



2102470 Học máy

Bài giảng: K-Means Clustering

Chương 3: Phân cụm

Ôn lại bài học trước

- Bạn có nhớ ? % ?

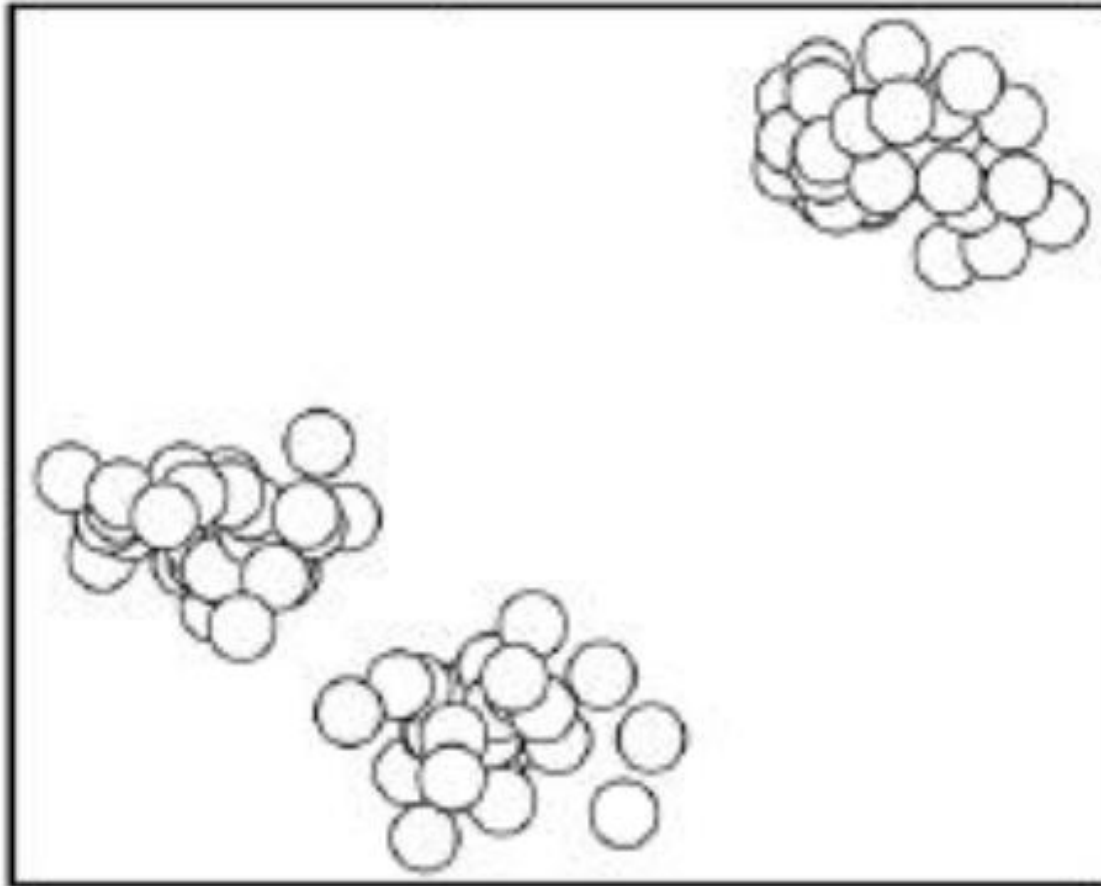
Nội dung chính

- 3.1 Khái niệm về phân cụm
- 3.2 Mô tả bài toán phân cụm
- 3.3 Hàm mục tiêu
- **3.4 K-Means**
- 3.5 Ví dụ về bài toán phân cụm

3.3 K-means

K-means

- Ví dụ: Có một tập dữ liệu đơn giản, cần thực hiện phân cụm



Trường hợp 1

- Giả sử chúng ta đã biết các tâm (centroids)

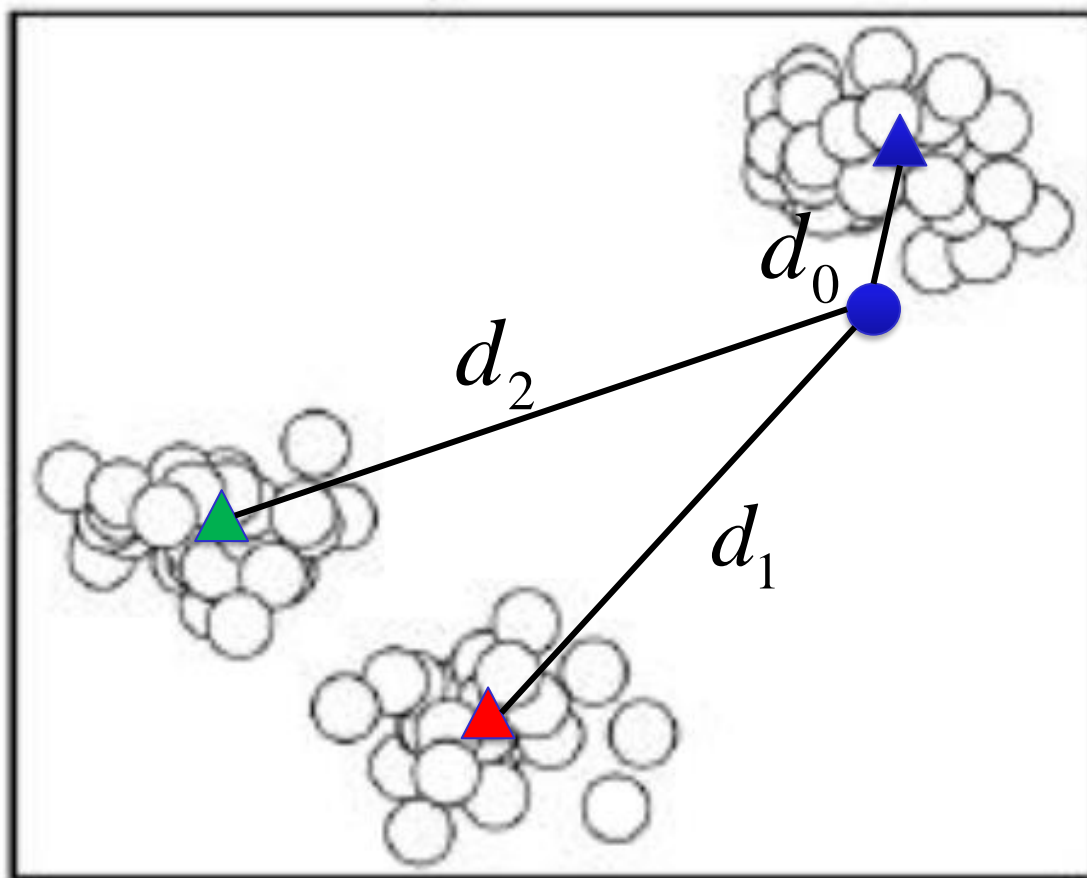
▲ cụm 0

▲ cụm 1

▲ cụm 2

$$k = 3$$

$$d_2 > d_1 > d_0$$



Trường hợp 1

- Dễ dàng đánh nhãn cho tất cả các mẫu trong tập dữ liệu
 - Bằng cách gán mẫu vào cụm mà “khoảng cách” từ mẫu tới cụm đó là gần nhất

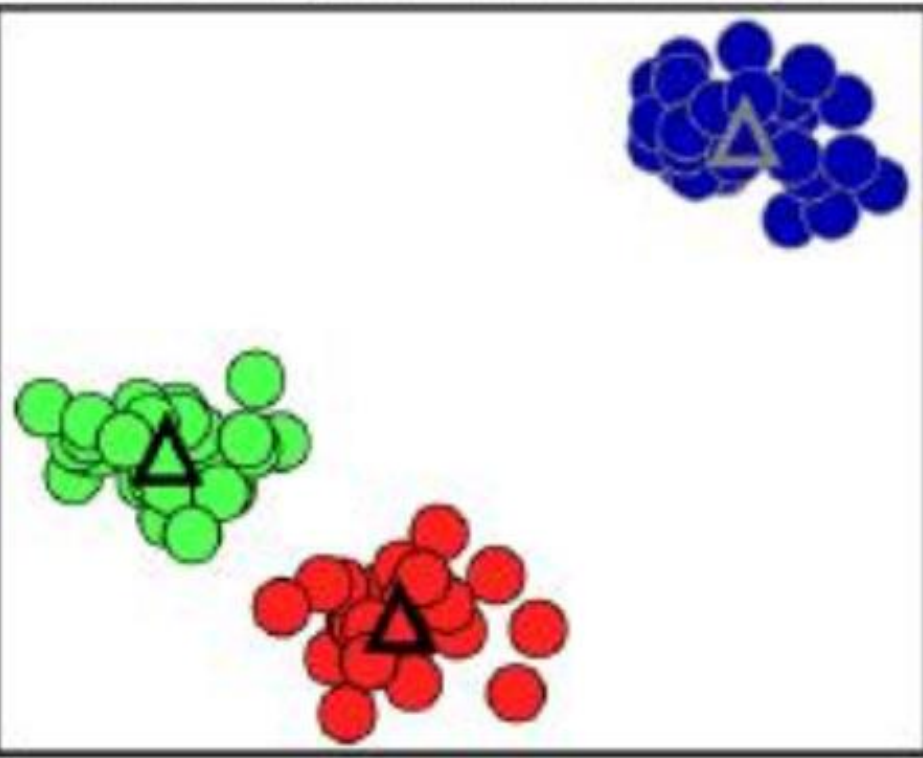


Làm sao xác định được các centroids?



Trường hợp 2

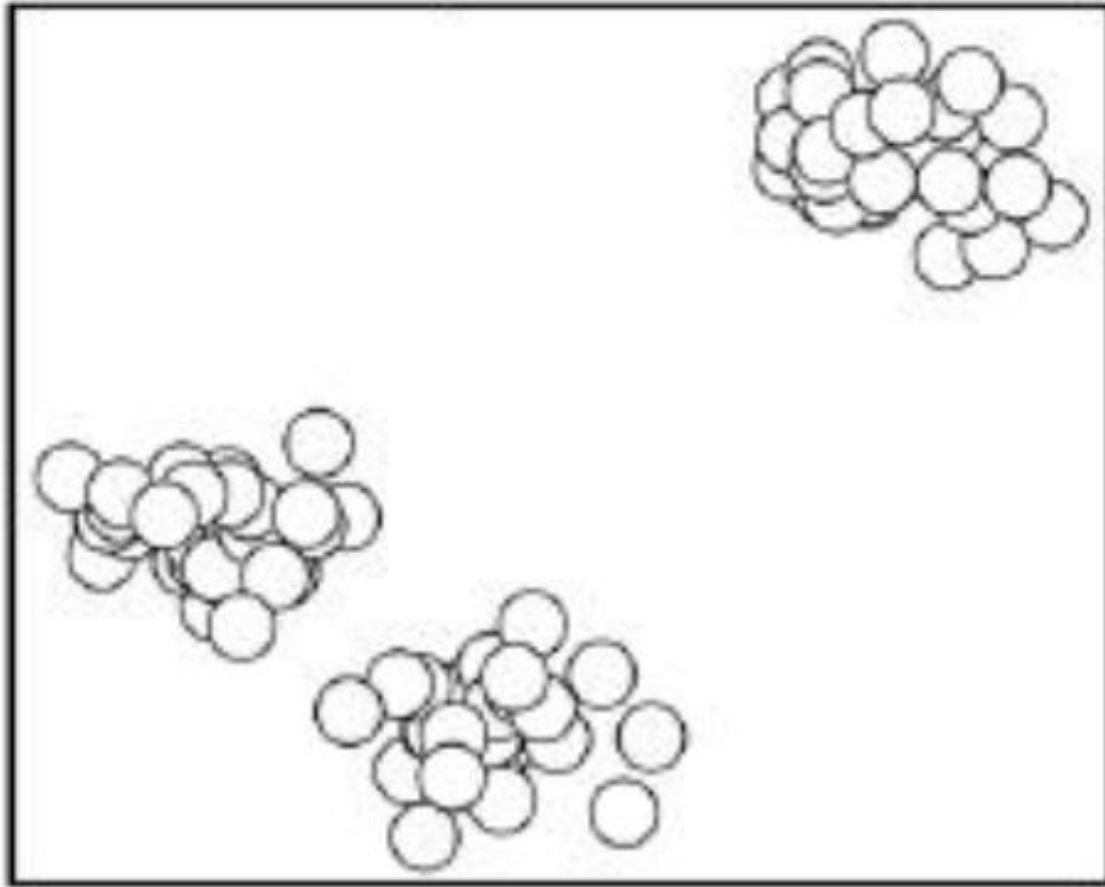
- Giả sử tất cả các mẫu đã được dán nhãn



- Dễ dàng tính được các tâm của các cụm
 - Thông qua việc tính trung bình của mẫu trong cụm đó

K-means

- Ví dụ: Có một tập dữ liệu đơn giản, cần thực hiện phân cụm



- Các mẫu không được đánh nhãn
- Không biết thông tin về các tâm của cụm

K-means

- Có k cụm

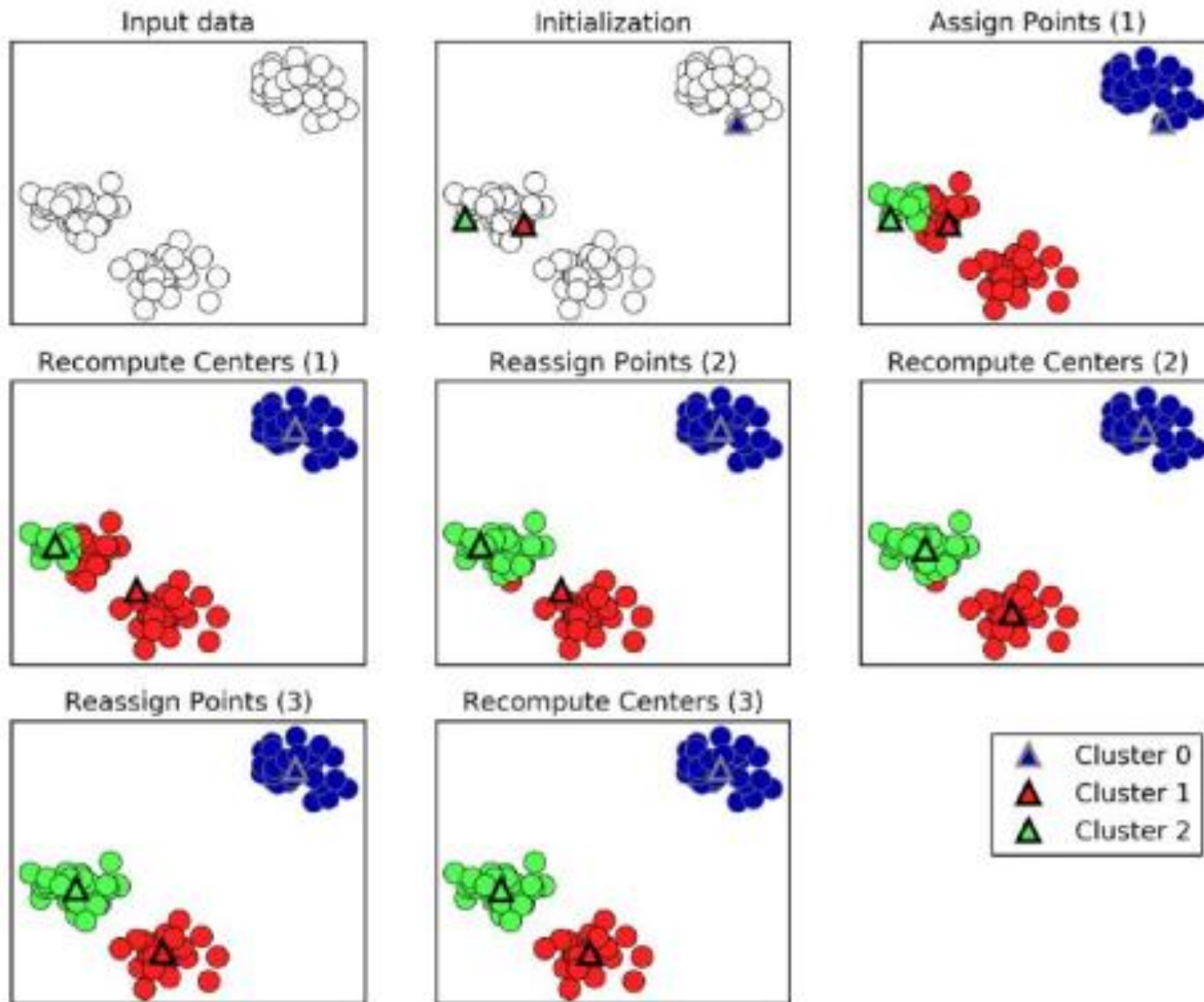
B1: Khởi tạo ngẫu nhiên k tâm

B2: Gán các mẫu vào cụm

B3: Cập nhật các tâm

B4: Lặp lại bước 2, 3 cho tới khi các tâm cố định

K-means



K-means – Ví dụ 1

- Ví dụ 1: Áp dụng k-means, $k = 2$



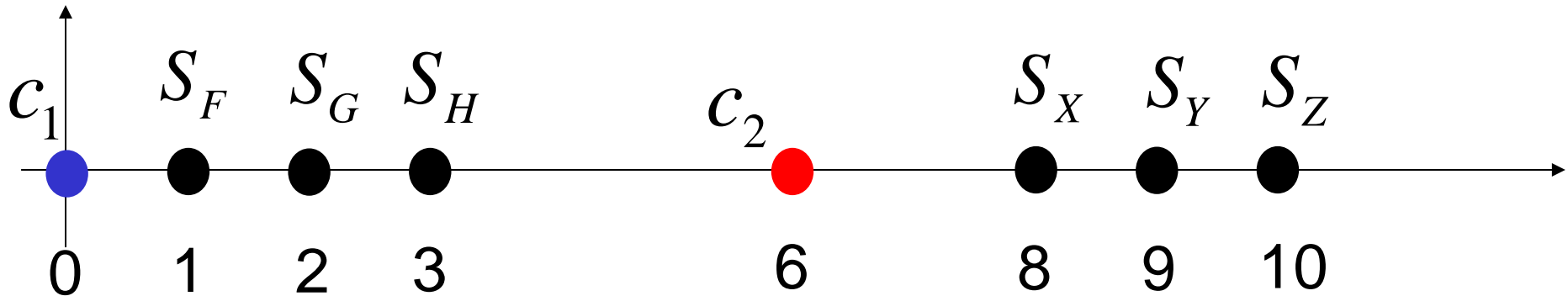
Bộ dữ liệu có $N = 6$ mẫu

$$\{S_F(1), S_G(2), S_F(3), S_X(8), S_Y(9), S_Z(10)\}$$



K-means – Ví dụ 1

- Ví dụ 1: Áp dụng k-means, $k = 2$



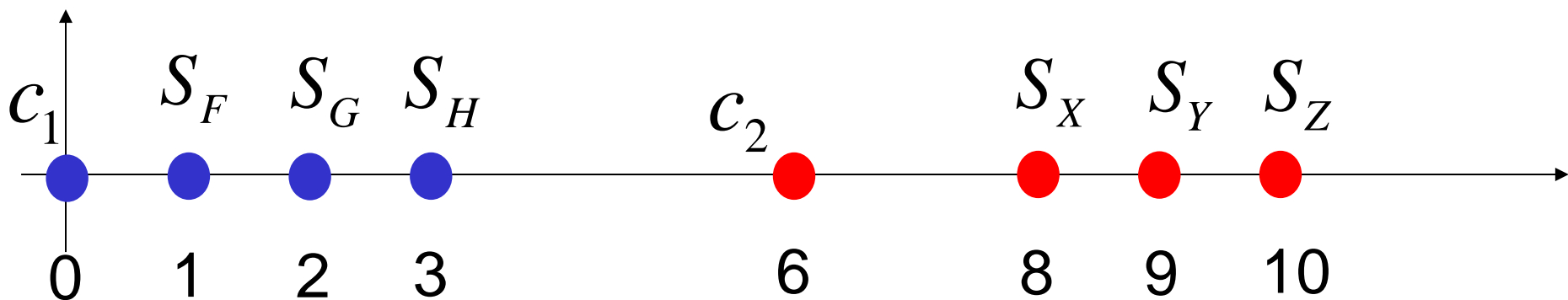
B1: Khởi tạo ngẫu nhiên $k = 2$ tâm

$$c_1 = 0$$

$$c_2 = 6$$

K-means – Ví dụ 1

- Ví dụ 1: Áp dụng k-means, $k = 2$

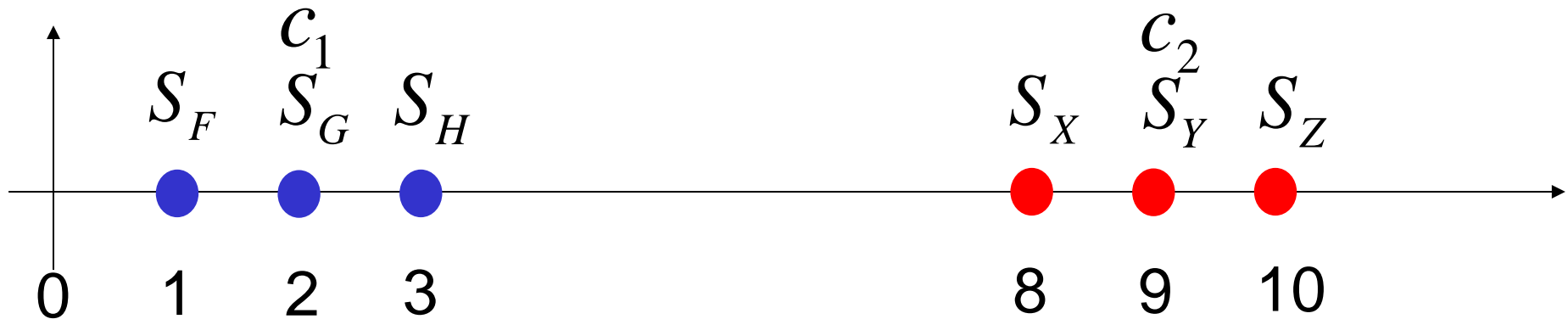


B2: Gán các mẫu vào cụm

$$\begin{aligned} d(S_F, c_1) &= \sqrt{(1-0)^2} = 1 & d(S_F, c_2) &= \sqrt{(1-6)^2} = 5 & S_F &\in C_1 \\ \vdots & & & & & \\ d(S_X, c_1) &= \sqrt{(8-0)^2} = 8 & d(S_X, c_2) &= \sqrt{(8-6)^2} = 2 & S_X &\in C_2 \end{aligned}$$

K-means – Ví dụ 1

- Ví dụ 1: Áp dụng k-means, $k = 2$



B3: Cập nhật các tâm

$$c_1 = 2$$

$$c_2 = 9$$

K-means – Ví dụ 1

- Ví dụ 1: Áp dụng k-means, $k = 2$



B4: Lặp lại bước 2, 3

$$\begin{aligned} d(S_F, c_1) &= \sqrt{(1-0)^2} = 1 & d(S_F, c_2) &= \sqrt{(1-6)^2} = 5 & S_F &\in C_1 \\ \vdots & & & & & \\ d(S_X, c_1) &= \sqrt{(8-0)^2} = 8 & d(S_X, c_2) &= \sqrt{(8-6)^2} = 2 & S_X &\in C_2 \end{aligned}$$

Có các tâm cố định => Dừng

K-means – Ví dụ 1

- Đánh giá

$$\begin{aligned} E &= \sum_{i=1}^{k=2} \sum_{S \in C_i} d(S, c_i)^2 \\ &= \left(d(S_F, c_1)^2 + d(S_G, c_1)^2 + d(S_H, c_1)^2 \right) \\ &\quad + \left(d(S_X, c_2)^2 + d(S_Y, c_2)^2 + d(S_Z, c_2)^2 \right) \\ &= 2 + 2 = 4 \end{aligned}$$

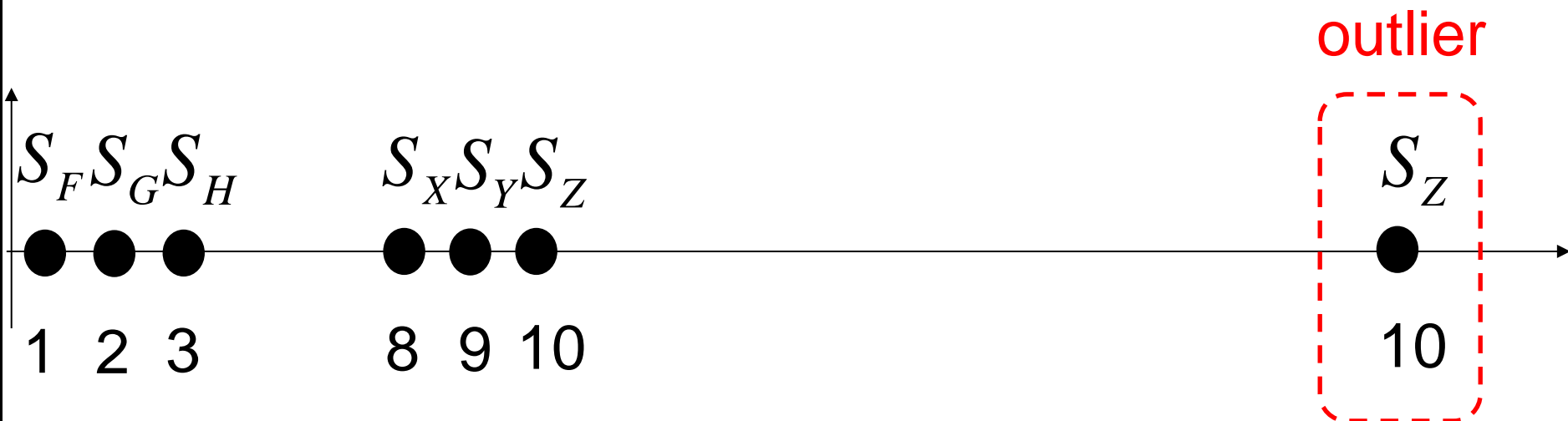
K-means – Ví dụ 2

- Ví dụ 2: Áp dụng k-means, $k = 2$

Bộ dữ liệu có $N = 7$ mẫu

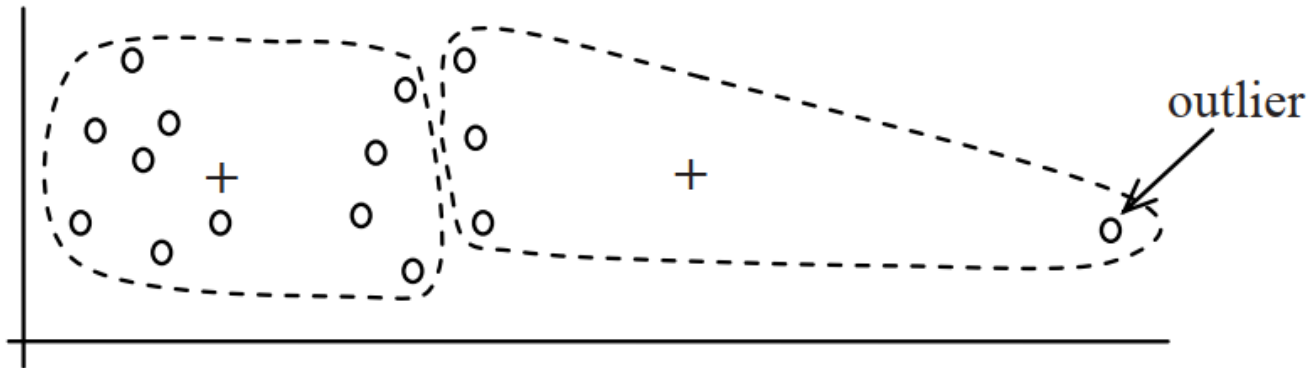


$$\{S_F(1), S_G(2), S_F(3), S_X(8), S_Y(9), S_Z(10), S_W(25)\}$$

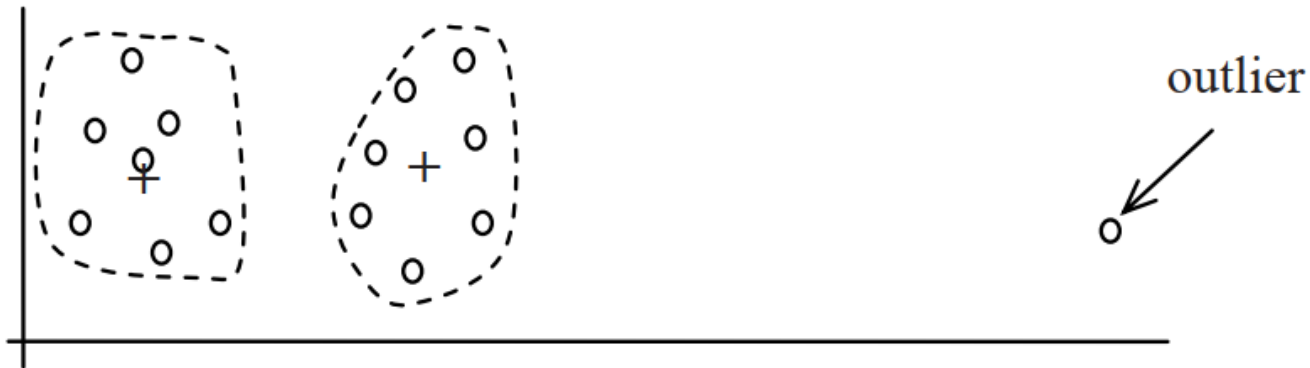


Outlier

- Chú ý đến ảnh hưởng của outlier [B10TLTK2]



(A): Undesirable clusters



(B): Ideal clusters

K-means – Ví dụ 3

- Ví dụ 3: Áp dụng K-means



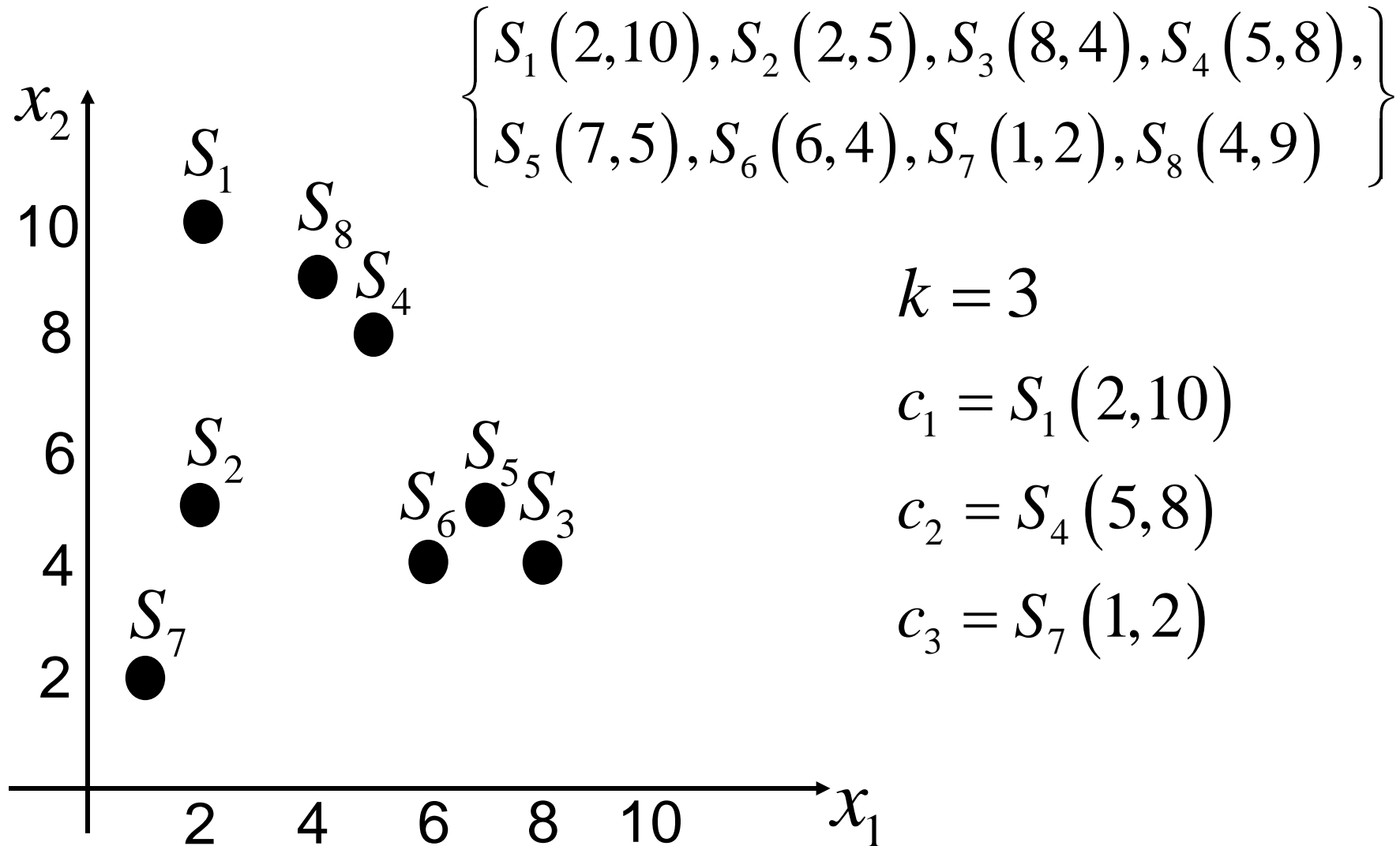
Bộ dữ liệu có $N = 8$ mẫu

$$\left\{ \begin{array}{l} S_1(2, 10), S_2(2, 5), S_3(8, 4), S_4(5, 8), \\ S_5(7, 5), S_6(6, 4), S_7(1, 2), S_8(4, 9) \end{array} \right\}$$

- (a) Trực quan hóa dữ liệu để chọn k cụm và khởi tạo ngẫu nhiên k tâm
- (b) Xác định 3 tâm mới sau lần thực hiện đầu tiên
- (c) Xác định 3 cụm sau khi kết thúc thuật toán

K-means – Ví dụ 3

- Ví dụ 3: (a) Trực quan hóa dữ liệu



K-means – Ví dụ 3

- Ví dụ 3: (a) Trực quan hóa dữ liệu, chọn k và các tâm

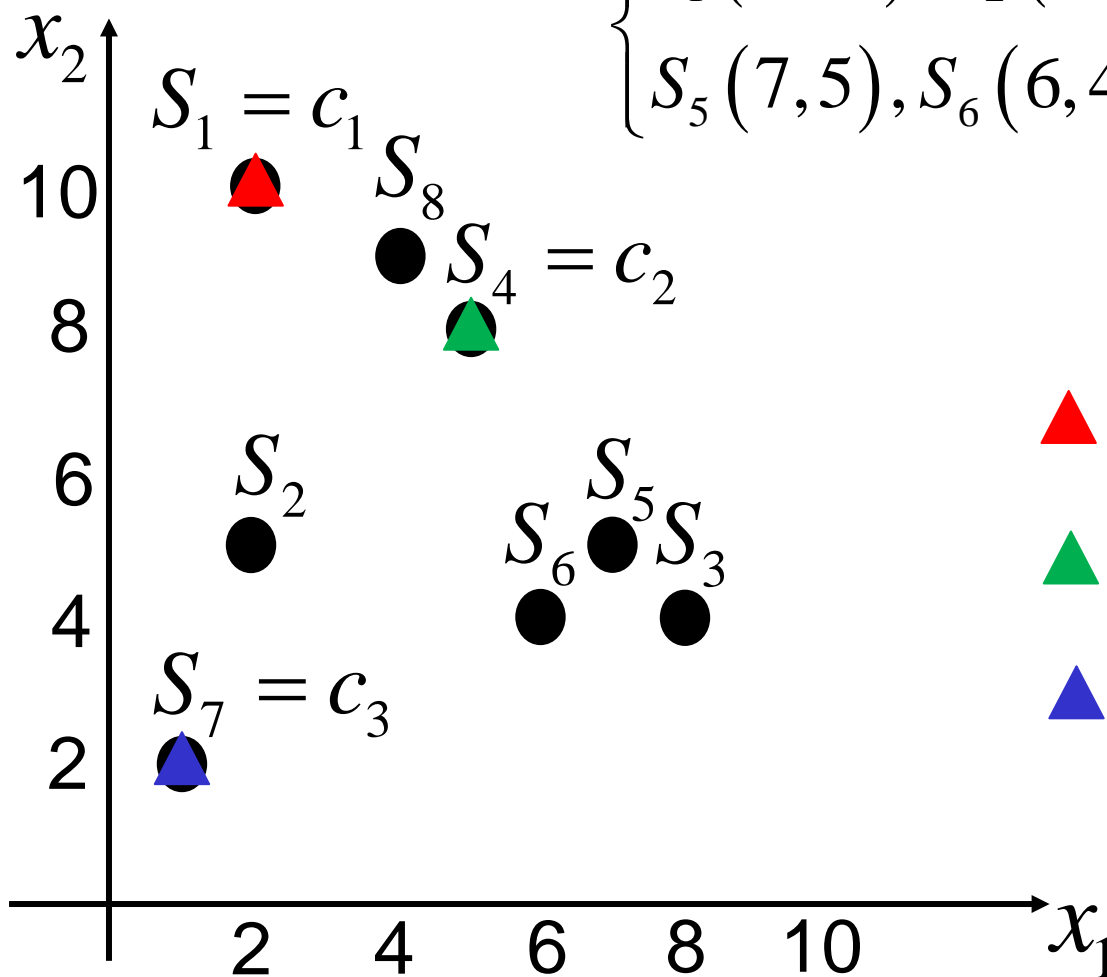
$$\left\{ S_1(2,10), S_2(2,5), S_3(8,4), S_4(5,8), \right. \\ \left. S_5(7,5), S_6(6,4), S_7(1,2), S_8(4,9) \right\}$$

$$k = 3$$

$$\blacktriangle c_1 = S_1(2,10)$$

$$\blacktriangle c_2 = S_4(5,8)$$

$$\blacktriangle c_3 = S_7(1,2)$$



K-means – Ví dụ 3

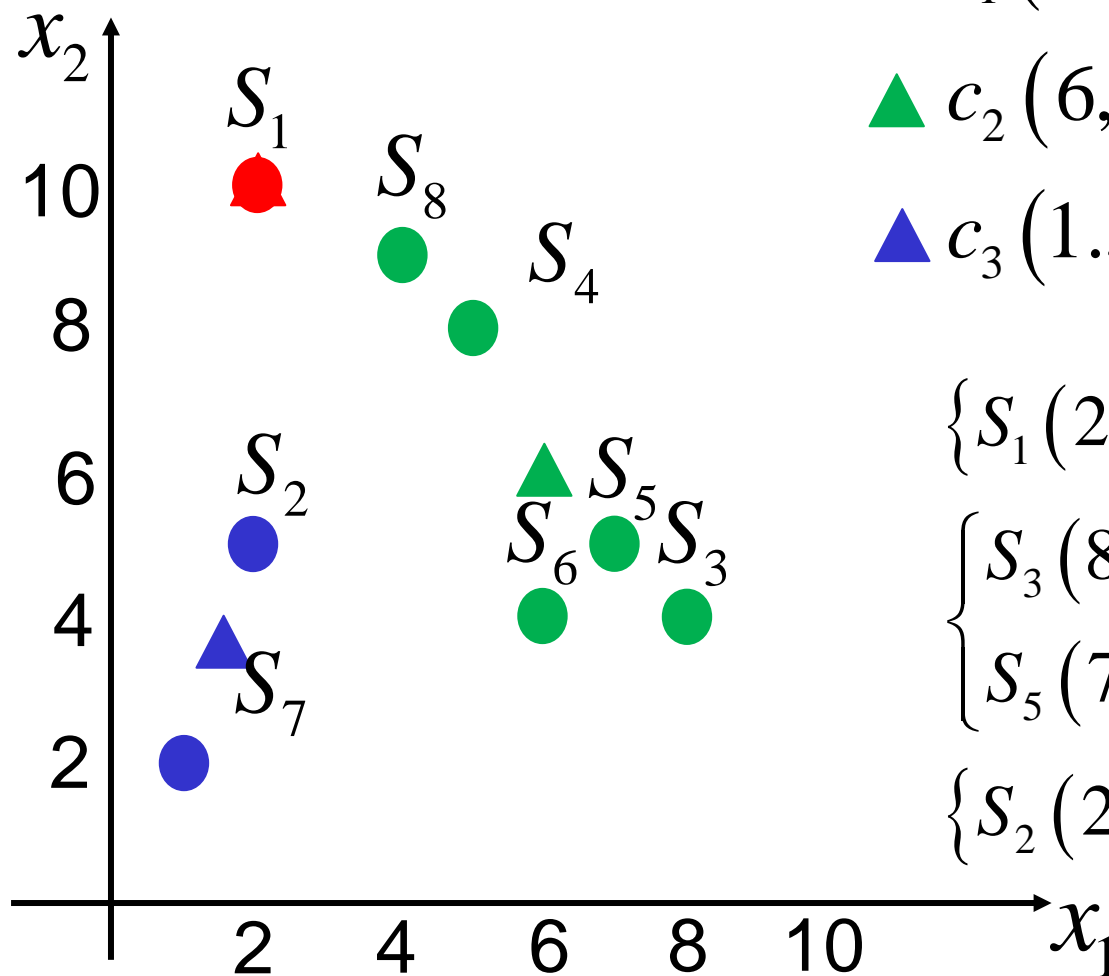
- Ví dụ 3: (b) Kết thúc lần thực hiện đầu tiên

$k = 3$

▲ $c_1(2, 10) = S_1$

▲ $c_2(6, 6)$

▲ $c_3(1.5, 3.5)$



$\{S_1(2, 10)\}$

$\left\{ S_3(8, 4), S_4(5, 8), \right. \\ \left. S_5(7, 5), S_6(6, 4), S_8(4, 9) \right\}$

$\{S_2(2, 5), S_7(1, 2)\}$

K-means – Ví dụ 3

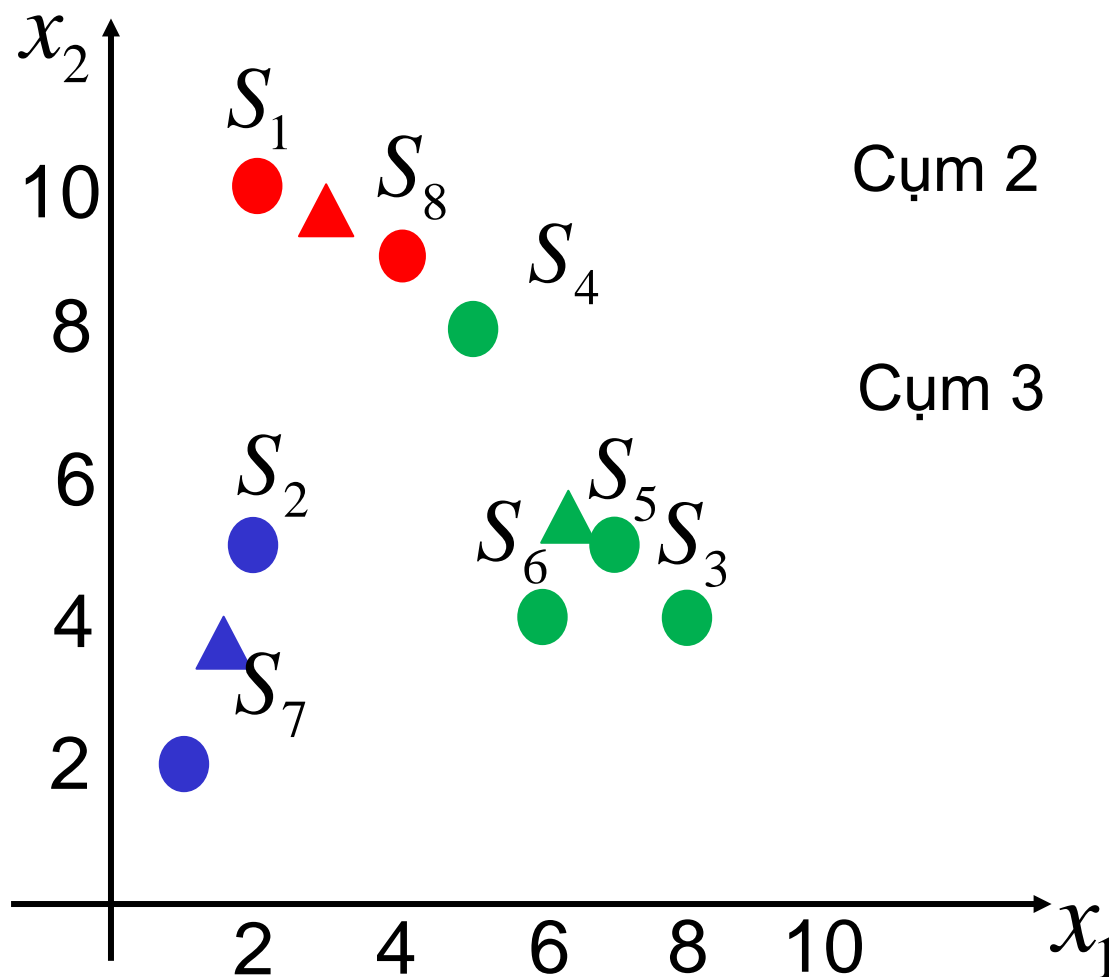
- Ví dụ 3: (c) Kết thúc thuật toán

$k = 3$

Cụm 1 $\{S_1(2,10), S_8(4,9)\}$

Cụm 2 $\left\{ S_3(8,4), S_4(5,8), \right. \\ \left. S_5(7,5), S_6(6,4) \right\}$

Cụm 3 $\{S_2(2,5), S_7(1,2)\}$



▲ $c_1(3, 9.5)$

▲ $c_2(6.5, 5.25)$

▲ $c_3(1.5, 3.5)$

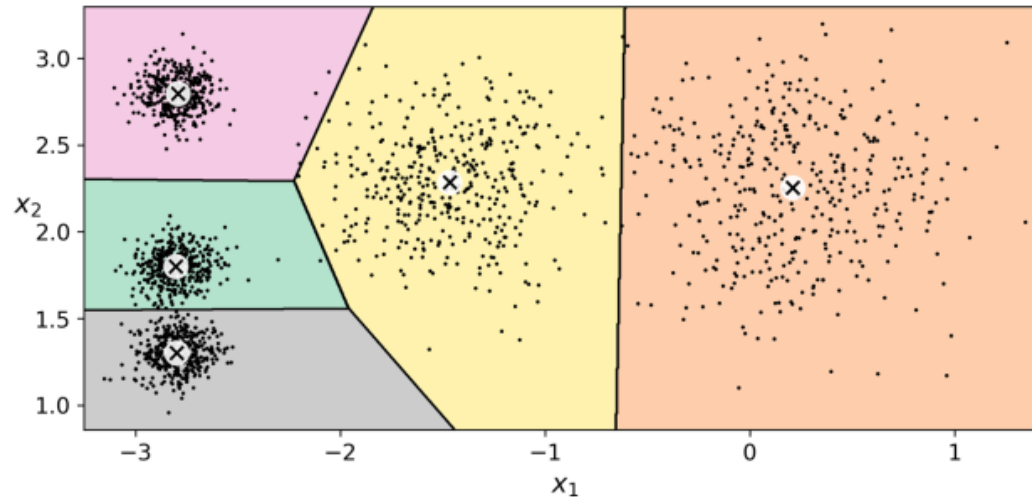
K-means

- Mặc dù thuật toán được đảm bảo để hội tụ, nó có thể không hội tụ đến kết quả đúng
 - Phụ thuộc vào thiết lập các tâm ban đầu
 - Lựa chọn số lượng tối ưu của các cụm

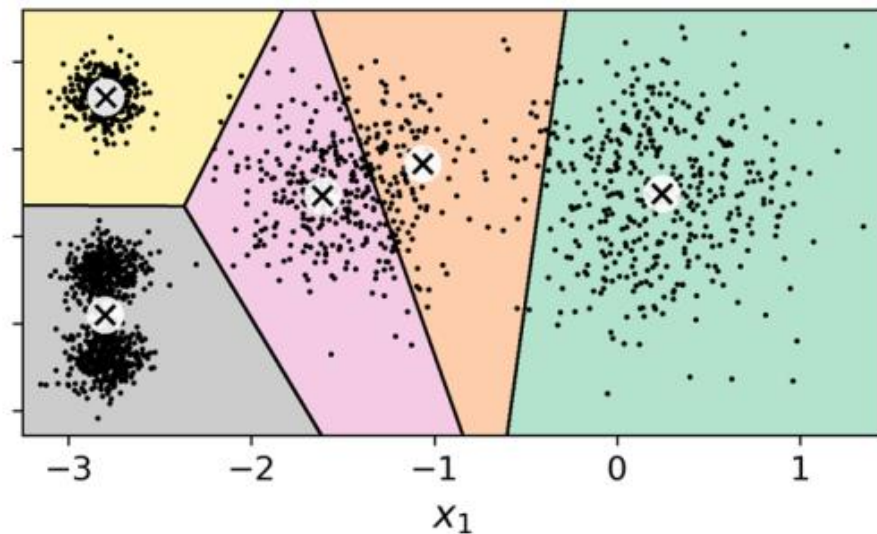
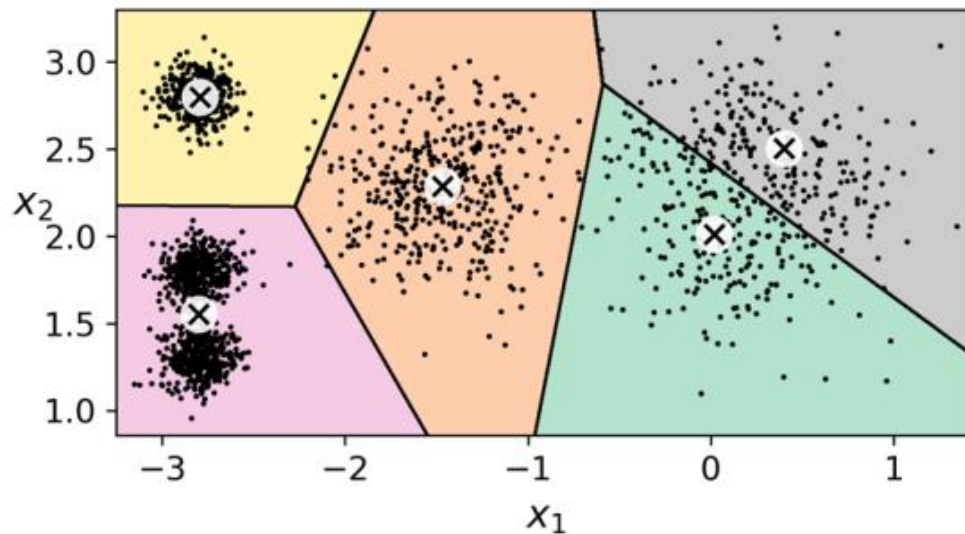
Xác lập tâm ban đầu

- Ảnh hưởng đến kết quả phân cụm

đúng

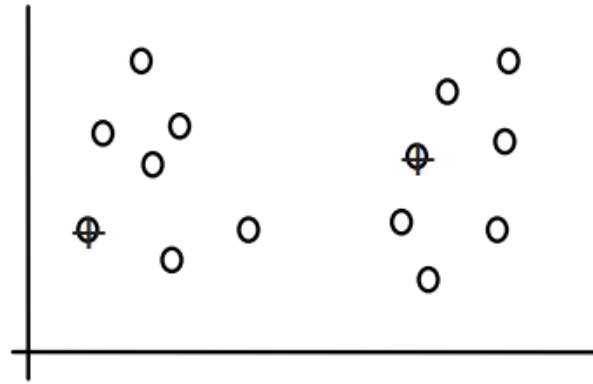


sai

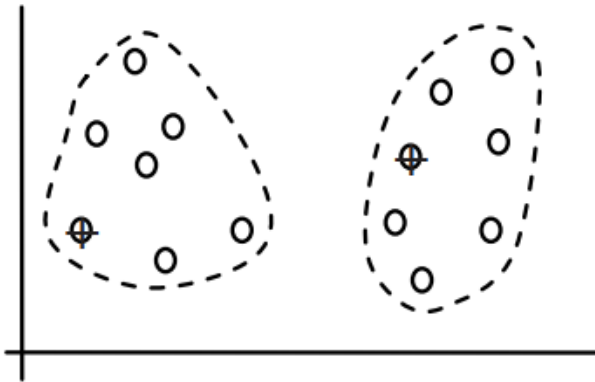


Xác lập tâm ban đầu

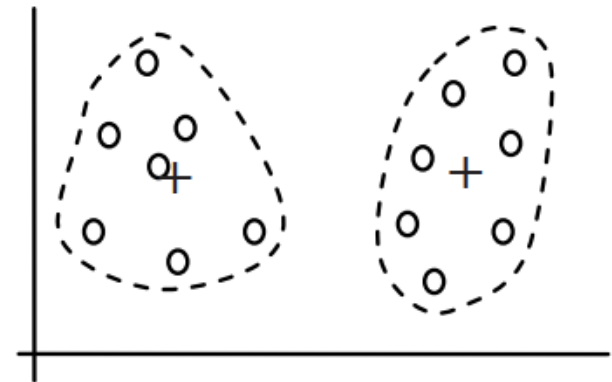
- Lựa chọn tâm ban đầu tốt [B10TLTK2]



(A). Random selection of k seeds (centroids)



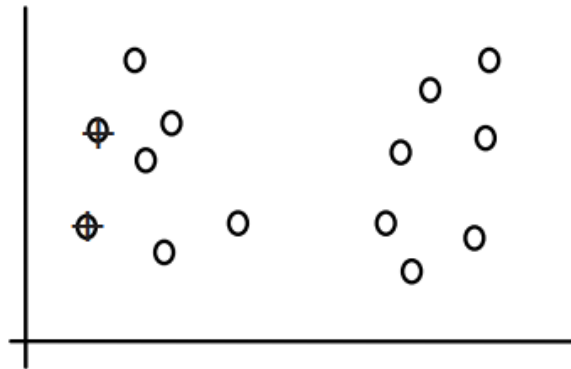
(B). Iteration 1



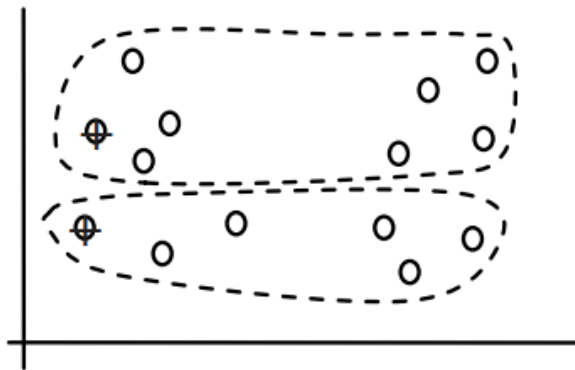
(C). Iteration 2

Xác lập tâm ban đầu

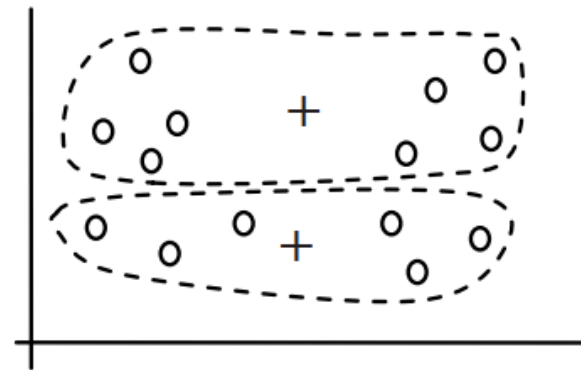
- Lựa chọn tâm ban đầu không tốt [B10TLTK2]



(A). Random selection of seeds (centroids)



(B). Iteration 1



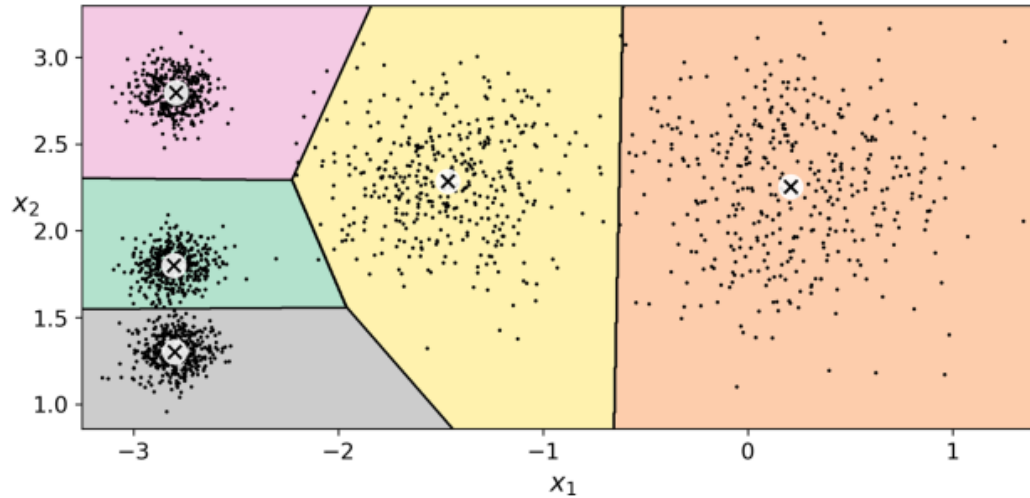
(C). Iteration 2

Số lượng tối ưu của các cụm

- Ảnh hưởng đến kết quả phân cụm

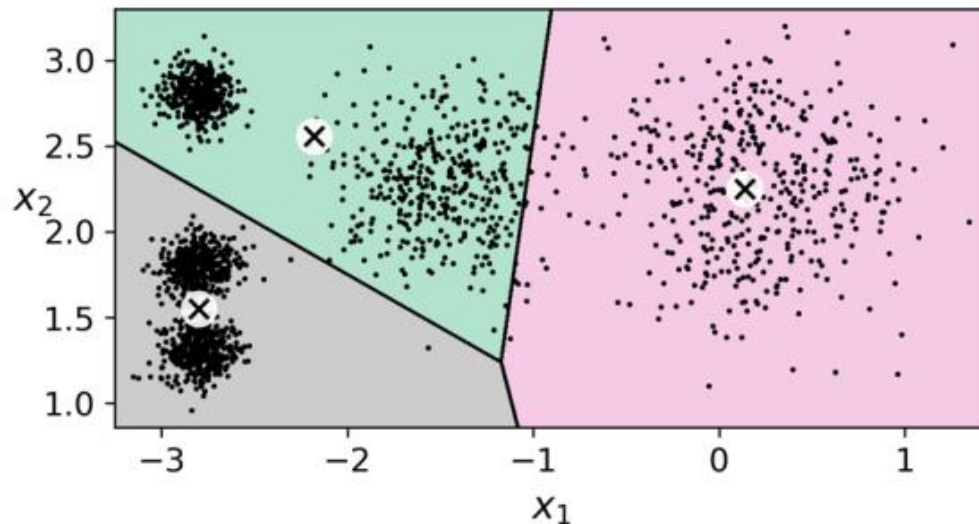
đúng

$k = 5$

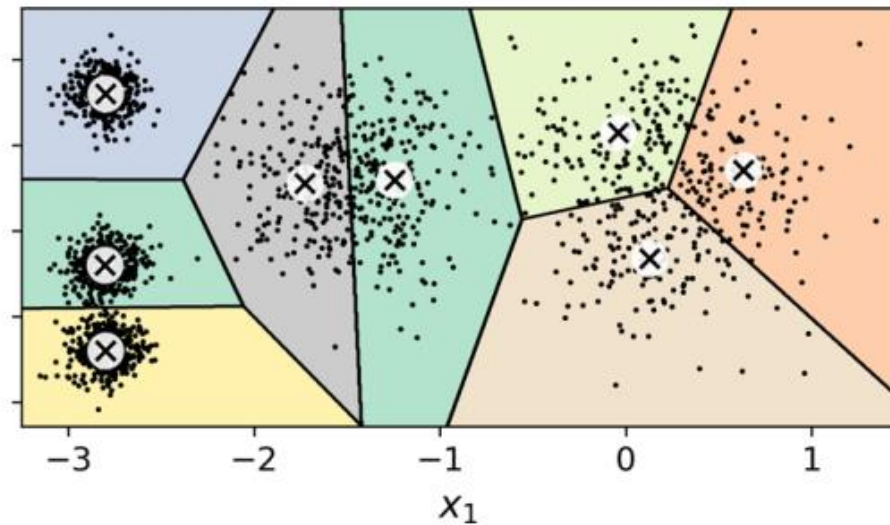


sai

$k = 3$



$k = 8$



Xác lập tâm ban đầu

- PP1: Nếu biết xấp xỉ vị trí tâm điểm ta có thể thiết lập trực tiếp chúng (thực hiện 1 thuật toán phân cụm khác trước đó)
 - Ví dụ trong thông qua tham số *init* và thiết lập *n_init* = 1 trong lớp Kmeans của scikit-learn (chứa danh sách các tâm điểm)

Xác lập tâm ban đầu

- PP2: Chạy thuật toán nhiều lần với các thiết lập ngẫu nhiên khác nhau và giữ lại giải pháp tốt nhất
 - Làm sao để biết được giải pháp nào là tốt nhất?
 - Có thể sử dụng thước đo hiệu quả để biết chính xác giải pháp nào là tốt nhất
 - model's inertia ~ khoảng cách bình phương trung bình giữa mỗi mẫu và tâm gần nhất của nó => càng nhỏ càng tốt

Xác lập tâm ban đầu

- PP3: Sử dụng một cách thiết lập ban đầu khác, ví dụ dùng thuật toán **k-means++**
 - Lựa chọn các tâm điểm xa nhau => thuật toán ít bị hội tụ về nghiệm cận tối ưu hơn

Xác lập tâm ban đầu

- k-means++

- B1: Lấy 1 tâm \mathbf{c}_1 , lựa chọn ngẫu nhiên từ tập dữ liệu
- B2: Xác định tiếp 1 tâm mới bằng cách chọn 1 mẫu \mathbf{x}_i có xác suất lớn nhất

$$p(\mathbf{x}_i) = \frac{d(\mathbf{x}_i)^2}{\sum_j d(\mathbf{x}_j)^2}$$

$d(\mathbf{x}_i)$ Khoảng cách giữa mẫu \mathbf{x}_i và tâm gần nhất đã được chọn

- B3: Lập lại B2 cho tới khi toàn bộ k tâm được chọn

K-means++

- Ví dụ 4: Áp dụng k-means++



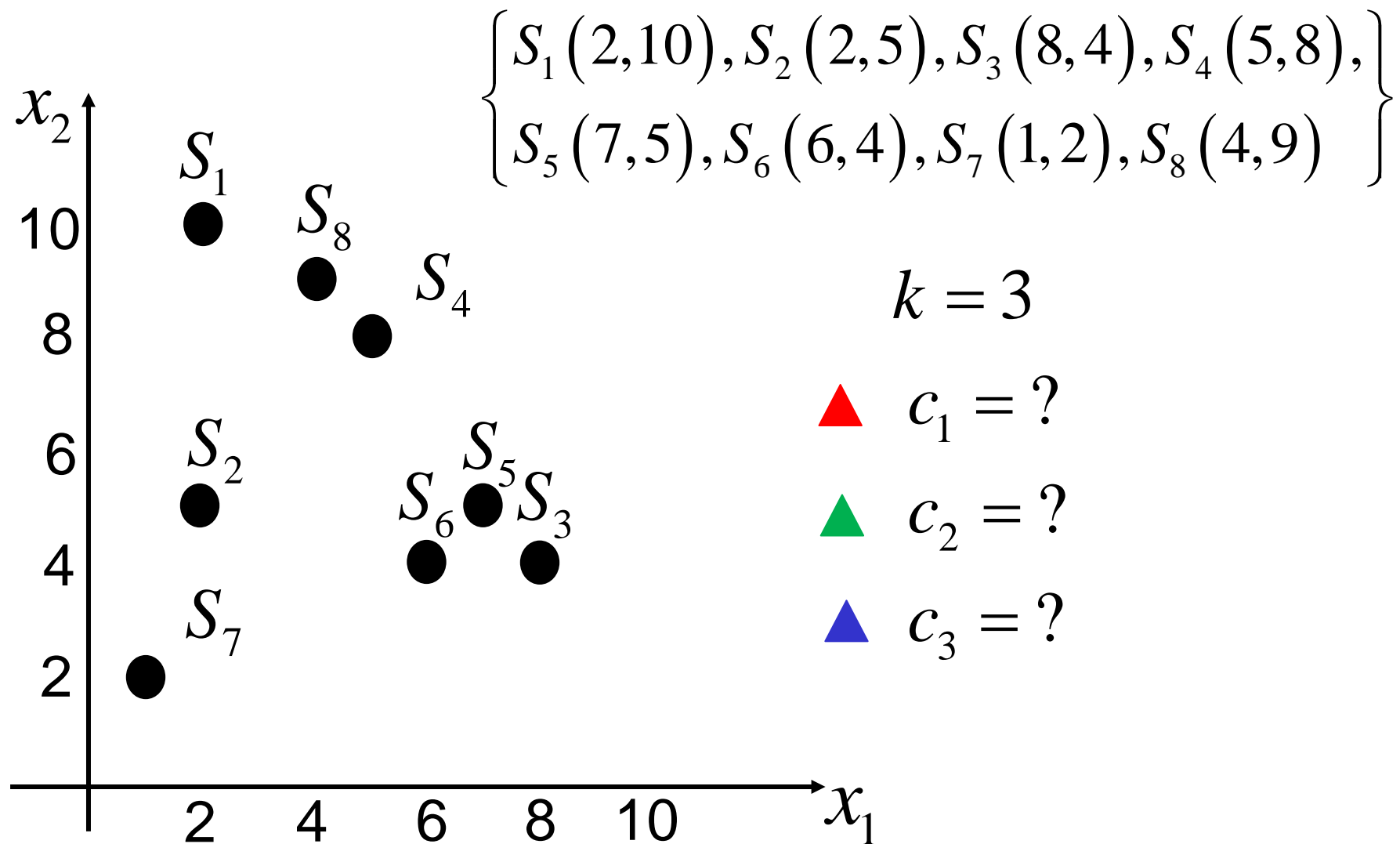
Bộ dữ liệu có $N = 8$ mẫu

$$\left\{ \begin{array}{l} S_1(2, 10), S_2(2, 5), S_3(8, 4), S_4(5, 8), \\ S_5(7, 5), S_6(6, 4), S_7(1, 2), S_8(4, 9) \end{array} \right\}$$

- (a) Trực quan hóa dữ liệu
- (b) Xác định 3 tâm ban đầu
- (c) Xác định 3 cụm sau khi kết thúc thuật toán

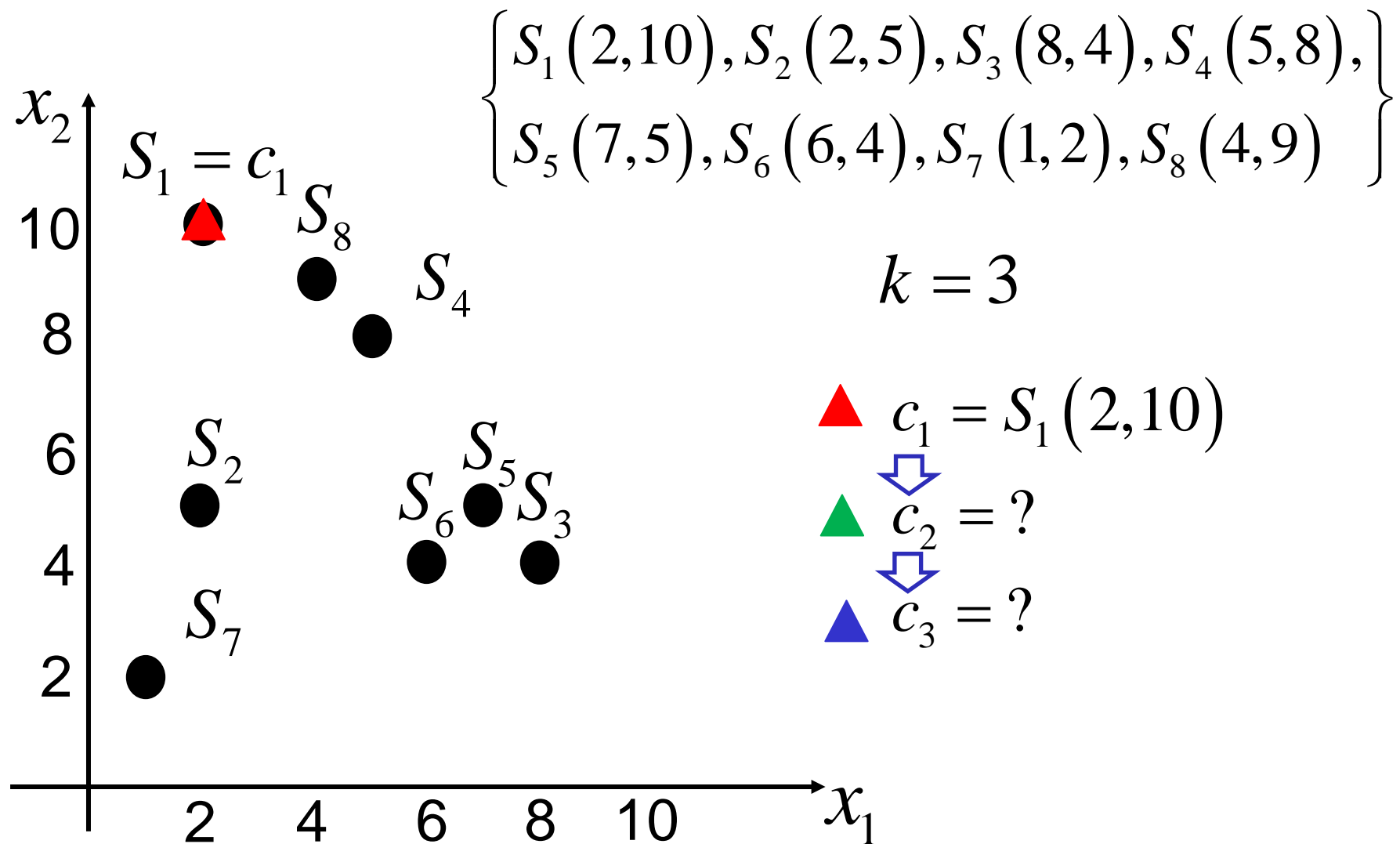
K-means++ – Ví dụ 4

- Ví dụ 4: (a) Trực quan hóa dữ liệu



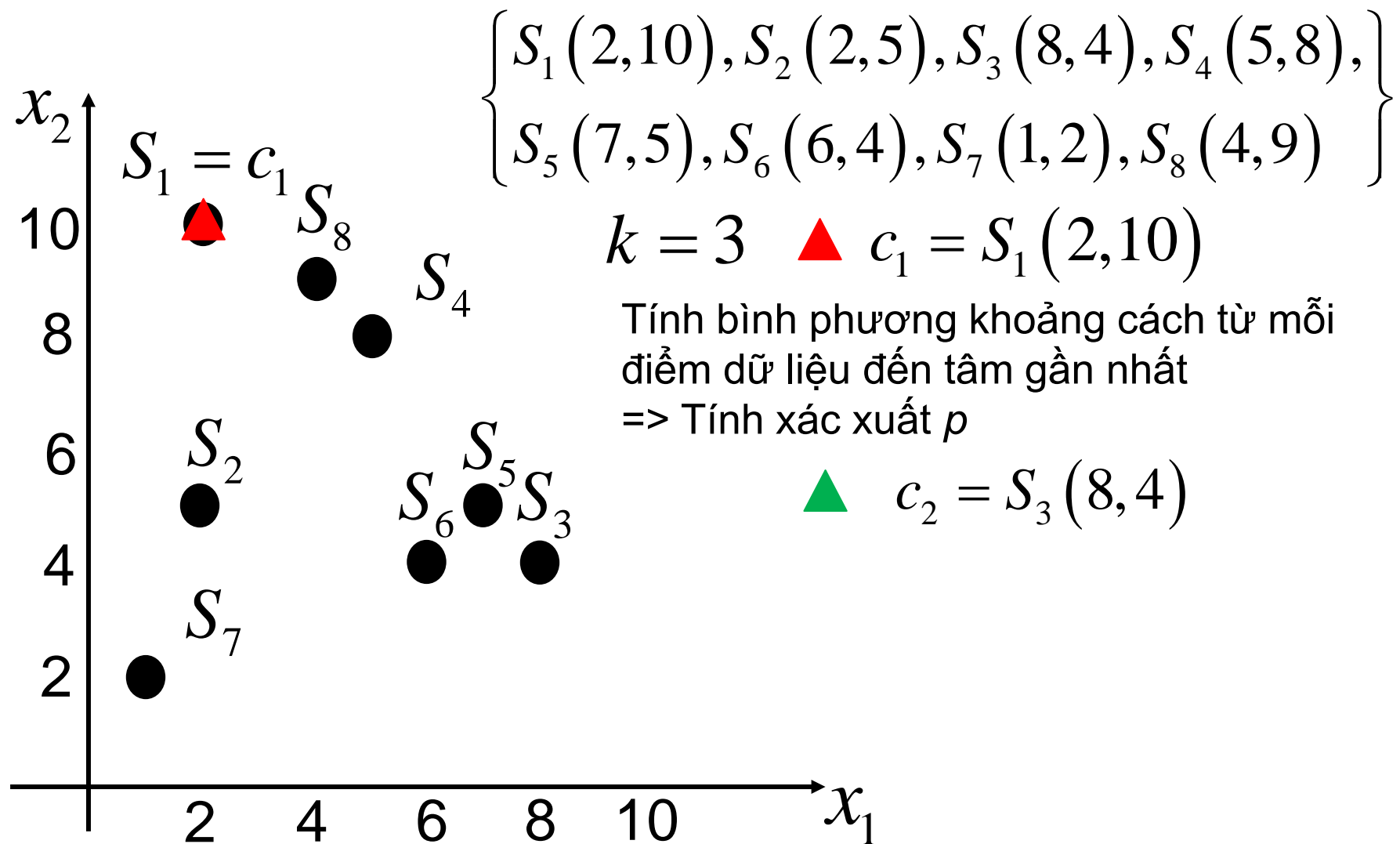
K-means++ – Ví dụ 4

- Ví dụ 4: (b) Xác định 3 tâm ban đầu



K-means++ – Ví dụ 4

- Ví dụ 4: (b) Xác định 3 tâm ban đầu



K-means++ - Ví dụ 4

- Ví dụ 4: (b) Xác định 3 tâm ban đầu

Mẫu	$d(S_i, c_1)^2$	$p(S_i)$
S_1	0	0
S_2	25	0.089
S_3	72	0.255
S_4	13	0.046
S_5	50	0.177
S_6	52	0.184
S_7	65	0.231
S_8	5	0.018

→ ▲ $c_2 = S_3 (8, 4)$


$$\sum_j d(S_j)^2 = 282$$

$$p(S_2) = \frac{d(S_2)^2}{\sum_j d(S_j)^2} = \frac{25}{282} = 0.089$$

K-means++ - Ví dụ 4

- Ví dụ 4: (b) Xác định 3 tâm ban đầu

Mẫu	$d(S_i, c_1)^2$	$d(S_i, c_2)^2$	$d(S_i)^2$	$p(S_i)$
S_1	0	72	0	0
S_2	25	36	25	0.245
S_3	72	0	0	0
S_4	13	25	13	0.125
S_5	50	2	2	0.020
S_6	52	4	4	0.039
S_7	65	53	53	0.520
S_8	5	45	5	0.049

⇒  $c_3 = S_7 (1, 2)$

$$p(S_2) = \frac{d(S_2)^2}{\sum_j d(S_j)^2} = \frac{25}{102} = 0.245$$

$$\sum_j d(S_j)^2 = 102$$

K-Means++ – Ví dụ 4

- Ví dụ 4: (c) Xác định 3 cụm khi thuật toán kết thúc

$$\left\{ S_1(2,10), S_2(2,5), S_3(8,4), S_4(5,8), \right. \\ \left. S_5(7,5), S_6(6,4), S_7(1,2), S_8(4,9) \right\}$$

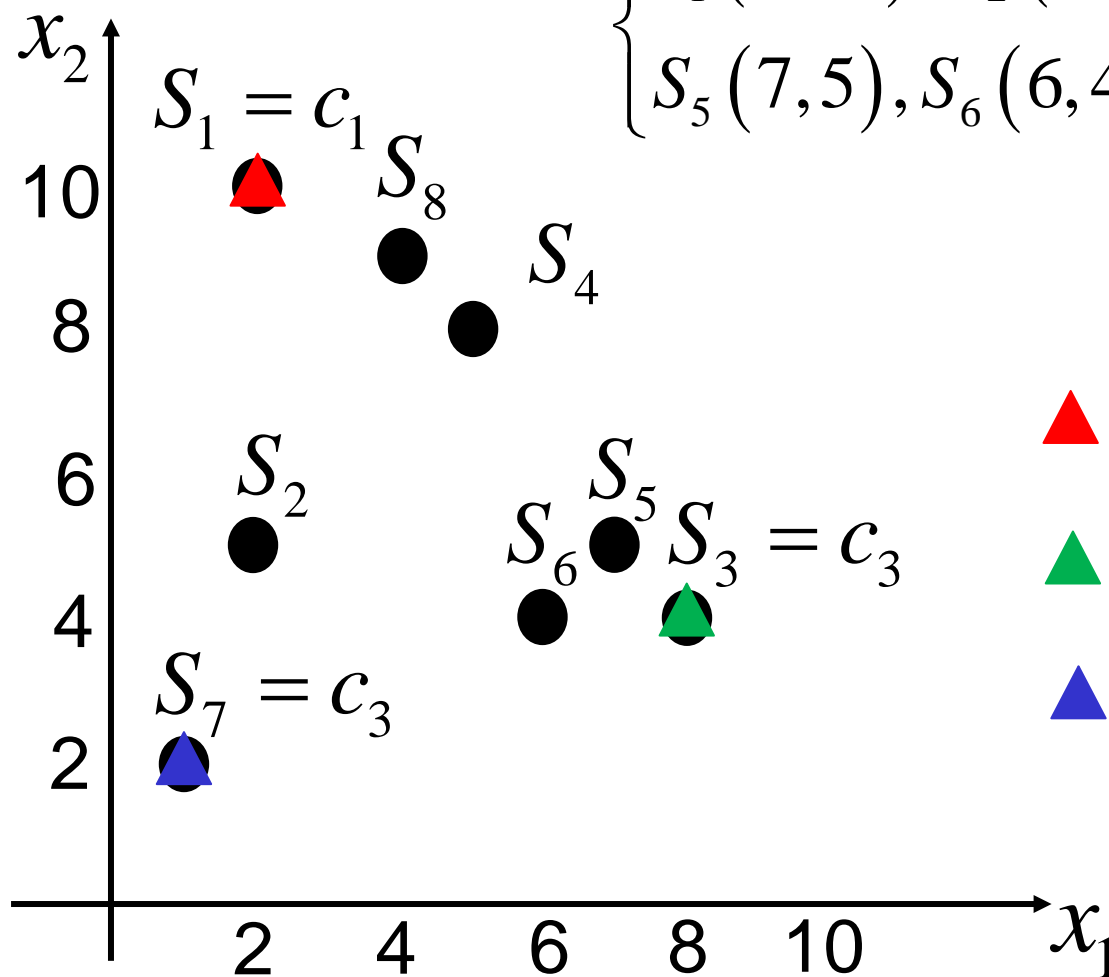
$$k = 3$$

$$\blacktriangle c_1 = S_1(2,10)$$

$$\blacktriangle c_2 = S_3(8,4)$$

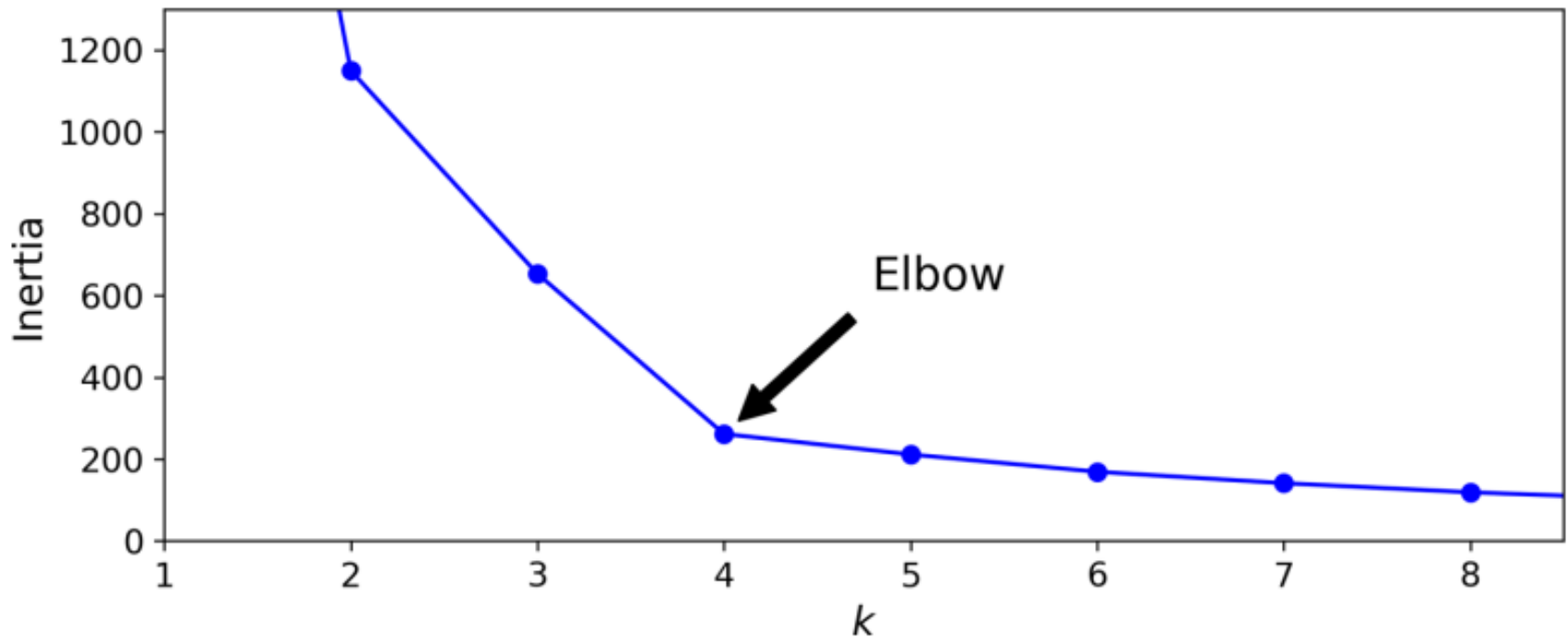
$$\blacktriangle c_3 = S_7(1,2)$$

Áp dụng k-means



Số lượng tối ưu của các cụm

- PP1: Sử dụng model's inertial, tính cho số lượng khác nhau của cụm k



$k = 4?$

Không thực sự chính xác

$k = 5?$

$k = 6?$

Số lượng tối ưu của các cụm

- PP2: Sử dụng silhouette score

- Hệ số silhouette của một mẫu

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

a : khoảng cách trung bình tới các điểm khác ở trong cùng một cụm

b : khoảng cách tới cụm gần nhất trung bình là khoảng cách trung bình tới các mẫu thuộc vào cụm gần nhất

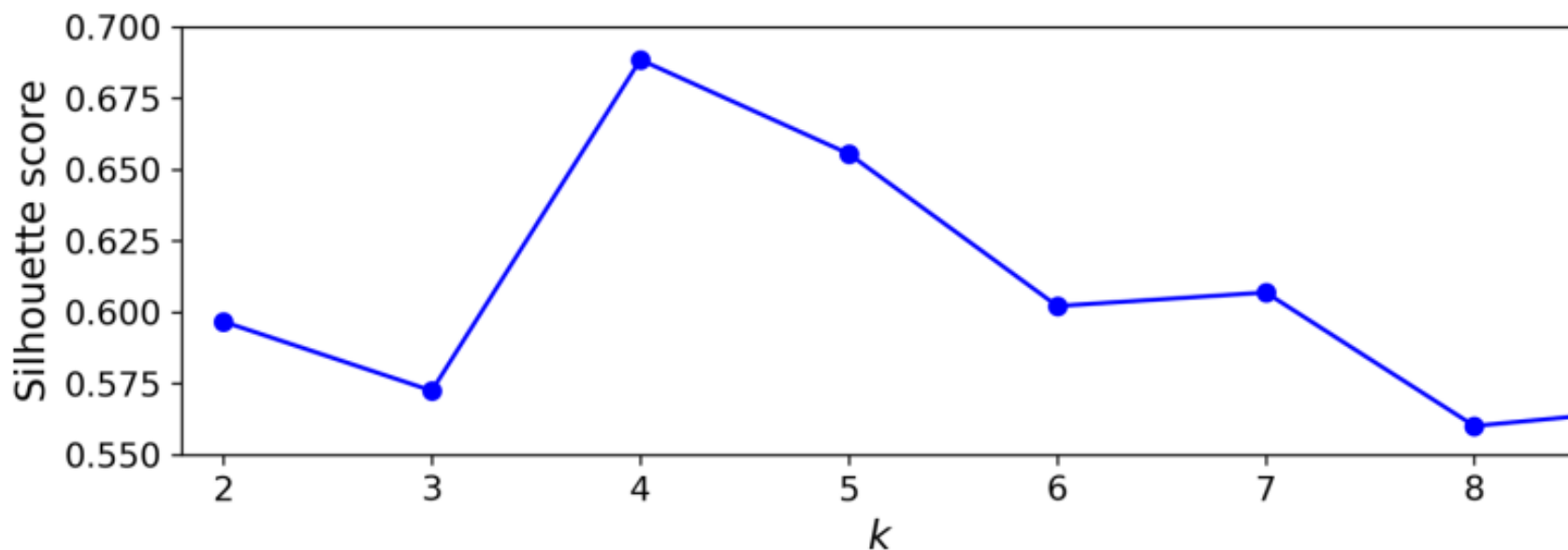
+ Gần với +1: mẫu này nằm bên trong cụm mà nó thuộc về và nằm xa các cụm khác

+ Gần với 0: mẫu này nằm gần với đường biên của cụm

+ Gần với -1: mẫu có thể đã được gán sai vào 1 cụm mà nó không thuộc về

Silhouette score

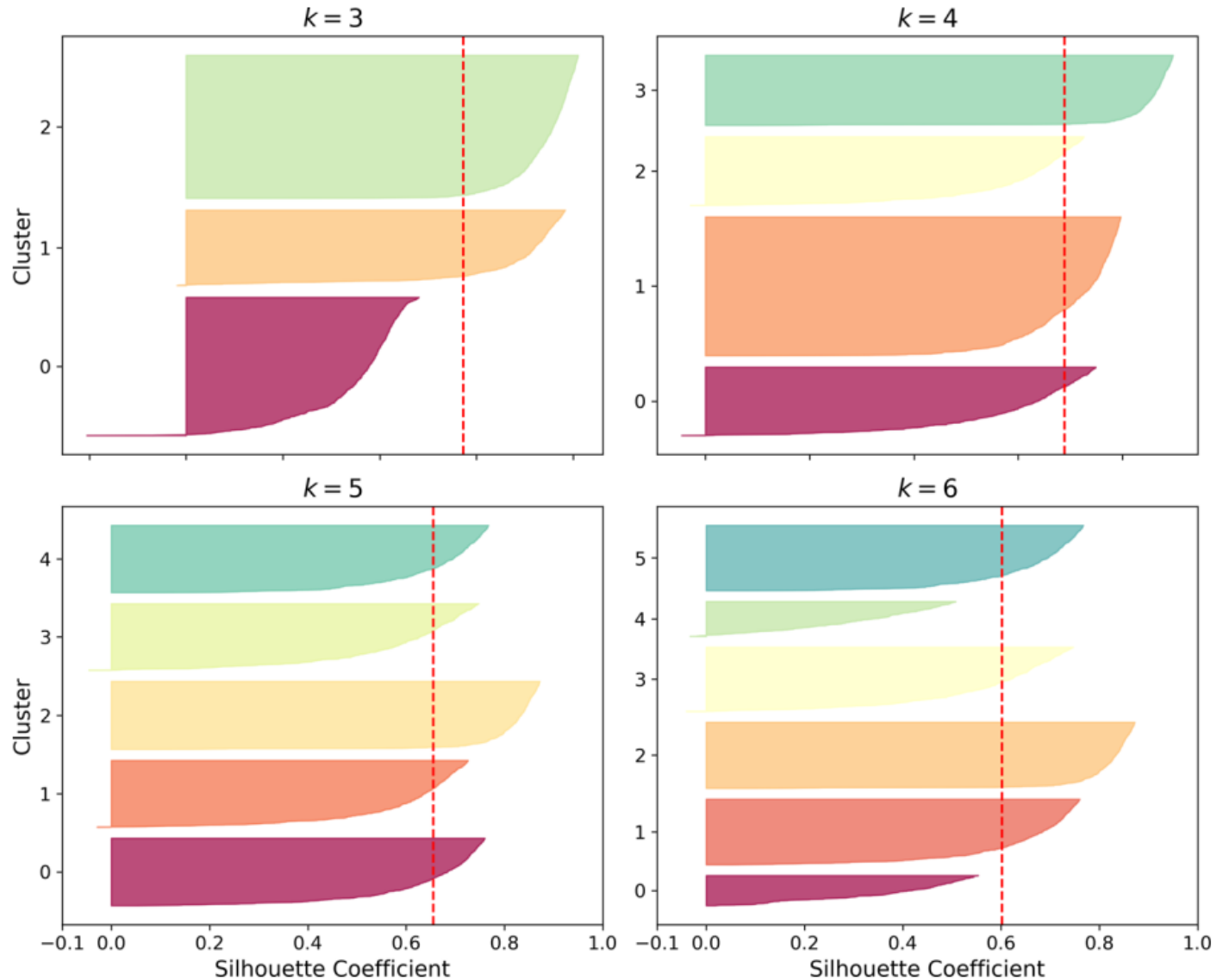
- Silhouette score = trung bình của các hệ số silhouette trên toàn bộ các mẫu



+ Biểu đồ silhouette: Biểu diễn tất cả các hệ số silhouette của các mẫu sắp xếp theo từng cụm

Kết hợp với 1 đường thẳng biểu diễn giá trị silhouette score

Sihouete diagrams



$k = 5?$

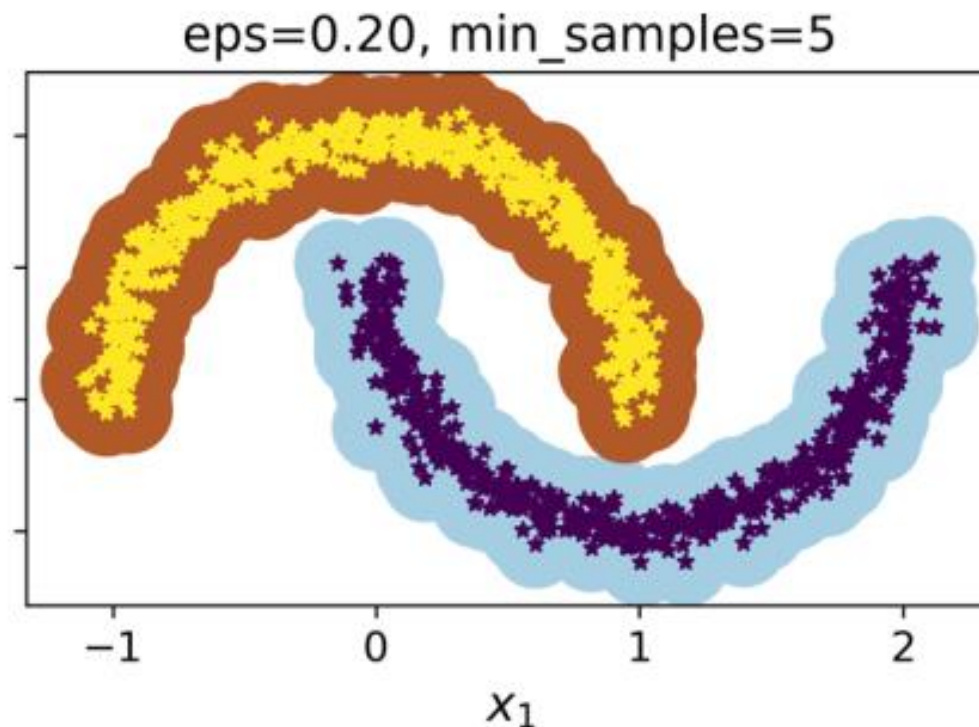
DBSCAN

- DBSCAN (Density-based spatial clustering of applications with noise)

Thuật toán này định nghĩa các cụm như là các vùng liên tục với mật độ cao

DBSCAN

- Thuật toán đơn giản nhưng mạnh mẽ, cho phép phân cụm cho bất kỳ hình dạng nào

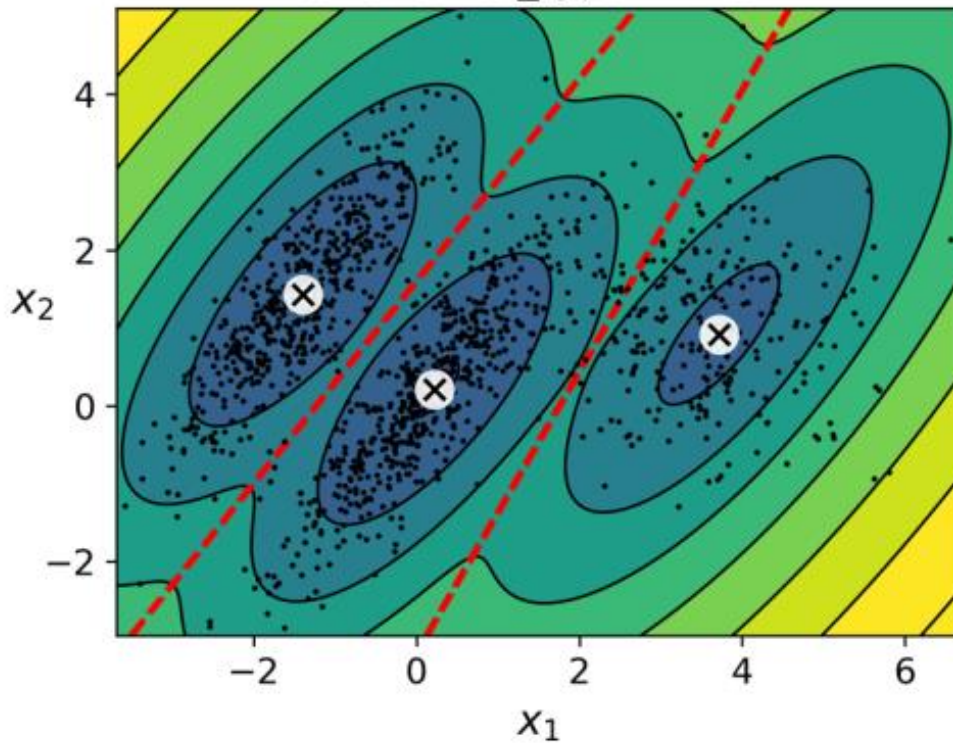


Gaussian mixture

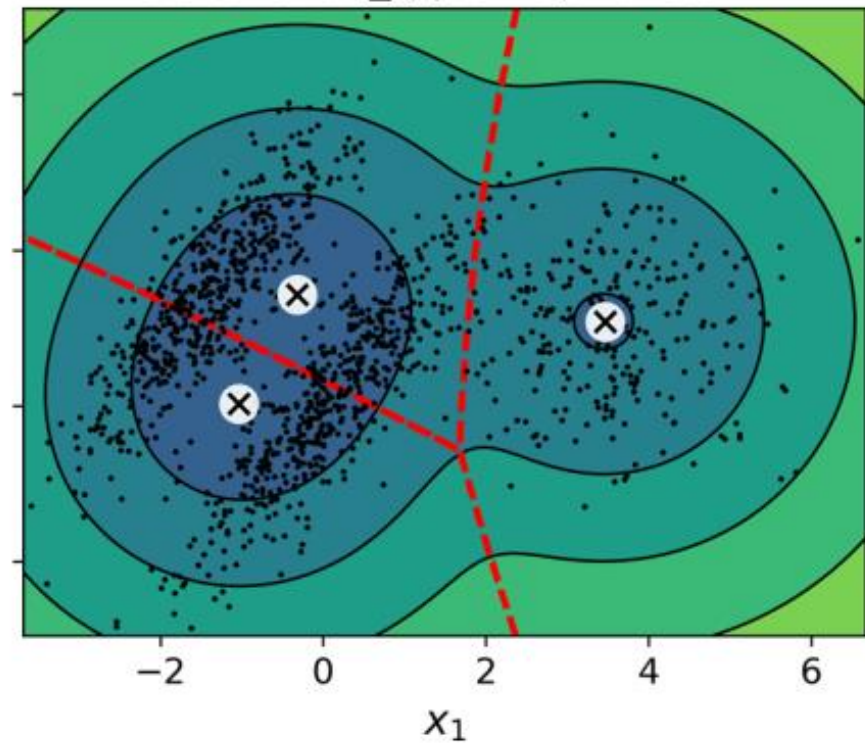
- GMM (Gaussian mixture model)
 - Là mô hình xác suất với giả sử rằng các mẫu được tạo ra từ sự pha trộn của một số phân phối Gaussian (có các tham số ta không biết)
 - Tất cả các mẫu tạo bởi cùng 1 phân phối Gaussian đơn hợp thành 1 cụm, thông thường có dạng elip

Gaussian mixture

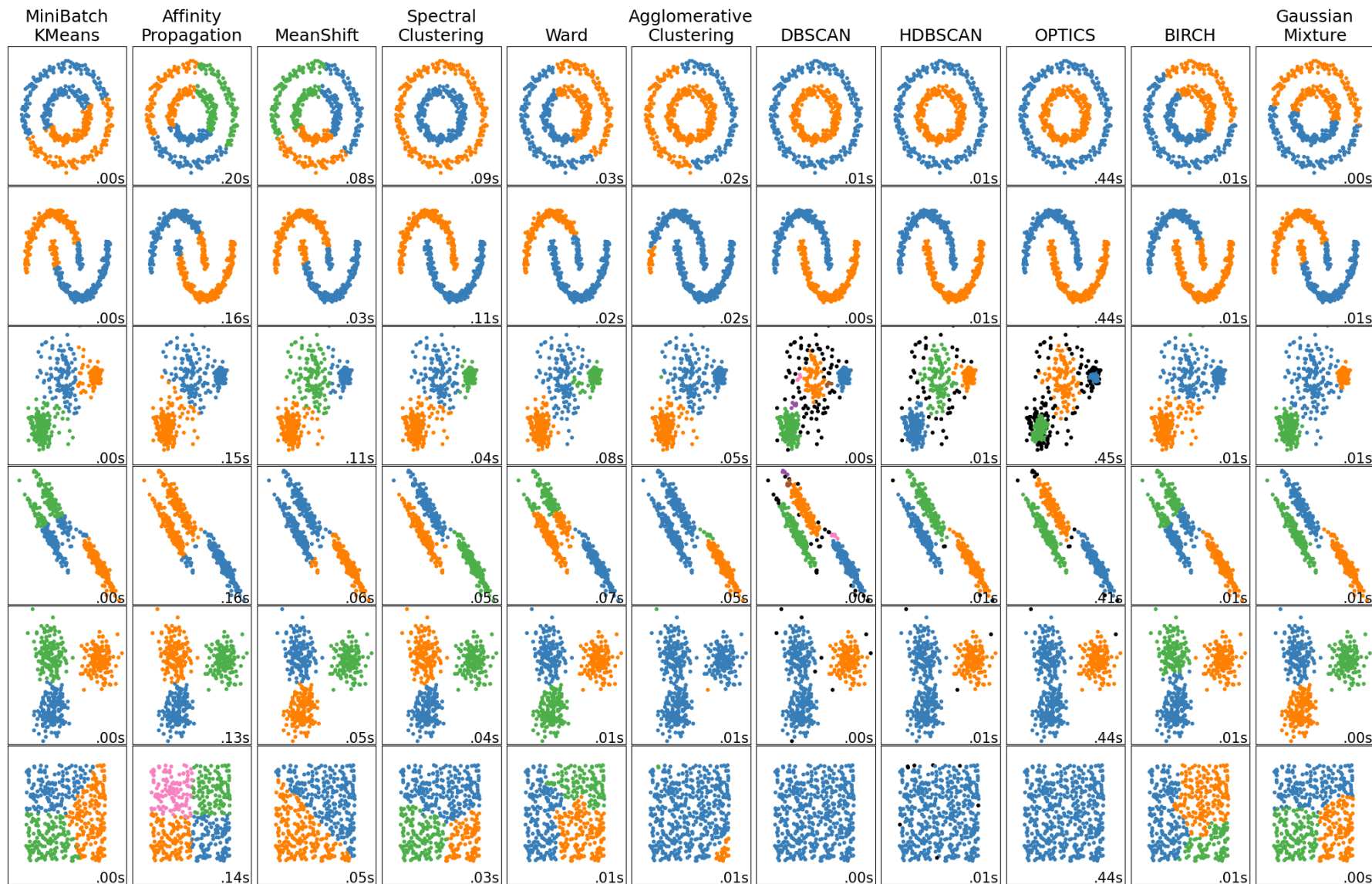
covariance_type="tied"



covariance_type="spherical"



Các thuật toán phân cụm khác



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Tổng kết

- Sinh viên hiểu và áp dụng được thuật toán k-means cho phân cụm
- Chú ý tới các yếu tố ảnh hưởng tới thuật toán như
 - Việc thiết lập các tâm ban đầu
 - Lựa chọn số lượng tối ưu của các cụm

Hoạt động sau buổi học

- Tìm hiểu thêm các thuật toán phân cụm khác

Chuẩn bị cho buổi học tiếp theo

- Tìm hiểu về các hàm để thực hiện phân cụm trong thư viện scikit-learn

Tài liệu tham khảo

- [B10TLTK1] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, 3rd Edition, 2011.
 - Chapter 10 Cluster Analysis: Basic Concepts and Methods
- [B10TLTK2] B. Liu, Web Data Mining – Exploring Hyperlinks, Contents, and Usage Data, Springer, 2nd Edition, 2011
 - Chapter 4 Unsupervised Learning