**香港中文大學**
**數學系**
**聚類演算法練習**

# 1 問題1：K-平均聚類(K-Means Clustering)

### 問題設定

我們應用採用歐幾里得距離(Euclidean distance) 測度的K-平均聚類演算法，將8 個二維數據點劃分為$K = 3$ 個不相交的聚類。

2D 樣本點數據集：

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

初始種子聚類中心（直接從數據集點中選擇）：

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

迭代前的初始聚類分配：

- 聚類1 初始中心：$A_1$
- 聚類2 初始中心：$A_4$
- 聚類3 初始中心：$A_7$

上三角成對歐幾里得距離矩陣（已修正的數學數值）：

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ | $A_8$ |
|---|---|---|---|---|---|---|---|---|
| $A_1$ | 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 |

我們首先執行一個完整的K-平均週期(epoch)，並在週期1 之後分析四個組成部分：

(a) 更新後的硬聚類(hard cluster) 成員資格

(b) 重新計算的新聚類形心(centroid) 座標

(c) 數據點與演變中形心的視覺化程式碼

(d) 正式的收斂停止準則定義

然後我們繼續迭代週期2 和週期3，直到聚類分配穩定下來（收斂）。

**解答**

兩個2D 點$a = (x_a, y_a)$ 和$b = (x_b, y_b)$ 的歐幾里得距離公式：

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

週期1 固定的初始種子形心：

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

**週期1：完整的距離計算與聚類分配**

計算每個樣本到所有三個初始種子的距離；將每個點分配到歐幾里得距離最小的聚類中。

**點**$A_1 = (2, 10)$

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

**聚類分配**：$A_1$ 屬於聚類1。

**點**$A_2 = (2, 5)$

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

**聚類分配**：$A_2$ 屬於聚類3。

**點**$A_3 = (8, 4)$

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

**聚類分配**：$A_3$ 屬於聚類2。

**點**$A_4 = (5, 8)$

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

**聚類分配**：$A_4$ 屬於聚類2。

**點**$A_5 = (7, 5)$

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

**聚類分配：**$A_5$ 屬於聚類2。

**點**$A_6 = (6, 4)$

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

**聚類分配：**$A_6$ 屬於聚類2。

**點**$A_7 = (1, 2)$

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

**聚類分配：**$A_7$ 屬於聚類3。

**點**$A_8 = (4, 9)$

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

**聚類分配：**$A_8$ 屬於聚類2。

**(a) 週期1 後的新聚類**

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

**(b) 週期1 後的新聚類形心**

包含$m$ 個數據點的聚類形心公式：

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

$C_1 = (2, 10)$ （單一成員$A_1$，座標無變化）

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

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

**(c) 視覺化程式碼**

請參閱Python 程式碼。

**(d) 收斂準則定義**

當**所有數據點在連續兩個週期中保持完全相同的聚類成員資格**時，K-平均演算法即達到收斂。如果在重新計算形心後沒有樣本切換聚類，後續的迭代就不會改變聚類分配或形心座標，演算法便會終止。對於此數據集，在週期3 之後達到了完全收斂。

**週期2：聚類分配與形心重新計算**

傳入週期2 的新種子形心（週期1 的輸出）：

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

**所有點的距離計算** $A_1 = (2, 10)$

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

**分配**：聚類1

$A_2 = (2, 5)$

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

**分配**：聚類3

$A_3 = (8, 4)$

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

**分配**：聚類2

$A_4 = (5, 8)$

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

**分配**：聚類2

$A_5 = (7, 5)$

$$d(A_5, \text{Seed}_1) = \sqrt{(2-7)^2 + (10-5)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$$

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

**分配**：聚類2

$A_6 = (6, 4)$

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

**分配**：聚類2

$A_7 = (1, 2)$

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

**分配**：聚類3

$A_8 = (4, 9)$

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

**分配**：聚類1

**週期2 後的聚類分配**

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

**週期2 新形心**

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

**週期3：聚類分配與形心重新計算**

週期3 的種子（週期2 輸出的形心）：

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

**所有點的距離計算** $A_1 = (2, 10)$

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

**分配**：聚類1

$A_2 = (2, 5)$

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

**分配**：聚類3

$A_3 = (8, 4)$

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

**分配**：聚類2

$A_4 = (5, 8)$

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

**分配**：聚類1

$A_5 = (7, 5)$

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

**分配**：聚類2

$A_6 = (6, 4)$

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

**分配**：聚類2

$A_7 = (1, 2)$

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

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

**分配**：聚類3

$A_8 = (4, 9)$

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

**分配**：聚類1

**週期3 後的聚類分配**

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

**週期3 新形心**

$$C_1 = \left(\frac{2+5+4}{3},\ \frac{10+8+9}{3}\right) = \left(\frac{11}{3},\ \frac{27}{3}\right) \approx (3.67,\ 9)$$
$$C_2 = \left(\frac{8+7+6}{3},\ \frac{4+5+4}{3}\right) = \left(\frac{21}{3},\ \frac{13}{3}\right) = (7,\ 4.33)$$
$$C_3 = (1.5,\ 3.5) \quad \text{（無變化）}$$

**最終收斂結果**

在完成週期3 後，我們使用週期3 的形心重新執行分配步驟，並發現**沒有任何點切換聚類成員資格**。聚類分配已完全穩定，因此K-平均演算法在週期3 達到收斂，最終的聚類分組如上所列。

# 2 問題2：具有距離閾值的最近鄰聚類(Nearest Neighbor Clustering)

### 問題設定

我們實作一個使用歐幾里得距離的順序最近鄰聚類演算法。對於每個新輸入的數據點，如果它與任何先前已聚類的點之間的最小歐幾里得距離嚴格小於固定的閾值$t$，則該點會加入現有的聚類。如果最小距離大於或等於$t$，則為該點建立一個新的單例聚類(singleton cluster)。

#### 數據集

該數據集包含8 個二維樣本點：

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

#### 距離測度定義

對於任意兩點$A_i = (x_i, y_i)$ 和$A_j = (x_j, y_j)$，歐幾里得距離為：

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

#### 聚類超參數閾值

$$t = 4$$

#### 正式的聚類程序規則

1. 初始化第一個聚類$K_1$，僅包含第一個點$A_1$。
2. 按照$A_2, A_3, \ldots, A_8$ 的順序迭代每個後續的點$A_i$：
   a. 計算從$A_i$ 到每個已分配至任何聚類的點的歐幾里得距離。
   b. 記錄最小距離$d_{\min}$ 以及包含最近現有點的聚類。
   c. 如果$d_{\min} < t$：將$A_i$ 分配至其最近鄰居所在的聚類。
   d. 如果$d_{\min} \geq t$：構建一個新的單例聚類$K_{\text{new}} = \{A_i\}$。
3. 處理完所有8 個數據點後終止，並輸出所有最終聚類$\{K_1, K_2, \ldots, K_k\}$。

### 完整的逐步聚類計算

我們依序處理各個點，並在每次迭代中明確計算所有所需的距離數值。

**步驟1：使用第一個點$A_1$ 進行初始化** 不存在先前已聚類的點；建立初始聚類。

$$K_1 = \{A_1\}$$

**步驟2：處理第二個點**$A_2 = (2, 5)$　**僅**存在一個已聚類的點：$A_1$。計算成對距離：

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

比較：$d_{\min} = 5 \geq t = 4$。必須建立新的聚類。

$$K_2 = \{A_2\}$$

目前聚類集合：$\{K_1 = \{A_1\},\ K_2 = \{A_2\}\}$

**步驟3：處理第三個點**$A_3 = (8, 4)$　**現**有已聚類的點：$A_1, A_2$。計算兩個距離：

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

最小距離：$d_{\min} \approx 6.08 \geq t = 4$。建立新聚類。

$$K_3 = \{A_3\}$$

目前聚類集合：$\{K_1 = \{A_1\},\ K_2 = \{A_2\},\ K_3 = \{A_3\}\}$

**步驟4：處理第四個點**$A_4 = (5, 8)$　**現**有已聚類的點：$A_1, A_2, A_3$。計算所有三個距離：

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

最小距離：$d_{\min} \approx 3.61 < t = 4$。最近的鄰居是$A_1 \in K_1$；將$A_4$ 加入$K_1$。

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

目前聚類集合：$\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3\}\}$

**步驟5：處理第五個點**$A_5 = (7, 5)$　**現**有已聚類的點：$A_1, A_2, A_3, A_4$。我們只需要最小距離值；計算到$A_3$（最接近的候選點）的距離：

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

到$A_1, A_2, A_4$ 的所有其他距離都大於1.41，因此$d_{\min} \approx 1.41 < t = 4$。最近的鄰居是$A_3 \in K_3$；將$A_5$ 加入$K_3$。

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

目前聚類集合：$\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3, A_5\}\}$

**步驟6：處理第六個點**$A_6 = (6, 4)$ **現**有已聚類的點：$A_1, A_2, A_3, A_4, A_5$。計算到聚類$K_3$成員（最接近的候選點）的距離：

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

全域最小距離為$d_{\min} \approx 1.41 < t = 4$。最近的鄰居是$A_5 \in K_3$；將$A_6$ 加入$K_3$。

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

目前聚類集合：$\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2\},\ K_3 = \{A_3, A_5, A_6\}\}$

**步驟7：處理第七個點**$A_7 = (1, 2)$ **現**有已聚類的點：$A_1, A_2, A_3, A_4, A_5, A_6$。最接近的現有點是$A_2 \in K_2$：

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

$d_{\min} \approx 3.16 < t = 4$。將$A_7$ 加入聚類$K_2$。

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

目前聚類集合：$\{K_1 = \{A_1, A_4\},\ K_2 = \{A_2, A_7\},\ K_3 = \{A_3, A_5, A_6\}\}$

**步驟8：處理第八個點**$A_8 = (4, 9)$ **現**有已聚類的點：$A_1, A_2, A_3, A_4, A_5, A_6, A_7$。最接近的現有點是$A_4 \in K_1$：

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

$d_{\min} \approx 1.41 < t = 4$。將$A_8$ 加入聚類$K_1$。

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

**最終聚類結果**

在處理完所有8 個數據點後，最終的三個不相交聚類為：

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

**結論**

在這個相同的數據集上，具有歐幾里得距離閾值$t = 4$ 的順序最近鄰聚類演算法，產生了與問題1 中收斂後的K-平均結果完全相同的聚類劃分。

# 3 問題3：K-中心點聚類演算法(K-Medoids Clustering Algorithm) (PAM: 圍繞中心點劃分)

### K-中心點的核心概念

K-中心點是一種無監督的劃分聚類演算法。與使用合成聚類形心的K-平均不同，K-中心點選擇中心點*(medoids)*（實際現有的數據點）作為聚類的代表。其目標是最小化總簇內距離（稱為聚類成本）。K-中心點的標準實作稱為PAM (Partitioning Around Medoids，圍繞中心點劃分)。

### 定義

- **中心點(Medoid)**：屬於原始數據集的代表點，它能使該聚類中所有其他點到它的距離總和最小化；即聚類中最中心的數據點。
- **聚類成本(Clustering Cost)**：每個非中心數據點到其被分配的聚類中心點的距離總和（即簇內距離總和）。

### PAM 演算法逐步程序

1. 隨機選擇$K$ 個不同的原始數據點作為初始的中心點集合。
2. 計算每個數據點到所有$K$ 個中心點的距離（支援的測度：曼哈頓距離(Manhattan)、歐幾里得距離等）。
3. 將每個點分配至對應於其最近中心點的聚類。
4. 透過加總所有簇內距離來計算總聚類成本$C_{\text{curr}}$。
5. 迭代所有可能的交換：挑選一個非中心點$D_j$ 並與一個當前的中心點$M_i$ 交換，形成一個暫時的新中心點集合。
6. 為交換後的中心點配置重新計算聚類成本$C_{\text{new}}$。
7. 如果$C_{\text{new}} < C_{\text{curr}}$，永久接受該交換並更新中心點；否則還原為原始的中心點集合。
8. 重複步驟4–7，直到沒有任何有效的交換能降低總成本（聚類配置穩定，達到收斂）。

**K-中心點數值例子（$K = 2$，曼哈頓距離）**

**包含10 個2D 點的數據集**

| 點標籤 | 座標$(x, y)$ |
|---|---|
| $A_1$ | $(2, 6)$ |
| $A_2$ | $(3, 8)$ |
| $A_3$ | $(4, 7)$ |
| $A_4$ | $(6, 2)$ |
| $A_5$ | $(6, 4)$ |
| $A_6$ | $(7, 3)$ |
| $A_7$ | $(7, 4)$ |
| $A_8$ | $(8, 5)$ |
| $A_9$ | $(7, 6)$ |
| $A_{10}$ | $(3, 4)$ |

使用的距離測度：點$P = (x_p, y_p)$ 和$Q = (x_q, y_q)$ 之間的曼哈頓距離：

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

目標聚類數量：$K = 2$。迭代1 選擇的初始中心點為：$M_1 = A_{10} = (3, 4)$，$M_2 = A_6 = (7, 3)$。

**迭代1：初始中心點集合**$M_1 = (3, 4),\ M_2 = (7, 3)$

### 3.0.1 完整距離與聚類分配表

| 點 | 座標 | $d(\cdot, M_1)$ | $d(\cdot, M_2)$ | 分配的聚類 |
|---|---|---|---|---|
| $A_1$ | $(2, 6)$ | $\|2-3\| + \|6-4\| = 1 + 2 = 3$ | $\|2-7\| + \|6-3\| = 5 + 3 = 8$ | 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 |

**迭代1 的聚類成員資格**

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

### 3.0.2 成本計算

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

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

**迭代2：暫時交換候選點**$M_2 \leftarrow A_7 = (7,4)$

試驗中心點集合：$M_1 = (3,4),\ M_2 = (7,4)$

#### 3.0.3 距離與分配表

| 點 | 座標 | $d(\cdot, M_1)$ | $d(\cdot, M_2)$ | 分配的聚類 |
|---|---|---|---|---|
| $A_1$ | $(2,6)$ | $1+2=3$ | $5+2=7$ | 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 |

**迭代2 試驗的聚類成員資格**

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

**成本計算**

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

**交換決定**

$20 < 22$，因此這次交換降低了總聚類成本。保留這個新的永久中心點集合：

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

**迭代3：暫時交換候選點**$M_2 \leftarrow A_5 = (6,4)$

試驗中心點集合：$M_1 = (3,4),\ M_2 = (6,4)$

#### 3.0.4 距離與分配表

| 點 | 座標 | $d(\cdot, M_1)$ | $d(\cdot, M_2)$ | 分配的聚類 |
|---|---|---|---|---|
| $A_1$ | $(2,6)$ | $1+2=3$ | $4+2=6$ | 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 |

**迭代3 試驗的聚類成員資格**

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

**成本計算**

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

**交換決定**

$22 > 20$，這次交換增加了總成本，因此拒絕該交換。保留先前的最佳中心點：

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

### 3.1 最終收斂的聚類結果

#### 3.1.1 收斂後的最佳中心點

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

**最終聚類劃分**

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

**演算法備註**

1. 與合成的K-平均形心不同，中心點始終是實際的數據點；這對異常值(outliers) 較為穩健(robust)。
2. PAM 反覆測試所有可能的中心點交換，並僅接受具有嚴格較低總簇內成本的配置。

3. 每次迭代都需要完整的成對距離計算和聚類重新分配，這導致計算複雜度很高；不適合用於極大的數據集。
4. 收斂條件：中心點與非中心點之間的任何交換都無法降低總聚類成本。