Ds tutorial

Question 1

Given the following dataset of points in a 2D space:

Dataset = {(1, 2), (2, 1), (2, 3), (3, 2), (7, 8), (8, 7), (8, 9), (9, 8)}

Assume (K = 2 \l) and the initial centroids are C_1 = (2, 2) and C_2 = (8, 8).

a)Perform three iterations of the K-Means clustering algorithm. Show all calculations, including:

- Assignment of points to clusters.
- Recalculation of centroids after each iteration.

b)Discuss one limitation of the K-Means clustering algorithm and suggest a way to overcome it.

Answer:

a)

Algorithm:

We will perform three iterations of the K-Means clustering algorithm for the given dataset with K = 2, using the initial centroids:

C_1 = (2,2)

C_2 = (8,8)

The Euclidean distance formula is used to assign points to the nearest cluster:

$d((x1,y\ 1),(x2\ ,y2))= \sqrt{(x2 -x1)^2 +(y2 -y1)^2}$

Iteration 1: Assignment of Points to Clusters

We compute the distances of each point from both centroids:

Point	Distance to C_1 (2,2)	Distance to C_2 (8,8)	Assigned Cluster
(1,2)	$\sqrt{(1)} = 1.0$	$\sqrt{(85)} \approx 9.22.$	C_1
(2,1)	$\sqrt{(1)} = 1.0$	$\sqrt{(85)} \approx 9.22.$	C_1
(2,3)	$\sqrt{(1)} = 1.0$	$\sqrt{(61)} \approx 7.81$	C_1
(3,2)	$\sqrt{(1)} = 1.0$	$\sqrt{(61)} \approx 7.81$	C_1
(7,8)	$\sqrt{(61)} \approx 7.81$	$\sqrt{(1)} = 1.0$	C_2
(8,7)	$\sqrt{(61)} \approx 7.81$	$\sqrt{(1)} = 1.0$	C_2
(8,9)	$\sqrt{(85)} \approx 9.22$	$\sqrt{(1)} = 1.0$	C_2
(9,8)	$\sqrt{(85)} \approx 9.22$	$\sqrt{(1)} = 1.0$	C_2

New clusters:

C_1 (Cluster 1): {(1,2), (2,1), (2,3), (3,2)}

C_2 (Cluster 2): {(7,8), (8,7), (8,9), (9,8)}

Recalculating Centroids

New centroid for C_1:

$C1'=((1+2+2+3)\div4,(2+1+3+2)\div4) = (8/4,8/4) =(2,2)$

New centroid for C_2:
$C2'$=((7+8+8+9)/4,(8+7+9+8)/4) = (32/4,32/4) = (8,8)

Since the centroids remain unchanged, the algorithm has converged after one iteration and further iterations will produce the same result.

b)
Limitation of K-Means Clustering & Solution
One limitation of K-Means is that it is sensitive to initial centroid placement, which can lead to poor clustering or local minima.

Solution:
A common way to overcome this issue is using K-Means++ initialization, which selects initial centroids in a way that maximizes the spread of points, reducing the likelihood of poor convergence.

Question 2:
Given the following dataset of points in a 2D space:
Dataset = {(1, 1), (1, 2), (2, 2), (3, 3), (7, 7), (8, 7), (8, 8), (9, 8)}

a) Perform hierarchical clustering using the single-linkage method. Show the step-by-step process, including:
- The distance matrix at each step.
- The merging of clusters.
- The dendrogram representation.
(10 Marks)

b) Compare the single-linkage method with the complete-linkage method in hierarchical clustering. Discuss one advantage and one disadvantage of each method. (10 Marks)

Answer:
a)
Step-by-Step Process of Single-Linkage Clustering
We will use the single-linkage method, where the distance between two clusters is defined as the minimum distance between any two points in the clusters.
Compute the Initial Distance Matrix
We use the Euclidean distance formula:

$d((x1,y1),(x2,y2)) = \sqrt{(x2-x1)^2+(y2-y1)^2}$

	(1,1)	(1,2)	(2,2)	(3,3)	(7,7)	(8,7)	(8,8)	(9,8)
(1,1)	0	1.0	1.41	2.83	8.49	9.22	9.90	10.63
(1,2)	1.0	0	1.0	2.24	7.81	8.54	9.22	9.90

(2,2)	1.41	1.0	0	1.41	7.07	7.81	8.49	9.22
(3,3)	2.83	2.24	1.41	0	5.66	6.40	7.07	7.81
(7,7)	8.49	7.81	7.07	5.66	0	1.0	1.41	2.0
(8,7)	9.22	8.54	7.81	6.40	1.0	0	1.0	1.41
(8,8)	9.90	9.22	8.49	7.07	1.41	1.0	0	1.0
(9,8)	10.63	9.90	9.22	7.81	2.0	1.41	1.0	0

Merge the Closest Clusters
1.Merge (1,1) and (1,2) (distance = 1.0)
2.Merge (8,7) and (8,8) (distance = 1.0)
3.Merge (1,1,1,2) with (2,2) (distance = 1.0)
4.Merge (7,7) with (8,7,8,8) (distance = 1.0)
5.Merge (7,7,8,7,8,8) with (9,8) (distance = 1.0)
6.Merge (1,1,1,2,2,2) with (3,3) (distance = 1.41)
7.Merge (1,1,1,2,2,2,3,3) with (7,7,8,7,8,8,9,8) (distance = 5.66)

Draw the Dendrogram
A dendrogram can be drawn where:

- The x-axis represents data points.
- The y-axis represents the distance at which clusters are merged.

The final two clusters merge at distance 5.66, forming one large cluster.

b)
Comparison of Single-Linkage and Complete-Linkage Methods

Method	Advantage	Disadvantage
Single-Linkage.	Preserves the shape of clusters well, useful for. identifying elongated clusters.	Can suffer from chaining, where distant points are included in the same cluster
Complete-Linkage	Produces compact, spherical clusters, reducing chaining issues.	Can break up natural clusters if the maximum distance is too large

Thus, single-linkage works well for non-spherical clusters, while complete-linkage is better for evenly distributed clusters.