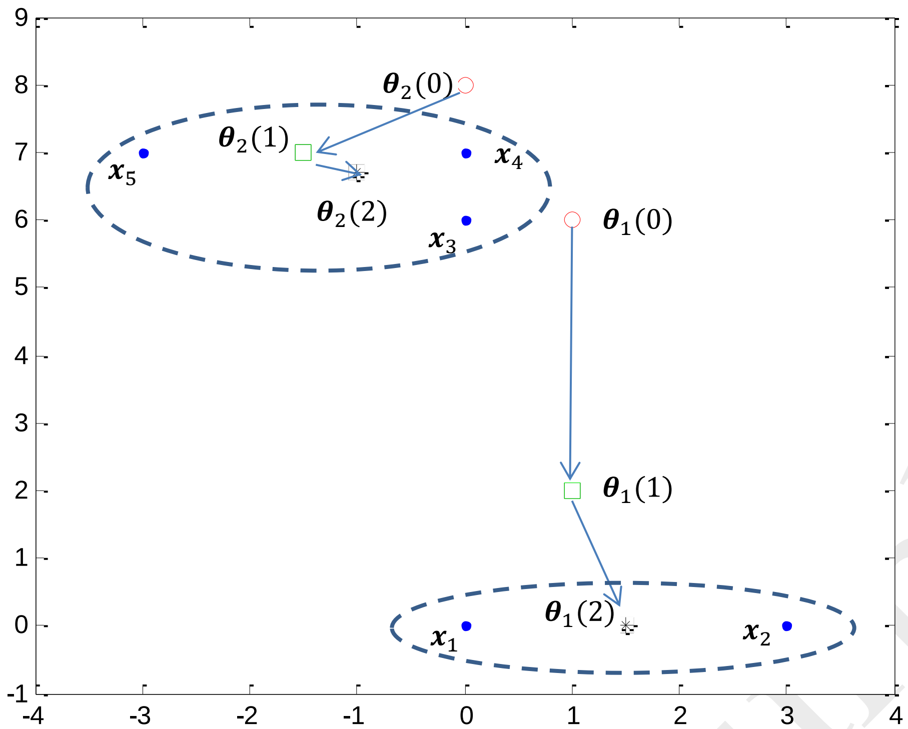
$$\boldsymbol{\theta}_1 \equiv \boldsymbol{\theta}_1(3) = \frac{1}{n_1} \cdot (\boldsymbol{x}_1 + \boldsymbol{x}_2) = \frac{1}{2} \cdot \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \end{bmatrix} \right) = \frac{1}{2} \cdot \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$$

and

$$\boldsymbol{\theta}_2 \equiv \boldsymbol{\theta}_2(3) = \frac{1}{n_2} \cdot (\boldsymbol{x}_3 + \boldsymbol{x}_4 + \boldsymbol{x}_5) = \frac{1}{3} \cdot \left(\begin{bmatrix} 0 \\ 6 \end{bmatrix} + \begin{bmatrix} 0 \\ 7 \end{bmatrix} + \begin{bmatrix} -3 \\ 7 \end{bmatrix} \right) = \frac{1}{3} \cdot \begin{bmatrix} -3 \\ 20 \end{bmatrix} = \begin{bmatrix} -1 \\ 20/3 \end{bmatrix}.$$

Since the values of $\boldsymbol{\theta}_j$'s, $j = 1,2$, remain unaltered for two successive iterations (2nd and 3rd), the algorithm terminates (equivalently, we can say that since the clustering of the data points remain unaltered for two successive iterations, the algorithm terminates). The resulting clusters are $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and $C_2 = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$ and the respective cluster representatives are $\boldsymbol{\theta}_1 = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$ and $\boldsymbol{\theta}_2 = \begin{bmatrix} -1 \\ 20/3 \end{bmatrix}$.

The evolution of the algorithm is shown in the following figure.



(b) Applying the k-means algorithm for this initialization scenario, we will see that all data points lie closer to $\boldsymbol{\theta}_1(0)$ than $\boldsymbol{\theta}_2(0)$ (alternatively, we can say that $\boldsymbol{\theta}_2(0)$ does not "win" on any data point). Thus, the k-means will return a single cluster containing all the data points, or, strictly speaking,
$C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$ and $C_2 = \emptyset$. Cluster C_1 is represented by

$$\boldsymbol{\theta}_1 \equiv \boldsymbol{\theta}_1(1) = \frac{1}{n_1} \cdot (\boldsymbol{x}_1, +\boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_4 + \boldsymbol{x}_5) = \frac{1}{5} \cdot \left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}3\\0\end{bmatrix} + \begin{bmatrix}0\\6\end{bmatrix} + \begin{bmatrix}0\\7\end{bmatrix} + \begin{bmatrix}-3\\7\end{bmatrix}\right) = \frac{1}{5} \cdot \begin{bmatrix}0\\20\end{bmatrix} = \begin{bmatrix}0\\4\end{bmatrix}.$$

(d) In general, the data set will be split into three clusters, provided that each representative "wins" on at least one data point.

Exercise 4 ***(Matrix theory-based hierarchical clustering)*****:** Consider the following dissimilarity matrix:

$$P_0 = \begin{bmatrix} 0 & 1 & 2 & 26 & 37 \\ 1 & 0 & 3 & 25 & 36 \\ 2 & 3 & 0 & 16 & 25 \\ 26 & 25 & 16 & 0 & 1.5 \\ 37 & 36 & 25 & 1.5 & 0 \end{bmatrix}$$

where the corresponding squared Euclidean distance is adopted. Run the seven agglomerative matrix-based clustering algorithms and determine the resulting clustering hierarchy, as well as the dissimilarity levels where the clusterings are produced.

Solution: As one can easily observe, the first three vectors, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_2$, are very close to each other and far away from the others. Likewise, $\boldsymbol{x}_4$ and $\boldsymbol{x}_5$ lie very close to each other and far away from the first three vectors.

For this problem all seven algorithms discussed before result in the same dendrogram. The only difference is that each clustering is formed at a different dissimilarity level. Of cource, the initial clustering is $\Re_0 = \{\{\boldsymbol{x}_1\},\{\boldsymbol{x}_2\},\{\boldsymbol{x}_3\},\{\boldsymbol{x}_4\},\{\boldsymbol{x}_5\}\}$. Let us first consider the single link algorithm. Since P_0 is symmetric, we consider only the upper diagonal elements. The smallest of these elements equals 1 and occurs at position (1,2) of P_0. Thus, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ come into the same cluster and $\Re_1 = \{\{\boldsymbol{x}_1,\boldsymbol{x}_2\},\{\boldsymbol{x}_3\},\{\boldsymbol{x}_4\},\{\boldsymbol{x}_5\}\}$ is produced. In the sequel, the dissimilarities among the newly formed cluster and the remaining ones have to be computed. This can be achieved via Eq. (2)[2]. The resulting proximity matrix, P_1, is

$$P_1 = \begin{bmatrix} 0 & 2 & 25 & 36 \\ 2 & 0 & 16 & 25 \\ 25 & 16 & 0 & 1.5 \\ 36 & 25 & 1.5 & 0 \end{bmatrix}$$

Its first row and column correspond to the cluster $\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$. The smallest of the upper diagonal elements of P_1 equals 1.5. This means that at the next stage, the clusters $\{\boldsymbol{x}_4\}$ and $\{\boldsymbol{x}_5\}$ will stick together into a single cluster, producing $\Re_2 = \{\{\boldsymbol{x}_1,\boldsymbol{x}_2\},\{\boldsymbol{x}_3\},\{\boldsymbol{x}_4,\boldsymbol{x}_5\}\}$. Employing Eq. (13.4), we obtain

$$P_2 = \begin{bmatrix} 0 & 2 & 25 \\ 2 & 0 & 16 \\ 25 & 16 & 0 \end{bmatrix}$$

where the first row (column) corresponds to $\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$, and the second and third rows (columns) correspond to $\{\boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_4,\boldsymbol{x}_5\}$, respectively. Proceeding as before, at the next stage $\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$ and $\{\boldsymbol{x}_3\}$ will get together in a single cluster and $\Re_3 = \{\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\},\{\boldsymbol{x}_4,\boldsymbol{x}_5\}\}$ is produced. The new proximity matrix, P_3, becomes

$$P_3 = \begin{bmatrix} 0 & 16 \\ 16 & 0 \end{bmatrix}$$

where the first and the second row (column) correspond to $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_4,\boldsymbol{x}_5\}$ clusters, respectively. Finally, $\Re_4 = \{\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5\}\}$ will be formed at dissimilarity level equal to 16.

Working in a similar fashion, we can apply the remaining six algorithms to P_0.[3]

However, care must be taken when we apply UPGMA, UPGMC, and Ward's method. In these cases, when a merging takes place the parameters a_i , a_j , b, and c in

$$d(C_q,C_s) = a_i d(C_i,C_s) + a_j(d(C_j,C_s) + b d(C_i,C_j) + c|d(C_i,C_s) - d(C_j,C_s)|$$

must be properly adjusted. The proximity levels at which each clustering is formed for each algorithm are shown in Table 1.

[2] All references are referred to the slides of the 9th lecture.

[3] Note that in the case of Ward's algorithm, the initial dissimilarity matrix should be $12 \cdot P_0$, due to the definition of the distance d'_{ij} between C_i and C_j is defined as

$$d'_{ij} = \frac{n_i n_j}{n_i + n_j} ||\boldsymbol{m}_i - \boldsymbol{m}_j||2$$

where $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ are the mean vectors associated with C_i and C_j, respectively.

The considered task is a nice problem with two well-defined compact clusters lying away from each other. The preceding example demonstrates that in such “easy” cases all algorithms work satisfactorily (as happens with most of the clustering algorithms proposed in the literature). The particular characteristics of each algorithm are revealed when more demanding situations are faced.

	SL	*CL*	*WPGMA*	*UPGMA*	*WPGMC*	*UPGMC*	*Ward*
$\mathfrak{R}_0$	0	0	0	0	0	0	0
$\mathfrak{R}_1$	1	1	1	1	1	1	0.5
$\mathfrak{R}_2$	1.5	1.5	1.5	1.5	1.5	1.5	0.75
$\mathfrak{R}_3$	2	3	2.5	2.5	2.25	2.25	1.5
$\mathfrak{R}_4$	16	37	25.75	27.5	24.69	26.46	31.75

Table 4.1: The Results Obtained with the Seven Algorithms Discussed when they are applied to the proximity matrix of Example 1.

Exercise 5 ***(graph theory-based hierarchical clustering)*****:** Consider the following dissimilarity matrix:

$$P(X) = \begin{bmatrix} 0 & 1 & 9 & 18 & 19 & 20 & 21 \\ 1 & 0 & 8 & 13 & 14 & 15 & 16 \\ 9 & 8 & 0 & 17 & 10 & 11 & 12 \\ 18 & 13 & 17 & 0 & 5 & 6 & 7 \\ 19 & 14 & 10 & 5 & 0 & 3 & 4 \\ 20 & 15 & 11 & 6 & 3 & 0 & 2 \\ 21 & 16 & 12 & 7 & 4 & 2 & 0 \end{bmatrix}$$

Let $h(k)$ be the **node degree property** with $k = 2$; that is, it is required that each node has at least two incident edges. Derive the associated dissimilarity dendrogram.

Figure 1 shows the $G(13)$ proximity graph produced by this dissimilarity matrix. Then the obtained threshold dendrogram is shown in Figure 2a. At dissimilarity level 1, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ form a single cluster. This happens because $\{\boldsymbol{x}_1\} \cup \{\boldsymbol{x}_2\}$ is complete at $G(1)$, despite the fact that property $h(2)$ is not satisfied (remember the disjunction between conditions (b1) and (b2) in Eq. (4)). Similarly, $\{\boldsymbol{x}_6\} \cup \{\boldsymbol{x}_7\}$ forms a cluster at dissimilarity level 2. The next clustering is formed at level 4, since $\{\boldsymbol{x}_5\} \cup \{\boldsymbol{x}_6, \boldsymbol{x}_7\}$ becomes complete in $G(4)$.

At level 6, $\boldsymbol{x}_4$, $\boldsymbol{x}_5$, $\boldsymbol{x}_6$, and $\boldsymbol{x}_7$ lie for the first time in the same cluster. Although this subgraph is not complete, it does satisfy $h(2)$. Finally, at level 9, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ come into the same cluster. Note that, although all nodes in the graph have node degree equal to 2, the final clustering will be formed at level 10 because at level 9 the graph is not connected. Assume now that $h(k)$ is the node connectivity property, with

$k = 2$; that is, all pairs of nodes in a connected subgraph are joined by at least two paths having no nodes in common. The dissimilarity dendrogram produced in this case is shown in Figure 2b. Finally, the dissimilarity dendrogram produced when the edge connectivity property with $k = 2$ is employed is shown in Figure 2c.

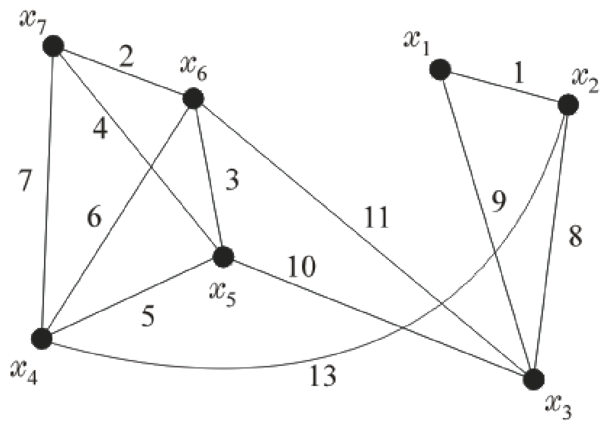


Figure 5.1: The proximity graph $G(13)$ derived by the dissimilarity matrix P given in Exercise 5.

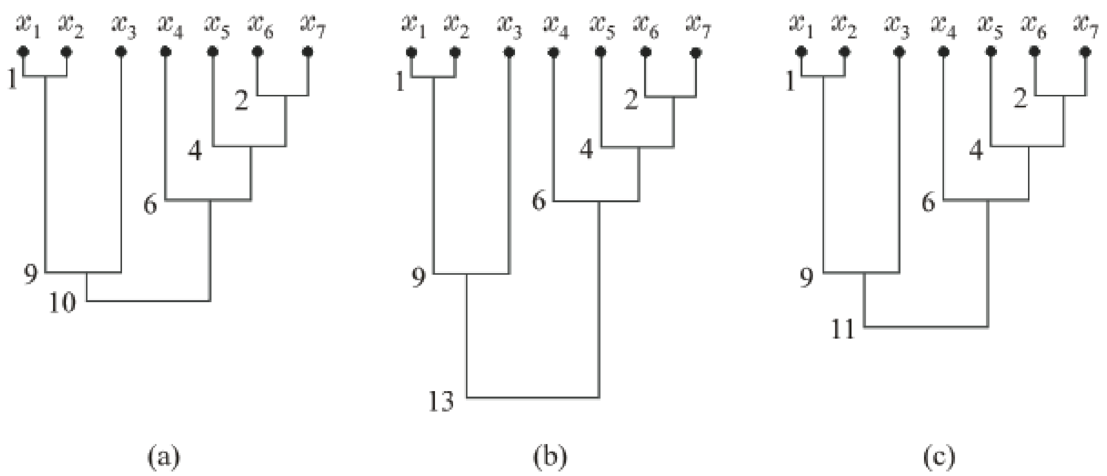


Figure 5.2: Dissimilarity dendrograms related to Exercise 5. (a) Dissimilarity dendrogram produced when $h(k)$ is the node degree property, with $k = 2$. (b) Dissimilarity dendrogram produced when $h(k)$ is the node connectivity property, with $k = 2$. (c) Dissimilarity dendrogram produced when $h(k)$ is the edge connectivity property, with $k = 2$.

Exercise 6 *(Spectral clustering)***:** Consider the data set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [0, 0]^T$, $\boldsymbol{x}_2 = [0, 1]^T$, $\boldsymbol{x}_3 = [5, 0]^T$, $\boldsymbol{x}_4 = [5, 1]^T$, $\boldsymbol{x}_5 = [4, 1]^T$. Perform spectral clustering using the 1-NN for the construction of the similarity graph (initial phase). The weight of the edges in the graph will be computed via the equation $s(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2)$.

Solution: <u>*Construction of the **similarity matrix**:*</u> It is $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\left|\left|\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0\\1\end{bmatrix}\right|\right|^2\right) = \exp(-1) = 0.3679 \approx 0.4$. In the same spirit, we compute the similarities between any pair of points and we end up with the following similarity matrix[4]

$$S = \begin{bmatrix} 1.0 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.4 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.4 & 0.1 \\ 0.0 & 0.0 & 0.4 & 1.0 & 0.4 \\ 0.0 & 0.0 & 0.1 & 0.4 & 1.0 \end{bmatrix}$$

<u>*Construction of the **similarity graph**:*</u> From the first row of matrix S, it follows that the nearest neighbor of $\boldsymbol{x}_1$ is $\boldsymbol{x}_2$,

the nearest neighbor of $\boldsymbol{x}_2$ is $\boldsymbol{x}_1$,

the nearest neighbor of $\boldsymbol{x}_3$ is $\boldsymbol{x}_4$,

the nearest neighbor of $\boldsymbol{x}_4$ is $\boldsymbol{x}_3$ (we could take $\boldsymbol{x}_5$ instead),

the nearest neighbor of $\boldsymbol{x}_5$ is $\boldsymbol{x}_4$.

Thus, the similarity graph $G = (V, E)$, consists of five vertices, i.e., $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5\}$, each one corresponding to a data point, while the set of edges is $E = \{e_{12}, e_{34}, e_{45}\}$. The associated ***weighted adjacency matrix*** is

$$W = \begin{bmatrix} 1.0 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.4 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.4 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 1.0 \end{bmatrix}$$

<u>*Construction of the **Laplacian matrix**:*</u> The degree matrix and the (unnormalized) Laplacian matrices are, respectively

$$D = \begin{bmatrix} 1.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.4 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.4 \end{bmatrix} \text{ and } L = D - W = \begin{bmatrix} +0.4 & -0.4 & 0.0 & 0.0 & 0.0 \\ -0.4 & +0.4 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & +0.4 & -0.4 & 0.0 \\ 0.0 & 0.0 & -0.4 & +0.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.4 & +0.4 \end{bmatrix}$$

<u>*Determination of the zero eigenvalues and the associated eigenvectors of L:*</u> It is

$$det(L - \lambda I) = \begin{vmatrix} 0.4-\lambda & -0.4 & 0.0 & 0.0 & 0.0 \\ -0.4 & 0.4-\lambda & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4-\lambda & -0.4 & 0.0 \\ 0.0 & 0.0 & -0.4 & 0.4-\lambda & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.4 & 0.4-\lambda \end{vmatrix} = -\lambda \begin{vmatrix} 1 & -0.4 & 0.0 & 0.0 & 0.0 \\ 1 & 0.4-\lambda & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4-\lambda & -0.4 & 0.0 \\ 0.0 & 0.0 & -0.4 & 0.4-\lambda & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.4 & 0.4-\lambda \end{vmatrix}$$

$$= (-\lambda)^2 \begin{vmatrix} 1 & -0.4 & 0.0 & 0.0 & 0.0 \\ 1 & 0.4-\lambda & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1 & -0.4 & 0.0 \\ 0.0 & 0.0 & 1 & 0.4-\lambda & 0.0 \\ 0.0 & 0.0 & 1 & -0.4 & 0.4-\lambda \end{vmatrix}$$

[4] The precision is up to the first decimal.

The equation $det(L-\lambda I)=0$ has **two zero** eigenvalues[5].

The eigenvectors $\boldsymbol{x}$ corresponding to a zero eigenvalue of L should satisfy $L\boldsymbol{x}=0\boldsymbol{x}$. It is clear the for the vectors $\boldsymbol{u}_1=[1,1,0,0,0]^T$, $\boldsymbol{u}_2=[0,0,1,1,1]^T$ it is

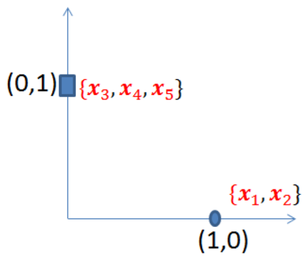
$$L\cdot\boldsymbol{u}_1=\begin{bmatrix}+0.4 & -0.4 & 0.0 & 0.0 & 0.0\\ -0.4 & +0.4 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & +0.4 & -0.4 & 0.0\\ 0.0 & 0.0 & -0.4 & +0.4 & 0.0\\ 0.0 & 0.0 & 0.0 & -0.4 & +0.4\end{bmatrix}\begin{bmatrix}1\\1\\0\\0\\0\end{bmatrix}=\begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix}=0\cdot\boldsymbol{u}_1$$

$$L\cdot\boldsymbol{u}_2=\begin{bmatrix}+0.4 & -0.4 & 0.0 & 0.0 & 0.0\\ -0.4 & +0.4 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & +0.4 & -0.4 & 0.0\\ 0.0 & 0.0 & -0.4 & +0.4 & 0.0\\ 0.0 & 0.0 & 0.0 & -0.4 & +0.4\end{bmatrix}\begin{bmatrix}0\\0\\1\\1\\1\end{bmatrix}=\begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix}=0\cdot\boldsymbol{u}_2$$

Thus, these are the eigenvectors associated with the zero eigenvalues.

Construction of the U matrix: It is

$$U=\begin{bmatrix}1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\\ 0 & 1\end{bmatrix}\begin{matrix}\equiv & \boldsymbol{y}_1\rightarrow & \boldsymbol{x}_1\\ \equiv & \boldsymbol{y}_2\rightarrow & \boldsymbol{x}_2\\ \equiv & \boldsymbol{y}_3\rightarrow & \boldsymbol{x}_3\\ \equiv & \boldsymbol{y}_4\rightarrow & \boldsymbol{x}_4\\ \equiv & \boldsymbol{y}_5\rightarrow & \boldsymbol{x}_5\end{matrix}$$



The i-th data vector is mapped to a vector in a new two-dim. space whose coordinates are the i-th associated coordinates of the two eigenvectors. Thus, the final clustering in the transformed space consists of the clusters $C'_1=\{\boldsymbol{y}_1,\boldsymbol{y}_2\}$ and $C'_2=\{\boldsymbol{y}_3,\boldsymbol{y}_4,\boldsymbol{y}_5\}$. Thus the clustering of the original data consists of the clusters

$C_1=\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$ and $C_2=\{\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5\}$.

Indicative questions

Question 1: Consider a clustering task where the involved N entities are represented in a two-dimensional feature space associated with the real-valued features x_1 and x_2. The x_1 values of the entities are ranged in [0,1], while the x_2 values of the entities are ranged in [0,1000] (assume also that the N x_1 values are uniformly arranged in [0,1] and the N x_2 values are uniformly arranged in [0,1000]). Propose a transformation of the original feature space, so that the distance between any two vectors to be equally influenced by both feature values.

Hint: If $a\le x\le b$, then $c\le c+\frac{x-a}{b-a}(d-c)\le d$

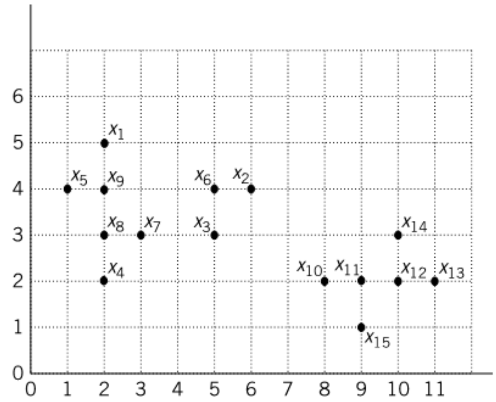
Note 1: You should be able to compute the overlap distance between two discrete-valued data vectors.

[5] The determinant left has no additional zero eigenvalues, since if we set λ=0 to it, all the columns of the determinant are independent.

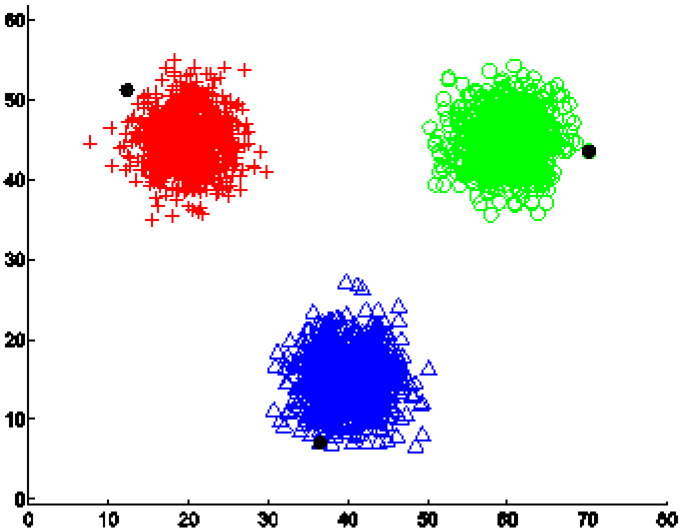
Question 2: What are the differences between the mean vector, the mean center and the median center of cluster?

Question 3: In the case where elongated clusters are formed by the data vectors of a data set, what kind of representatives you would use in the establishment a relevant clustering algorithm? Is there any case where the adoption of point representatives would work satisfactory in this case?

Question 4: In the data set the is depicted graphically below run the BSAS algorithm for $\Theta = 2$ and $q = 2$ starting from $\boldsymbol{x}_1$ and then from $\boldsymbol{x}_{15}$. What is the resulting clusterings for the two cases? Are they identical?



Question 5: Consider the following data set. What will be the result of (a) the k-means, (b) the fuzzy c-means and (c) the possibilistic algorithms for (a) three, (b) four and (c) five representatives, starting from different initial conditions?



Question 6: Derive a cost function optimization clustering algorithm based on a certain cost function $J(\Theta, U)$, where Θ is the set of the parameter vectors associated with the clusters and U is a matrix whose (i,j) quantifies the relationship between the i-th vector and the j-th cluster.

Hint: An iterative two-step procedure will be followed: At the first step U will be updated for fixed Θ and at the second step Θ will be updated for fixed U. The kind of representatives will be decided based on the shape of the clusters formed by the data.

Question 7: Propose a way for estimating the true number of clsuters, using an algorithm that (a) does not take as input the number of clusters and (b) it does require knowledge of the number of clusters.

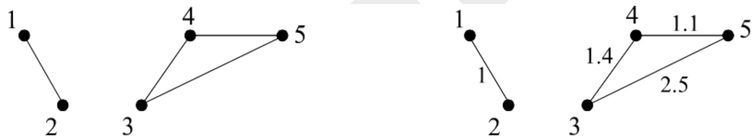
Hint: Remember the plots with (a) the large "plateau" and (b) the significant "knee".

Question 8: Is it possible for the k-medoids algorithm to be derived using tools from mathematical analysis (e.g., gradients etc)?

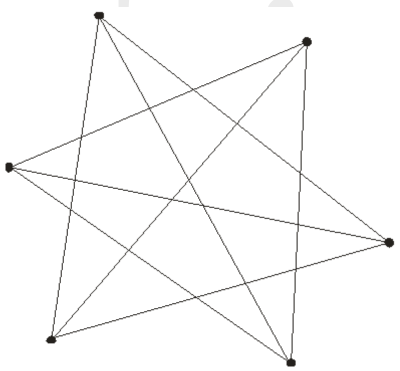
Question 9: Determine the threshold graph $G(3)$ and the dissimilarity graph $G_p(3)$ for the data set whose associated proximity matrix is

$$P(X) = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0 \end{bmatrix}$$

Answer:



Question 10: Determine the node connectivity, the edge connectivity and the node degree of the graph.



Hint: See the slides.

Question 11: Propose ways for determining the natural clustering from a hierarchy of clusterings that best describes the data.

Hint: See slides (e.g., intrinsic and extrinsic methods)

Question 12: What is the main approach in dealing with large data sets, in the hierarchical clustering algorithms case?

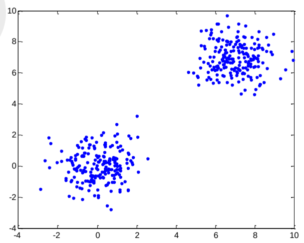
Hint: Sampling.

Question 13: If the neighboring edges of a given edge (of weight 30) in the MST associated with a data set have weights $2, 3, 5, 1$ and $q = 2$, is the above edge “unusually large”?

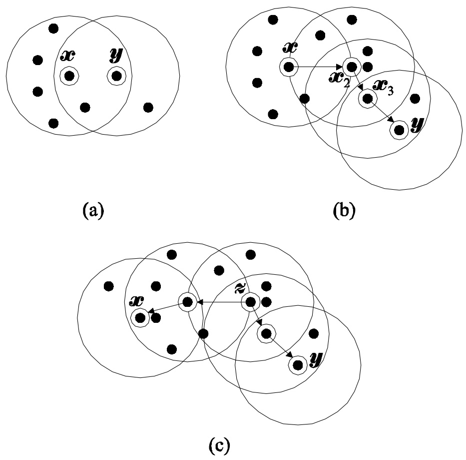
Hint: See slides (graph-algorithm based on the MST).

Question 14: How many data points are in the region of influence defined by two data points (in the algorithm based on regions of influence).

Question 15: Consider the two-cluster task shown below. Consider the basic competitive and the leaky learning scheme and assume that two representatives are considered. The first one lies in the middle of the two clsuters and the other one far away from both. What will be the result of each one of the clusters?



Question 16: Is the point $\boldsymbol{x}$ (a) directly density reachable, (b) density reachable and (c) density connected with $\boldsymbol{y}$ (DBSCAN $q = 5$)?



Question 17: We apply different clustering algorithms on a certain data set and we end up with K different clusterings. We would like to combine all of them in order to take a single representative clustering. Define the corresponding co-association matrix and run a hierarchical algorithm.

Question 18: Subspace clustering algorithms have been designed for identifying clusters that live in the same subspace of the original feature space (yes/no).

Answer: No

Question 19: Cosnider the Rand measure that measures the agreement between two clustering structures. When it takes its minimum and maximum values and what are they?

Answer: Maximum value (1) – perfect match, Minimum value (0) – perfect mismatch