THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics
Exercises on Clustering Algorithms

1 Problem 1: K-Means Clustering

Problem Formulation

We apply the K-means clustering algorithm with the Euclidean distance metric to partition 8 two-dimensional data points into $K = 3$ disjoint clusters.

Dataset of 2D sample points:

$$A_1 = (2, 10), \quad A_2 = (2, 5), \quad A_3 = (8, 4), \quad A_4 = (5, 8),$$
$$A_5 = (7, 5), \quad A_6 = (6, 4), \quad A_7 = (1, 2), \quad A_8 = (4, 9)$$

Initial seed cluster centers (selected directly from dataset points):

$$\text{Seed}_1 = A_1 = (2, 10), \quad \text{Seed}_2 = A_4 = (5, 8), \quad \text{Seed}_3 = A_7 = (1, 2)$$

Initial cluster assignment before iteration:

- Cluster 1 initial center: A_1
- Cluster 2 initial center: A_4
- Cluster 3 initial center: A_7

Upper-triangular pairwise Euclidean distance matrix (corrected mathematical values):

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	$\sqrt{25}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A_2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A_3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A_4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A_5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A_6						0	$\sqrt{29}$	$\sqrt{29}$
A_7							0	$\sqrt{58}$
A_8								0

We first execute one full epoch of K-means and analyze four components after Epoch 1:

(a) Updated hard cluster memberships

(b) Recalculated new cluster centroid coordinates

(c) Visualization code for points and evolving centroids

(d) Formal convergence stopping criterion definition

We then continue iterating through Epoch 2 and Epoch 3 until cluster assignments stabilize (convergence).

Solution

Euclidean distance formula for two 2D points $a = (x_a, y_a)$ and $b = (x_b, y_b)$:

$$d(a, b) = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$$

Initial seed centroids fixed for Epoch 1:

$$\text{Seed}_1 = (2, 10), \quad \text{Seed}_2 = (5, 8), \quad \text{Seed}_3 = (1, 2)$$

Epoch 1: Full Distance Calculations & Cluster Assignment

Compute distance from each sample to all three initial seeds; assign each point to the cluster with minimal Euclidean distance.

Point $A_1 = (2, 10)$

$$\begin{aligned} d(A_1, \text{Seed}_1) &= \sqrt{(2-2)^2 + (10-10)^2} = \sqrt{0+0} = 0 \\ d(A_1, \text{Seed}_2) &= \sqrt{(5-2)^2 + (8-10)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61 \\ d(A_1, \text{Seed}_3) &= \sqrt{(1-2)^2 + (2-10)^2} = \sqrt{1+64} = \sqrt{65} \approx 8.06 \end{aligned}$$

Cluster Assignment: A_1 belongs to Cluster 1.

Point $A_2 = (2, 5)$

$$\begin{aligned} d(A_2, \text{Seed}_1) &= \sqrt{(2-2)^2 + (5-10)^2} = \sqrt{0+25} = \sqrt{25} = 5 \\ d(A_2, \text{Seed}_2) &= \sqrt{(5-2)^2 + (8-5)^2} = \sqrt{9+9} = \sqrt{18} \approx 4.24 \\ d(A_2, \text{Seed}_3) &= \sqrt{(1-2)^2 + (2-5)^2} = \sqrt{1+9} = \sqrt{10} \approx 3.16 \end{aligned}$$

Cluster Assignment: A_2 belongs to Cluster 3.

Point $A_3 = (8, 4)$

$$\begin{aligned} d(A_3, \text{Seed}_1) &= \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.49 \\ d(A_3, \text{Seed}_2) &= \sqrt{(5-8)^2 + (8-4)^2} = \sqrt{9+16} = \sqrt{25} = 5 \\ d(A_3, \text{Seed}_3) &= \sqrt{(1-8)^2 + (2-4)^2} = \sqrt{49+4} = \sqrt{53} \approx 7.28 \end{aligned}$$

Cluster Assignment: A_3 belongs to Cluster 2.

Point $A_4 = (5, 8)$

$$\begin{aligned} d(A_4, \text{Seed}_1) &= \sqrt{(2-5)^2 + (10-8)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61 \\ d(A_4, \text{Seed}_2) &= \sqrt{(5-5)^2 + (8-8)^2} = \sqrt{0+0} = 0 \\ d(A_4, \text{Seed}_3) &= \sqrt{(1-5)^2 + (2-8)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21 \end{aligned}$$

Cluster Assignment: A_4 belongs to Cluster 2.

Point $A_5 = (7, 5)$

$$d(A_5, \text{Seed}_1) = \sqrt{(2-7)^2 + (10-5)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$$
$$d(A_5, \text{Seed}_2) = \sqrt{(5-7)^2 + (8-5)^2} = \sqrt{4+9} = \sqrt{13} \approx 3.61$$
$$d(A_5, \text{Seed}_3) = \sqrt{(1-7)^2 + (2-5)^2} = \sqrt{36+9} = \sqrt{45} \approx 6.71$$

Cluster Assignment: A_5 belongs to Cluster 2.

Point $A_6 = (6, 4)$

$$d(A_6, \text{Seed}_1) = \sqrt{(2-6)^2 + (10-4)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$$
$$d(A_6, \text{Seed}_2) = \sqrt{(5-6)^2 + (8-4)^2} = \sqrt{1+16} = \sqrt{17} \approx 4.12$$
$$d(A_6, \text{Seed}_3) = \sqrt{(1-6)^2 + (2-4)^2} = \sqrt{25+4} = \sqrt{29} \approx 5.39$$

Cluster Assignment: A_6 belongs to Cluster 2.

Point $A_7 = (1, 2)$

$$d(A_7, \text{Seed}_1) = \sqrt{(2-1)^2 + (10-2)^2} = \sqrt{1+64} = \sqrt{65} \approx 8.06$$
$$d(A_7, \text{Seed}_2) = \sqrt{(5-1)^2 + (8-2)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$$
$$d(A_7, \text{Seed}_3) = \sqrt{(1-1)^2 + (2-2)^2} = \sqrt{0+0} = 0$$

Cluster Assignment: A_7 belongs to Cluster 3.

Point $A_8 = (4, 9)$

$$d(A_8, \text{Seed}_1) = \sqrt{(2-4)^2 + (10-9)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$$
$$d(A_8, \text{Seed}_2) = \sqrt{(5-4)^2 + (8-9)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$$
$$d(A_8, \text{Seed}_3) = \sqrt{(1-4)^2 + (2-9)^2} = \sqrt{9+49} = \sqrt{58} \approx 7.62$$

Cluster Assignment: A_8 belongs to Cluster 2.

(a) New Clusters After Epoch 1

$$\text{Cluster } 1 = \{A_1\}, \quad \text{Cluster } 2 = \{A_3, A_4, A_5, A_6, A_8\}, \quad \text{Cluster } 3 = \{A_2, A_7\}$$

(b) New Cluster Centroids After Epoch 1

Centroid formula for a cluster with m data points:

$$C = \left(\frac{1}{m} \sum_{i=1}^{m} x_i,\ \frac{1}{m} \sum_{i=1}^{m} y_i \right)$$

$C_1 = (2, 10)$ (single member A_1, no coordinate change)

$$C_2 = \left(\frac{8+5+7+6+4}{5},\ \frac{4+8+5+4+9}{5} \right) = \left(\frac{30}{5},\ \frac{30}{5} \right) = (6, 6)$$

$$C_3 = \left(\frac{2+1}{2},\ \frac{5+2}{2} \right) = (1.5,\ 3.5)$$

(c) Visualization Code

See Python codes.

(d) Convergence Criterion Definition

K-means converges when **all data points retain identical cluster memberships across two consecutive epochs**. If no sample switches clusters after recalculating centroids, subsequent iterations cannot alter cluster assignments or centroid coordinates, and the algorithm terminates. For this dataset, full convergence is reached after Epoch 3.

Epoch 2: Cluster Assignment and Centroid Recalculation

New seed centroids passed into Epoch 2 (output of Epoch 1):

$$\text{Seed}_1 = (2, 10), \quad \text{Seed}_2 = (6, 6), \quad \text{Seed}_3 = (1.5, 3.5)$$

Distance Calculations for All Points $A_1 = (2, 10)$

$$\begin{aligned} d(A_1, \text{Seed}_1) &= \sqrt{(2-2)^2 + (10-10)^2} = 0 \\ d(A_1, \text{Seed}_2) &= \sqrt{(6-2)^2 + (6-10)^2} = \sqrt{16+16} = \sqrt{32} \approx 5.66 \\ d(A_1, \text{Seed}_3) &= \sqrt{(1.5-2)^2 + (3.5-10)^2} = \sqrt{0.25+42.25} = \sqrt{42.5} \approx 6.52 \end{aligned}$$

Assignment: Cluster 1

$A_2 = (2, 5)$

$$\begin{aligned} d(A_2, \text{Seed}_1) &= \sqrt{(2-2)^2 + (5-10)^2} = \sqrt{25} = 5 \\ d(A_2, \text{Seed}_2) &= \sqrt{(6-2)^2 + (6-5)^2} = \sqrt{16+1} = \sqrt{17} \approx 4.12 \\ d(A_2, \text{Seed}_3) &= \sqrt{(1.5-2)^2 + (3.5-5)^2} = \sqrt{0.25+2.25} = \sqrt{2.5} \approx 1.58 \end{aligned}$$

Assignment: Cluster 3

$A_3 = (8, 4)$

$$\begin{aligned} d(A_3, \text{Seed}_1) &= \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.49 \\ d(A_3, \text{Seed}_2) &= \sqrt{(6-8)^2 + (6-4)^2} = \sqrt{4+4} = \sqrt{8} \approx 2.83 \\ d(A_3, \text{Seed}_3) &= \sqrt{(1.5-8)^2 + (3.5-4)^2} = \sqrt{42.25+0.25} = \sqrt{42.5} \approx 6.52 \end{aligned}$$

Assignment: Cluster 2

$A_4 = (5, 8)$

$$\begin{aligned} d(A_4, \text{Seed}_1) &= \sqrt{(2-5)^2 + (10-8)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61 \\ d(A_4, \text{Seed}_2) &= \sqrt{(6-5)^2 + (6-8)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24 \\ d(A_4, \text{Seed}_3) &= \sqrt{(1.5-5)^2 + (3.5-8)^2} = \sqrt{12.25+20.25} = \sqrt{32.5} \approx 5.70 \end{aligned}$$

Assignment: Cluster 2

$A_5 = (7, 5)$

$$d(A_5, \text{Seed}_1) = \sqrt{(2-7)^2 + (10-5)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$$
$$d(A_5, \text{Seed}_2) = \sqrt{(6-7)^2 + (6-5)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$$
$$d(A_5, \text{Seed}_3) = \sqrt{(1.5-7)^2 + (3.5-5)^2} = \sqrt{30.25+2.25} = \sqrt{32.5} \approx 5.70$$

Assignment: Cluster 2

$A_6 = (6, 4)$

$$d(A_6, \text{Seed}_1) = \sqrt{(2-6)^2 + (10-4)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$$
$$d(A_6, \text{Seed}_2) = \sqrt{(6-6)^2 + (6-4)^2} = \sqrt{0+4} = \sqrt{4} = 2$$
$$d(A_6, \text{Seed}_3) = \sqrt{(1.5-6)^2 + (3.5-4)^2} = \sqrt{20.25+0.25} = \sqrt{20.5} \approx 4.53$$

Assignment: Cluster 2

$A_7 = (1, 2)$

$$d(A_7, \text{Seed}_1) = \sqrt{(2-1)^2 + (10-2)^2} = \sqrt{1+64} = \sqrt{65} \approx 8.06$$
$$d(A_7, \text{Seed}_2) = \sqrt{(6-1)^2 + (6-2)^2} = \sqrt{25+16} = \sqrt{41} \approx 6.40$$
$$d(A_7, \text{Seed}_3) = \sqrt{(1.5-1)^2 + (3.5-2)^2} = \sqrt{0.25+2.25} = \sqrt{2.5} \approx 1.58$$

Assignment: Cluster 3

$A_8 = (4, 9)$

$$d(A_8, \text{Seed}_1) = \sqrt{(2-4)^2 + (10-9)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$$
$$d(A_8, \text{Seed}_2) = \sqrt{(6-4)^2 + (6-9)^2} = \sqrt{4+9} = \sqrt{13} \approx 3.61$$
$$d(A_8, \text{Seed}_3) = \sqrt{(1.5-4)^2 + (3.5-9)^2} = \sqrt{6.25+30.25} = \sqrt{36.5} \approx 6.04$$

Assignment: Cluster 1

Cluster Assignments After Epoch 2

$$\text{Cluster } 1 = \{A_1, A_8\}, \quad \text{Cluster } 2 = \{A_3, A_4, A_5, A_6\}, \quad \text{Cluster } 3 = \{A_2, A_7\}$$

Epoch 2 New Centroids

$$C_1 = \left(\frac{2+4}{2},\ \frac{10+9}{2}\right) = (3,\ 9.5)$$
$$C_2 = \left(\frac{8+5+7+6}{4},\ \frac{4+8+5+4}{4}\right) = \left(\frac{26}{4},\ \frac{21}{4}\right) = (6.5,\ 5.25)$$
$$C_3 = \left(\frac{2+1}{2},\ \frac{5+2}{2}\right) = (1.5,\ 3.5) \quad \text{(unchanged membership)}$$

Epoch 3: Cluster Assignment and Centroid Recalculation

Seeds for Epoch 3 (Epoch 2 output centroids):

$$\text{Seed}_1 = (3, 9.5), \quad \text{Seed}_2 = (6.5, 5.25), \quad \text{Seed}_3 = (1.5, 3.5)$$

Distance Calculations for All Points $A_1 = (2, 10)$

$$d(A_1, \text{Seed}_1) = \sqrt{(3-2)^2 + (9.5-10)^2} = \sqrt{1 + 0.25} = \sqrt{1.25} \approx 1.12$$
$$d(A_1, \text{Seed}_2) = \sqrt{(6.5-2)^2 + (5.25-10)^2} = \sqrt{20.25 + 22.56} = \sqrt{42.81} \approx 6.54$$
$$d(A_1, \text{Seed}_3) = \sqrt{(1.5-2)^2 + (3.5-10)^2} = \sqrt{0.25 + 42.25} = \sqrt{42.5} \approx 6.52$$

Assignment: Cluster 1

$A_2 = (2, 5)$

$$d(A_2, \text{Seed}_1) = \sqrt{(3-2)^2 + (9.5-5)^2} = \sqrt{1 + 20.25} = \sqrt{21.25} \approx 4.61$$
$$d(A_2, \text{Seed}_2) = \sqrt{(6.5-2)^2 + (5.25-5)^2} = \sqrt{20.25 + 0.0625} = \sqrt{20.3125} \approx 4.51$$
$$d(A_2, \text{Seed}_3) = \sqrt{(1.5-2)^2 + (3.5-5)^2} = \sqrt{0.25 + 2.25} = \sqrt{2.5} \approx 1.58$$

Assignment: Cluster 3

$A_3 = (8, 4)$

$$d(A_3, \text{Seed}_1) = \sqrt{(3-8)^2 + (9.5-4)^2} = \sqrt{25 + 30.25} = \sqrt{55.25} \approx 7.43$$
$$d(A_3, \text{Seed}_2) = \sqrt{(6.5-8)^2 + (5.25-4)^2} = \sqrt{2.25 + 1.5625} = \sqrt{3.8125} \approx 1.95$$
$$d(A_3, \text{Seed}_3) = \sqrt{(1.5-8)^2 + (3.5-4)^2} = \sqrt{42.25 + 0.25} = \sqrt{42.5} \approx 6.52$$

Assignment: Cluster 2

$A_4 = (5, 8)$

$$d(A_4, \text{Seed}_1) = \sqrt{(3-5)^2 + (9.5-8)^2} = \sqrt{4 + 2.25} = \sqrt{6.25} = 2.5$$
$$d(A_4, \text{Seed}_2) = \sqrt{(6.5-5)^2 + (5.25-8)^2} = \sqrt{2.25 + 7.5625} = \sqrt{9.8125} \approx 3.13$$
$$d(A_4, \text{Seed}_3) = \sqrt{(1.5-5)^2 + (3.5-8)^2} = \sqrt{12.25 + 20.25} = \sqrt{32.5} \approx 5.70$$

Assignment: Cluster 1

$A_5 = (7, 5)$

$$d(A_5, \text{Seed}_1) = \sqrt{(3-7)^2 + (9.5-5)^2} = \sqrt{16 + 20.25} = \sqrt{36.25} \approx 6.02$$
$$d(A_5, \text{Seed}_2) = \sqrt{(6.5-7)^2 + (5.25-5)^2} = \sqrt{0.25 + 0.0625} = \sqrt{0.3125} \approx 0.56$$
$$d(A_5, \text{Seed}_3) = \sqrt{(1.5-7)^2 + (3.5-5)^2} = \sqrt{30.25 + 2.25} = \sqrt{32.5} \approx 5.70$$

Assignment: Cluster 2

$A_6 = (6, 4)$

$$d(A_6, \text{Seed}_1) = \sqrt{(3-6)^2 + (9.5-4)^2} = \sqrt{9 + 30.25} = \sqrt{39.25} \approx 6.26$$
$$d(A_6, \text{Seed}_2) = \sqrt{(6.5-6)^2 + (5.25-4)^2} = \sqrt{0.25 + 1.5625} = \sqrt{1.8125} \approx 1.35$$
$$d(A_6, \text{Seed}_3) = \sqrt{(1.5-6)^2 + (3.5-4)^2} = \sqrt{20.25 + 0.25} = \sqrt{20.5} \approx 4.53$$

Assignment: Cluster 2

$A_7 = (1, 2)$

$$d(A_7, \text{Seed}_1) = \sqrt{(3-1)^2 + (9.5-2)^2} = \sqrt{4 + 56.25} = \sqrt{60.25} \approx 7.76$$

$$d(A_7, \text{Seed}_2) = \sqrt{(6.5-1)^2 + (5.25-2)^2} = \sqrt{30.25 + 10.5625} = \sqrt{40.8125} \approx 6.39$$
$$d(A_7, \text{Seed}_3) = \sqrt{(1.5-1)^2 + (3.5-2)^2} = \sqrt{0.25 + 2.25} = \sqrt{2.5} \approx 1.58$$

Assignment: Cluster 3

$A_8 = (4, 9)$

$$d(A_8, \text{Seed}_1) = \sqrt{(3-4)^2 + (9.5-9)^2} = \sqrt{1 + 0.25} = \sqrt{1.25} \approx 1.12$$
$$d(A_8, \text{Seed}_2) = \sqrt{(6.5-4)^2 + (5.25-9)^2} = \sqrt{6.25 + 14.0625} = \sqrt{20.3125} \approx 4.51$$
$$d(A_8, \text{Seed}_3) = \sqrt{(1.5-4)^2 + (3.5-9)^2} = \sqrt{6.25 + 30.25} = \sqrt{36.5} \approx 6.04$$

Assignment: Cluster 1

Cluster Assignments After Epoch 3

$$\text{Cluster 1} = \{A_1, A_4, A_8\}, \quad \text{Cluster 2} = \{A_3, A_5, A_6\}, \quad \text{Cluster 3} = \{A_2, A_7\}$$

Epoch 3 New Centroids

$$C_1 = \left(\frac{2+5+4}{3},\ \frac{10+8+9}{3}\right) = \left(\frac{11}{3},\ \frac{27}{3}\right) \approx (3.67,\ 9)$$
$$C_2 = \left(\frac{8+7+6}{3},\ \frac{4+5+4}{3}\right) = \left(\frac{21}{3},\ \frac{13}{3}\right) = (7,\ 4.33)$$
$$C_3 = (1.5,\ 3.5) \quad \text{(no change)}$$

Final Convergence Result

After completing Epoch 3, we re-run the assignment step with the Epoch 3 centroids and find **no points switch cluster memberships**. The cluster assignments stabilize fully, so the K-means algorithm converges at Epoch 3 with the final cluster groups listed above.

2 Problem 2: Nearest Neighbor Clustering with Distance Threshold

Problem Formulation

We implement a sequential nearest neighbor clustering algorithm using Euclidean distance. Each incoming data point joins an existing cluster if its minimal Euclidean distance to any previously clustered point is strictly smaller than a fixed threshold t. If the minimal distance is greater than or equal to t, a new singleton cluster is created for the point.

Dataset

The dataset contains 8 two-dimensional sample points:

$$A_1 = (2, 10), \quad A_2 = (2, 5), \quad A_3 = (8, 4), \quad A_4 = (5, 8),$$
$$A_5 = (7, 5), \quad A_6 = (6, 4), \quad A_7 = (1, 2), \quad A_8 = (4, 9)$$

Distance Metric Definition

For two arbitrary points $A_i = (x_i, y_i)$ and $A_j = (x_j, y_j)$, Euclidean distance is:

$$d(A_i, A_j) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Clustering Hyperparameter Threshold

$$t = 4$$

Formal Clustering Procedure Rules

1. Initialize the first cluster K_1 containing only the first point A_1.
2. Iterate over each subsequent point A_i in the order $A_2, A_3, \ldots, A_8$:
    a. Compute Euclidean distance from A_i to every point already assigned to any cluster.
    b. Record the minimal distance $d_{\min}$ and the cluster containing the nearest existing point.
    c. If $d_{\min} < t$: assign A_i to the cluster of its nearest neighbor.
    d. If $d_{\min} \geq t$: construct a new singleton cluster $K_{\text{new}} = \{A_i\}$.
3. Terminate after processing all 8 data points and output all final clusters $\{K_1, K_2, \ldots, K_k\}$.

Full Step-by-Step Clustering Calculations

We process points sequentially and compute all required distance values explicitly at each iteration.

Step 1: Initialize with first point A_1 No prior clustered points exist; create initial cluster.

$$K_1 = \{A_1\}$$

Step 2: Process second point $A_2 = (2, 5)$ Only one clustered point exists: A_1. Compute pairwise distance:

$$\begin{aligned} d(A_1, A_2) &= \sqrt{(2-2)^2 + (10-5)^2} \\ &= \sqrt{0^2 + 5^2} \\ &= \sqrt{25} = 5 \end{aligned}$$

Comparison: $d_{\min} = 5 \geq t = 4$. A new cluster must be created.

$$K_2 = \{A_2\}$$

Current cluster set: $\{K_1 = \{A_1\},\ K_2 = \{A_2\}\}$

Step 3: Process third point $A_3 = (8, 4)$ Existing clustered points: A_1, A_2. Calculate both distances:

$$d(A_3, A_1) = \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{6^2 + (-6)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.49$$
$$d(A_3, A_2) = \sqrt{(8-2)^2 + (4-5)^2} = \sqrt{6^2 + (-1)^2} = \sqrt{36+1} = \sqrt{37} \approx 6.08$$

Minimum distance: $d_{\min} \approx 6.08 \geq t = 4$. Create new cluster.

$$K_3 = \{A_3\}$$

Current cluster set: $\{K_1 = \{A_1\},\ K_2 = \{A_2\},\ K_3 = \{A_3\}\}$

Step 4: Process fourth point $A_4 = (5, 8)$ Existing clustered points: A_1, A_2, A_3. Compute all three distances:

$$d(A_4, A_1) = \sqrt{(5-2)^2 + (8-10)^2} = \sqrt{3^2 + (-2)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61$$
$$d(A_4, A_2) = \sqrt{(5-2)^2 + (8-5)^2} = \sqrt{3^2 + 3^2} = \sqrt{9+9} = \sqrt{18} \approx 4.24$$
$$d(A_4, A_3) = \sqrt{(5-8)^2 + (8-4)^2} = \sqrt{(-3)^2 + 4^2} = \sqrt{9+16} = \sqrt{25} = 5$$

Minimum distance: $d_{\min} \approx 3.61 < t = 4$. Nearest neighbor is $A_1 \in K_1$; add A_4 to K_1.

$$K_1 = \{A_1, A_4\}$$

Current cluster set: $\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3\}\}$

Step 5: Process fifth point $A_5 = (7, 5)$ Existing clustered points: A_1, A_2, A_3, A_4. We only need the minimal distance value; compute distance to A_3 (the closest candidate):

$$d(A_5, A_3) = \sqrt{(7-8)^2 + (5-4)^2} = \sqrt{(-1)^2 + 1^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$$

All other distances to A_1, A_2, A_4 are larger than 1.41, so $d_{\min} \approx 1.41 < t = 4$. Nearest neighbor is $A_3 \in K_3$; add A_5 to K_3.

$$K_3 = \{A_3, A_5\}$$

Current cluster set: $\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3, A_5\}\}$

Step 6: Process sixth point $A_6 = (6, 4)$ Existing clustered points: A_1, A_2, A_3, A_4, A_5. Compute distances to cluster K_3 members (closest candidates):

$$d(A_6, A_3) = \sqrt{(6-8)^2 + (4-4)^2} = \sqrt{(-2)^2 + 0^2} = \sqrt{4} = 2$$
$$d(A_6, A_5) = \sqrt{(6-7)^2 + (4-5)^2} = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} \approx 1.41$$

The global minimal distance is $d_{\min} \approx 1.41 < t = 4$. Nearest neighbor is $A_5 \in K_3$; add A_6 to K_3.

$$K_3 = \{A_3, A_5, A_6\}$$

Current cluster set: $\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3, A_5, A_6\}\}$

Step 7: Process seventh point $A_7 = (1, 2)$ Existing clustered points: $A_1, A_2, A_3, A_4, A_5, A_6$. The closest existing point is $A_2 \in K_2$:

$$d(A_7, A_2) = \sqrt{(1-2)^2 + (2-5)^2} = \sqrt{(-1)^2 + (-3)^2} = \sqrt{1+9} = \sqrt{10} \approx 3.16$$

$d_{\min} \approx 3.16 < t = 4$. Add A_7 to cluster K_2.

$$K_2 = \{A_2, A_7\}$$

Current cluster set: $\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2, A_7\},\ K_3 = \{A_3, A_5, A_6\}\}$

Step 8: Process eighth point $A_8 = (4, 9)$ Existing clustered points: $A_1, A_2, A_3, A_4, A_5, A_6, A_7$. The closest existing point is $A_4 \in K_1$:

$$d(A_8, A_4) = \sqrt{(4-5)^2 + (9-8)^2} = \sqrt{(-1)^2 + 1^2} = \sqrt{2} \approx 1.41$$

$d_{\min} \approx 1.41 < t = 4$. Add A_8 to cluster K_1.

$$K_1 = \{A_1, A_4, A_8\}$$

Final Clustering Result

After processing all 8 data points, the three final disjoint clusters are:

$$K_1 = \{A_1,\ A_4,\ A_8\}$$
$$K_2 = \{A_2,\ A_7\}$$
$$K_3 = \{A_3,\ A_5,\ A_6\}$$

Conclusion

The sequential nearest neighbor clustering algorithm with Euclidean distance threshold $t = 4$ produces identical cluster partitions to the converged K-means result from Problem 1 on this identical dataset.

3 Problem 3: K-Medoids Clustering Algorithm (PAM: Partitioning Around Medoids)

Core Concept of K-Medoids

K-Medoids is an unsupervised partitioning clustering algorithm. Unlike K-Means, which uses synthetic cluster centroids, K-Medoids selects *medoids*—actual existing data points—as cluster representatives. The objective is to minimize total intra-cluster distance (called clustering cost). The standard implementation of K-Medoids is known as PAM (Partitioning Around Medoids).

Definitions

- **Medoid**: A representative point belonging to the original dataset that minimizes the sum of distances to all other points in its cluster; the most central data point of a cluster.
- **Clustering Cost**: Total sum of distances from every non-medoid data point to its assigned cluster medoid (sum of intra-cluster distances).

PAM Algorithm Step-by-Step Procedure

1. Randomly select K distinct raw data points as the initial set of medoids.
2. For every data point, compute distance to all K medoids (supported metrics: Manhattan, Euclidean, etc.).
3. Assign each point to the cluster corresponding to its nearest medoid.
4. Calculate total clustering cost C_{curr} by summing all intra-cluster distances.
5. Iterate over all possible swaps: pick one non-medoid point D_j and exchange it with one current medoid M_i to form a temporary new medoid set.
6. Recalculate the clustering cost C_{new} for the swapped medoid configuration.
7. If $C_{\text{new}} < C_{\text{curr}}$, permanently accept the swap and update medoids; otherwise revert to the original medoid set.
8. Repeat steps 4–7 until no valid swap reduces total cost (cluster configuration stabilizes, convergence reached).

Numerical K-Medoids Example ($K = 2$, Manhattan Distance)

Dataset of 10 2D Points

Point Label	Coordinates (x, y)
A_1	$(2, 6)$
A_2	$(3, 8)$
A_3	$(4, 7)$
A_4	$(6, 2)$
A_5	$(6, 4)$
A_6	$(7, 3)$
A_7	$(7, 4)$
A_8	$(8, 5)$
A_9	$(7, 6)$
A_{10}	$(3, 4)$

Distance metric used: Manhattan distance between point $P = (x_p, y_p)$ and $Q = (x_q, y_q)$:

$$d(P, Q) = |x_q - x_p| + |y_q - y_p|$$

Target number of clusters: $K = 2$. Initial medoids selected for Iteration 1: $M_1 = A_{10} = (3, 4)$, $M_2 = A_6 = (7, 3)$.

Iteration 1: Initial Medoid Set $M_1 = (3, 4),\ M_2 = (7, 3)$

3.0.1 Full Distance & Cluster Assignment Table

Point	Coordinates	$d(\cdot, M_1)$	$d(\cdot, M_2)$	Assigned Cluster
A_1	$(2, 6)$	$\|2-3\| + \|6-4\| = 1 + 2 = 3$	$\|2-7\| + \|6-3\| = 5 + 3 = 8$	1
A_2	$(3, 8)$	$\|3-3\| + \|8-4\| = 0 + 4 = 4$	$\|3-7\| + \|8-3\| = 4 + 5 = 9$	1
A_3	$(4, 7)$	$\|4-3\| + \|7-4\| = 1 + 3 = 4$	$\|4-7\| + \|7-3\| = 3 + 4 = 7$	1
A_4	$(6, 2)$	$\|6-3\| + \|2-4\| = 3 + 2 = 5$	$\|6-7\| + \|2-3\| = 1 + 1 = 2$	2
A_5	$(6, 4)$	$\|6-3\| + \|4-4\| = 3 + 0 = 3$	$\|6-7\| + \|4-3\| = 1 + 1 = 2$	2
A_6	$(7, 3)$	$\|7-3\| + \|3-4\| = 4 + 1 = 5$	$\|7-7\| + \|3-3\| = 0 + 0 = 0$	2
A_7	$(7, 4)$	$\|7-3\| + \|4-4\| = 4 + 0 = 4$	$\|7-7\| + \|4-3\| = 0 + 1 = 1$	2
A_8	$(8, 5)$	$\|8-3\| + \|5-4\| = 5 + 1 = 6$	$\|8-7\| + \|5-3\| = 1 + 2 = 3$	2
A_9	$(7, 6)$	$\|7-3\| + \|6-4\| = 4 + 2 = 6$	$\|7-7\| + \|6-3\| = 0 + 3 = 3$	2
A_{10}	$(3, 4)$	$\|3-3\| + \|4-4\| = 0 + 0 = 0$	$\|3-7\| + \|4-3\| = 4 + 1 = 5$	1

Cluster Membership for Iteration 1

- Cluster 1 (Medoid $M_1 = (3, 4)$): $\{A_1, A_2, A_3, A_{10}\}$
- Cluster 2 (Medoid $M_2 = (7, 3)$): $\{A_4, A_5, A_6, A_7, A_8, A_9\}$

3.0.2 Cost Calculation

$$\begin{aligned}
\text{Cost}_1 &= d(A_1, M_1) + d(A_2, M_1) + d(A_3, M_1) + d(A_{10}, M_1) \\
&= 3 + 4 + 4 + 0 = 11 \\
\text{Cost}_2 &= d(A_4, M_2) + d(A_5, M_2) + d(A_6, M_2) + d(A_7, M_2) + d(A_8, M_2) + d(A_9, M_2) \\
&= 2 + 2 + 0 + 1 + 3 + 3 = 11
\end{aligned}$$

Total $\text{Cost}_{\text{Iter1}} = \text{Cost}_1 + \text{Cost}_2 = 11 + 11 = \boxed{22}$

Iteration 2: Temporary Swap Candidate $M_2 \leftarrow A_7 = (7, 4)$

Trial medoid set: $M_1 = (3, 4),\ M_2 = (7, 4)$

3.0.3 Distance & Assignment Table

Point	Coordinates	$d(\cdot, M_1)$	$d(\cdot, M_2)$	Assigned Cluster
A_1	$(2, 6)$	$1 + 2 = 3$	$5 + 2 = 7$	1
A_2	$(3, 8)$	$0 + 4 = 4$	$4 + 4 = 8$	1
A_3	$(4, 7)$	$1 + 3 = 4$	$3 + 3 = 6$	1
A_4	$(6, 2)$	$3 + 2 = 5$	$1 + 2 = 3$	2
A_5	$(6, 4)$	$3 + 0 = 3$	$1 + 0 = 1$	2
A_6	$(7, 3)$	$4 + 1 = 5$	$0 + 1 = 1$	2
A_7	$(7, 4)$	$4 + 0 = 4$	$0 + 0 = 0$	2
A_8	$(8, 5)$	$5 + 1 = 6$	$1 + 1 = 2$	2
A_9	$(7, 6)$	$4 + 2 = 6$	$0 + 2 = 2$	2
A_{10}	$(3, 4)$	$0 + 0 = 0$	$4 + 0 = 4$	1

Cluster Membership for Iteration 2 Trial

- Cluster 1: $\{A_1, A_2, A_3, A_{10}\}$
- Cluster 2: $\{A_4, A_5, A_6, A_7, A_8, A_9\}$

Cost Calculation

$$\begin{aligned} \text{Cost}_1 &= 3 + 4 + 4 + 0 = 11 \\ \text{Cost}_2 &= 3 + 1 + 1 + 0 + 2 + 2 = 9 \\ \text{Total Cost}_{\text{Iter2}} &= 11 + 9 = \boxed{20} \end{aligned}$$

Swap Decision

$20 < 22$, so this swap reduces total clustering cost. The new permanent medoid set is retained:

$$M_1 = (3, 4), \quad M_2 = (7, 4)$$

Iteration 3: Temporary Swap Candidate $M_2 \leftarrow A_5 = (6, 4)$

Trial medoid set: $M_1 = (3, 4),\ M_2 = (6, 4)$

3.0.4 Distance & Assignment Table

Point	Coordinates	$d(\cdot, M_1)$	$d(\cdot, M_2)$	Assigned Cluster
A_1	$(2,6)$	$1+2=3$	$4+2=6$	1
A_2	$(3,8)$	$0+4=4$	$3+4=7$	1
A_3	$(4,7)$	$1+3=4$	$2+3=5$	1
A_4	$(6,2)$	$3+2=5$	$0+2=2$	2
A_5	$(6,4)$	$3+0=3$	$0+0=0$	2
A_6	$(7,3)$	$4+1=5$	$1+1=2$	2
A_7	$(7,4)$	$4+0=4$	$1+0=1$	2
A_8	$(8,5)$	$5+1=6$	$2+1=3$	2
A_9	$(7,6)$	$4+2=6$	$1+2=3$	2
A_{10}	$(3,4)$	$0+0=0$	$3+0=3$	1

Cluster Membership for Iteration 3 Trial

- Cluster 1: $\{A_1, A_2, A_3, A_{10}\}$
- Cluster 2: $\{A_4, A_5, A_6, A_7, A_8, A_9\}$

Cost Calculation

$$\begin{aligned} \text{Cost}_1 &= 3+4+4+0 = 11 \\ \text{Cost}_2 &= 2+0+2+1+3+3 = 11 \\ \text{Total Cost}_{\text{Iter3}} &= 11+11 = \boxed{22} \end{aligned}$$

Swap Decision

$22 > 20$, this swap increases total cost, so the swap is rejected. The previous optimal medoids are preserved:

$$M_1 = (3,4), \quad M_2 = (7,4)$$

3.1 Final Converged Clustering Result

3.1.1 Optimal Medoids after Convergence

$$M_1 = (3,4) \quad (A_{10}), \qquad M_2 = (7,4) \quad (A_7)$$

Final Cluster Partitions

- Cluster 1 (Medoid A_{10}): $\{A_1(2,6),\ A_2(3,8),\ A_3(4,7),\ A_{10}(3,4)\}$
- Cluster 2 (Medoid A_7): $\{A_4(6,2),\ A_5(6,4),\ A_6(7,3),\ A_7(7,4),\ A_8(8,5),\ A_9(7,6)\}$

Algorithm Remarks

1. Medoids are always actual data points, unlike synthetic K-Means centroids; robust to outliers.

2. PAM iteratively tests all possible medoid swaps and only accepts configurations with strictly lower total intra-cluster cost.
3. Each iteration requires full pairwise distance computation and cluster reassignment, leading to high computational complexity; not suitable for extremely large datasets.
4. Convergence condition: no swap between medoid and non-medoid reduces total clustering cost.