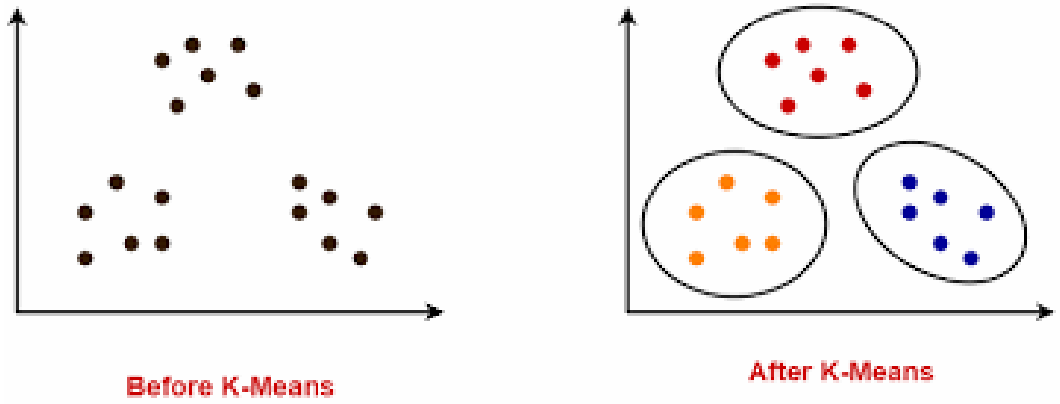
**Câu 1:**

**a) Trình bày ngắn gọn thuật toán k-means.**

Thuật toán K-Means là một phương pháp học máy không giám sát phổ biến, được sử dụng để phân cụm (clustering) các điểm dữ liệu thành K nhóm khác biệt. Mục tiêu là giảm thiểu tổng khoảng cách bình phương giữa các điểm dữ liệu và tâm cụm (centroid) mà chúng thuộc về.

- Giúp xác định các nhóm tự nhiên trong các tập dữ liệu chưa được gắn nhãn
- Hoạt động bằng cách nhóm các điểm dựa trên khoảng cách đến các trung tâm cụm
- Thường được sử dụng trong phân khúc khách hàng, nén hình ảnh và khám phá mẫu
- Hữu ích khi bạn cần cấu trúc từ dữ liệu thô, chưa được tổ chức




b) **Cho tập điểm: $x_1$ = {1, 3}, $x_2$ = {1.5, 3.2 }, $x_3$ = {1.3, 2.8}, $x_4$ = {3, 1}, Dùng k-means để gom cụm với k = 2.**

$$x_1 = (1,3) \qquad x_2 = (1.5,3.2) \qquad x_3 = (1.3,2.8) \qquad x_4 = (3,1)$$

Chọn ngẫu nhiên hai điểm bất kỳ làm tâm cụm:

Tâm cụm 1 ($C_1$): $x_1$ = (1,3)

Tâm cụm 2 ($C_2$): $x_4$ = (3, 1)

**Bước 1: Khoảng cách tới tâm ban đầu**

| Điểm ($x_i$) | $d(x_i, C_1)$ | $d(x_i, C_2)$ | Gán cụm |
|---|---|---|---|
| $x_1$(1,3) | $\sqrt{(1-1)^2+(3-3)^2}=0$ | $\sqrt{(1-3)^2+(3-1)^2}=\sqrt{8}\approx 2,828$ | 1 |
| $x_2$ = | $\sqrt{(1.5-1)^2+(3.2-3)^2}\approx 0,539$ | $\sqrt{(1.5-3)^2+(3.2-1)^2}\approx 2,662$ | 1 |

| | | | |
|---|---|---|---|
| {1.5, 3.2 } | | | |
| $x_3$ = {1.3, 2.8} | $\sqrt{(1.3-1)^2+(2.8-3)^2} \approx 0,36$ | $\sqrt{(1.3-3)^2+(2.8-1)^2} \approx 2,476$ | 1 |
| $x_4$ = {3, 1} | $\sqrt{(3-1)^2+(1-3)^2} \approx 2,828$ | $\sqrt{(3-3)^2+(1-1)^2} = 0$ | 2 |

Kết quả sau bước 1:

Cụm 1: $x_1$, $x_2$, $x_3$

Cụm 2: $x_4$

**Bước 2: Cập nhật tâm cụm:**

Tâm cụm mới $C_1^{'}$ (trung bình của $x_1$, $x_2$, $x_3$)

$$C_1^{'} = \left( \frac{1+1,5+1,3}{3}, \frac{3+3.2+2.8}{3} \right) \approx (1.27,3)$$

Tâm cụm mới $C_2^{'}$ (trung bình của $x_4$)

$$C_2^{'} = (3,1)$$

**Bước 3:** Gán cụm:

| Điểm ($x_i$) | $d(x_i, C_1^{'})$ | $d(x_i, C_2^{'})$ | Gán cụm |
|---|---|---|---|
| $x_1$ = (1,3) | 0,27 | 2,83 | 1 |
| $x_2$ = {1.5, 3.2 } | 0,3 | 2,66 | 1 |
| $x_3$ = {1.3, 2.8} | 0,2 | 2,48 | 1 |
| $x_4$ = {3, 1} | 2,83 | 0 | 2 |

Thuật toán K-Means đã hội tụ và phân chia các điểm dữ liệu thành hai cụm như sau:

Cụm 1: $\{x_1(1,3), x_2(1.5,3.2), x_3(1.3,2.8)\}$

Cụm 2: $\{x_4(3,1)\}$

**c)** Một xe đón khách về bến xe Miền Đông của công ty Mai Linh muốn đón n khách hàng. Do thời gian đón khách ít nên công ty muốn gom khách về k điểm để tiện việc đón. Giả sử n = 5 và k = 2. Năm khách hàng đang ở các tọa độ:

A(1,1), B(3,1), C(3,3), D(4,2), E(1,3). Anh/chị hãy cho biết nên nhóm khách nào tới điểm đón nào để việc đưa đón là thuận tiện nhất. Cho biết tọa độ của 2 điểm cần đón khách? Giả sử độ đo khoảng cách được sử dụng là khoảng cách Euclidean.

Bước 1: Khởi tạo 2 cụm

Tâm 1: $M_1$ = A(1,1)

Tâm 2: $M_2$ = D(4,2)

Bước 2: Gán khách hàng vào cụm gần nhất (lần 1)

| Khách hàng | *d*(X, $M_1$) | *d*(X, $M_2$) | Gán cụm |
|---|---|---|---|
| A(1,1) | 0 | 3,16 | 1 |
| B(3,1) | 2 | 1,41 | 2 |
| C(3,3) | 2,83 | 1,41 | 2 |
| D(4,2) | 3,16 | 0 | 2 |
| E(1,3) | 2 | 3,16 | 1 |

Kết quả sau lần 1:

Cụm 1: {A(1,1), E(1,3)}

Cụm 2: {B(3,1), C(3,3), D(4,2)}

Bước 3: Cập nhật tâm cụm:

Tâm mới $M_1^{'} = \left(\frac{1+1}{2}, \frac{1+3}{2}\right) = (1,2)$

Tâm mới $M_2^{'} = \left(\frac{3+3+4}{3}, \frac{1+3+2}{3}\right) = \left(\frac{10}{3}, \frac{6}{3}\right) = (3.33,2)$

Bước 2: Gán khách hàng vào cụm gần nhất (lần 2)

| Khách hàng | *d*(X, $M_1^{'}$) | *d*(X, $M_2^{'}$) | Gán cụm |
|---|---|---|---|
| A(1,1) | 1 | 1,66 | 1 |
| B(3,1) | 2,24 | 1,05 | 2 |
| C(3,3) | 2,24 | 1,05 | 2 |
| D(4,2) | 3 | 0,67 | 2 |
| E(1,3) | 1 | 1,66 | 1 |

Kết quả sau lần 2:

Cụm 1: {A(1,1), E(1,3)}

Cụm 2: {B(3,1), C(3,3), D(4,2)}

Vì kết quả gán cụm không thay đổi so với lần trước, thuật toán hội tụ và dừng lại.

**Câu 2**. Cho bảng dữ liệu sau:

| Subject | A | B |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

a) Áp dụng thuật toán k-means để phân cụm với k=2

Bước 1: Khởi tạo các tâm cụm

Tâm $C_1$: (1.0,1.0)

Tâm $C_2$: (5.0,7.0)

Bước 2: Phân cụm (lần 1)

| Subject | $d(x, M_1)$ | $d(x, M_2)$ | Gán cụm |
|---|---|---|---|
| 1 | 0 | 7.21 | 1 |
| 2 | 1.12 | 5.81 | 1 |
| 3 | 3.61 | 3.61 | Chọn 1 |
| 4 | 7.21 | 0 | 2 |
| 5 | 4.1 | 2.5 | 2 |
| 6 | 4.53 | 2.06 | 2 |
| 7 | 3.81 | 2.7 | 2 |

**Cụm 1:** {1, 2, 3}

**Cụm 2:** {4, 5, 6, 7}

Bước 3: Cập nhật tâm cụm

$C_1'$ = (1.83, 2.33)

$C_2'$ = (4.13, 5.38)

Bước 4: Phân cụm

| **Subject** | **d(x,C1′)** | **d(x,C2′)** | **Cụm** |
|---|---|---|---|
| 1 | 1.63 | 4.88 | **1** |
| 2 | 0.49 | 3.99 | **1** |
| 3 | 2.15 | 1.61 | **2** |

| | | | |
|---|---|---|---|
| 4 | 5.39 | 1.93 | **2** |
| 5 | 3.19 | 0.65 | **2** |
| 6 | 3.69 | 0.44 | **2** |
| 7 | 2.65 | 0.94 | **2** |

Bước 5: Cập nhật tâm cụm

$C_1''$ = (1.25, 1.5)

$C_2''$ = (3.9, 5.1)

Bước 6: Phân cụm

| Subject | d(x,C1″) | d(x,C2″) | Cụm |
|---|---|---|---|
| 1 | 0.56 | 4.88 | **1** |
| 2 | 0.56 | 3.86 | **1** |
| 3 | 3.01 | 1.28 | **2** |
| 4 | 6.55 | 2.08 | **2** |
| 5 | 3.87 | 0.47 | **2** |
| 6 | 4.39 | 0.61 | **2** |
| 7 | 3.32 | 0.69 | **2** |

Kết quả cuối cùng

Cụm 1: Subject 1, Subject 2

Cụm 2: Subject 3, Subject 4, Subject 5, Subject 6, Subject 7

b) Áp dụng thuật toán Agglomerative Hierarchical Clustering để phân cụm dùng Complete Link, Single Link.

Ta có ma trận khoảng cách

| **Điểm** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| **1** | 0 | 1.12 | 3.61 | 7.21 | 4.72 | 5.32 | 4.30 |
| **2** | 1.12 | 0 | 2.50 | 6.10 | 3.61 | 4.24 | 3.20 |
| **3** | 3.61 | 2.50 | 0 | 3.61 | 1.12 | 1.80 | 0.71 |
| **4** | 7.21 | 6.10 | 3.61 | 0 | 2.50 | 2.06 | 2.92 |
| **5** | 4.72 | 3.61 | 1.12 | 2.50 | 0 | 1.00 | 0.50 |
| **6** | 5.32 | 4.24 | 1.80 | 2.06 | 1.00 | 0 | 1.12 |
| **7** | 4.30 | 3.20 | 0.71 | 2.92 | 0.50 | 1.12 | 0 |

---

- Single Link (khoảng cách giữa cụm = min khoảng cách giữa 2 điểm thuộc hai cụm)

Khoảng cách nhỏ nhất trong ma trận điểm ban đầu là d(5,7) = 0.5 nên gộp 5 và 7 -> {5,7}

Ta xét

d({5,7},3) = min(d(5,3)=1.118034, d(7,3)=0.707107) = 0.707107

d({5,7},6) = min(d(5,6)=1.0, d(7,6)=1.118034) = 1.0

d({5,7},4) = min(d(5,4)=2.5, d(7,4)=2.915476) = 2.5

Vì d{5,7} đến {1},{2} khoảng cách lớn nên bỏ qua. Khoảng cách nhỏ nhất bây giờ là d(3, {5,7}) = 0.707107-> {3,5,7}

Tiếp tục xét

d({3,5,7},6) = min(d(3,6)=1.802776, d(5,6)=1.0, d(7,6)=1.118034) = 1.0

d({3,5,7},4) = min(3.605551, 2.5, 2.915476) = 2.5 (min là 2.5 từ 5→4)

d({3,5,7},1), d(...,2) đều lớn hơn 1.0 nên bỏ qua.

Khoảng cách nhỏ nhất hiện tại là d(6, {3,5,7}) = 1.0-> {3,5,6,7}

Cặp còn lại có khoảng cách nhỏ: d(1,2)=1.118034. Nên ta gộp {1} và {2} => {1,2} với khoảng cách ghép = 1.118034.

Bây giờ còn 3 cụm: {1,2}, {3,5,6,7}, {4}. Vì k=2 nên ta cần xuống 2 cụm.

d({3,5,6,7},4) = min(d(3,4)=3.605551, d(5,4)=2.5, d(6,4)=2.061553, d(7,4)=2.915476) = 2.061553 (từ 6→4)

d({1,2}, {3,5,6,7}) lớn hơn

Gộp {4} với {3,5,6,7} vì khoảng cách nhỏ nhất => {3,4,5,6,7}. Khoảng cách ghép = 2.061553.

Kết quả cuối (k=2):

- Cụm 1: {1,2}
- Cụm 2: {3,4,5,6,7}

- Complete Link (khoảng cách giữa cụm = max khoảng cách giữa 2 điểm thuộc hai cụm)

Khoảng nhỏ nhất ban đầu  là d(5,7)=0.5 (min và max cùng có 0.5 vì cả cụm đơn).
→ Gộp {5} và {7} => {5,7}. Khoảng cách ghép là 0.5.

Khoảng cách nhỏ tiếp theo trong toàn bộ ma trận là d(1,2)=1.118034 -> {1,2}

d_complete({5,7},3) = max(d(5,3)=1.118034, d(7,3)=0.707107) = 1.118034

d_complete({5,7},6) = max(1.0, 1.118034) = 1.118034

d_complete({5,7},4) = max(2.5, 2.915476) = 2.915476

Gộp {3} với {5,7} => {3,5,7} với khoảng cách ghép là 1.118034

Xét d_complete giữa {3,5,7} và {6}: d(3,6)=1.802776, d(5,6)=1.0, d(7,6)=1.118034 → max = 1.802776 → Gộp {6} vào => {3,5,6,7}. Khoảng cách ghép là 1.802776.

Ta được 3 cụm {1,2} và {3,5,6,7} và {4}

d_complete({3,5,6,7},4) = max(d(3,4)=3.605551, d(5,4)=2.5, d(6,4)=2.061553, d(7,4)=2.915476) = 3.605551.

d_complete({1,2},{3,5,6,7}) lớn hơn.

Vì vậy gộp {4} vào {3,5,6,7} -> {3,4,5,6,7} với khoảng cách ghép là 3.605551

Kết quả cuối (k=2):

- Cụm 1: {1,2}
- Cụm 2: {3,4,5,6,7}