Week 7: Revision

Q2. What is the Euclidean distance between the two vectors $\begin{bmatrix} 8 \\ -1 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 11 \\ 1 \\ -4 \end{bmatrix}$?

Your answer must be accurate to at least 3 decimal places.

$$Ed = \sqrt{(8-4)^2 + (-1-11)^2 + (1-1)^2 + (0--4)^2} = \sqrt{4^2 + (-12)^2 + 0^2 + 4^2}$$

$$= \sqrt{16 + 144 + 0 + 16} = \sqrt{176} = 4\sqrt{11} = 13.266$$

Q3. Let $x_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $x_2 = \begin{bmatrix} 7 \\ -3 \end{bmatrix}$ and $x_3 = \begin{bmatrix} -3 \\ 7 \end{bmatrix}$

Suppose the centroid of the cluster $\{x_1, x_2, x_3\}$ is $\begin{bmatrix} a \\ b \end{bmatrix}$

The centroid of $\{x_1, x_2, x_3\} = \frac{1}{3}\begin{bmatrix} 2+7+(-3) \\ -1-3+7 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

Q4. Let $\omega_1 = \left\{ \begin{bmatrix} -3 \\ 6 \end{bmatrix}, \begin{bmatrix} 1 \\ 9 \end{bmatrix} \right\}$ and $\omega_2 = \left\{ \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 7 \\ 4 \end{bmatrix} \right\}$

What is the average linkage between the two clusters? Your answer must be accurate to at least 3 decimal places.

$$d_{avg}(\omega_1, \omega_2) = \frac{1}{|\omega_1| \cdot |\omega_2|} \sum_{x_1 \in \omega_1,} \sum_{x_2 \in \omega_2} \|x_1 - x_2\|$$

	$\begin{bmatrix} 5 \\ 6 \end{bmatrix}$	$\begin{bmatrix} 7 \\ 4 \end{bmatrix}$
$\begin{bmatrix} -3 \\ 6 \end{bmatrix}$	8	$2\sqrt{26}$
$\begin{bmatrix} 1 \\ 9 \end{bmatrix}$	5	$\sqrt{61}$

$\therefore d_{avg} = \frac{1}{2 \cdot 2}(8 + 2\sqrt{26} + 5 + \sqrt{61}) = 7.752$

$d = \sqrt{(-3-5)^2 + (6-6)^2} = \sqrt{8^2} = 8$

$d = \sqrt{(-3-7)^2 + (6-4)^2} = \sqrt{(-10)^2 + (2)^2} = \sqrt{100+4} = \sqrt{104} = 2\sqrt{26}$

$d = \sqrt{(1-5)^2 + (9-6)^2} = \sqrt{(-4)^2 + (3)^2} = \sqrt{16+9} = \sqrt{25} = 5$

$d = \sqrt{(1-7)^2 + (9-4)^2} = \sqrt{(-6)^2 + (5)^2} = \sqrt{36+25} = \sqrt{61}$

Q1:- K-Means algorithm is a randomized algorithm, while Agglomerative Hierarchical Clustering is a deterministic algorithm.

- The output of K-Means algorithm is at most k clusters.

- In Agglomerative Hierarchical Clustering, we need to choose a linkage function that measures the distance between two clusters.
- Using single linkage has higher tendency to generate widespread cluster.
- The complete linkage of two clusters is the maximum distance between any observation of one cluster and any observation of the other cluster.

Q5: A dataset has 8 observations in $\mathbb{R}^2$. The observations are

$x_1 = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$ $x_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$ $x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $x_4 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$

$x_5 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $x_6 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$ $x_7 = \begin{bmatrix} 7 \\ 4 \end{bmatrix}$ $x_8 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

Run Agglomerative Hierarchical Clustering on this dataset by using single linkage. Draw the dendrogram. If 4 clusters should be output, what are the 4 clusters?

$\{x_1, x_4, x_7\}, \{x_2, x_6\}, \{x_3, x_5\}, \{x_8\}.$

Week 7: Quiz

Q4. Let $\omega_1 = \left\{ \begin{bmatrix} -3 \\ 6 \end{bmatrix}, \begin{bmatrix} 1 \\ 10 \end{bmatrix} \right\}$ and $\omega_2 = \left\{ \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 7 \\ 4 \end{bmatrix}, \begin{bmatrix} -3 \\ 3 \end{bmatrix} \right\}$.

What is the centroid linkage between the two clusters? Your answer must be accurate to at least 3 decimal places.

$$\text{Centroid of } \omega_1 = \begin{bmatrix} \frac{-3+1}{2} \\ \frac{6+10}{2} \end{bmatrix} = \begin{bmatrix} -1 \\ 8 \end{bmatrix}$$

$$\text{Centroid of } \omega_2 = \begin{bmatrix} \frac{5+7-3}{3} \\ \frac{5+4+3}{3} \end{bmatrix} = \begin{bmatrix} \frac{9}{3} \\ \frac{12}{3} \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$= \sqrt{(-1-3)^2 + (8-4)^2} = \sqrt{(-4)^2 + 4^2} = \sqrt{16+16}$$

$$= 5.657$$

- The centroid linkage between the two clusters is the distance between their centroids.

Q5. A dataset has 6 observations in $\mathbb{R}^3$. the observations are

$$x_1 = \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad x_3 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \quad x_4 = \begin{bmatrix} 7 \\ 3 \\ -1 \end{bmatrix} \quad x_5 = \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} \quad x_6 = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$$

Run Agglomerative Hierarchical Clustering on this dataset by using complete linkage. Draw the dendrogram.

- The height of the root of the dendrogram is ____
- One of the children nodes corresponds to a cluster that contains x_1.

If the centroid of this cluster is $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$, then a= __, b= __ & c= ____

Q1. - Unsupervised learning deals with unlabelled datasets.

- One technique for unsupervised learning is clustering.
- If we do not know how many clusters we desire at the beginning, we should use agglomerative hierarchical clustering algorithm.

Q2. For any two clusters; when we compute their single linkage, complete linkage and average linkage, which of the following inequality is true?

single linkage ≤ average linkage ≤ complete linkage

Q3. When K-Means algorithm is applied to a dataset with n observations, which of the followings are true?

The two nearest observations must belong to the same cluster on the final outputs.