Exam: K-Means algorithm

Q5. b) Consider the following dataset $\mathcal{D}$ with 6 datapoints. Each datapoint is in $\mathbb{R}^2$.

$$\mathcal{D} = \left\{ x_1 = \begin{bmatrix}0\\0\end{bmatrix}, x_2 = \begin{bmatrix}1\\1\end{bmatrix}, x_3 = \begin{bmatrix}2\\2\end{bmatrix}, x_4 = \begin{bmatrix}6\\4\end{bmatrix}, x_5 = \begin{bmatrix}-3\\0\end{bmatrix}, x_6 = \begin{bmatrix}3\\5\end{bmatrix} \right\}$$

- We run the K-means algorithm with Euclidean distance on this dataset, where $K=2$.
- Suppose that initially x_1, x_4, x_6 are assigned to cluster 1, while x_2, x_3, x_5 are assigned to cluster 2.
- Compute the clusters formed immediately after the first iteration of the algorithm. Show steps.

Step① Assign each observations to clusters

Observation	x_1	x_2	x_3	x_4	x_5	x_6
Cluster	1	2	2	1	2	1

Step② Compute the centroid of each cluster

$$\text{Centroid of cluster 1} = \frac{1}{3}\begin{bmatrix}0+6+3\\0+4+5\end{bmatrix} = \frac{1}{3}\begin{bmatrix}9\\9\end{bmatrix} = \begin{bmatrix}3\\3\end{bmatrix}$$

$$\text{Centroid of cluster 2} = \frac{1}{3}\begin{bmatrix}1+2+(-3)\\1+2+0\end{bmatrix} = \frac{1}{3}\begin{bmatrix}0\\3\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix}$$

Step③ Re-assign it to the cluster whose centroid is the closest to the observation.

- If there is a tie, priority is given to staying with the current cluster.
- We compute the square of the Euclidean distance between an observation and a centroid

Clusters:	Cluster 1	Cluster 2	Re-assign to cluster
Centroids:	$\begin{bmatrix}3\\3\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	
$x_1 = \begin{bmatrix}0\\0\end{bmatrix}$	$9+9=18$	1	2
$x_2 = \begin{bmatrix}1\\1\end{bmatrix}$	$(3-1)^2+(3-1)^2=4+4=8$	$1^2=1$	2
$x_3 = \begin{bmatrix}2\\2\end{bmatrix}$	$(3-2)^2+(3-2)^2=1^2+1^2=2$	$2^2+1^2=5$	1

$$\begin{bmatrix}3\\3\end{bmatrix} \qquad \begin{bmatrix}0\\1\end{bmatrix}$$

$x_4 = \begin{bmatrix}6\\4\end{bmatrix}$ $\quad 3^2+1^2 = 10 \quad 6^2+3^2 = 36+9 = 45 \quad 1$

$x_5 = \begin{bmatrix}-3\\0\end{bmatrix}$ $\quad 6^2+3^2 = 36+9 = 45 \quad 3^2+1^2 = 9+1 = 10 \quad 2$

$x_6 = \begin{bmatrix}3\\5\end{bmatrix}$ $\quad 2^2 = 4 \quad 3^2+4^2 = 9+16 = 25 \quad 1$

Observation	x_1	x_2	x_3	x_4	x_5	x_6
Cluster	2	2	1	1	2	1

The cluster assignment changes. As there is change, we need

Centroid of cluster 1 $= \frac{1}{3}\begin{bmatrix}2+6+3\\2+0+5\end{bmatrix} = \frac{1}{3}\begin{bmatrix}11\\7\end{bmatrix} = \begin{bmatrix}11/3\\7/3\end{bmatrix}$

Centroid of cluster 2 $= \frac{1}{3}\begin{bmatrix}0+1+(-3)\\0+1+0\end{bmatrix} = \frac{1}{3}\begin{bmatrix}-2\\1\end{bmatrix} = \begin{bmatrix}-2/3\\1/3\end{bmatrix}$

c) An important observation was made about the K-means algorithm with Euclidean distances:

- For any dataset $\{x_1, x_2, \ldots, x_n\}$ and
- for any initial assignments to clusters
- the algorithm always reduces the value of $\sum_{i=1}^{n} \|x_i - C(i)\|^2$
- after each iteration,

where, • $C(i)$ denote the centroid of the cluster that datapoint x_i belongs to,

• and $\|x_i - C(i)\|$ denotes the Euclidean distance between x_i and $C(i)$

Explain why this important observation is true.

- K-Means algorithm is an iterative algorithm.
- In each iteration, it improves the total distance between each observation and the centroid of its cluster
- When each observation is close to the centroid of its cluster, the observations in the same cluster are close to each other.

Exam paper: CS5100J Data Analysis

K-Means Algorithm

Q5. b) $D = \left\{ x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, x_3 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, x_4 = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, x_5 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}, x_6 = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \right\}$

- Run the K-Means algorithm with Euclidean distances
- where $K = 2$
- x_1, x_4, x_6 cluster 1 and x_2, x_3, x_5 cluster 2

1. Compute centroid of each cluster

$$\bar{x}1 = \frac{1}{3}\begin{bmatrix} 0+6+3 \\ 0+4+5 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 9 \\ 9 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

$$\bar{x}2 = \frac{1}{3}\begin{bmatrix} 1+2+(-3) \\ 1+2+0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

2. Compute Euclidean distance between an observation and a centroid

	$\begin{bmatrix} 3 \\ 3 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	Re-assign to cluster
$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\sqrt{3^2+3^2} = \sqrt{18}$	$\sqrt{0+1^2} = \sqrt{1}$	2
$x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\sqrt{8}$	$\sqrt{1}$	2
$x_3 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$	$\sqrt{2}$	$\sqrt{5}$	1
$x_4 = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$	$\sqrt{10}$	$\sqrt{36+9} = \sqrt{45}$	1
$x_5 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$	$\sqrt{(-3-3)^2+(0-3)^2} = \sqrt{45}$	$\sqrt{10}$	2
$x_6 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$	$\sqrt{4}$	$\sqrt{25}$	1

c) The algorithm always reduces the value $\sum_{i=1}^{n} \|x_i - C(i)\|^2$

- $C(i)$ centroid of the cluster

Explain why this important observation is true.

d) Explain why the k-means algorithm with Euclidea distance always terminates.
- for any dataset and
- for any initial assignments to clusters.

- k-means must terminate
- potential function for each cluster assignment C
- let C_j be its j^{th} cluster, and V_j the centroid of C_j

$$\Phi(C) = \sum_{C_j} \sum_{x \in C_j} \|x - V_j\|^2$$

$\Phi(C)$ = the total distance between each observation and its cluster's centroid.

- the value of $\Phi(C)$ strictly decreases after each iteration, except for the last iteration.
- consequently, k-means must terminate on any input.

- Since there are only finitely many possible cluster assignments, so k-means must terminate.

5. K-Means algorithms

a) The K-Means algorithm with Euclidean distance
- is a very popular and widely used method for data clustering.

What is the basic assumption on the distribution of the data in this K-Means clustering?

In K-Means algorithm in Euclidean distance measure there are total two assumptions made:
1. Clusters are Spherical in shape and
2. Clusters are of similar in sizes.
3. Data points in one cluster are not well separated from data points of other clusters &
4. there is wide variation in density among the data points.

b) Answer the following questions in the context of the K-Means algorithm.
- What are the inputs?
- which parameters are usually specified by the user?
- what objective function does the K-Means algorithm minimise?

The inputs are datasets with observations in $\mathbb{R}^d$.
And the parameter K specified by the user.

The K-Means algorithm minimise the total distance between each observation and its cluster's centroid.