Trần Chí Vỹ - 3123580065

Câu 1. K-Means

a) Trình bày ngắn gọn thuật toán K-Means

Thuật toán **K-Means** là một phương pháp phân cụm (clustering) dựa trên khoảng cách, nhằm chia N điểm dữ liệu thành K cụm, sao cho mỗi điểm dữ liệu thuộc về cụm có tâm cụm (centroid) gần nhất.

1. **Khởi tạo:** Chọn số cụm K và khởi tạo K tâm cụm ban đầu $\mu_1, \ldots, \mu_K$ (thường là ngẫu nhiên hoặc chọn từ các điểm dữ liệu).
2. **Bước Gán nhãn (E-Step):** Gán mỗi điểm dữ liệu x_i vào cụm C_j có tâm cụm μ_j gần nó nhất (sử dụng khoảng cách Euclidean).

$$C_j = \{x_i \mid d(x_i, \mu_j) \leq d(x_i, \mu_k) \text{ với mọi } k \neq j\}$$

3. **Bước Cập nhật tâm cụm (M-Step):** Tính toán lại tâm cụm mới μ_j^{new} bằng cách lấy giá trị trung bình của tất cả các điểm thuộc cụm đó.

$$\mu_j^{\text{new}} = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

4. **Lặp:** Lặp lại Bước 2 và 3 cho đến khi các tâm cụm không còn thay đổi đáng kể hoặc đạt số lần lặp tối đa.

---

b) Áp dụng K-Means cho tập điểm x_1, x_2, x_3, x_4 với $k = 2$

Dữ liệu: $x_1(1,3), x_2(1.5, 3.2), x_3(1.3, 2.8), x_4(3,1)$. $K = 2$. **Khởi tạo:** Chọn x_1 và x_4 làm tâm cụm ban đầu.

$$\mu_1^{(0)} = x_1 = (1,3), \quad \mu_2^{(0)} = x_4 = (3,1)$$

Lần lặp 1

1. **Gán nhãn:** Tính khoảng cách $d(x_i, \mu_j)$ (Khoảng cách Euclidean).

Điểm (x_i)	$d(x_i, \mu_1^{(0)})$	$d(x_i, \mu_2^{(0)})$	**Nhãn cụm**
$x_1(1,3)$	**0**	$\sqrt{(1-3)^2 + (3-1)^2} = 2.83$	C1
$x_2(1.5, 3.2)$	**0.54**	$\sqrt{(1.5-3)^2 + (3.2-1)^2} = 2.66$	C1
$x_3(1.3, 2.8)$	**0.36**	$\sqrt{(1.3-3)^2 + (2.8-1)^2} = 2.40$	C1
$x_4(3,1)$	2.83	**0**	C2

   Phân cụm: $C_1 : \{x_1, x_2, x_3\}, \quad C_2 : \{x_4\}$

2. **Cập nhật tâm cụm:**

$$\mu_1^{(1)} = \left(\frac{1 + 1.5 + 1.3}{3}, \frac{3 + 3.2 + 2.8}{3}\right) = (1.27, 3)$$

$$\mu_2^{(1)} = \left(\frac{3}{1}, \frac{1}{1}\right) = (3,1)$$

Lần lặp 2

1. **Gán nhãn:** Dùng $\mu_1^{(1)} \approx (1.27, 3)$ và $\mu_2^{(1)} = (3, 1)$.

Điểm (x_i)	$d(x_i, \mu_1^{(1)})$	$d(x_i, \mu_2^{(1)})$	**Nhãn cụm**
$x_1(1, 3)$	**0.27**	2.83	C1
$x_2(1.5, 3.2)$	**0.30**	2.66	C1
$x_3(1.3, 2.8)$	**0.20**	2.40	C1
$x_4(3, 1)$	2.65	**0**	C2

Phân cụm: $C_1 : \{x_1, x_2, x_3\}, \quad C_2 : \{x_4\}$

Kết quả: Phân cụm không thay đổi. Thuật toán dừng lại.

- **Cụm 1:** $\{x_1, x_2, x_3\}$
- **Cụm 2:** $\{x_4\}$

---

c) Bài toán gom khách hàng Mai Linh

Dữ liệu: $A(1,1), B(3,1), C(3,3), D(4,2), E(1,3)$. $K = 2$. **Khởi tạo:** Chọn A và C làm tâm cụm ban đầu.

$$\mu_1^{(0)} = A = (1,1), \quad \mu_2^{(0)} = C = (3,3)$$

Lần lặp 1

1. **Gán nhãn:**

KH	$d(x_i, \mu_1^{(0)})$	$d(x_i, \mu_2^{(0)})$	**Nhãn cụm**
$A(1,1)$	**0**	2.83	C1
$B(3,1)$	**2.00**	2.00	C1 (Chọn C1)
$C(3,3)$	2.83	**0**	C2
$D(4,2)$	3.16	**1.41**	C2
$E(1,3)$	**2.00**	2.00	C1 (Chọn C1)

Phân cụm: $C_1 : \{A, B, E\}, \quad C_2 : \{C, D\}$

2. **Cập nhật tâm cụm:**

$$\mu_1^{(1)} = \left(\frac{1+3+1}{3}, \frac{1+1+3}{3}\right) = \left(\frac{5}{3}, \frac{5}{3}\right) \approx (1.67, 1.67)$$

$$\mu_2^{(1)} = \left(\frac{3+4}{2}, \frac{3+2}{2}\right) = (3.5, 2.5)$$

Lần lặp 2

1. **Gán nhãn:** Dùng $\mu_1^{(1)} \approx (1.67, 1.67)$ và $\mu_2^{(1)} = (3.5, 2.5)$.

KH	$d(x_i, \mu_1^{(1)})$	$d(x_i, \mu_2^{(1)})$	**Nhãn cụm**
$A(1,1)$	**0.94**	2.69	C1
$B(3,1)$	1.58	**1.03**	C2
$C(3,3)$	1.58	**0.71**	C2
$D(4,2)$	2.50	**0.71**	C2
$E(1,3)$	**1.58**	2.50	C1

Phân cụm: $C_1 : \{A, E\}, \quad C_2 : \{B, C, D\}$

2. **Cập nhật tâm cụm:**

$$\mu_1^{(2)} = \left(\frac{1+1}{2}, \frac{1+3}{2}\right) = (1, 2)$$

$$\mu_2^{(2)} = \left(\frac{3+3+4}{3}, \frac{1+3+2}{3}\right) = \left(\frac{10}{3}, 2\right) \approx (3.33, 2)$$

Lần lặp 3

1. **Gán nhãn:** Dùng $\mu_1^{(2)} = (1, 2)$ và $\mu_2^{(2)} \approx (3.33, 2)$.

KH	$d(x_i, \mu_1^{(2)})$	$d(x_i, \mu_2^{(2)})$	**Nhãn cụm**
$A(1,1)$	**1.00**	2.36	C1
$B(3,1)$	2.24	**0.33**	C2
$C(3,3)$	2.24	**1.05**	C2
$D(4,2)$	3.00	**0.67**	C2
$E(1,3)$	**1.00**	2.36	C1

Phân cụm: $C_1 : \{A, E\}, \quad C_2 : \{B, C, D\}$

Kết quả: Phân cụm không thay đổi. Thuật toán dừng lại.

Kết luận:

- **Nhóm khách 1:** $\{A(1,1), E(1,3)\}$. **Điểm đón (tâm cụm):** $\mu_1 = (1, 2)$.
- **Nhóm khách 2:** $\{B(3,1), C(3,3), D(4,2)\}$. **Điểm đón (tâm cụm):** $\mu_2 = (\frac{10}{3}, 2) \approx (3.33, 2)$.

Câu 2. Phân cụm với K-Means và Hierarchical Clustering

Bảng dữ liệu:

Subject	**A** (x)	**B** (y)	**Tọa độ**
1	1.0	1.0	(1.0, 1.0)
2	1.5	2.0	(1.5, 2.0)
3	3.0	4.0	(3.0, 4.0)
4	5.0	7.0	(5.0, 7.0)
5	3.5	5.0	(3.5, 5.0)
6	4.5	5.0	(4.5, 5.0)
7	3.5	4.5	(3.5, 4.5)

a) Áp dụng thuật toán K-Means để phân cụm với $k = 2$

Khởi tạo: Chọn Subject 1 và Subject 4 làm tâm cụm ban đầu.

$$\mu_1^{(0)} = (1.0, 1.0), \quad \mu_2^{(0)} = (5.0, 7.0)$$

Lần lặp 1

1. **Gán nhãn:**

Subject	$d(x_i, \mu_1^{(0)})$	$d(x_i, \mu_2^{(0)})$	**Nhãn cụm**
1	**0**	6.40	C1
2	**1.12**	5.10	C1
3	**3.61**	3.61	C1 ($\approx$)
4	6.40	**0**	C2
5	4.03	**2.50**	C2
6	4.03	**2.06**	C2
7	3.81	**2.83**	C2

Phân cụm: $C_1 : \{1, 2, 3\}, \quad C_2 : \{4, 5, 6, 7\}$

2. **Cập nhật tâm cụm:**

$$\mu_1^{(1)} = \left(\frac{1.0 + 1.5 + 3.0}{3}, \frac{1.0 + 2.0 + 4.0}{3} \right) \approx (1.83, 2.33)$$

$$\mu_2^{(1)} = \left(\frac{5.0 + 3.5 + 4.5 + 3.5}{4}, \frac{7.0 + 5.0 + 5.0 + 4.5}{4} \right) \approx (4.13, 5.38)$$

Lần lặp 2

1. **Gán nhãn:** Dùng $\mu_1^{(1)} \approx (1.83, 2.33)$ và $\mu_2^{(1)} \approx (4.13, 5.38)$.

Subject	$d(x_i, \mu_1^{(1)})$	$d(x_i, \mu_2^{(1)})$	**Nhãn cụm**
1	**1.57**	5.38	C1
2	**0.47**	4.28	C1
3	2.04	**1.78**	C2
4	5.64	**1.84**	C2
5	3.15	**0.74**	C2
6	3.77	**0.53**	C2
7	2.74	**1.08**	C2

Phân cụm: $C_1 : \{1, 2\}, \quad C_2 : \{3, 4, 5, 6, 7\}$

2. **Cập nhật tâm cụm:**

$$\mu_1^{(2)} = \left(\frac{1.0 + 1.5}{2}, \frac{1.0 + 2.0}{2} \right) = (1.25, 1.5)$$

$$\mu_2^{(2)} = \left(\frac{3.0 + 5.0 + 3.5 + 4.5 + 3.5}{5}, \frac{4.0 + 7.0 + 5.0 + 5.0 + 4.5}{5} \right) = (3.9, 5.1)$$

Lần lặp 3

1. **Gán nhãn:** Dùng $\mu_1^{(2)} = (1.25, 1.5)$ và $\mu_2^{(2)} = (3.9, 5.1)$.

Subject	$d(x_i, \mu_1^{(2)})$	$d(x_i, \mu_2^{(2)})$	**Nhãn cụm**
1	**0.56**	5.07	C1
2	**0.56**	4.01	C1
3	3.15	**1.39**	C2
4	6.27	**2.16**	C2
5	3.51	**0.41**	C2
6	3.71	**0.61**	C2
7	3.08	**0.78**	C2

Phân cụm: $C_1 : \{1, 2\}, \quad C_2 : \{3, 4, 5, 6, 7\}$

Kết quả: Phân cụm không thay đổi. Thuật toán dừng lại.

- **Cụm 1:** $\{1, 2\}$
- **Cụm 2:** $\{3, 4, 5, 6, 7\}$

---

b) Áp dụng thuật toán Agglomerative Hierarchical Clustering

Ma trận khoảng cách ban đầu (D_0) (Khoảng cách Euclidean):

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1.12	3.61	6.40	4.03	4.03	3.81
2	1.12	0	2.50	5.10	3.04	3.54	2.69
3	3.61	2.50	0	3.61	1.12	1.58	**0.71**
4	6.40	5.10	3.61	0	2.50	2.06	2.83
5	4.03	3.04	1.12	2.50	0	1.00	**0.50**
6	4.03	3.54	1.58	2.06	1.00	0	0.86
7	3.81	2.69	**0.71**	2.83	**0.50**	0.86	0

Khoảng cách nhỏ nhất: $d(5,7) = 0.50$. Hợp nhất 1: **(5, 7)**.

I. Complete Linkage (Khoảng cách Max: $\mathbf{d_C(C_A, C_B) = \max d(x_i, x_j)}$)

1. **Hợp nhất 1: (5, 7)** tại 0.50.
2. **Hợp nhất 2: (5, 7) và 6.**
   $d_C((5,7),6) = \max(d(5,6), d(7,6)) = \max(1.00, 0.86) = \mathbf{1.00}$.
   Hợp nhất: **(5, 6, 7)** tại 1.00.
3. **Hợp nhất 3: 1 và 2.**
   $d_C(1,2) = \mathbf{1.12}$.
   Hợp nhất: **(1, 2)** tại 1.12.
4. **Hợp nhất 4: 3 và (5, 6, 7).**
   $d_C(3,(5,6,7)) = \max(d(3,5), d(3,6), d(3,7)) = \max(1.12, 1.58, 0.71) = \mathbf{1.58}$.
   Hợp nhất: **(3, 5, 6, 7)** tại 1.58.
5. **Hợp nhất 5: 4 và (3, 5, 6, 7).**
   $d_C(4,(3,5,6,7)) = \max(d(4,3), d(4,5), d(4,6), d(4,7)) = \max(3.61, 2.50, 2.06, 2.83) = 3.61$.
6. **Hợp nhất 6: (1, 2) và (3, 4, 5, 6, 7).**
   $d_C((1,2),(3,4,5,6,7)) = \max(d(1,3), d(1,4), \dots, d(2,7)) = \max(3.61, 6.40, \dots) = \mathbf{6.40}$.

Kết quả Complete Link (K=2): Ta còn 2 cụm sau Lần hợp nhất 4.

- **Cụm 1:** $\{1, 2\}$
- **Cụm 2:** $\{3, 4, 5, 6, 7\}$

II. Single Linkage (Khoảng cách Min: $\mathbf{d_S(C_A, C_B) = \min d(x_i, x_j)}$)

1. **Hợp nhất 1: (5, 7)** tại 0.50.
2. **Hợp nhất 2: (5, 7) và 3.**
   $d_S((5,7),3) = \min(d(5,3), d(7,3)) = \min(1.12, 0.71) = \mathbf{0.71}$.
   Hợp nhất: **(3, 5, 7)** tại 0.71.
3. **Hợp nhất 3: (3, 5, 7) và 6.**
   $d_S((3,5,7),6) = \min(d(3,6), d(5,6), d(7,6)) = \min(1.58, 1.00, 0.86) = \mathbf{0.86}$.
   Hợp nhất: **(3, 5, 6, 7)** tại 0.86.
4. **Hợp nhất 4: 1 và 2.**
   $d_S(1,2) = \mathbf{1.12}$.
   Hợp nhất: **(1, 2)** tại 1.12.
5. **Hợp nhất 5: (3, 5, 6, 7) và 4.**
   $d_S((3,5,6,7),4) = \min(d(3,4), d(5,4), d(6,4), d(7,4)) = \min(3.61, 2.50, 2.06, 2.83) = \mathbf{2.06}$.
   Hợp nhất: **(3, 4, 5, 6, 7)** tại 2.06.

6. **Hợp nhất 6: (1, 2) và (3, 4, 5, 6, 7).**
   $d_S((1,2),(3,4,5,6,7)) = \min(d(1,\text{C2}), d(2,\text{C2})) = d(2,3) = \mathbf{2.50}$.

Kết quả Single Link (K=2): Ta còn 2 cụm sau Lần hợp nhất 5.

- **Cụm 1:** $\{1,2\}$
- **Cụm 2:** $\{3,4,5,6,7\}$
