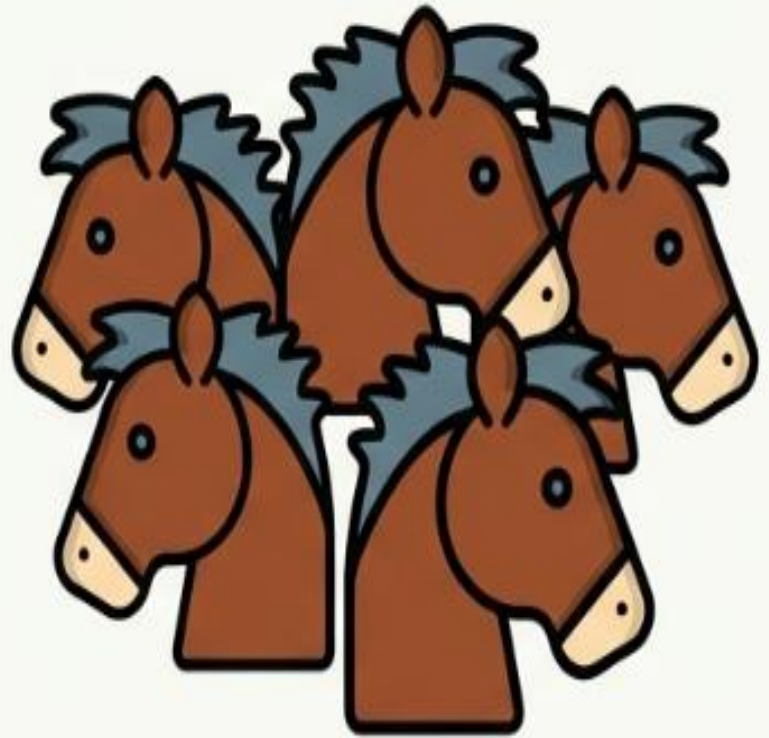


K-MEANS ++

## Giới thiệu



Ý tưởng cốt lõi của K-MEANS++

- Là một cải tiến của K-means
- Là thuật toán học không giám sát
- Để nhóm các điểm dữ liệu tương tự lại với nhau dựa trên sự tương đồng của chúng
- Ý tưởng: khởi tạo các tâm cụm theo cách thông minh hơn so với việc khởi tạo ngẫu nhiên được sử dụng bởi K-means

## Quy trình xây dựng mô hình K-MEANS++

- Bước 2: Khởi tạo tâm cụm
  - Bước 2.1: Khởi tạo tâm cụm đầu tiên bằng cách chọn ngẫu nhiên một điểm dữ liệu từ tập dữ liệu
  - Bước 2.2: Đối với mỗi điểm dữ liệu chưa được chọn làm tâm cụm, hãy tính khoảng cách giữa các điểm đến tâm cụm gần nhất
  - Bước 2.3: Chọn tâm cụm tiếp theo từ các điểm dữ liệu còn lại với xác suất tỷ lệ thuận với bình phương khoảng cách đến tâm cụm gần nhất.
  - Bước 2.4: Lặp lại bước 2 và 3 cho đến khi tất cả các tâm cụm K đã được khởi tạo.

## Quy trình xây dựng mô hình K-MEANS++

- Bước 1: Chọn số cụm k
  - Bước 2: Khởi tạo trọng tâm ban đầu
  - Bước 3: Gán các điểm dữ liệu vào các cụm
  - Bước 4: Cập nhật trọng tâm
  - Bước 5: Kiểm tra hội tụ
- 
- Lưu ý: Thuật toán lặp lại 2 bước 3 và 4 cho tới khi các trọng tâm không thay đổi hoặc sự thay đổi nhỏ hơn một ngưỡng cho trước.

## THỰC HÀNH

- Bước 1: Chọn  $k=2$
- Bước 2:
  - 2.1. Khởi tạo tâm đầu tiên:  $\mu^0 = (1,2)$
  - 2.2. Tính khoảng cách giữa các điểm đến tâm

$$\bullet d_{01} = (1 - 1)^2 + (2 - 2)^2 = 0$$

$$\bullet d_{02} = (1 - 1)^2 + (2 - 4)^2 = 4$$

$$\bullet d_{03} = (1 - 3)^2 + (2 - 4)^2 = 8$$

$$\bullet d_{04} = (1 - 5)^2 + (2 - 7)^2 = 41$$

$$\bullet d_{05} = (1 - 3)^2 + (2 - 8)^2 = 40$$

$$\Rightarrow S = 0 + 4 + 8 + 41 + 40 = 93$$

Điểm toán	Điểm văn
1	2
1	4
3	4
5	7
3	8

- Xác suất chọn mỗi điểm làm tâm cụm:

$$\bullet p_1 = 0/93$$

$$\bullet p_2 = 4/93$$

$$\bullet p_3 = 8/93$$

$$\bullet p_4 = 41/93$$

$$\bullet p_5 = 40/93$$

$\Rightarrow$  Tâm cụm tiếp theo là  $\mu^1 = (5,7)$

# Hierarchical Clustering

## Ý tưởng cốt lõi của Hierarchical Clustering

- Thay vì cố gắng chia không gian thành một số cụ thể các nhóm, thuật toán xây dựng một cây phân cấp (dendrogram) cho phép ta quan sát các mức cắt (cut) khác nhau để nhận các tập con mong muốn.

## Ý tưởng cốt lõi của Hierarchical Clustering

- Hierarchical clustering thuộc nhóm unsupervised learning. Có hai hướng tiếp cận chính:
  - Agglomerative (Bottom-up): Bắt đầu từ mỗi điểm là một cụm riêng lẻ, rồi lặp lại hợp nhất hai cụm gần nhất cho tới khi chỉ còn một cụm chung.
  - Divisive (Top-down): Bắt đầu từ toàn bộ dữ liệu là một cụm, rồi chia tách dần thành các cụm con.

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 1: Khởi tạo
- Bước 2: Tính ma trận khoảng cách giữa mọi cặp cụm
- Bước 3: Tìm hai cụm gần nhất
- Bước 4: Hợp nhất
- Bước 5: Cập nhật ma trận khoảng cách
- Bước 6: Lặp lại
- Bước 7: Truy ngược để lấy k cụm hoặc dừng lại khi có số cụm hợp lý

# Quy trình xây dựng mô hình Hierarchical clustering

- Bước 1: Khởi tạo

- Mỗi điểm  $x_i$  bắt đầu là một cụm riêng lẻ. Gọi tập cụm hiện tại là  $C = \{c_1, c_2, \dots, c_N\}$  với  $c_i = \{x_i\}$ .

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 2: Tính ma trận khoảng cách giữa mọi cặp cụm
  - Tính  $d(c_i, c_j)$  cho mọi cặp cụm (ban đầu chính là khoảng cách giữa hai điểm). Lưu kết quả vào một ma trận khoảng cách ( $N \times N$ ).

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 3: Tìm hai cụm gần nhất
  - Chọn cặp cụm  $c_i, c_j$  có  $d(c_i, c_j)$  nhỏ nhất.

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 4: Hợp nhất
  - Hợp nhất  $c_i, c_j$  thành cụm mới  $c_k = c_i \cup c_j$ . Ghi lại bước hợp nhất này và khoảng cách hợp nhất (độ cao trên dendrogram).

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 5: Cập nhật ma trận khoảng cách
  - Xóa hàng/ cột tương ứng  $c_i$  và  $c_j$ ; thêm hàng/ cột cho cụm  $c_k$ .
  - Cách tính  $d(c_k, c_h)$  phụ thuộc vào **linkage**:
    - Single linkage:  $d(c_k, c_h) = \min \left( d(c_i, c_h), d(c_j, c_h) \right)$
    - Complete linkage:  $d(c_k, c_h) = \max \left( d(c_i, c_h), d(c_j, c_h) \right)$

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 6: Lặp lại
  - Lặp lại Bước 3–5 tới khi chỉ còn một cụm hoặc đạt được số cụm mong muốn.

## Quy trình xây dựng mô hình Hierarchical clustering

- Bước 7: Truy ngược dendrogram để lấy k cụm
  - Sau khi có dendrogram, cắt ở một mức cao thích hợp để lấy số cụm k mong muốn hoặc dựa trên ngưỡng khoảng cách hợp nhất.

# THỰC HÀNH

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

## THỰC HÀNH

- Bước 1: Khởi tạo
  - 5 cụm: {p1}, {p2}, {p3}, {p4}, {p5}.

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

## THỰC HÀNH

- Bước 2: Tính ma trận khoảng cách giữa mọi cặp cụm

	P1	P2	P3	P4	p5
P1	0	1	1.41	9.9	10.6
P2	1	0	1	9.2	9.9
P3	1.41	1	0	8.5	9.2
P4	9.9	9.2	8.5	0	1
P5	10.6	9.9	9.2	1	0

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

## THỰC HÀNH

- Bước 3: Tìm hai cụm gần nhất
  - $(p1,p2)=1$  và  $(p4,p5)=1$ .

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

	P1	P2	P3	P4	p5
P1	0	1	1.41	9.9	10.6
P2	1	0	1	9.2	9.9
P3	1.41	1	0	8.5	9.2
P4	9.9	9.2	8.5	0	1
P5	10.6	9.9	9.2	1	0

## THỰC HÀNH

- Bước 4: Hợp nhất
  - Hợp nhất (p1,p2) thành  $p_{12}$

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	8

	P1	P2	P3	P4	p5
P1	0	1	1.41	9.9	10.6
P2	1	0	1	9.2	9.9
P3	1.41	1	0	8.5	9.2
P4	9.9	9.2	8.5	0	1
P5	10.6	9.9	9.2	1	0

# THỰC HÀNH

## • Bước 5: Cập nhật ma trận khoảng cách

	P <sub>12</sub>	P3	P4	p5
P <sub>12</sub>	0	1	9.2	9.9
P3	1	0	8.5	9.2
P4	9.2	8.5	0	1
P5	9.9	9.2	1	0

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

## • Bước 5: Cập nhật ma trận khoảng cách

- Xóa hàng/ cột tương ứng  $c_i$  và  $c_j$ ; thêm hàng/ cột cho cụm  $c_k$ .
- Cách tính  $d(c_k, c_h)$  phụ thuộc vào linkage:
  - Single linkage:  $d(c_k, c_h) = \min(d(c_i, c_h), d(c_j, c_h))$
  - Complete linkage:  $d(c_k, c_h) = \max(d(c_i, c_h), d(c_j, c_h))$

	P1	P2	P3	P4	p5
P1	0	1	1.41	9.9	10.6
P2	1	0	1	9.2	9.9
P3	1.41	1	0	8.5	9.2
P4	9.9	9.2	8.5	0	1
P5	10.6	9.9	9.2	1	0

## THỰC HÀNH

- Bước 6: Lặp lại
  - Lặp lại Bước 3–5 tới khi chỉ còn một cụm hoặc đạt được số cụm mong muốn.

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

- Bước 7: Truy ngược dendrogram để lấy k cụm

## THỰC HÀNH

- Bước 3: Tìm hai cụm gần nhất
  - $(p_{12}, p_3)=1$  và  $(p_4, p_5)=1$ .

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

	P <sub>12</sub>	P3	P4	p5
P <sub>12</sub>	0	1	9.2	9.9
P3	1	0	8.5	8.5
P4	8.5	8.5	0	1
P5	9.9	8.5	1	0

THỰC HÀNH

- Bước 4: Hợp nhất
  - Hợp nhất ( $p_{12}, p_3$ ) thành  $p_{123}$

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

	$P_{12}$	$P_3$	$P_4$	$p_5$
$P_{12}$	0	1	9.2	9.9
$P_3$	1	0	8.5	9.2
$P_4$	9.2	8.5	0	1
$P_5$	9.9	9.2	1	0

# THỰC HÀNH

## • Bước 5: Cập nhật ma trận khoảng cách

	P <sub>123</sub>	P4	p5
P <sub>123</sub>	0	8.5	9.2
P4	8.5	0	1
P5	9.2	1	0

	P <sub>12</sub>	P3	P4	p5
P <sub>12</sub>	0	1	9.2	9.9
P3	1	0	8.5	9.2
P4	9.2	8.5	0	1
P5	9.9	9.2	1	0

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

- Bước 5: Cập nhật ma trận khoảng cách
  - Xóa hàng/ cột tương ứng  $c_i$  và  $c_j$ ; thêm hàng/ cột cho cụm  $c_k$ .
  - Cách tính  $d(c_k, c_h)$  phụ thuộc vào **linkage**:
    - Single linkage:  $d(c_k, c_h) = \min(d(c_i, c_h), d(c_j, c_h))$
    - Complete linkage:  $d(c_k, c_h) = \max(d(c_i, c_h), d(c_j, c_h))$

## THỰC HÀNH

- Bước 6: Lặp lại
  - Lặp lại Bước 3–5 tới khi chỉ còn một cụm hoặc đạt được số cụm mong muốn.

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

- Bước 7: Truy ngược dendrogram để lấy k cụm

## THỰC HÀNH

- Bước 3: Tìm hai cụm gần nhất
  - $(p4, p5)=1$ .

	P <sub>123</sub>	P4	p5
P <sub>123</sub>	0	8.5	9.2
P4	8.5	0	1
P5	9.2	1	0

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

## THỰC HÀNH

- Bước 4: Hợp nhất
  - Hợp nhất ( $p_4, p_5$ ) thành  $p_{45}$

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

	$P_{123}$	P4	p5
$P_{123}$	0	8.5	9.2
P4	8.5	0	1
P5	9.2	1	0

# THỰC HÀNH

## • Bước 5: Cập nhật ma trận khoảng cách

	P <sub>123</sub>	P45
P <sub>123</sub>	0	8.5
P45	8.5	0

	P <sub>123</sub>	P4	p5
P <sub>123</sub>	0	8.5	9.2
P4	8.5	0	1
P5	9.2	1	0

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

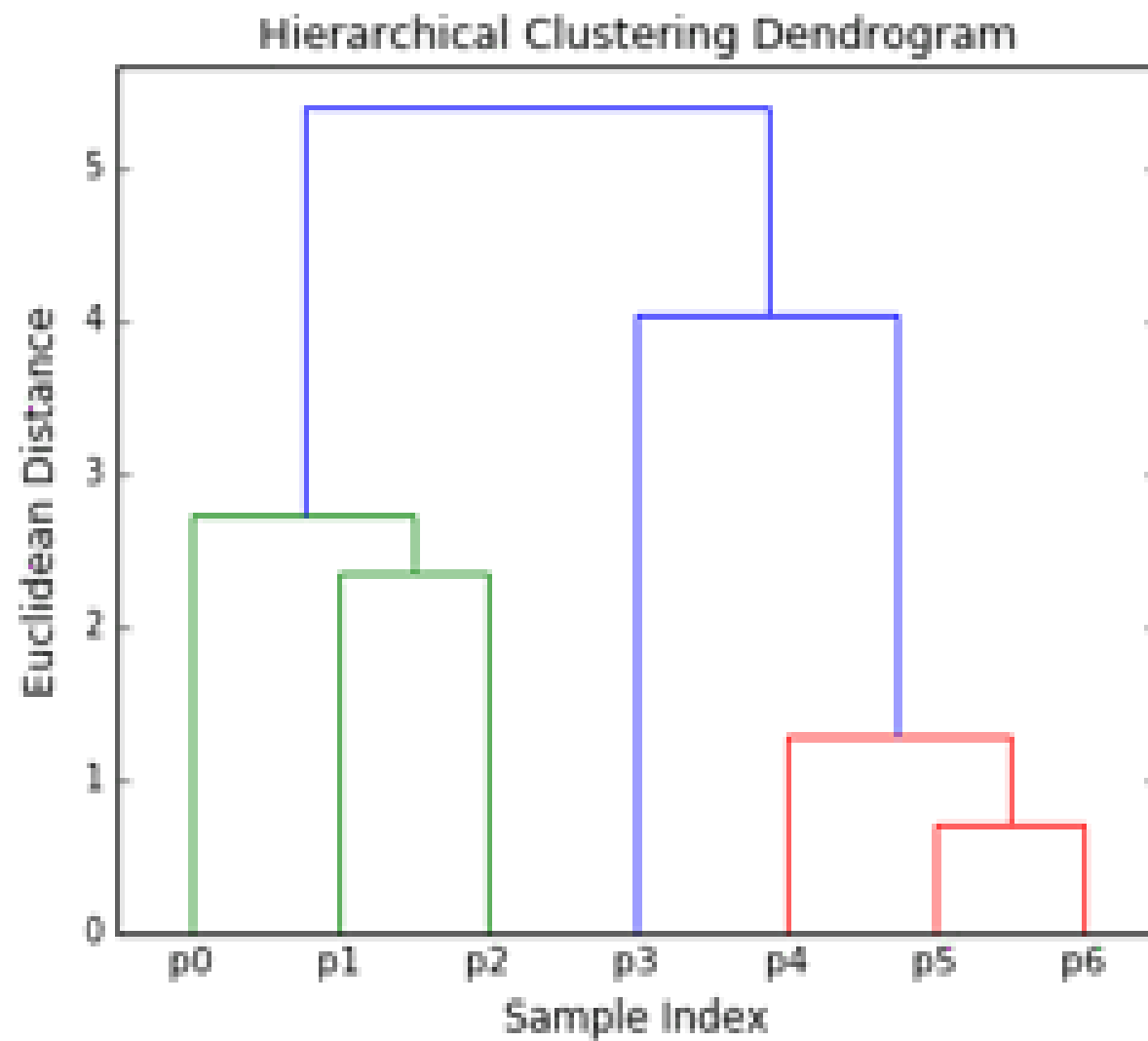
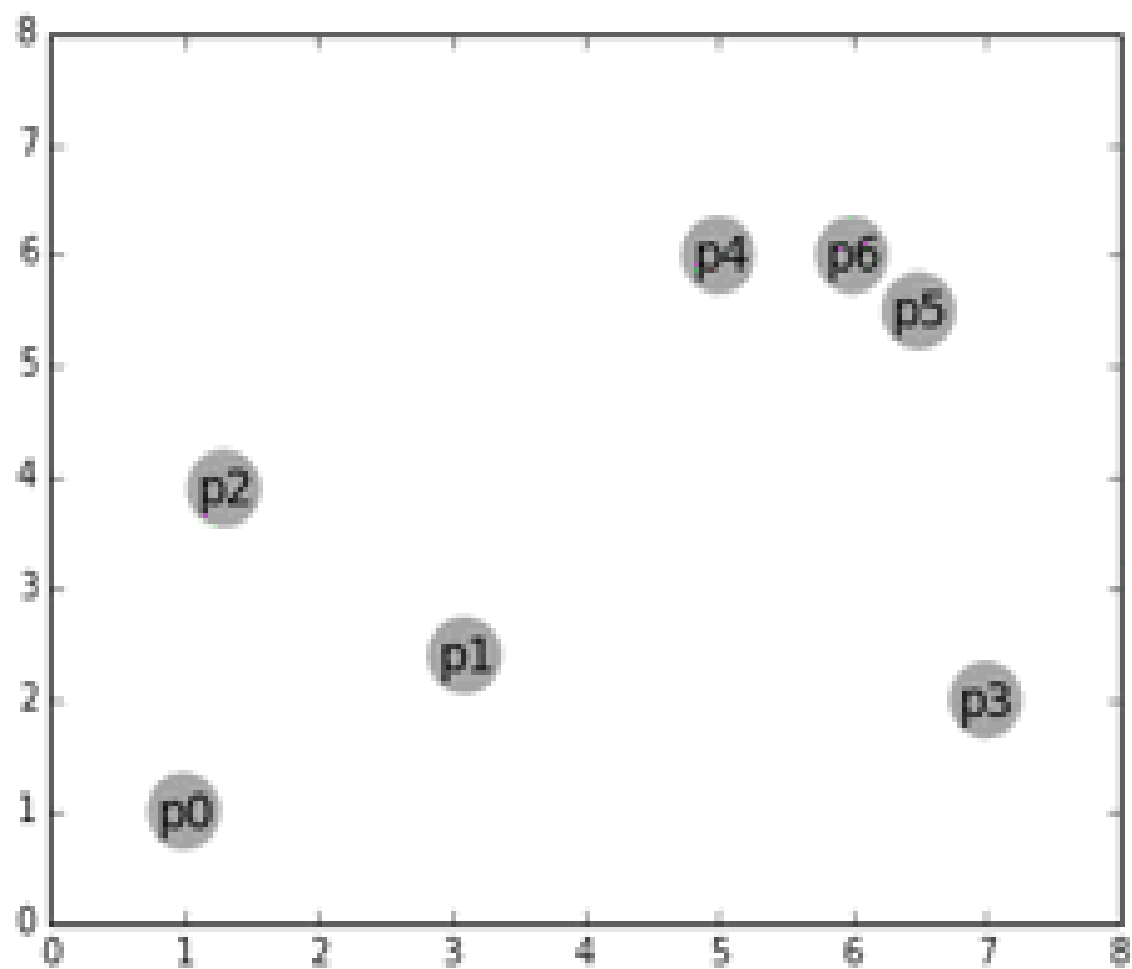
- Bước 5: Cập nhật ma trận khoảng cách
  - Xóa hàng/ cột tương ứng  $c_i$  và  $c_j$ ; thêm hàng/ cột cho cụm  $c_k$ .
  - Cách tính  $d(c_k, c_h)$  phụ thuộc vào **linkage**:
    - Single linkage:  $d(c_k, c_h) = \min(d(c_i, c_h), d(c_j, c_h))$
    - Complete linkage:  $d(c_k, c_h) = \max(d(c_i, c_h), d(c_j, c_h))$

## THỰC HÀNH

- Bước 7: Truy ngược dendrogram để lấy k cụm
  - Sau khi có dendrogram, cắt ở một mức cao thích hợp để lấy số cụm k mong muốn hoặc dựa trên ngưỡng khoảng cách hợp nhất.

Điểm	X	Y
P1	1	1
P2	1	2
P3	2	2
P4	8	8
P5	8	9

Điểm	Feature 1	Feature 2	Feature 3
$p_1$	2.1	3.1	1.6
$p_2$	3.2	3.6	2.1
$p_3$	3.6	3.1	2.6
$p_4$	7.9	8.1	7.6
$p_5$	8.6	8.7	8.2
$p_6$	9.1	8.1	8.6
$p_7$	1.2	2.1	1.7



Silhouette Score

## Silhouette score

- Dùng để đánh giá mức độ tốt của việc phân cụm (clustering).
- Score đo xem:
  - Các điểm có gần cụm của chính mình hay không
  - Và có xa cụm gần nhất khác hay không
- Giá trị Silhouette:
  - +1 → điểm nằm rất đúng cụm
  - 0 → điểm nằm ở ranh giới giữa các cụm
  - -1 → điểm nằm sai cụm (gần cụm khác hơn)
- Silhouette trung bình của tất cả điểm → đánh giá chất lượng phân cụm.

## Công thức tính Silhouette score

- Giả sử điểm  $i$  nằm trong cụm  $C$ .
- Bước 1 — Tính  $a(i)$ : khoảng cách trung bình từ  $i$  đến các điểm trong cùng cụm
- gọi là *intra-cluster distance* (độ kết dính cụm)

$$a(i) = \frac{1}{|C| - 1} \sum_{j \in C, j \neq i} d(i, j)$$

## Công thức tính Silhouette score

- Giả sử điểm  $i$  nằm trong cụm  $C$ .
- Bước 2 — Tính  $b(i)$ : khoảng cách trung bình từ  $i$  đến cụm “láng giềng gần nhất”
- gọi là *nearest-cluster distance*
- Đối với mỗi cụm  $C' \neq C$ :

$$d(i, C') = \frac{1}{|C'|} \sum_{j \in C'} d(i, j)$$

- Cụm nào cho giá trị nhỏ nhất sẽ là cụm láng giềng gần nhất.

$$b(i) = \min_{C' \neq C} d(i, C')$$

## Công thức tính Silhouette score

- Giả sử điểm  $i$  nằm trong cụm  $C$ .
- Bước 3 — Tính Silhouette cho điểm  $i$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- $s(i) \rightarrow 1$  (rất đúng cụm)
- $s(i) \rightarrow 0$  (đứng ranh giới)
- $s(i) < 0$  (cụm bị lẫn)

Công thức tính Silhouette score

- Silhouette Score chung cho cả mô hình

$$S = \frac{1}{n} \sum_{i=1}^n s(i)$$

=> Càng gần 1 là càng tốt

Điểm	Feature 1	Feature 2	Feature 3
$p_1$	2.1	3.1	1.6
$p_2$	3.2	3.6	2.1
$p_3$	3.6	3.1	2.6
$p_4$	7.9	8.1	7.6
$p_5$	8.6	8.7	8.2
$p_6$	9.1	8.1	8.6
$p_7$	1.2	2.1	1.7

Ví dụ

- Giả sử ta có 2 cụm:
  - Cụm A: {p1, p2}
  - Cụm B: {p3, p4}
- Ta xét điểm p1 trong cụm A.

Ví dụ

- Giả sử ta có 2 cụm:
  - Cụm A: {p1, p2}
  - Cụm B: {p3, p4}
- Ta xét điểm p1 trong cụm A.
- **Bước 1: Tính  $a(p1)$** 
  - Giả sử:
    - $d(p1, p2) = 1$
    - $a(p1) = 1$

Ví dụ

- Giả sử ta có 2 cụm:
  - Cụm A: {p1, p2}
  - Cụm B: {p3, p4}
- Ta xét điểm p1 trong cụm A.

- **Bước 1: Tính  $a(p1)$**

- Giả sử:
  - $d(p1, p2) = 1$
  - $a(p1) = 1$

- **Bước 2: Tính  $b(p1)$**

- Khoảng cách trung bình từ p1 đến cụm B:
- Giả sử:
  - $d(p1, p3) = 8$
  - $d(p1, p4) = 9$

$$b(p1) = \frac{8 + 9}{2} = 8.5$$

- **Bước 3: Tính silhouette**

$$s(p1) = \frac{8.5 - 1}{8.5} = 0.88$$

- p1 được phân cụm rất tốt.