Học máy không giám sát

Nguyễn Thanh Tùng, **Trần Thị Ngân**
Khoa Công nghệ thông tin – Đại học Thủy lợi
tungnt@tlu.edu.vn, ***ngantt@tlu.edu.vn***

Học máy không giám sát

- ***Học không giám sát:*** tập các công cụ thống kê xử lý dữ liệu chỉ có biến đầu vào, không có biến đích
  - Ta chỉ có các biến X mà không có các nhãn Y
  - Mục tiêu: phát hiện các mẫu/các đặc tính của dữ liệu
    - vd. trực quan hóa hoặc diễn giải dữ liệu nhiều chiều

HỌC CÓ GIÁM SÁT VS. KHÔNG GIÁM SÁT

Học có giám sát

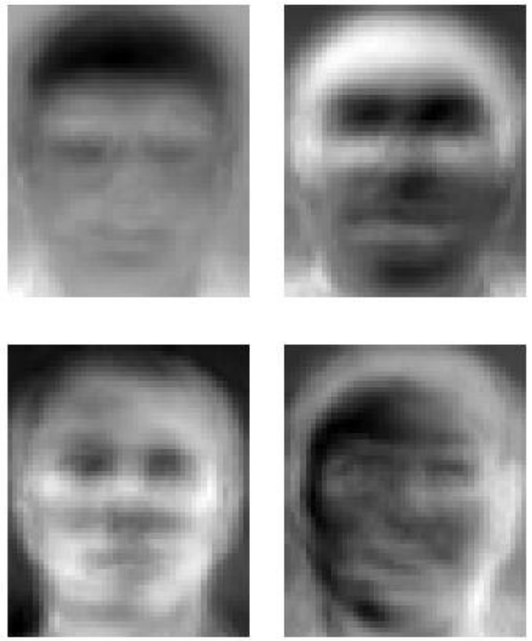
- Input: Cả X và Y đều đã biết
- Output: Dự đoán một kết quả (phân lớp/dự báo)

Học không giám sát

- Input: Chỉ biết X, không biết Y
- Output: Tạo ra một biến hoặc nhãn mới, (phân cụm/giảm chiều)

Học không giám sát

- Ví dụ ứng dụng:
  - Cho một tập các tài liệu văn bản, cần xác định tập các tài liệu có chung chủ đề như thể thao, chính trị, ca nhạc,..
  - Cho các ảnh khuôn mặt có số chiều cao (high dimension), tìm một biểu diễn đơn giản/thu gọn của các ảnh này để đưa vào bộ phân lớp nhận dạng khuôn mặt



(AT&T Laboratories Cambridge)

Học không giám sát

- **Tại sao học không giám sát luôn là thách thức lớn?**
  - Phân tích khám phá dữ liệu (Exploratory data analysis) – mục tiêu không được định nghĩa rõ ràng
  - Khó đánh giá hiệu năng – không biết được đáp án đúng (“right answer” unknown)
  - Xử lý dữ liệu với số chiều lớn

Sắp xếp các đối tượng vào các nhóm (clustering)

Có bao nhiêu clusters

Có bao nhiêu clusters?

2 clusters

4 clusters

6 clusters

Tuỳ thuộc vào "resolution" !

Học không giám sát

- **Hai cách tiếp cận:**
  - *Phân tích cụm (Cluster analysis)*
    - Xác định các nhóm mẫu đồng nhất (có các đặc tính chung)
  - *Giảm chiều dữ liệu (Dimensionality Reduction)*
    - Tìm cách biểu diễn với số chiều thấp hơn dựa trên tính chất và trực quan hóa dữ liệu

Các lĩnh vực ứng dụng

- ✓ **Marketing:** Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn;
- ✓ **Biology:** Phân nhóm động vật và thực vật dựa vào các thuộc tính của chúng;

Các lĩnh vực ứng dụng

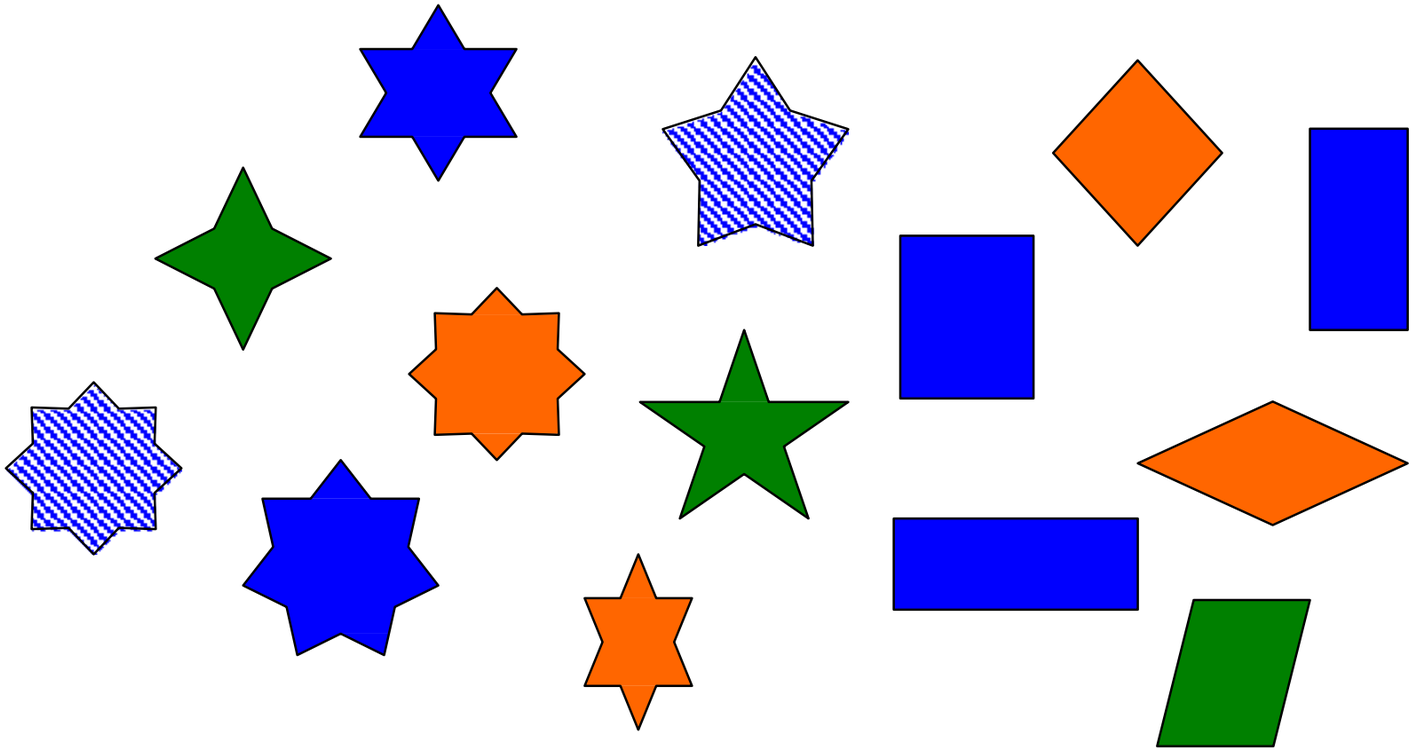
- ✓ **Libraries:** Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…;
- ✓ **Insurance, Finance:** Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds);
- ✓ **WWW:** Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog);…

Phân tích cụm & K-means

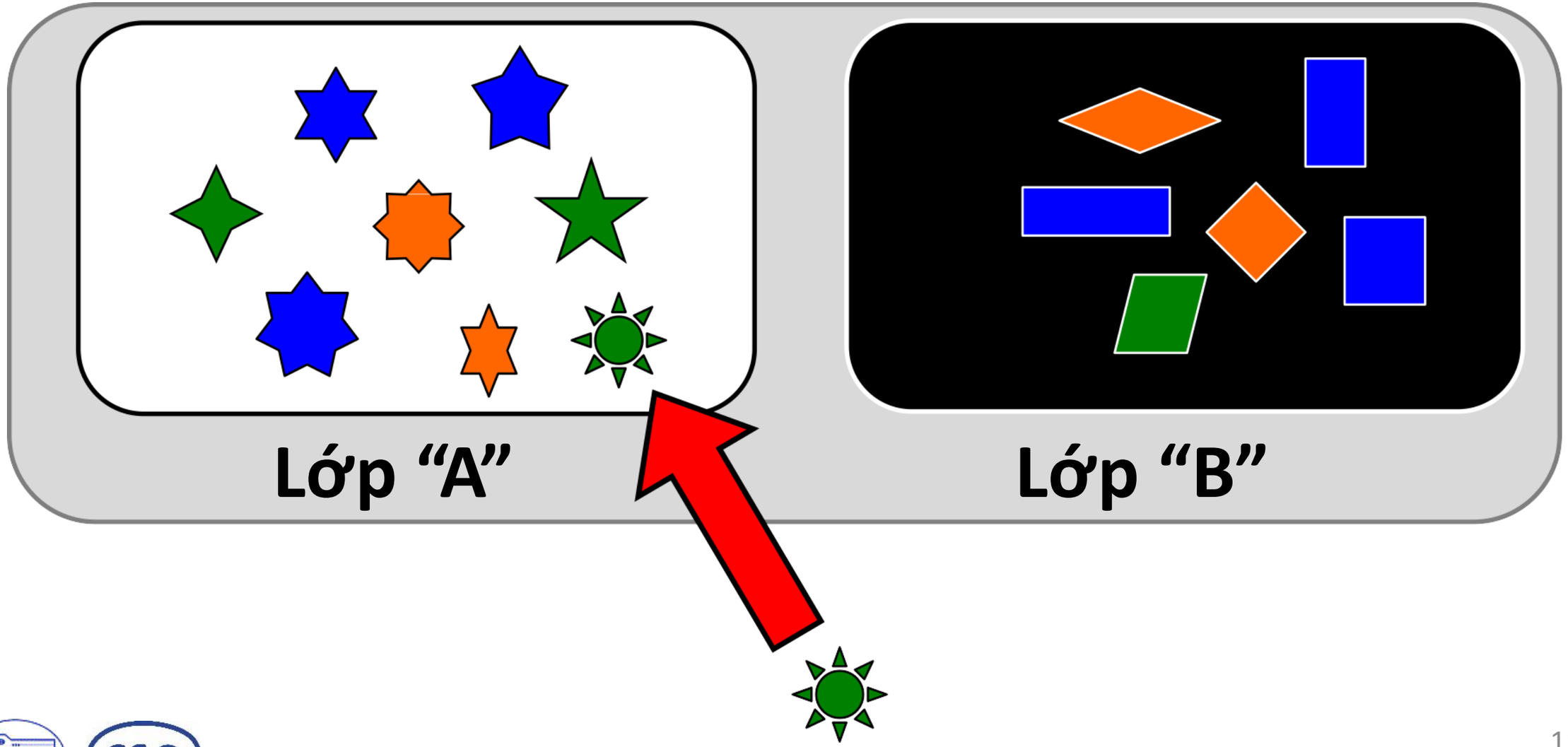
Phân tích cụm (Phân cụm)

- *Phân cụm (clustering):* là tập các phương pháp nhằm tìm ra các nhóm con trong dữ liệu
  - Các mẫu có đặc điểm chung trong cùng 1 nhóm nhưng khác với các mẫu ở ngoài nhóm
  - Việc gom nhóm là phân tích cấu trúc dữ liệu nội tại, điều này khác với phân lớp

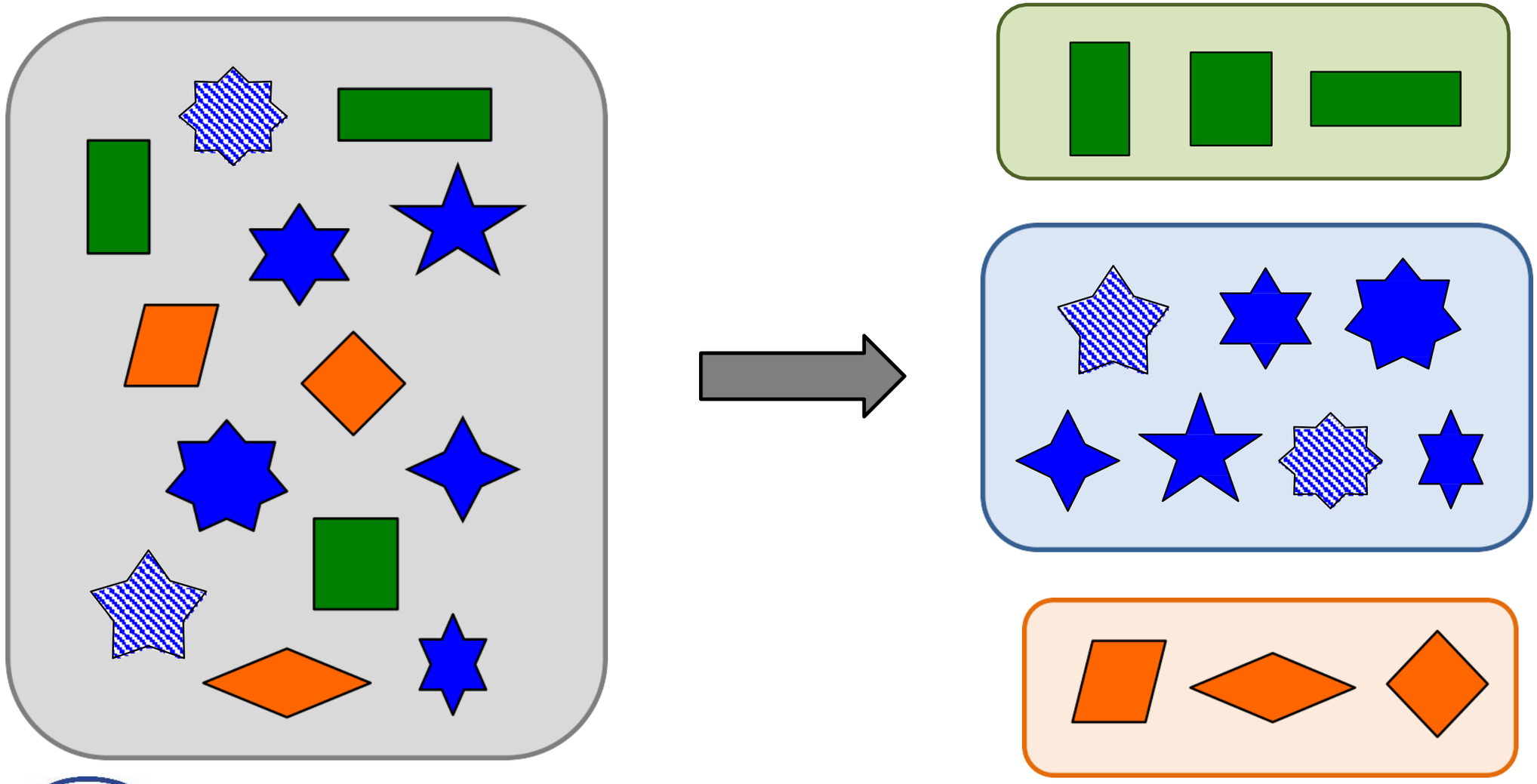
Phân cụm vs. Phân lớp



Phân lớp



Phân cụm



Phân cụm là gì?

➢ Là quá trình phân chia 1 tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn:

- Các đối tượng trong 1 cụm "tương tự" nhau.
- Các đối tượng khác cụm thì "không tương tự" nhau.

➢ Mục đích: giải quyết vấn đề tìm kiếm, phát hiện các cụm, các mẫu dữ liệu trong 1 tập hợp ban đầu các dữ liệu không có nhãn.

Các mô hình phân cụm

Hai mô hình phân cụm thông dụng:

- Phương pháp dựa trên tâm cụm (Centroid-based)
- Phương pháp phân cấp (Hierarchical)

Phân cụm

Các mô hình khác:

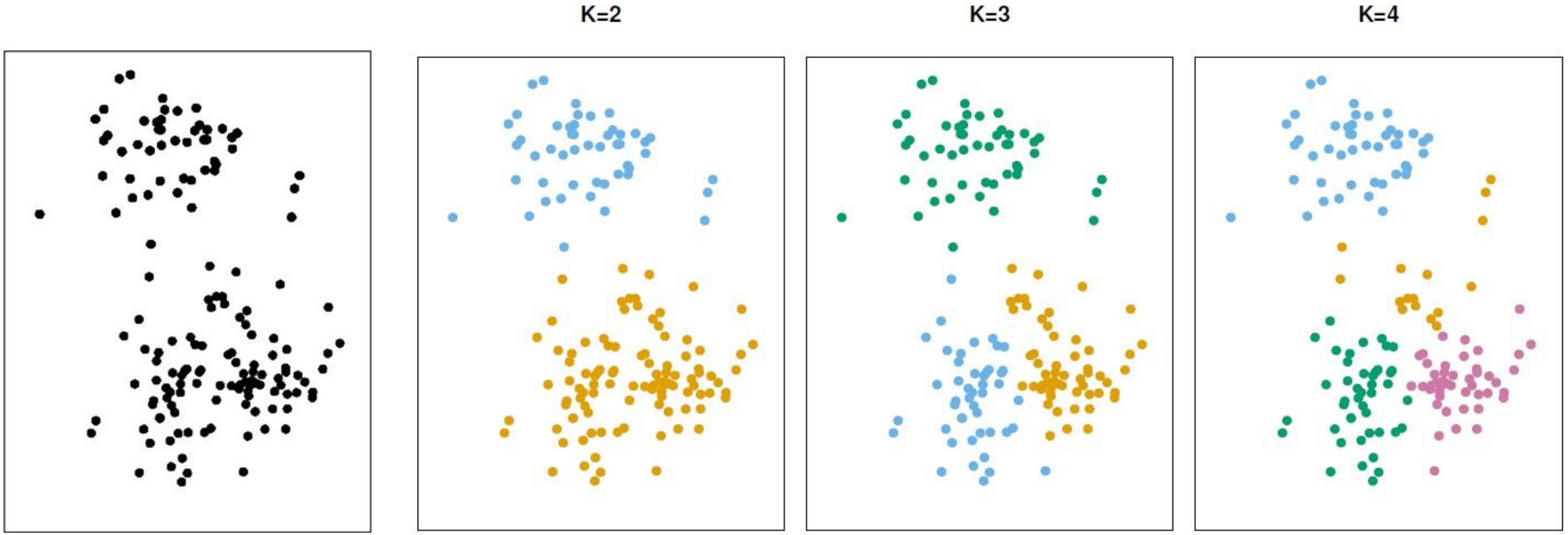
- Phân cụm dựa trên mô hình (Model-based)
  - Mỗi cụm được thể hiện bằng một phân bố thống kê tham số
  - Dữ liệu là một hỗn hợp các phân bố
- Phân cụm mờ
  - Cứng (Hard): Các mẫu được chia thành các cụm riêng biệt
  - Mờ (fuzzy): Các mẫu có thể thuộc nhiều hơn 1 cụm với một độ thuộc nhất định.

Phân cụm K–means

- Gom nhóm dữ liệu thành K cụm riêng biệt
  - Mỗi cụm K được định nghĩa bởi 1 véc tơ tâm cụm (centroid - là giá trị trung bình của tất cả các đối tượng trong cụm)
  - Mỗi đối tượng gán cho 1 cụm đơn (tâm cụm gần nhất)
  - Yêu cầu số lượng cụm đầu vào K
  - "Phân cụm tốt" là cực tiểu sự biến đổi giữa các cụm

    "Tính tương tự (Similarity)" trong K-means được xác định theo khoảng cách Euclidean

Phân cụm K–means



Đo lường khoảng cách (distance)

- Euclidean $\sqrt{[\Sigma (y - x)^2}$
- Squared Euclidean $\Sigma (y - x)^2$
- City-Block $\Sigma \mid y - x \mid$
- Chebychev $\max \mid y - x \mid$
- Cosine $\cos r_{xy}$

Khoảng cách Euclide

Ví dụ: Cho 2 điểm X(2, 0) và Y(-2, -2). Tính khoảng cách Euclide giữa 2 điểm này

Ta có:

$$\text{Euclidean}(X,Y) = \sqrt{[(-2-2)^2 + (-2-0)^2]}$$

$$= \sqrt{(4^2 + 2^2)} = \sqrt{20} = 4.47$$

Khoảng cách giữa X và Y = 4.47

Squared Euclidean

Ví dụ: Cho 2 điểm X(2, 0) và Y(-2, -2). Tính khoảng cách Squared Euclidean giữa 2 điểm này

Ta có:

$$\text{Squared\_Euclidean}(X,Y) = (-2-2)^2 + (-2-0)^2$$

$$= 4^2 + 2^2 = 20$$

Khoảng cách giữa X và Y = 4.47

Các cách tính toán distance khác

	X	Y	
(1)	4	4	v_1
(2)	8	4	v_2
(3)	15	8	v_3
(4)	24	4	v_4
(5)	24	12	v_5

Chuẩn 1:

$$\|v_3 - v_2\|_1 = \delta_1(v_3, v_2)$$

$$= |15-8| + |8-4|$$

$$= 7 + 4 = 11$$

Chuẩn 2:

$$\|v_3 - v_2\|_2 = \delta_2(v_3, v_2)$$

$$= [(15-8)^2 + (8-4)^2]^{1/2}$$

$$= 65^{1/2} \sim 8.062$$

Chuẩn vô cùng:

$$\|v_3 - v_2\|_\infty = \delta_\infty(v_3, v_2)$$

$$= \max(|15-8|, |8-4|)$$

$$= \max(7,4) = 7$$

Ví dụ với 5 điểm: Euclidean distance

Cho 5 điểm với các tọa độ tương ứng. Tính khoảng cách Euclide giữa 5 điểm này

	X	Y
(1)	4	4
(2)	8	4
(3)	15	8
(4)	24	4
(5)	24	12

	(1)	(2)	(3)	(4)	(5)
(1)	-	**4.0**	11.7	20.0	21.5
(2)		-	8.1	16.0	17.9
(3)			-	9.8	9.8
(4)				-	8.0
(5)					-

Ma trận tương tự (Similarity matrix)

(1)	(2)	(3)	(4)	(5)	
*	4.0	11.7	20.0	21.5	(1)
	*	8.1	16.0	17.9	(2)
		*	9.8	9.8	(3)
			*	8.0	(4)
				*	(5)

Mục đích của phân cụm

- Xác định được bản chất của việc nhóm các đối tượng trong 1 tập dữ liệu không có nhãn.
- Phân cụm không dựa trên 1 tiêu chuẩn chung nào, mà dựa vào tiêu chí mà người dùng cung cấp trong từng trường hợp.

Phân cụm K–means

- Các tâm cụm cực tiểu sự biến đổi giữa các cụm

$$J = \frac{1}{n} \sum_{i=1}^{K} \sum_{x \in c_i} |x - \mu_i|^2 \longrightarrow \text{MIN}$$

  - Các tâm cụm (trung tâm của cụm): $\mu_i = \frac{1}{|c_i|} \sum_{x \in c_i} x$

- Bài toán cực tiểu hóa này là tối ưu tổ hợp

  Giải pháp cho cực tiểu hóa địa phương ta sử dụng phương pháp lặp

Thuật toán K–means

Input

- Tập mẫu $X = \{x_i | i = 1, 2, \ldots, N\}$, $x_i \in R^d$
- Số cụm: K

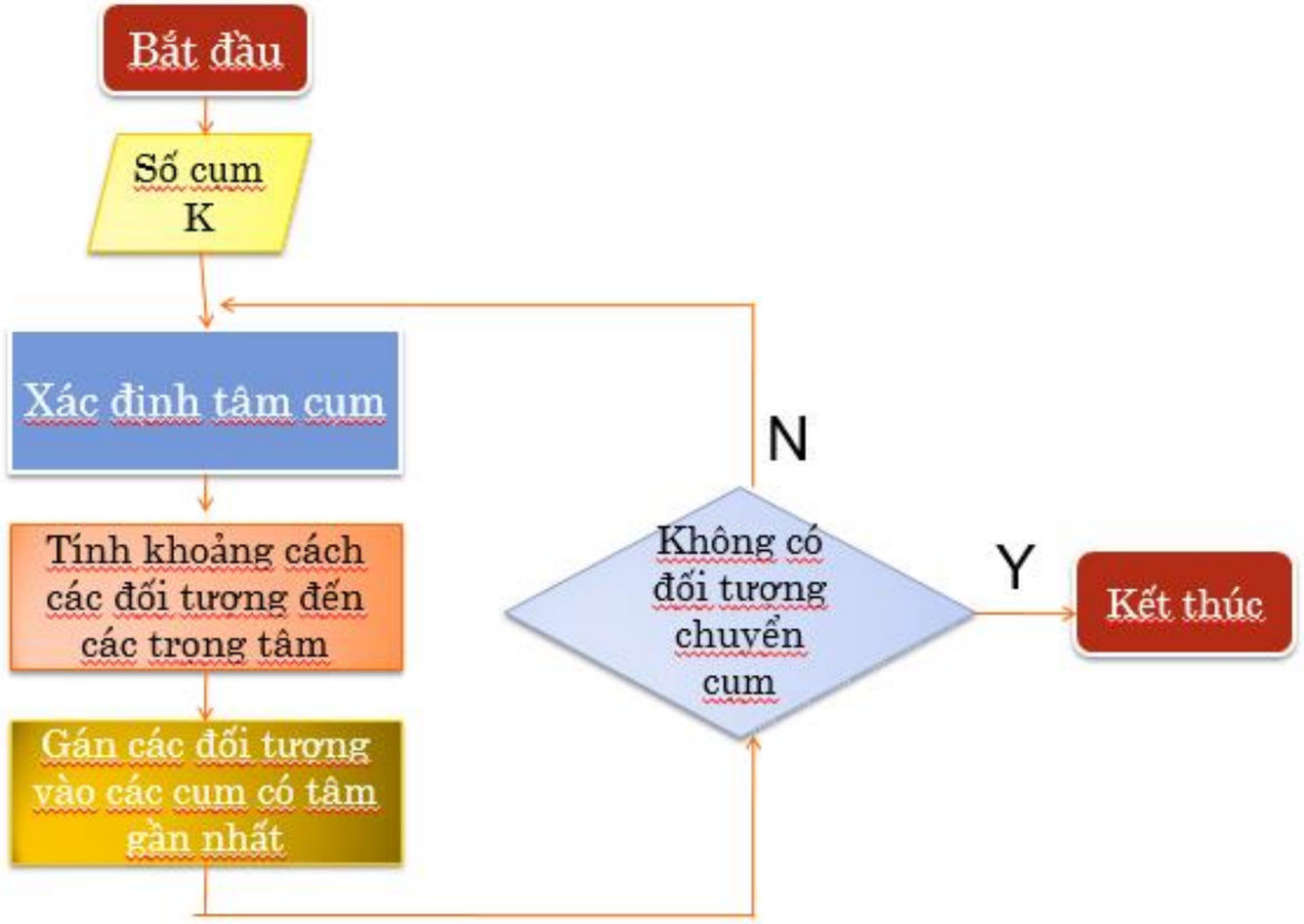
Output

Các cụm C_k ($k = 1 \div K$) tách rời và hàm mục tiêu J đạt cực tiểu

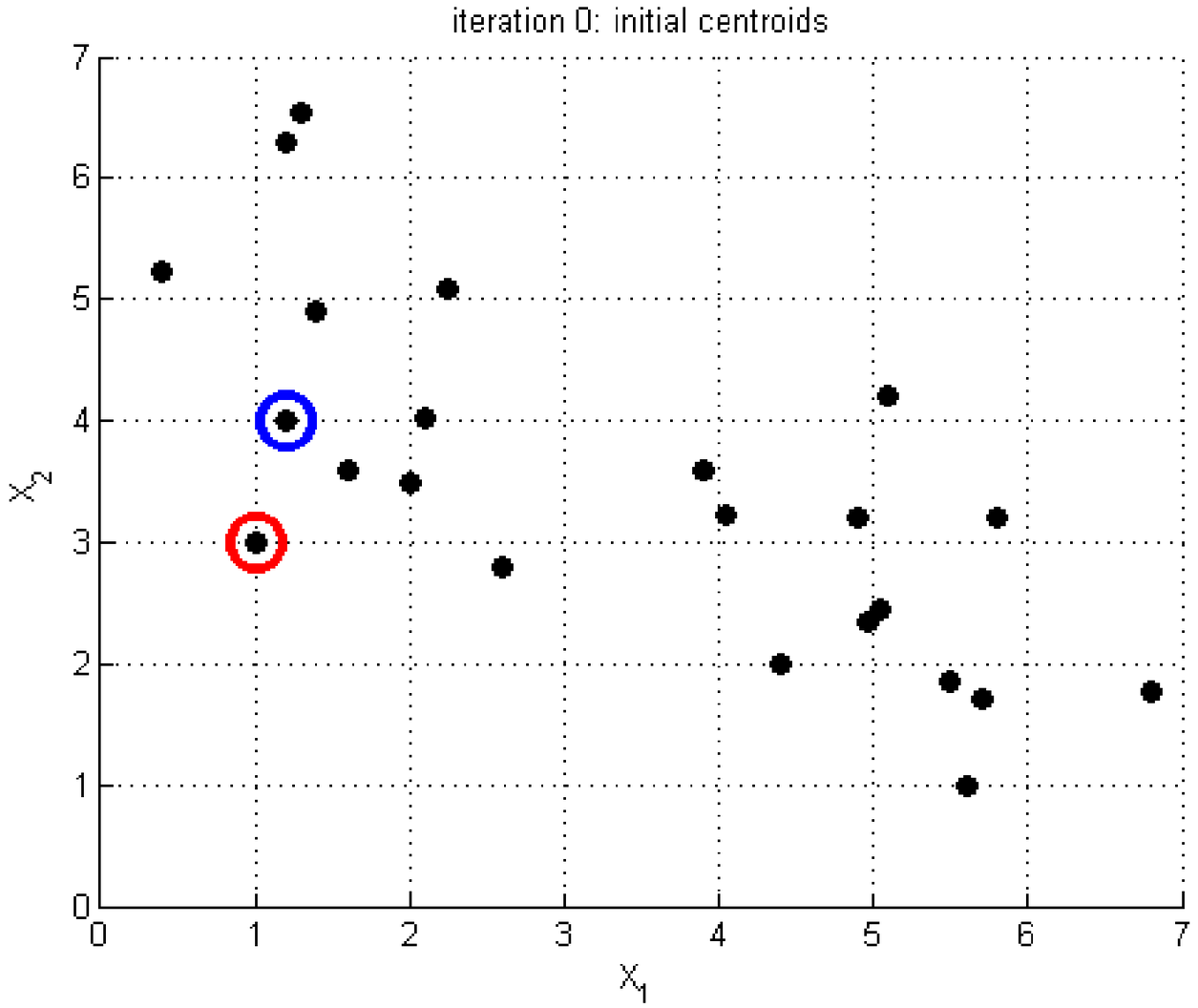
Thuật toán K--means

1) Khởi tạo: Chọn **ngẫu nhiên** K tâm cụm
2) Tính toán khoảng cách từ các đối tượng đến các tâm để phân hoạch dữ liệu (bằng cách gán mỗi đối tượng vào cụm mà nó gần tâm nhất)
3) Tính lại các tâm cụm mới trong mỗi cụm
4) Lặp lại 2 và 3 cho đến khi "thỏa mãn điều kiện" (khi các tâm cụm ổn định và các đối tượng không dịch chuyển giữa các cụm)

Thuật toán K--means

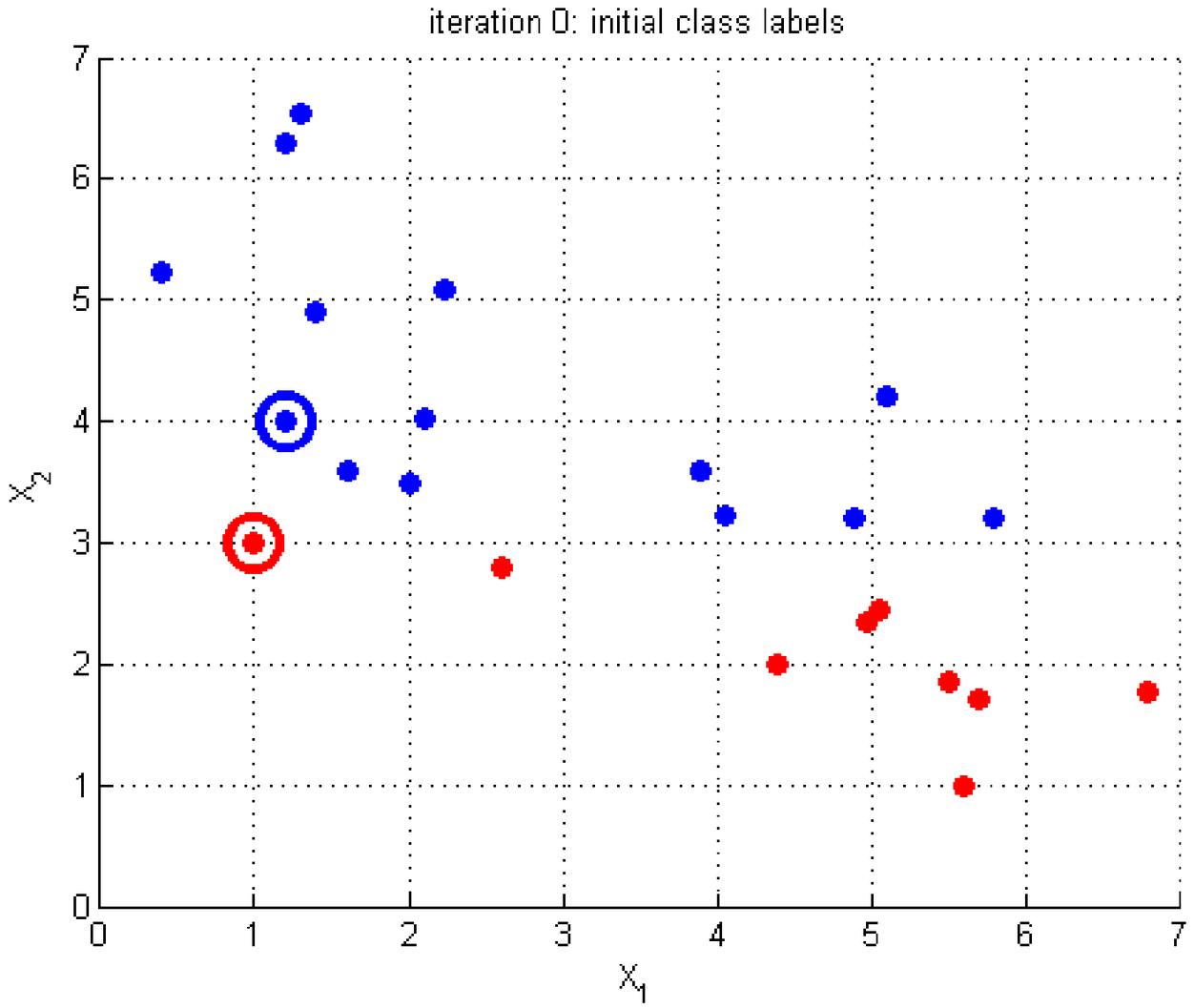


Thuật toán K–means



Khởi tạo tâm cụm

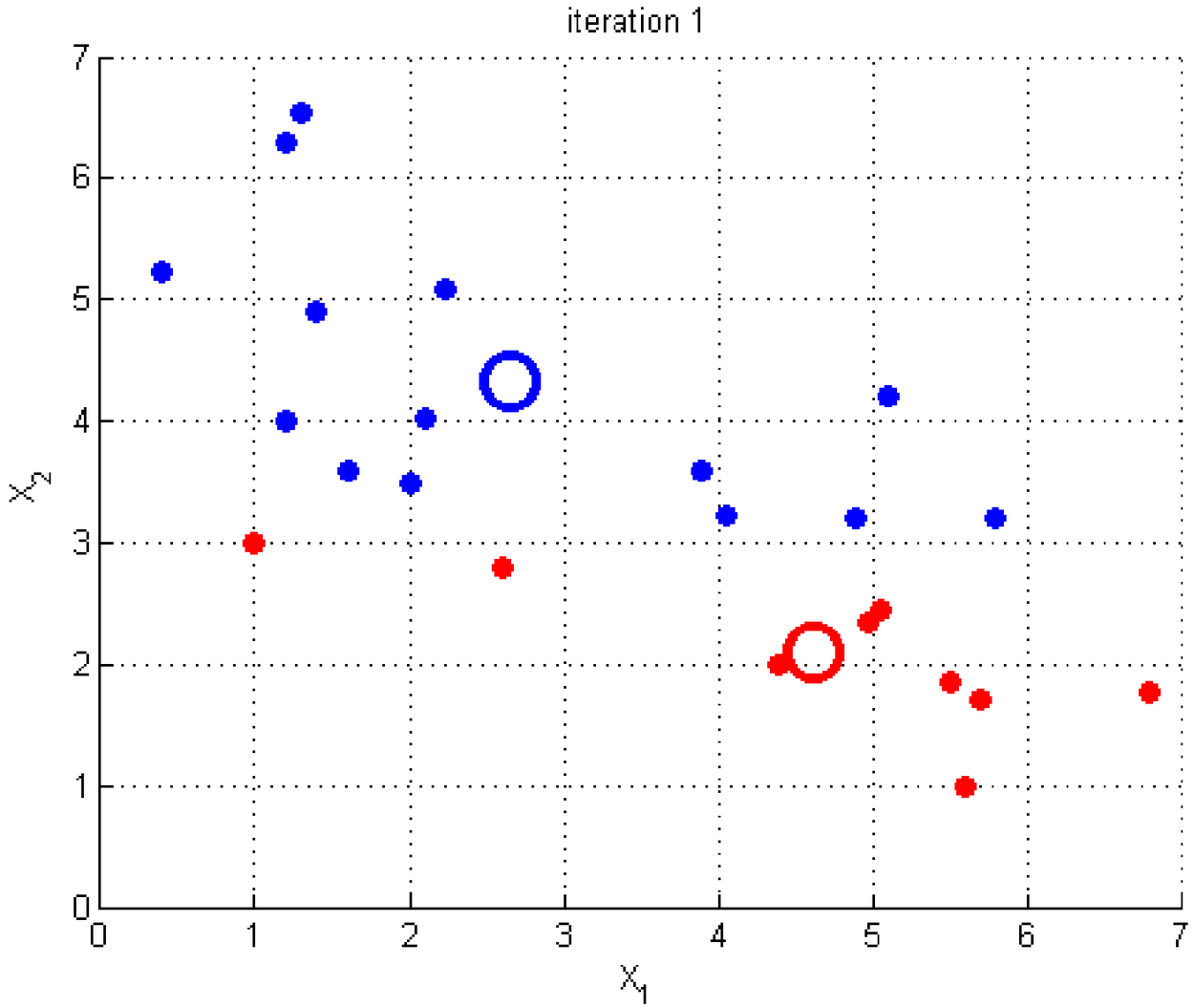
Thuật toán K–means



Khởi tạo tâm cụm

Gán các cụm ban đầu

Thuật toán K–means

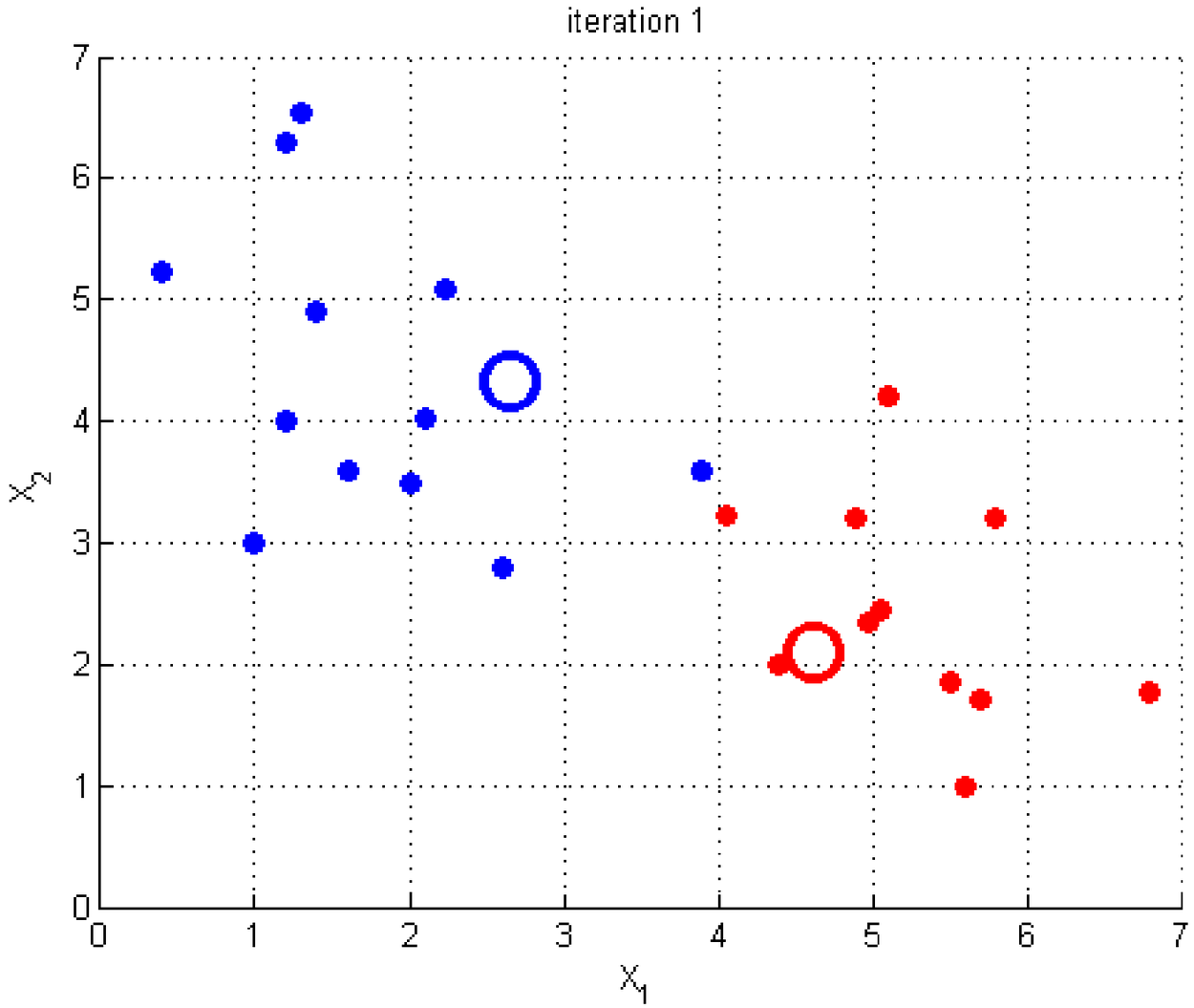


Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Thuật toán K–means



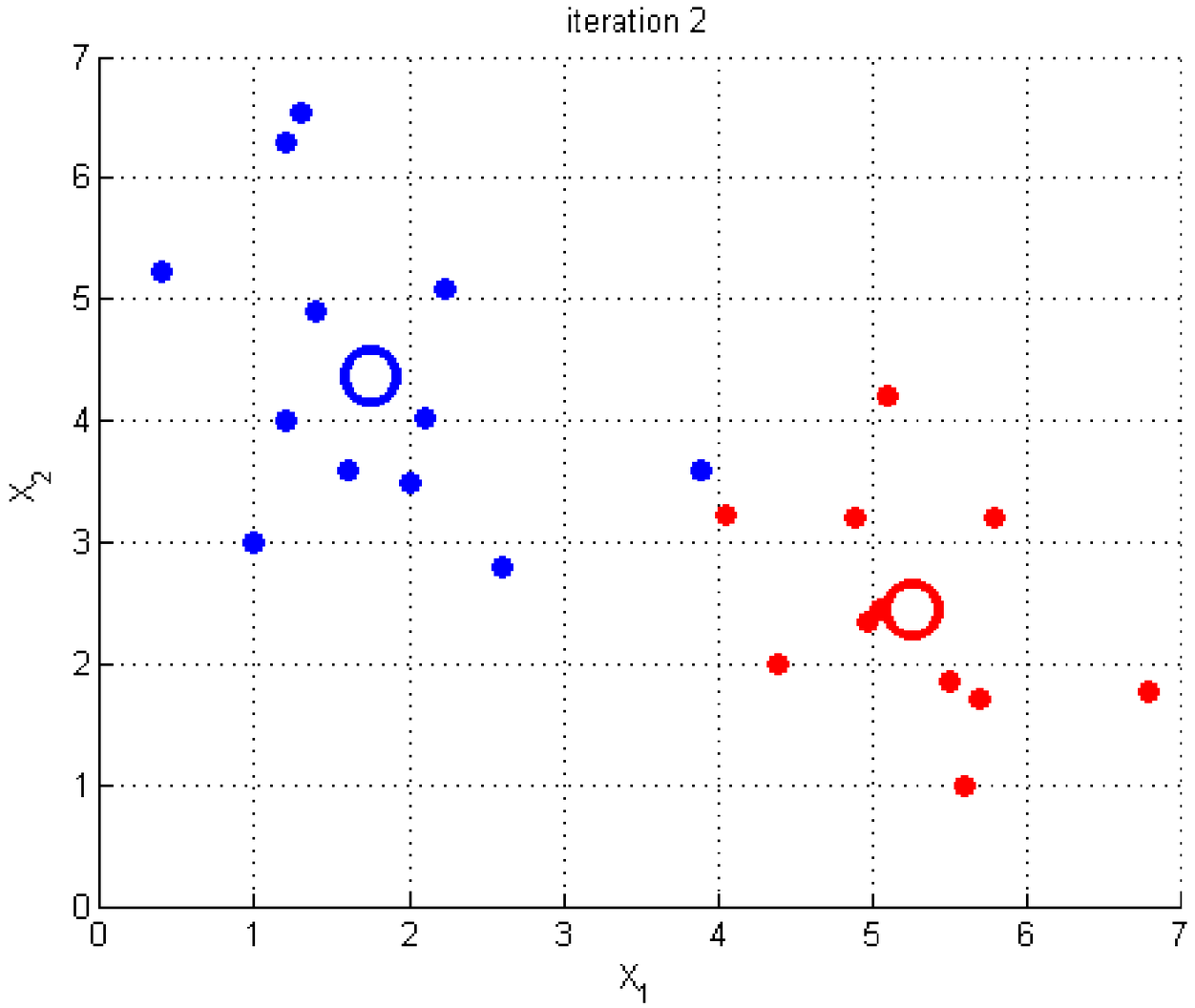
Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

Thuật toán K–means



Khởi tạo tâm cụm

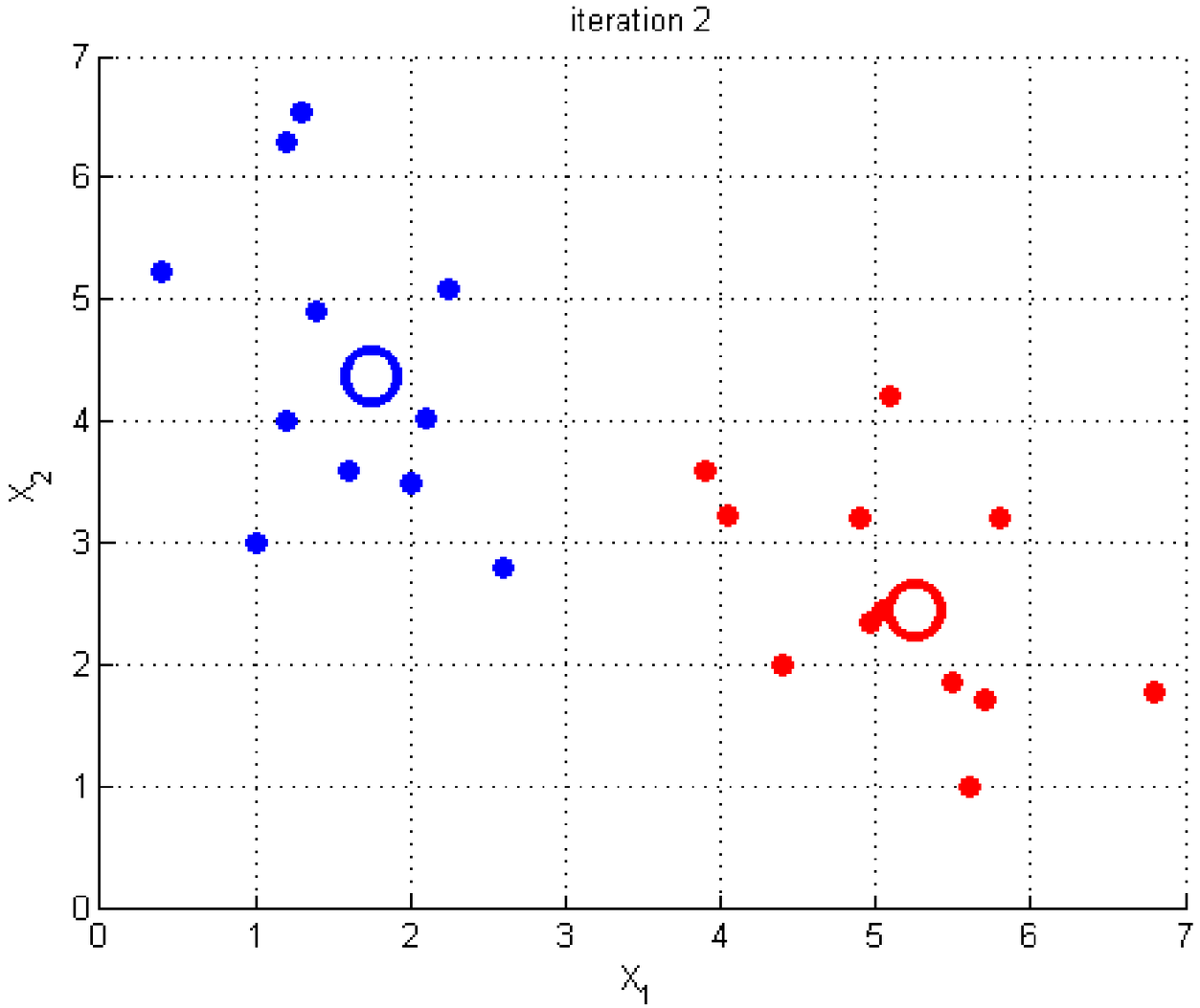
Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Thuật toán K–means



Khởi tạo tâm cụm

Gán các cụm ban đầu

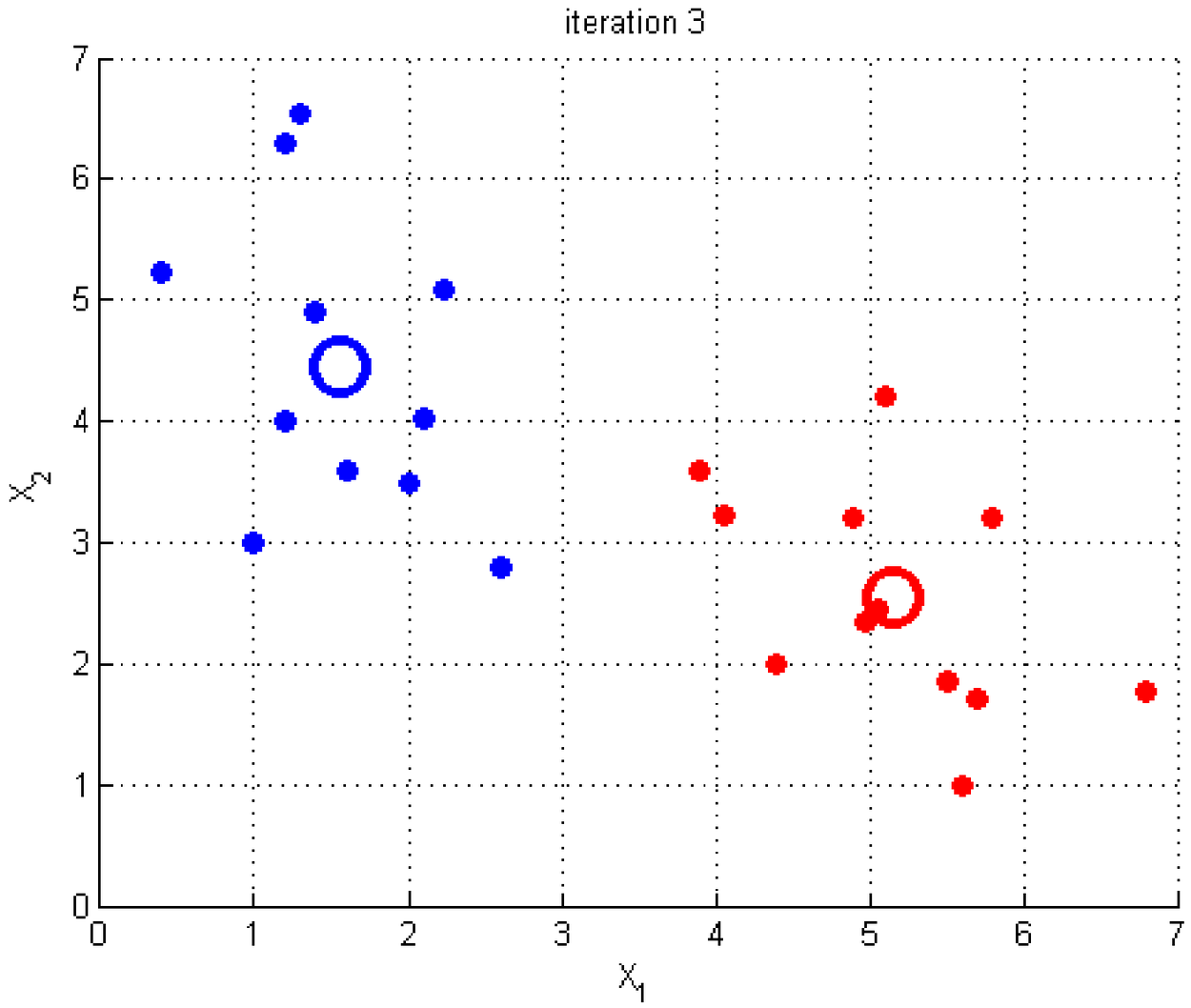
Cập nhật các tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Gán lại các cụm

Thuật toán K–means



Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

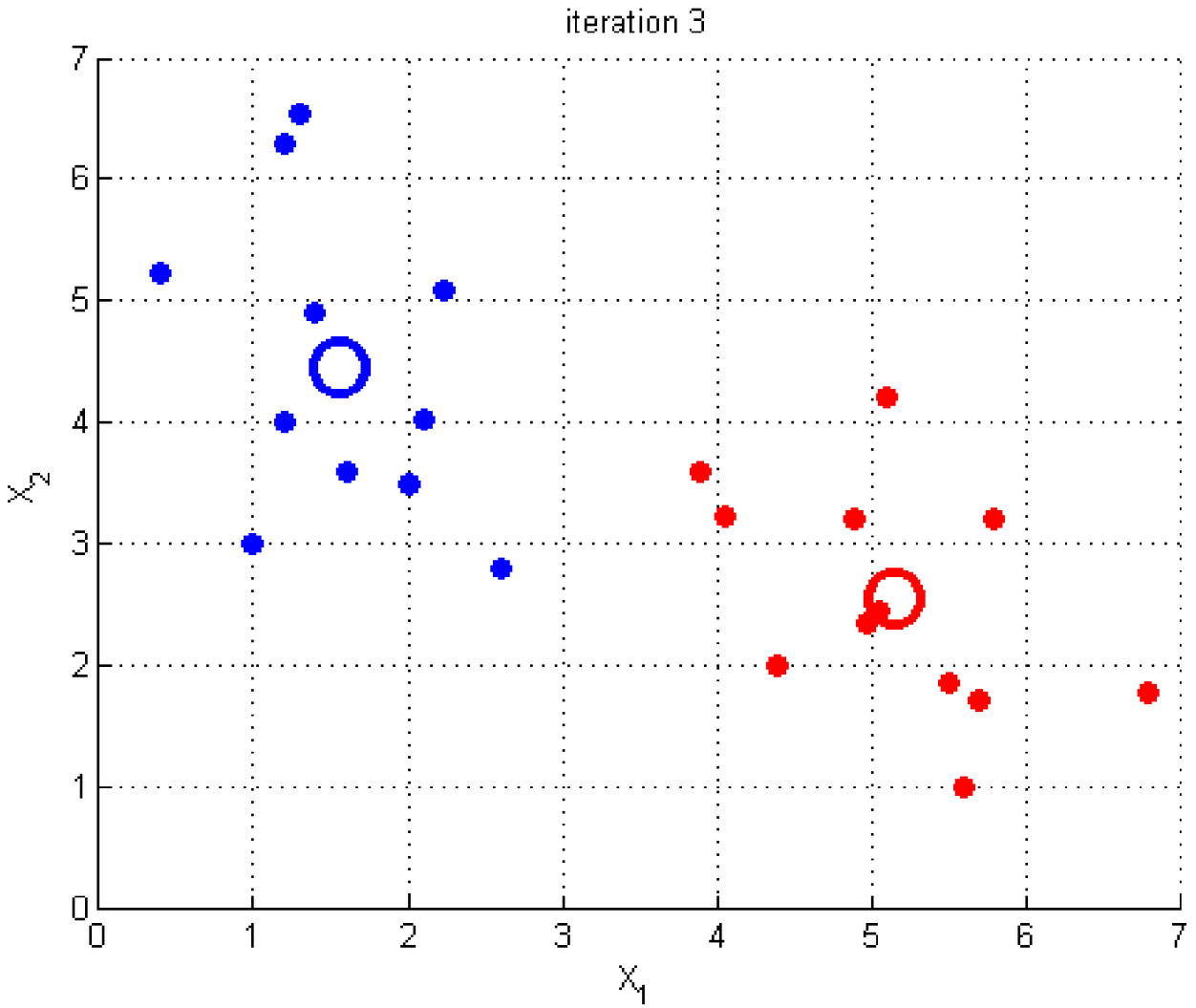
Gán lại các cụm

Cập nhật tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Thuật toán K–means



Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

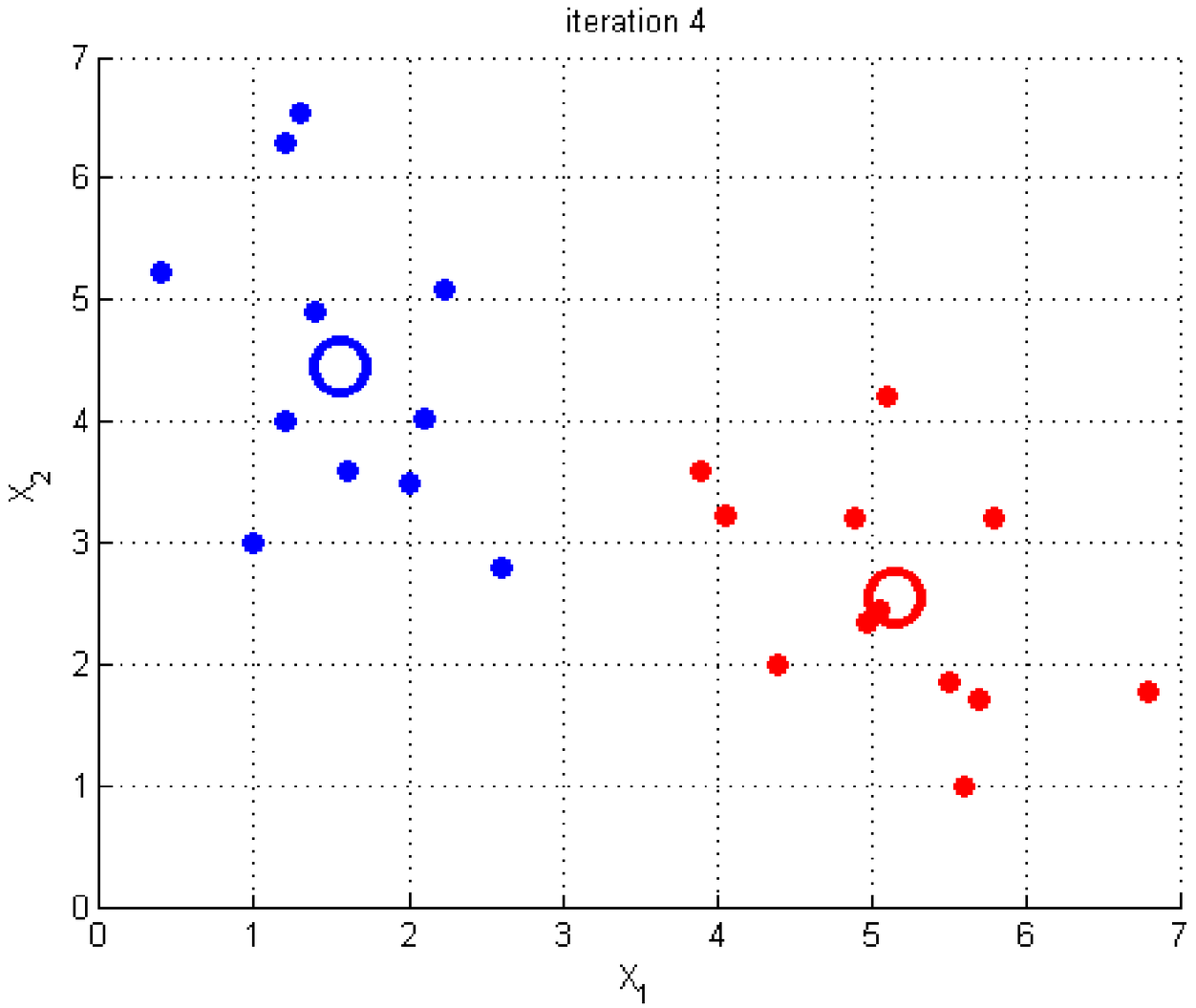
Cập nhật tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Gán lại các cụm

Thuật toán K–means



Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Gán lại các cụm

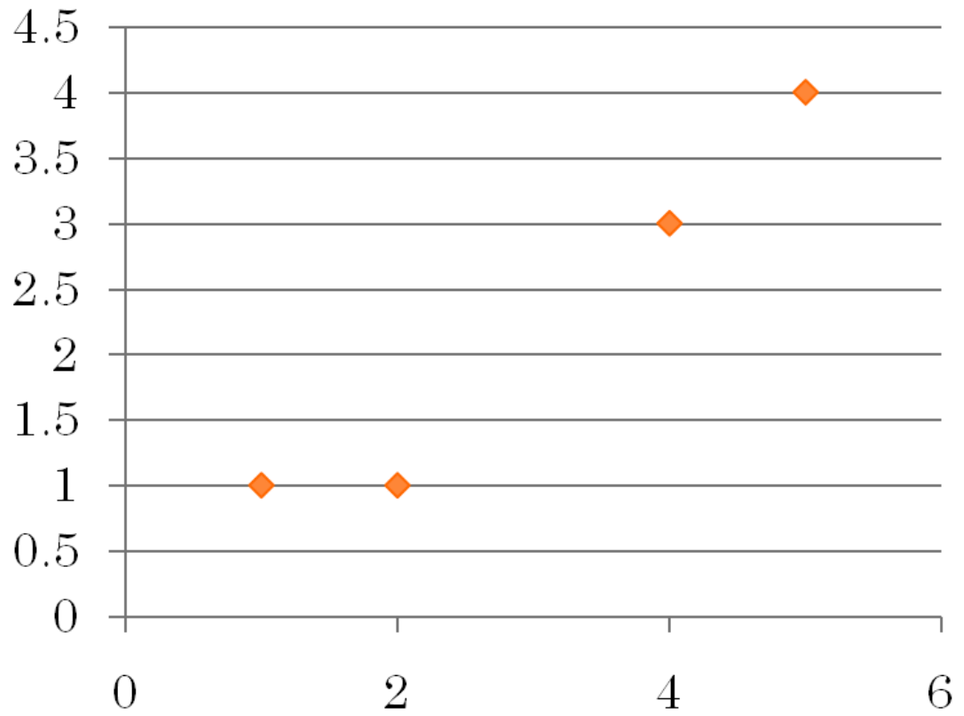
Cập nhật tâm cụm

Gán lại các cụm

Thỏa mãn điều kiện

VÍ DỤ: KHỞI TẠO TÂM C1 = A, C2 = B. ÁP DỤNG K-MEANS CHO DỮ LIỆU SAU

Đối tượng	Thuộc tính 1 (X)	Thuộc tính 2 (Y)
A	1	1
B	2	1
C	4	3
D	5	4



Thuật toán K–means: VÍ DỤ

Áp dụng thuật toán K-means với K = 3 để phân cụm dữ liệu về tổng chi tiêu (spend) và độ tuổi (age) của 20 khách hàng như sau:

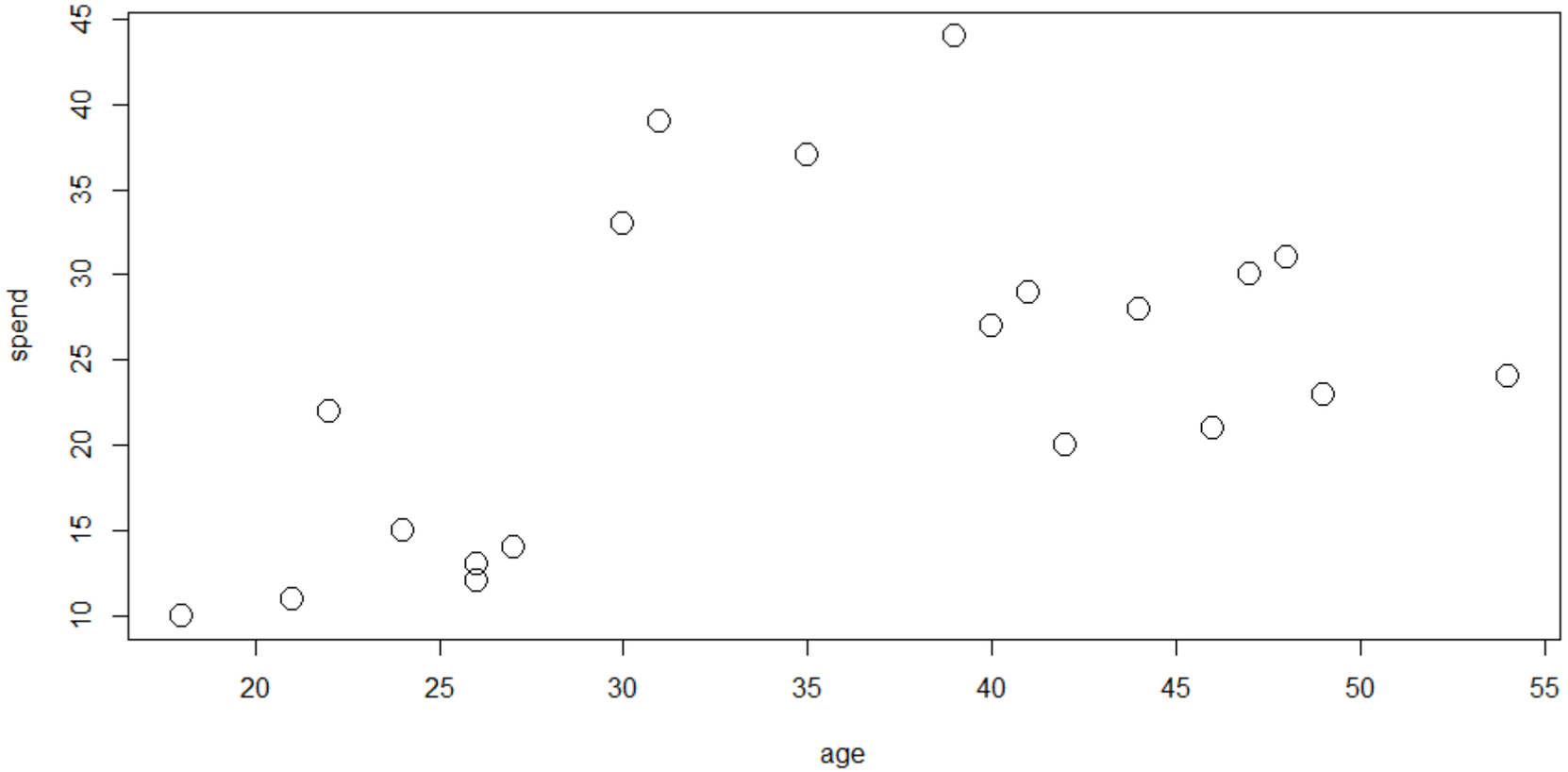
```
age = c(18, 21, 22, 24, 26, 26, 27, 30, 31,
35, 39, 40, 41,   42, 44, 46, 47, 48, 49, 54)
spend = c(10, 11, 22, 15, 12, 13, 14, 33, 39,
  37, 44, 27, 29, 20, 28, 21, 30, 31, 23, 24)
```

Thuật toán K–means: VÍ DỤ

```
age = c(18, 21, 22, 24, 26, 26, 27, 30, 31, 35, 39, 40, 41,  42, 44, 46, 47, 48, 49, 54)
spend = c(10, 11, 22, 15, 12, 13, 14, 33, 39, 37, 44, 27, 29, 20, 28, 21, 30, 31,  23, 24)
df1=data.frame(age,spend)
```

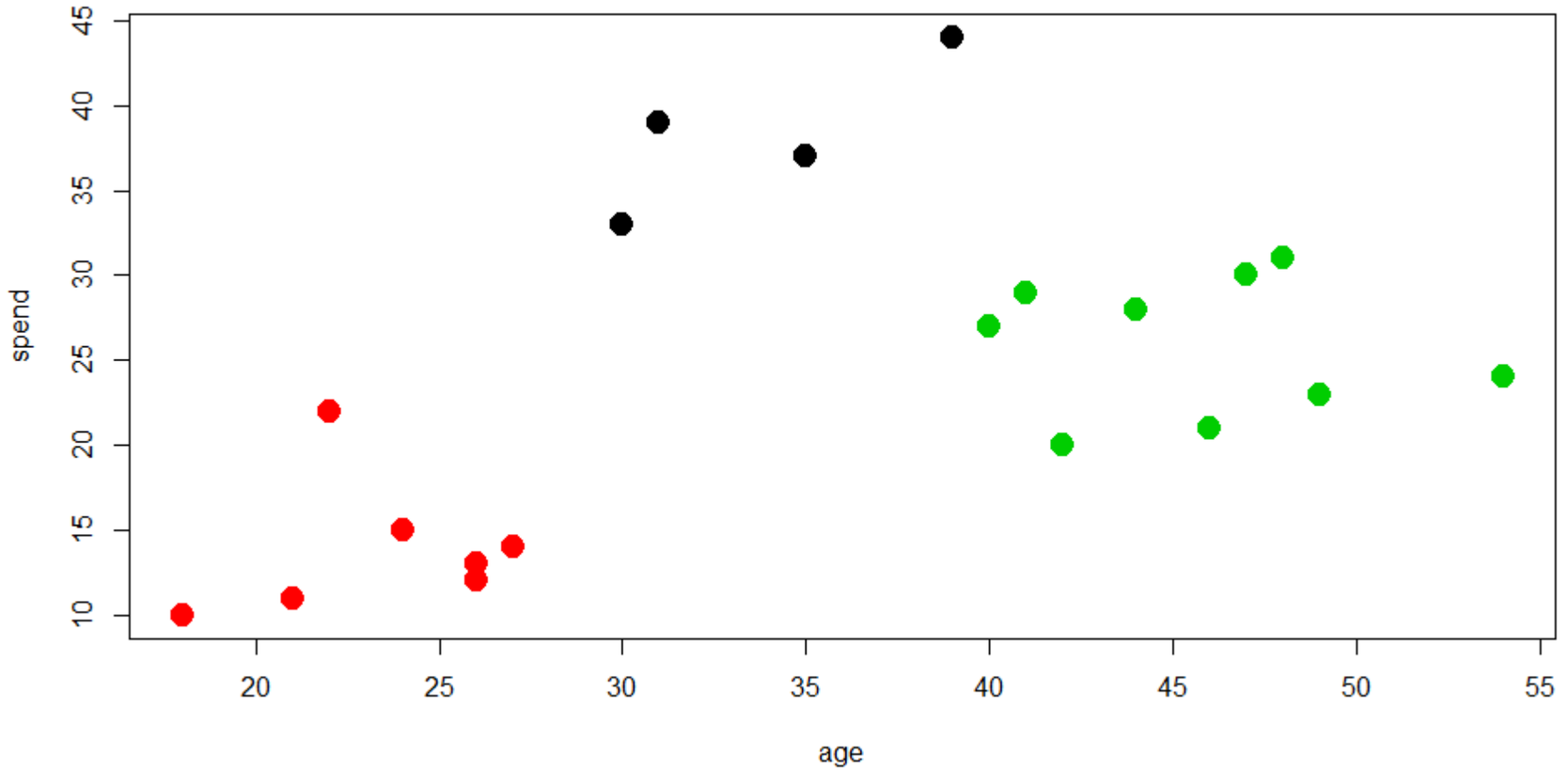
Thuật toán K–means: VÍ DỤ

```
plot(age,spend,xlab="age", ylab="spend")
```



Thuật toán K–means: VÍ DỤ

```
km.res1 <- kmeans(df1, 3, nstart = 25)
plot(age,spend,pch=19,col=km.res1$cluster)
```



Thuật toán K--means: VÍ DỤ

Kiểm tra từng thông tin:

km.res1$size

km.res1$centers

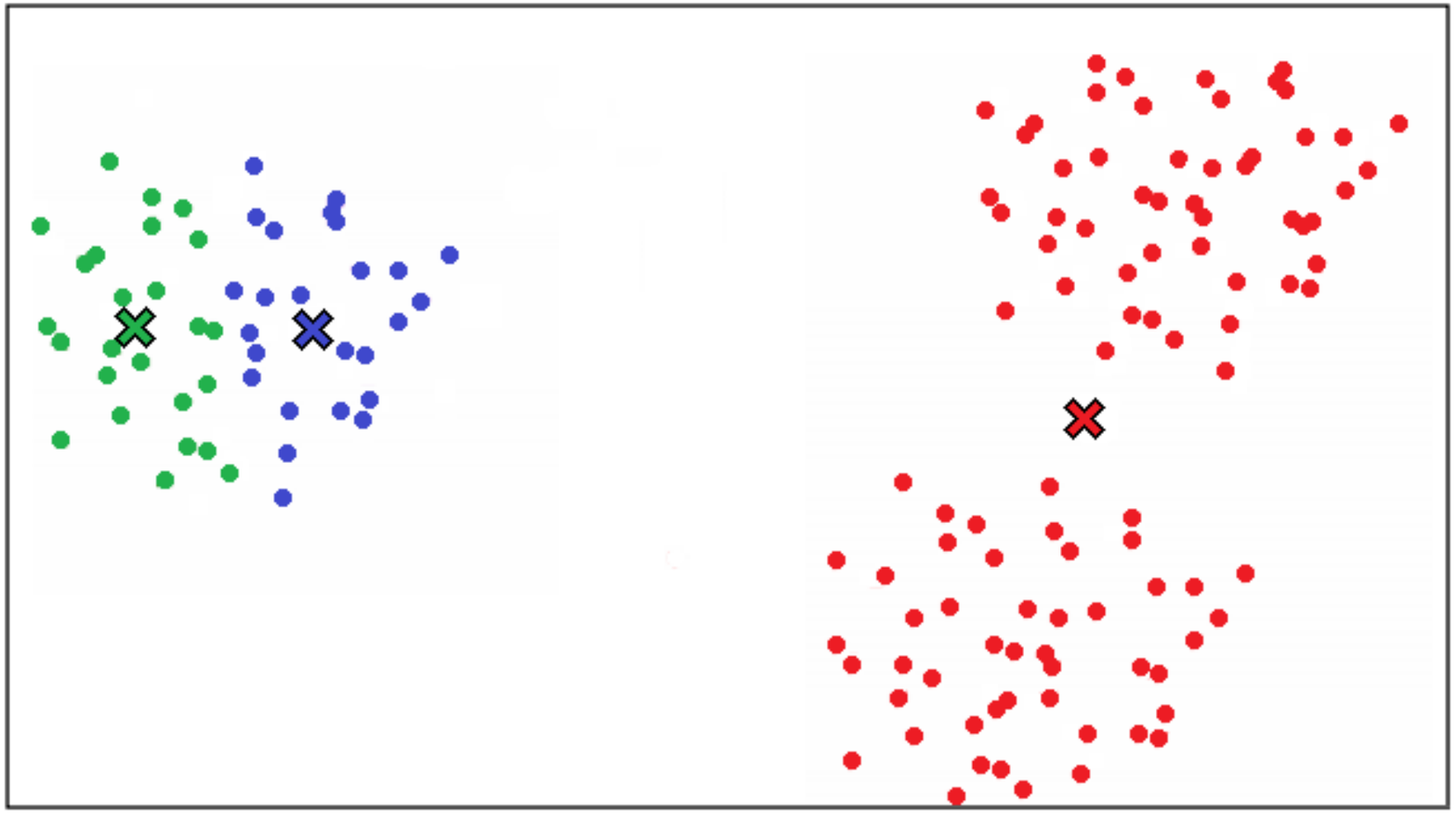
km.res1$withinss

km.res1$betweenss

km.res1$tot.withinss

Thuật toán K–means

- Khởi tạo không tốt dẫn đến kết quả phân cụm kém

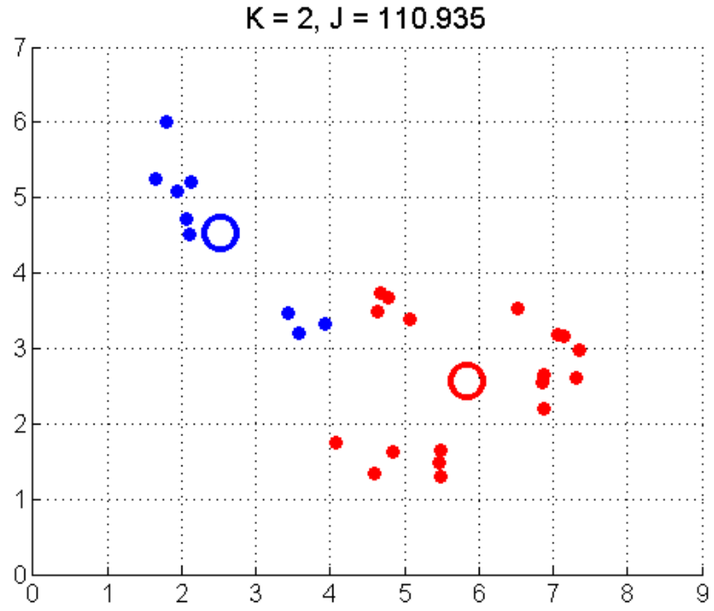


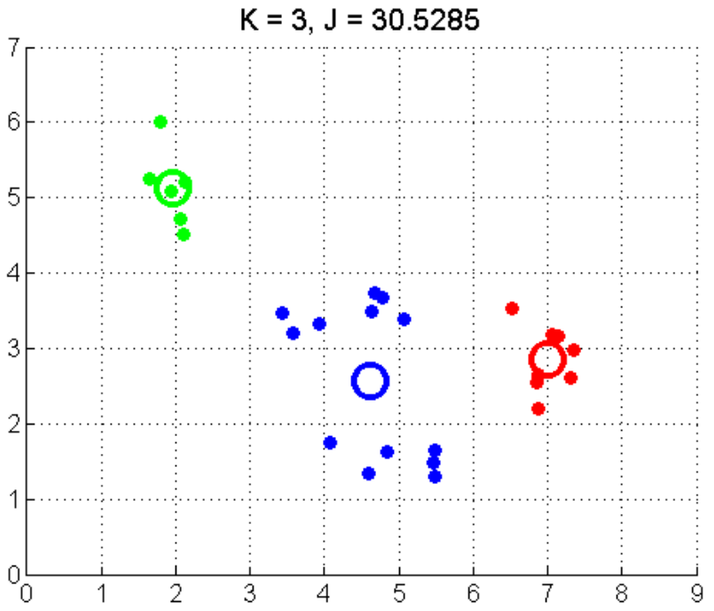
Một số cách khởi tạo tâm cụm

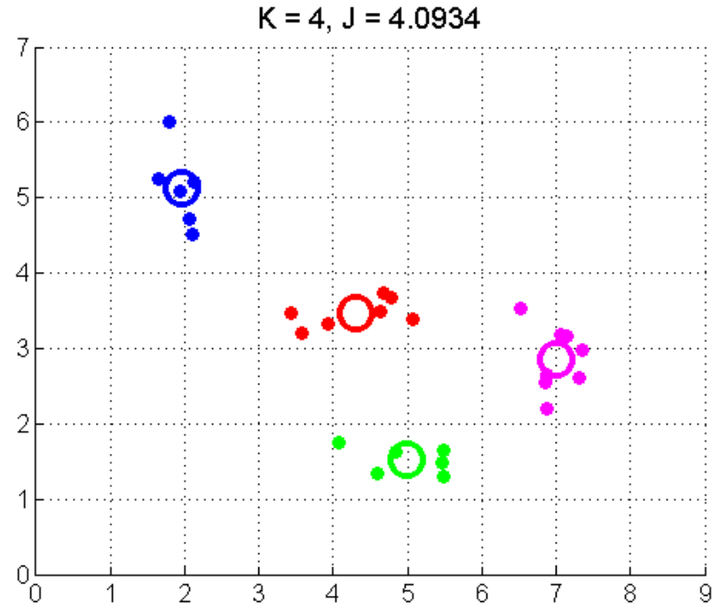
- Chọn ngẫu nhiên trong K đối tượng
- Phân hoạch ngẫu nhiên dữ liệu
- Chọn K điểm xa nhau “far apart”
- Khởi tạo bằng cách sử dụng kết quả của phương pháp phân cụm khác

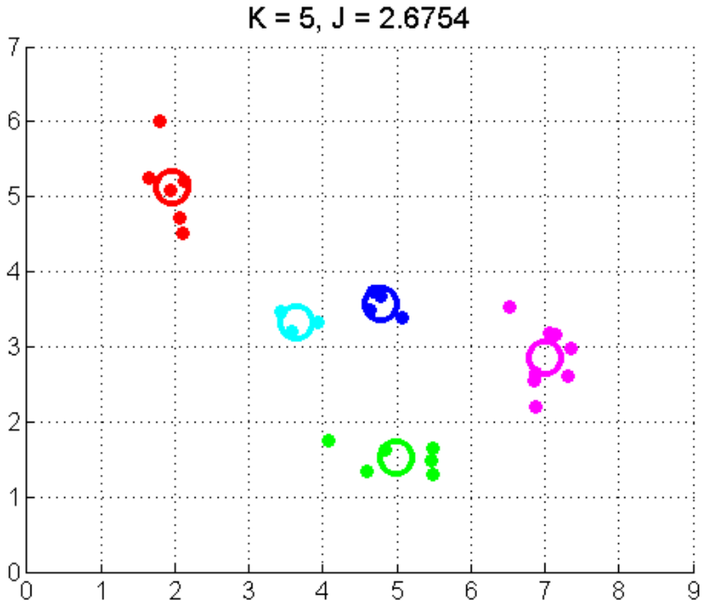
Bao nhiêu cụm?

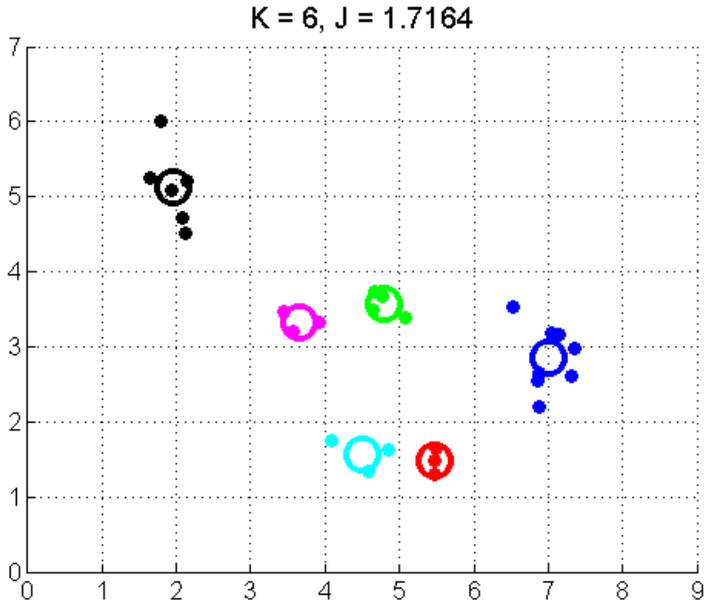
- K-means yêu cầu đầu vào là K (number of clusters)
  - Ta cần hiểu về bài toán ứng dụng để chọn K
  - Ngược lại, việc chọn K được xác định từ dữ liệu


K = 2, J = 110.935


K = 3, J = 30.5285


K = 4, J = 4.0934

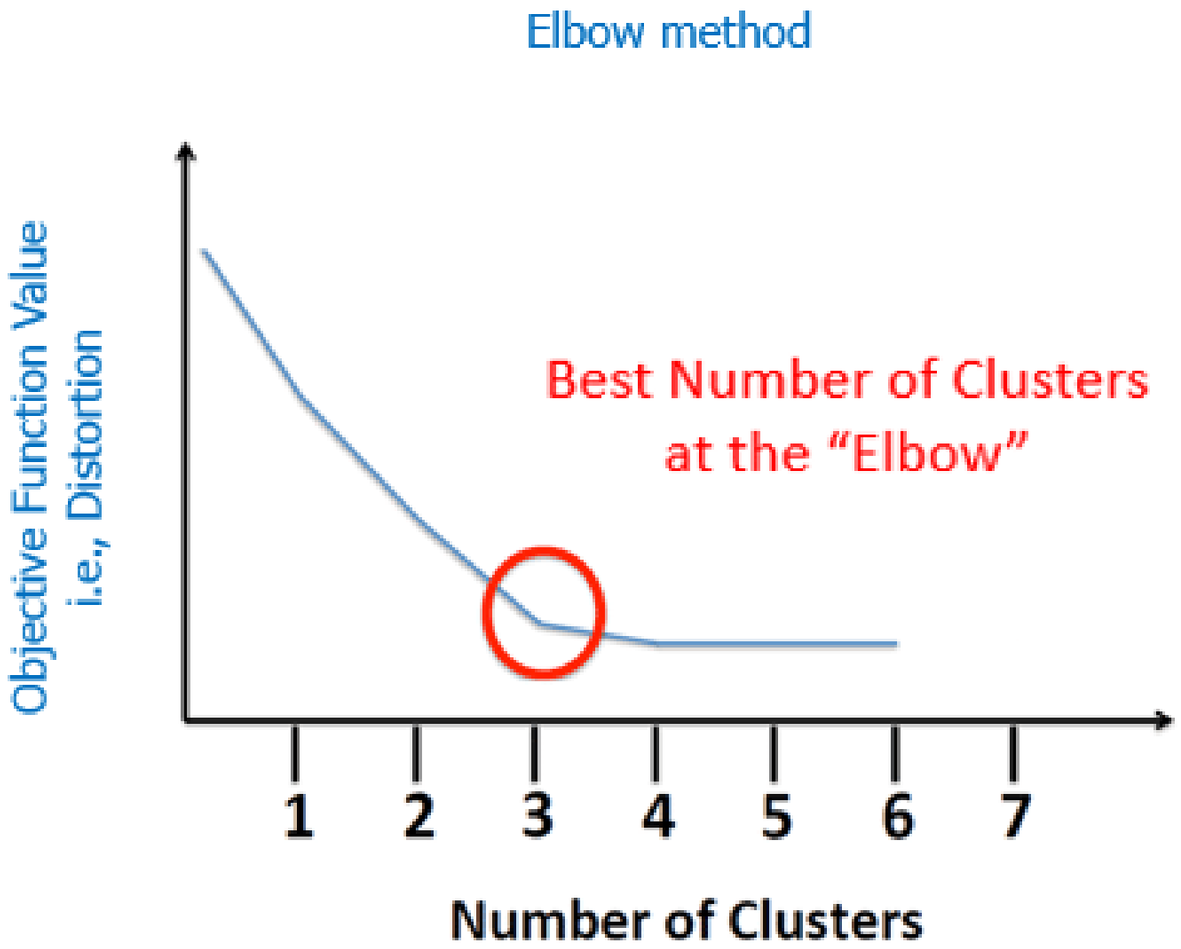

K = 5, J = 2.6754


K = 6, J = 1.7164

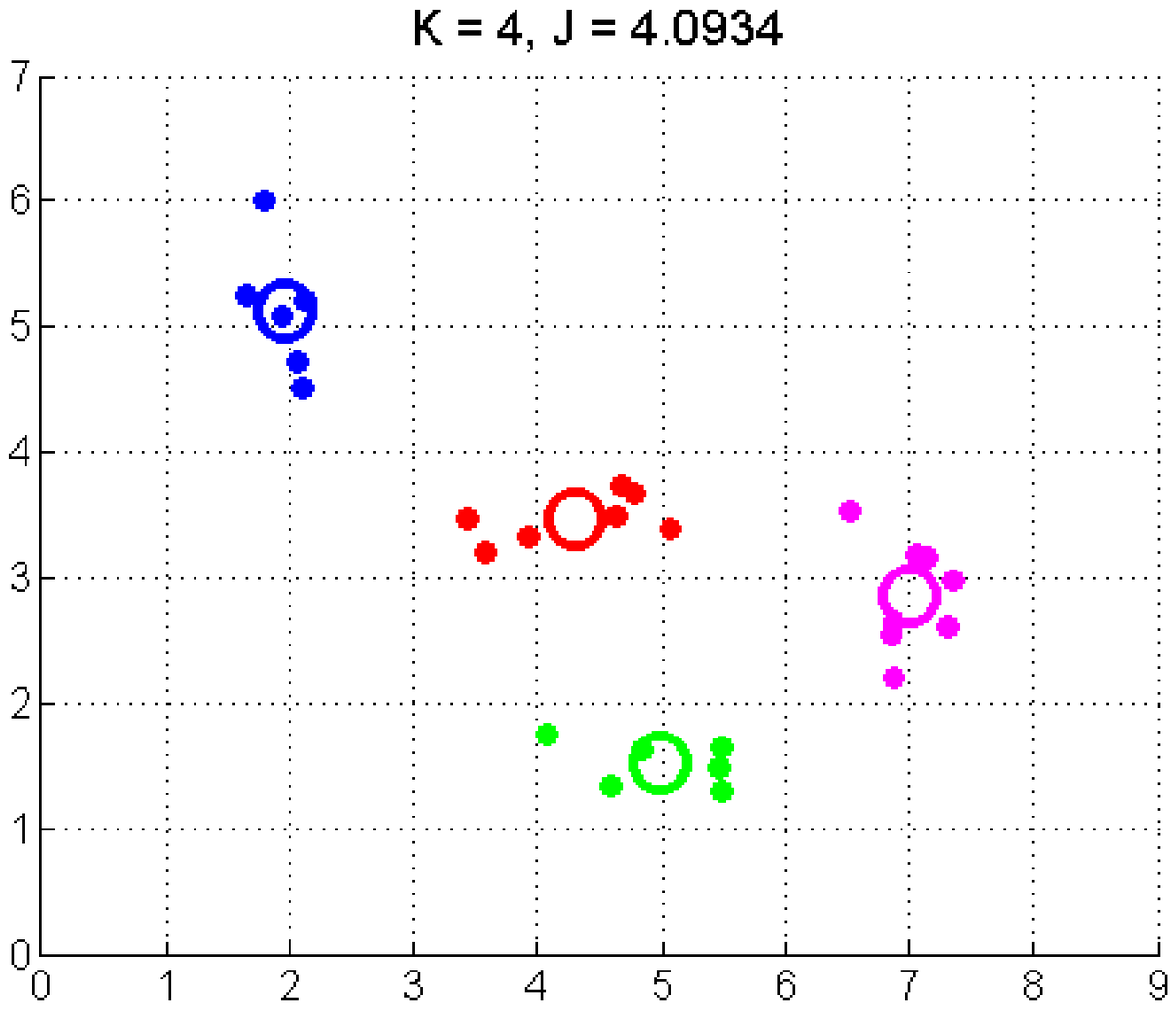
Bao nhiêu cụm?

- Không thể tính được giá trị K để cực tiểu mục tiêu J
  - J giảm đồng thời với tăng K
- Phương pháp dựa trên kinh nghiệm (Heuristic):
  - Với mỗi giá trị ứng viên của K,
  - Tính toán phân cụm bằng K–means M lần, tìm mục tiêu nhỏ nhất J_K
  - Tìm điểm "khuỷu tay (elbow)" trong đường mục tiêu (K vs J_K)

Bao nhiêu cụm? - Exploratory



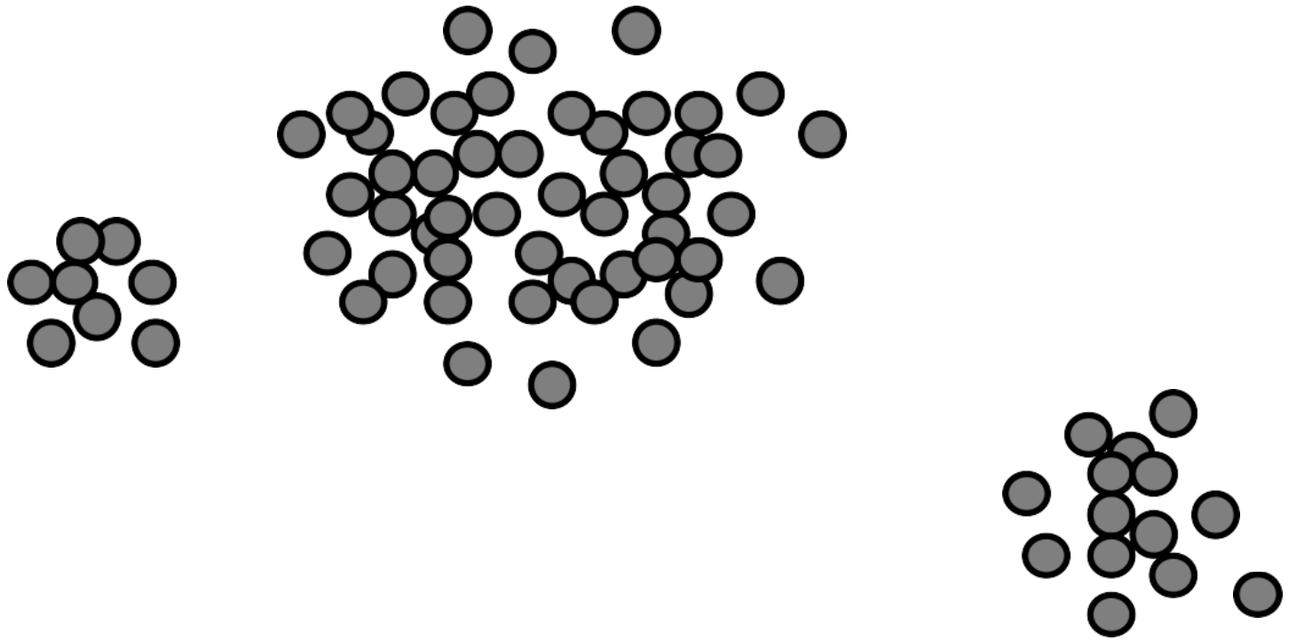
Bao nhiêu cụm?



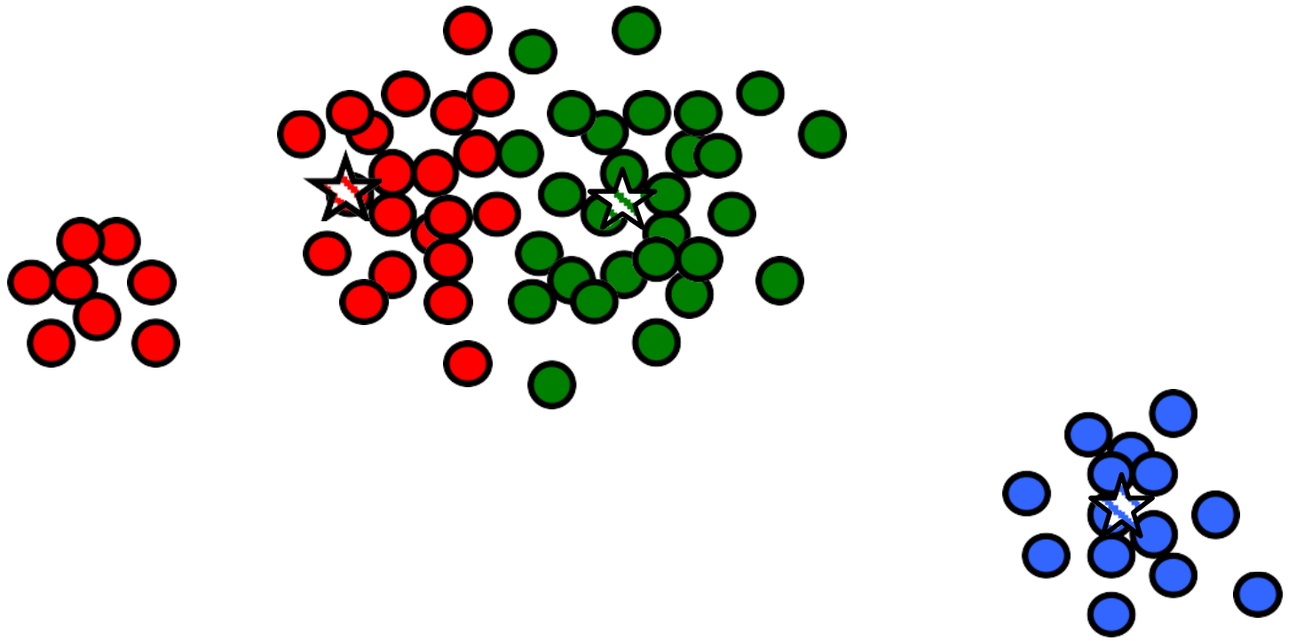
Thuật toán K–means

- Ưu điểm
  - Dễ cài đặt
  - Luôn hội tụ với số lần lặp ít
  - Có thể triển khai trên những tập dữ liệu với số chiều lớn
- Nhược điểm
  - Giá trị K là tham số đầu vào (khó xác định tối ưu)
  - Thuật toán lặp trả về cực tiểu địa phương*

Thuật toán K–means



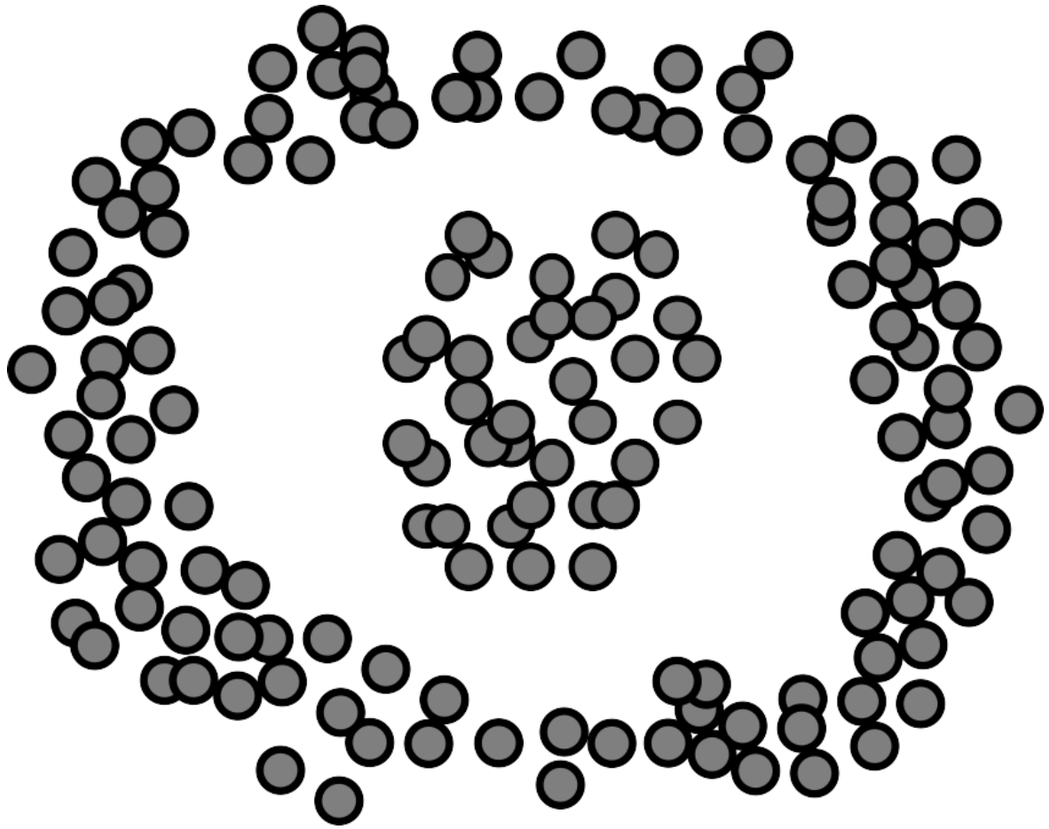
Thuật toán K–means



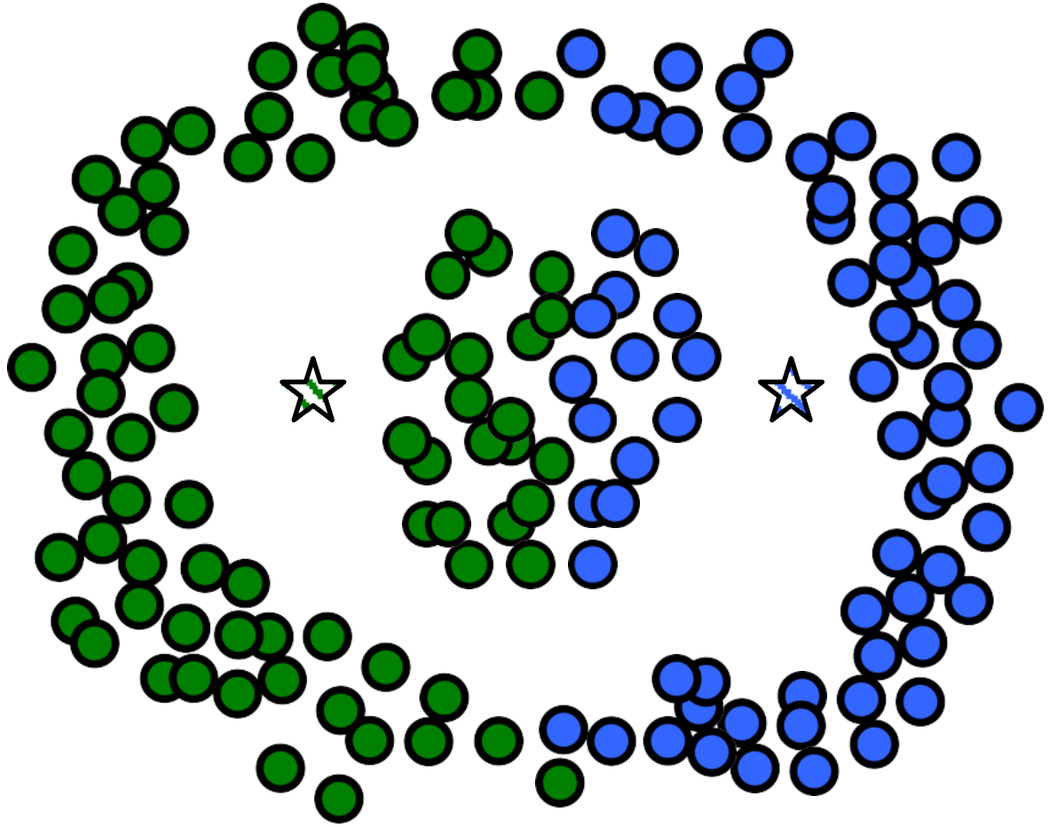
Thuật toán K–means

- Ưu điểm
  - Dễ cài đặt
  - Luôn hội tụ với số lần lặp ít
  - Có thể triển khai trên những tập dữ liệu với số chiều lớn
- Nhược điểm
  - Giá trị K là tham số đầu vào (khó xác định tối ưu)
  - Thuật toán lặp trả về cực tiểu địa phương*
  - Giả thiết tất cả các cụm hình cầu và có kích thước xấp xỉ nhau *

Thuật toán K–means



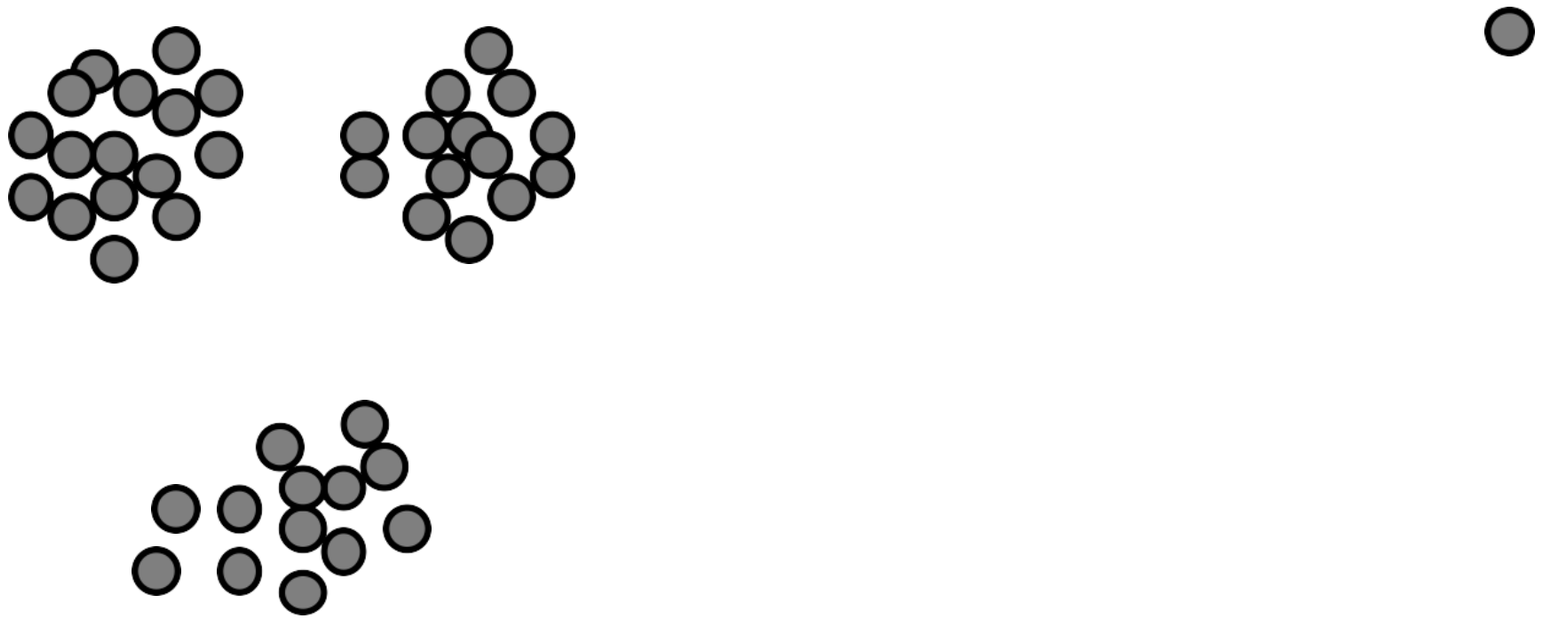
Thuật toán K–means



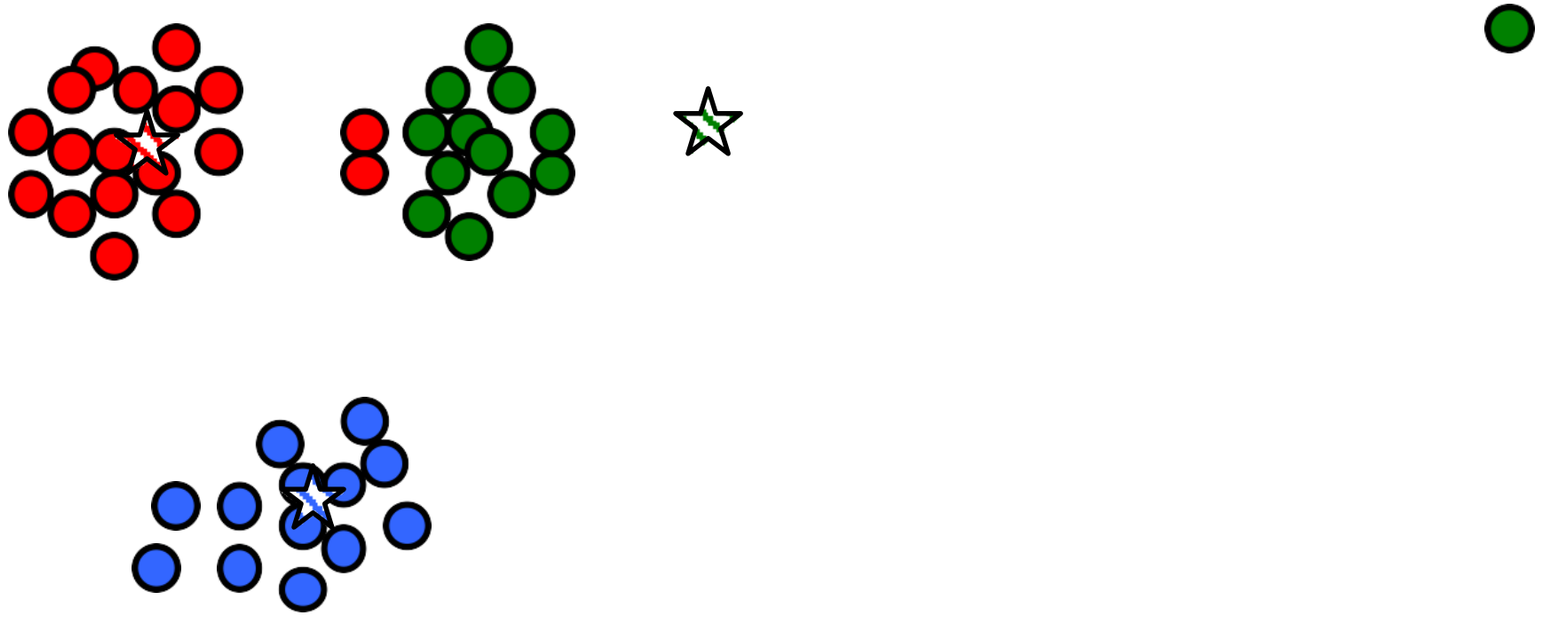
Thuật toán K–means

- Ưu điểm
  - Dễ cài đặt
  - Luôn hội tụ với số lần lặp ít
  - Có thể triển khai trên những tập dữ liệu với số chiều lớn
- Nhược điểm
  - Giá trị K là tham số đầu vào (khó xác định tối ưu)
  - Thuật toán lặp trả về cực tiểu địa phương*
  - Giả thiết tất cả các cụm hình cầu và có kích thước xấp xỉ nhau *
  - Nhạy với các phần tử ngoại lai*

Thuật toán K–means



Thuật toán K–means



Thuật toán K–means

- Ưu điểm
  - Dễ cài đặt
  - Luôn hội tụ với số lần lặp ít
  - Có thể triển khai trên những tập dữ liệu với số chiều lớn
- Nhược điểm
  - Giá trị K là tham số đầu vào (khó xác định tối ưu)
  - Thuật toán lặp trả về cực tiểu địa phương*
  - Giả thiết tất cả các cụm hình cầu và có kích thước xấp xỉ nhau *
  - Nhạy với các phần tử ngoại lai*
  - ***một số nhược điểm được khắc phục bằng vài biến thể của K-means**

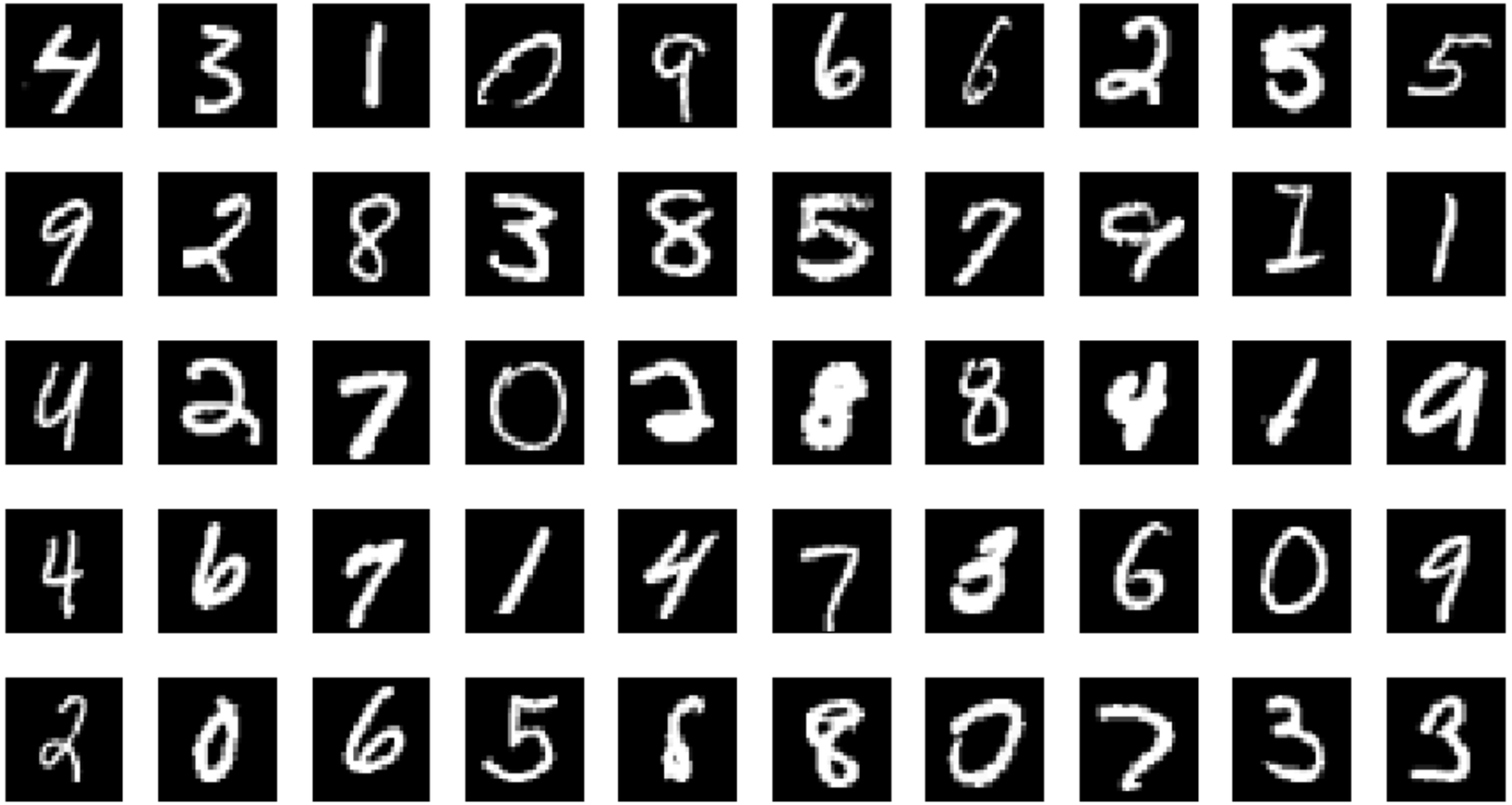
Thuật toán K–means

- Khắc phục nhược điểm
  - Khởi tạo K không tốt → ta chạy thuật toán nhiều lần
  - K-medians: Tâm cụm được tính bằng giá trị trung vị thay cho giá trị trung bình của K-means
  - K-medoids
    - Yêu cầu: "tâm cụm" phải là 1 trong các điểm dữ liệu
    - xử lý tốt hơn các phần tử ngoại lai
    - linh hoạt hơn – có thể dùng nhiều độ đo
    - *nhưng* thời gian tính toán lâu hơn vì phải tính các tâm cụm

Thuật toán K–means

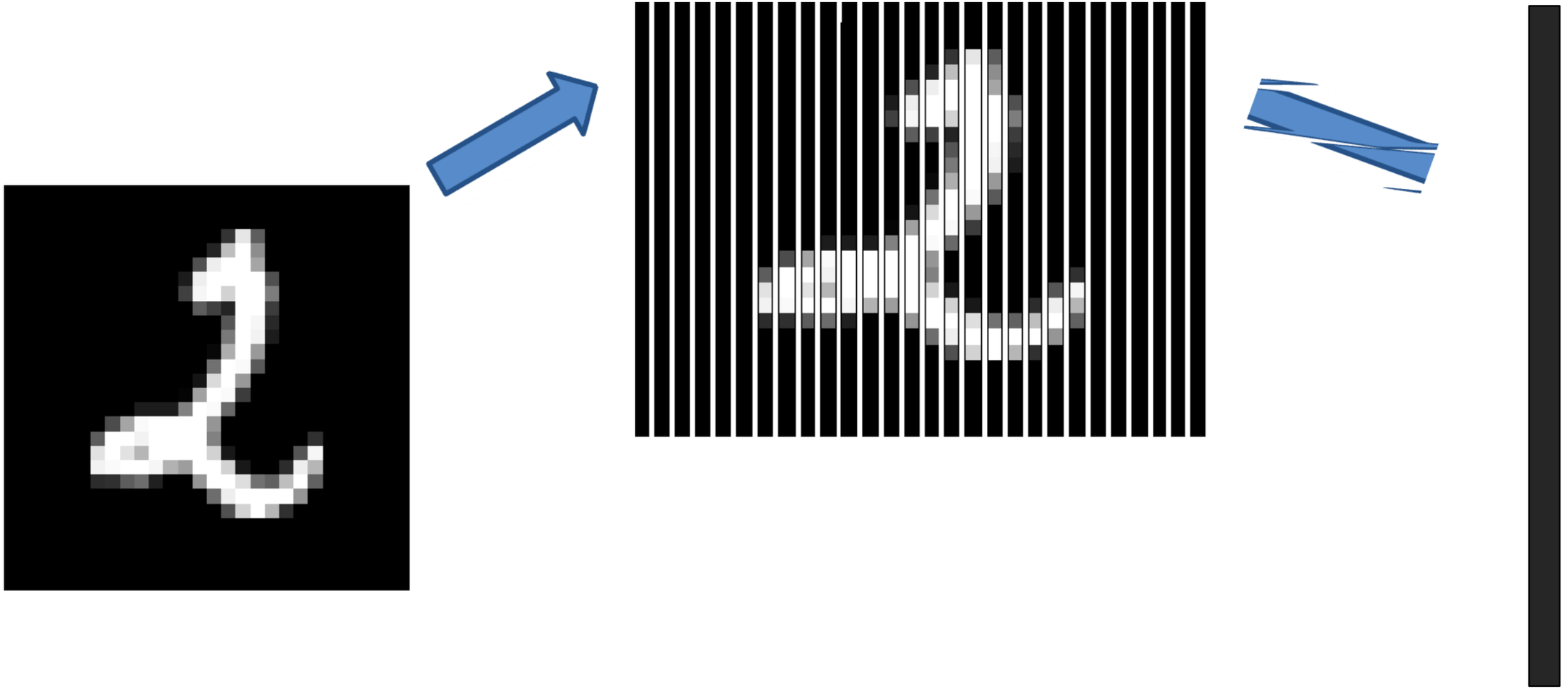
- Chúng ta thực hiện thuật toán với dữ liệu có 2–3 thuộc tính (rất dễ để minh họa). Tuy nhiên, trong thực tế, ta thường gặp nhiều hơn 2 thuộc tính khi phân tích dữ liệu
- Phân cụm sẽ khó khăn hơn rất nhiều khi gặp số chiều lớn

Phân cụm chữ viết tay



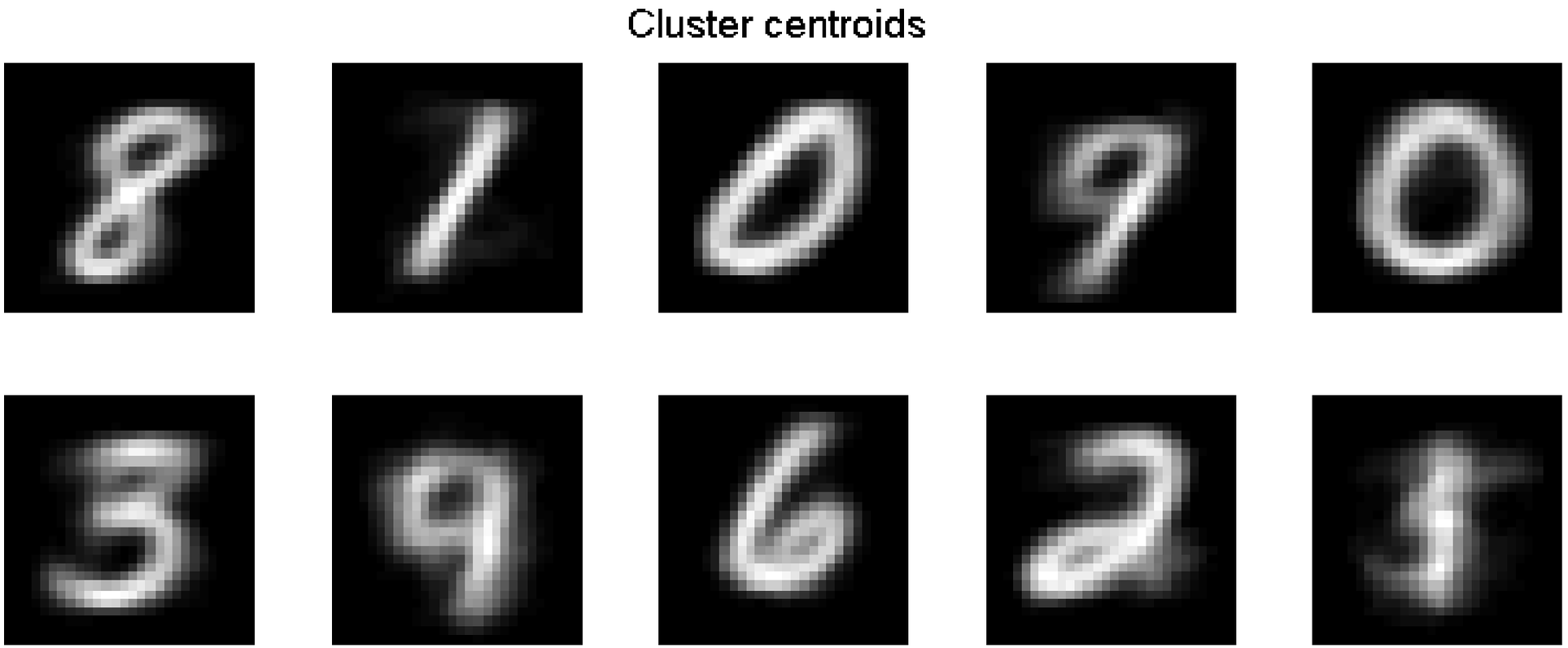
MNIST dataset: http://cis.jhu.edu/~sachin/digit/digit.html

Phân cụm chữ viết tay

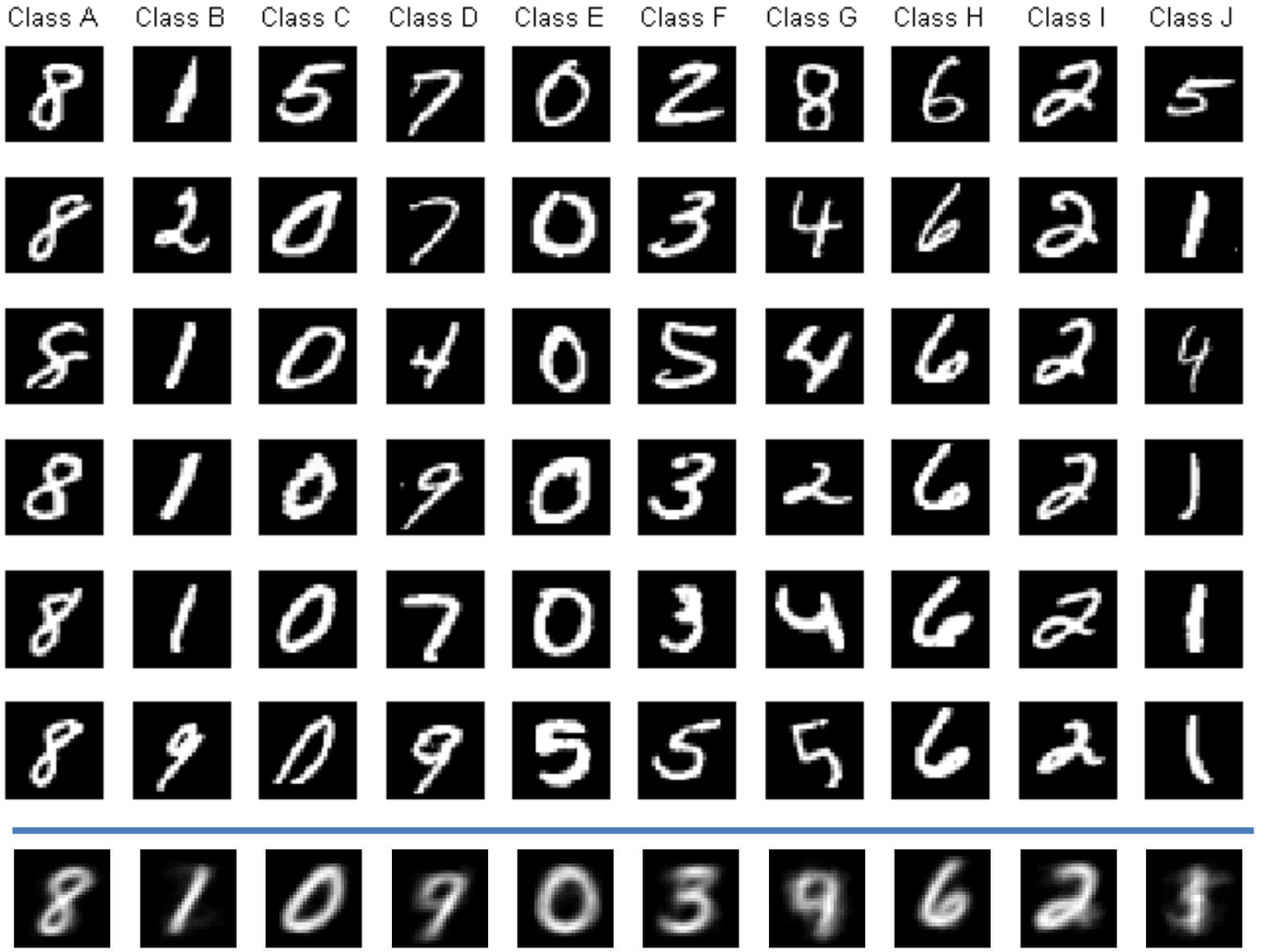


Phân cụm chữ viết tay

- Áp dụng K-means, sử dụng $K = 10$



Phân cụm chữ viết tay



Phương pháp phân cấp

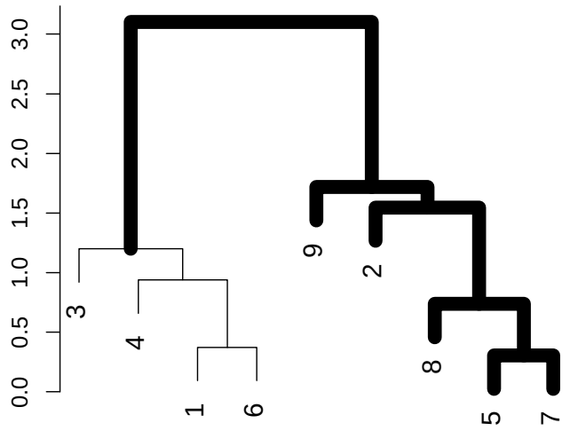
- Phương pháp phân cấp (phân cụm cây)
  - Các cụm dựa trên khoảng cách giữa các mẫu
  - Hiển thị theo phân cấp mà không theo cách phân hoạch dữ liệu

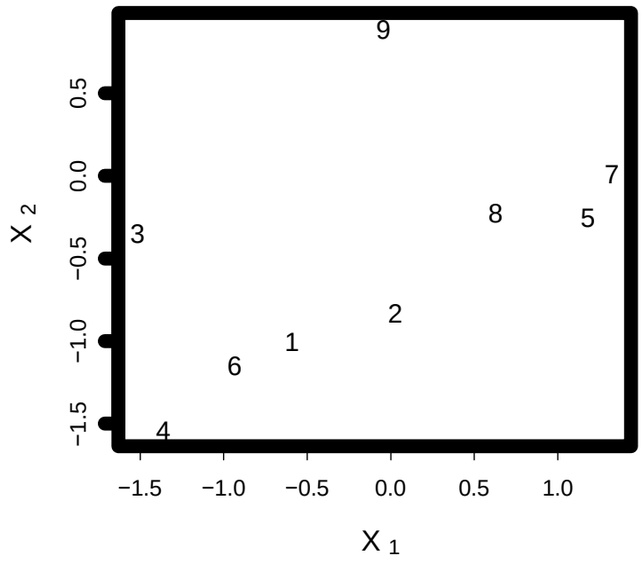
Phân cụm phân cấp

- Phân cụm theo phương pháp K-Means yêu cầu chọn tham số đầu vào là số lượng cụm K
- Nếu ta không muốn làm theo cách trên, ta có thể dùng phương pháp phân cụm phân cấp
- Phân cụm phân cấp có ưu điểm là hiển thị các quan sát (mẫu) dạng hình cây nên dễ hình dung, được gọi là phân cụm theo cấu trúc cây (Dendogram)

Phân cụm phân cấp

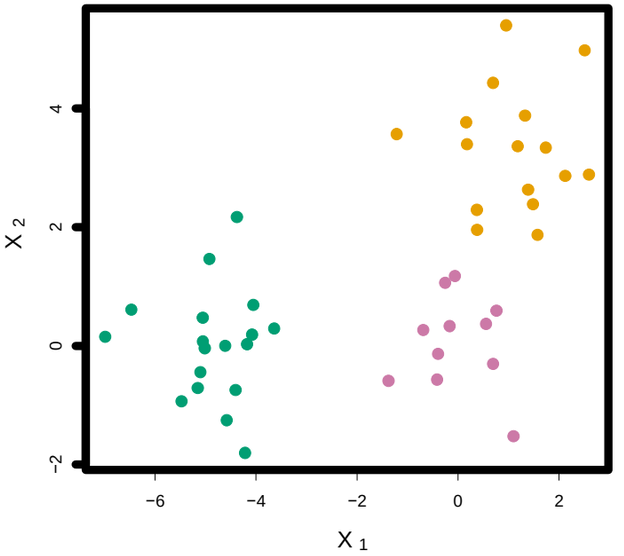
- Đầu tiên nhập các điểm gần nhau nhất (5 và 7)
- Độ cao của việc hợp nhất (theo trục dọc) phản ánh độ tương tự của các điểm
- Sau khi các điểm được hợp nhất, chúng được xem như 1 mẫu để tiếp tục tiến hành giải thuật

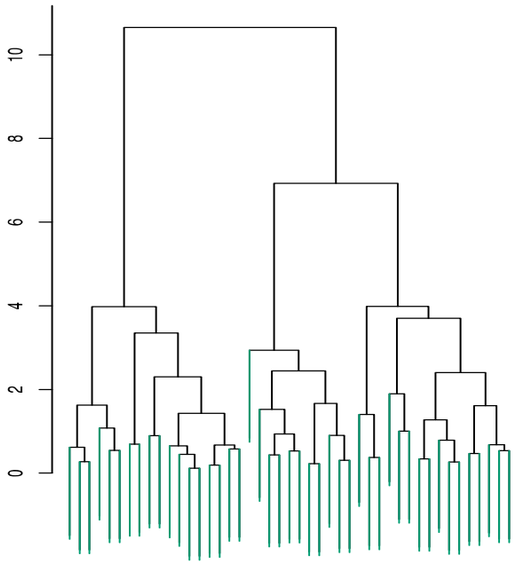




Diễn giải phương pháp phân cấp

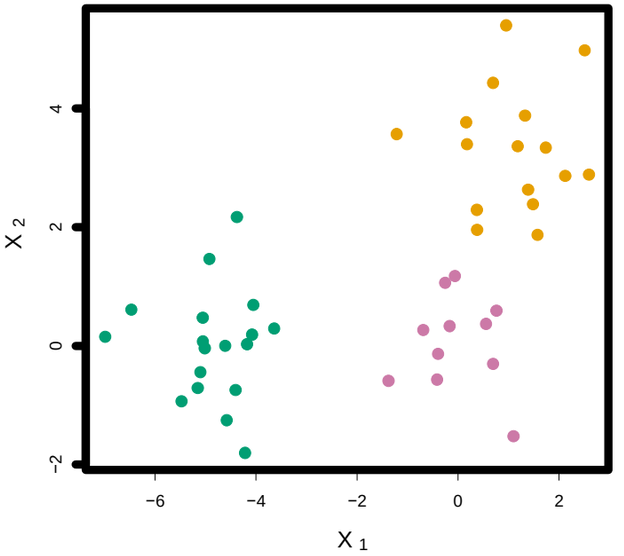
- Mỗi "lá" của cây phân cấp biểu diễn một trong 45 mẫu
- Phần đáy của cây, mỗi mẫu là 1 lá riêng biệt. Tuy nhiên, càng lên cao các lá sẽ hợp nhất với nhau. Việc này thể hiện các mẫu có độ tương tự với các mẫu khác.

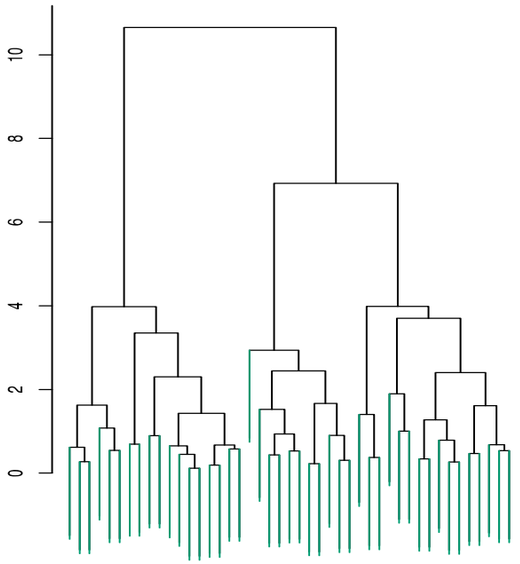




Diễn giải phương pháp phân cấp

- Khi di chuyển cao lên phần ngọn của cây, số lượng mẫu đã được hợp nhất. Trước đó (phần dưới của cây) với 2 mẫu hợp nhất, chúng có chung đặc tính (gần) với nhau.

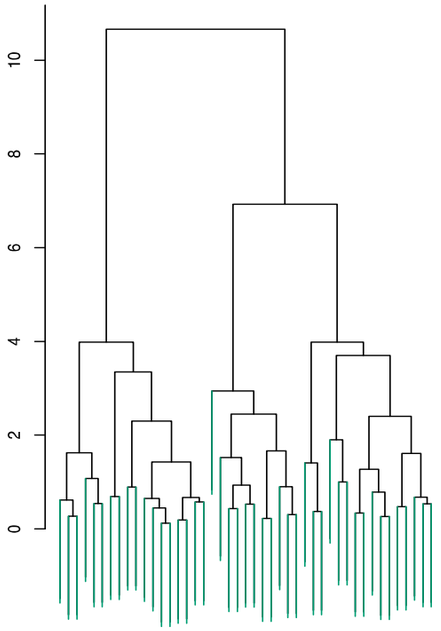




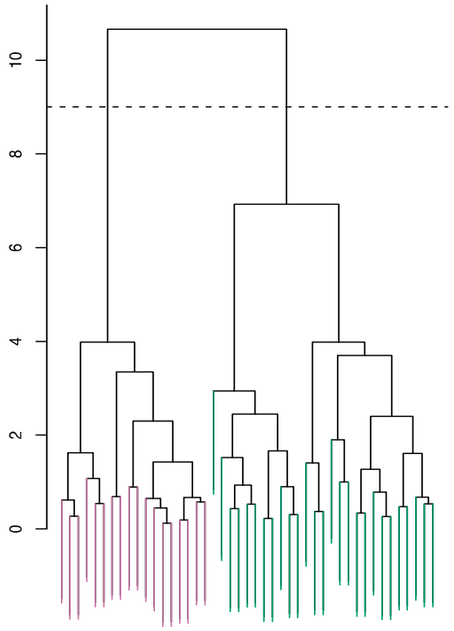
Lựa chọn các cụm

Để chọn các cụm ta kẻ đường thẳng ngang cây phân cấp

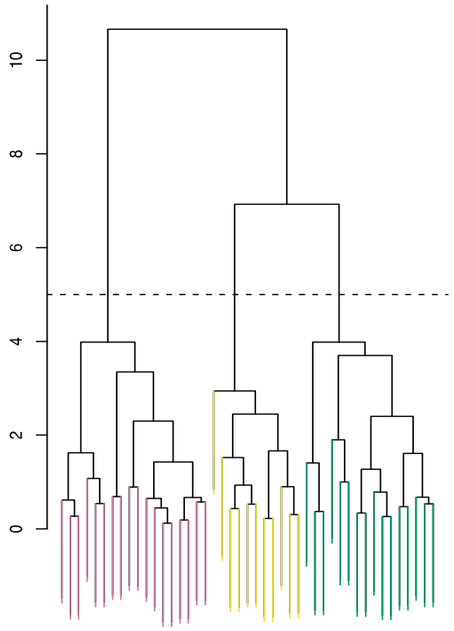
Ta có thể chọn số lượng cụm tùy thuộc vào vị trí đường kẻ



One Cluster



Two Clusters



Three Clusters

Giải thuật (trộn các cụm)

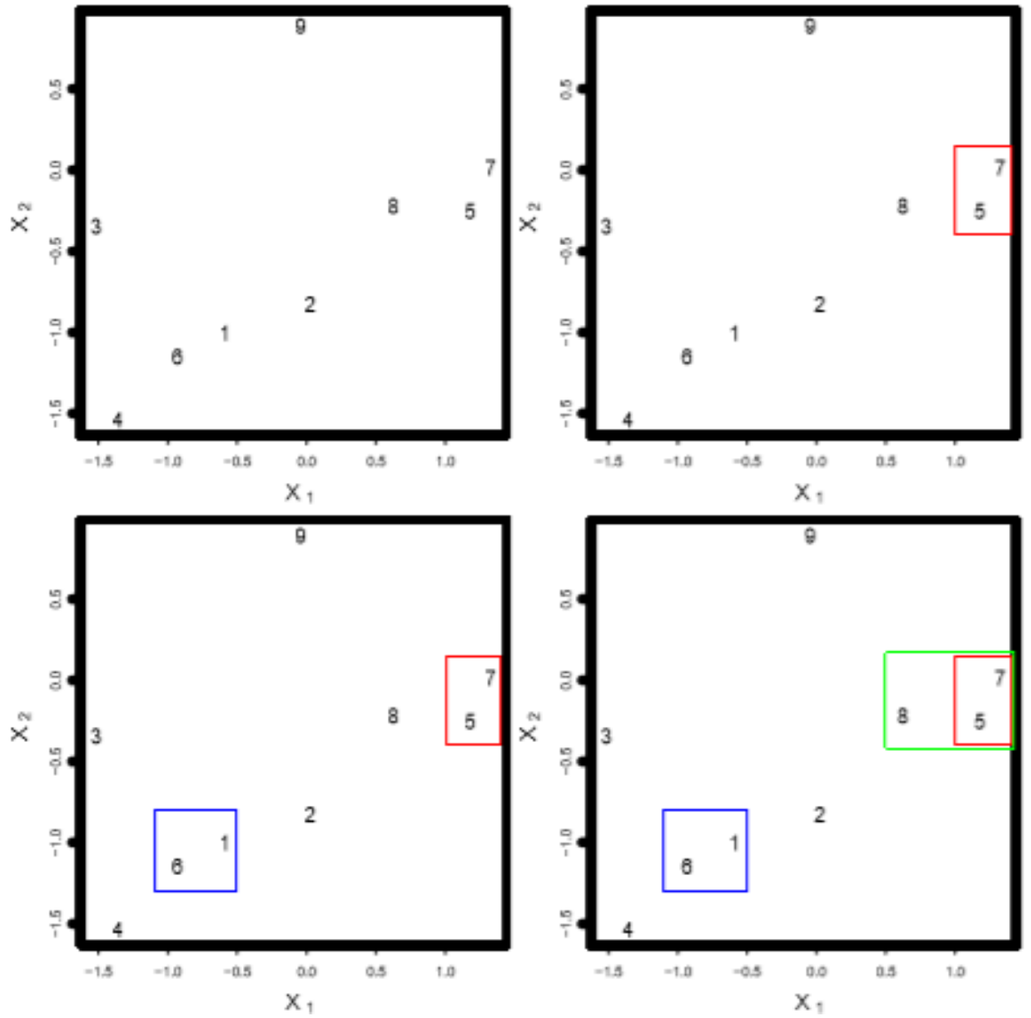
Phân cụm bằng cấu trúc cây:

- Khởi tạo với mỗi điểm là 1 cụm riêng biệt (n cụm), chính là 1 nút trong dendrogram
- Tính toán độ tương tự (gần) giữa các điểm/cụm
- Hợp nhất 2 cụm mà chúng có độ tương tự cao nhất, ta còn lại n-1 cụm
- Hợp nhất 2 cụm tiếp theo có độ tương tự cao nhất, ta còn lại n-2 cụm
- Quá trình trên tiếp tục cho đến khi chỉ còn 1 cụm (là nút gốc trong dendrogram)

Giải thuật (trộn các cụm)

Ví dụ

Bắt đầu với 9 cụm
Hợp nhất 5 và 7
Hợp nhất 6 và 1
Hợp nhất cụm (5,7) với 8.
Quá trình tiếp tục cho đến khi tất cả các cụm được hợp nhất.



Ta định nghĩa sự khác biệt ntn?

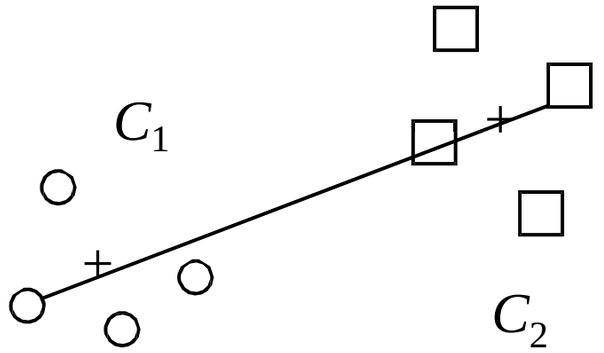
Việc triển khai phương pháp phân cấp cần giải quyết vấn đề khá hiển nhiên, đó là làm sao để định nghĩa sự khác biệt (dissimilarity) hoặc mối liên kết (linkage) giữa cụm hợp nhất (5, 7) và cụm 8?

Có 4 lựa chọn:

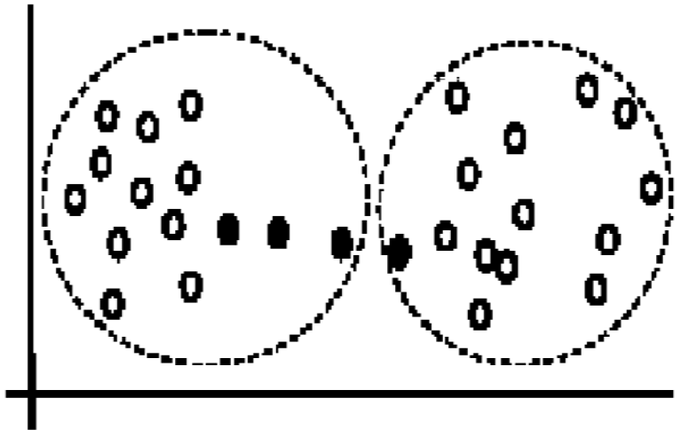
- Liên kết đầy (Complete Linkage)
- Liên kết đơn (Single Linkage)
- Liên kết trung bình giữa các nhóm (Average Linkage)
- Liên kết tâm (Centroid Linkage)

Các phương pháp liên kết

Liên kết đầy: Khoảng cách giữa 2 cụm là khoảng cách lớn nhất giữa 2 mẫu tương ứng của 2 cụm đó



- Nhạy cảm (gặp lỗi phân cụm) đối với các ngoại lai (outliers)
- Có xu hướng sinh ra các cụm có dạng “bụi cây” (clumps)

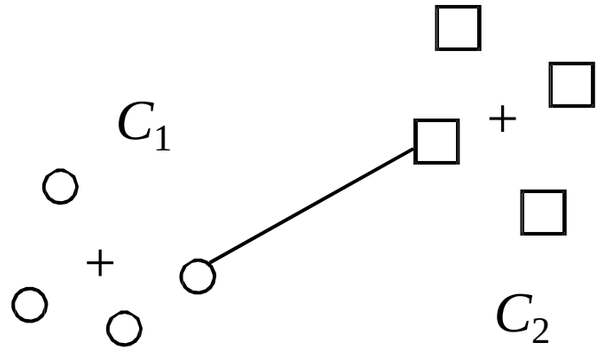


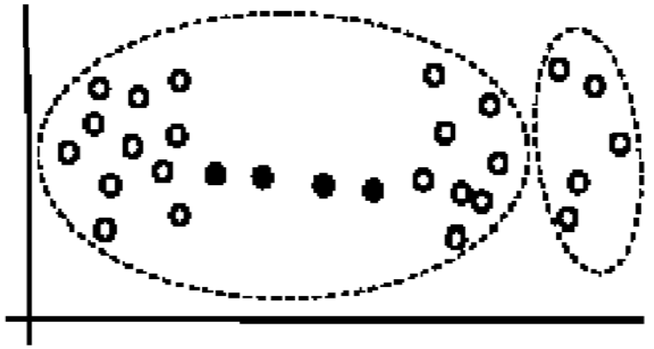
[Liu, 2006]

Các phương pháp liên kết

Liên kết đơn: Khoảng cách giữa 2 cụm là khoảng cách nhỏ nhất giữa các mẫu (các thành viên) của 2 cụm đó.

Có xu hướng sinh ra các cụm có dạng "chuỗi dài" (long chain)





[Liu, 2006]

Các phương pháp liên kết

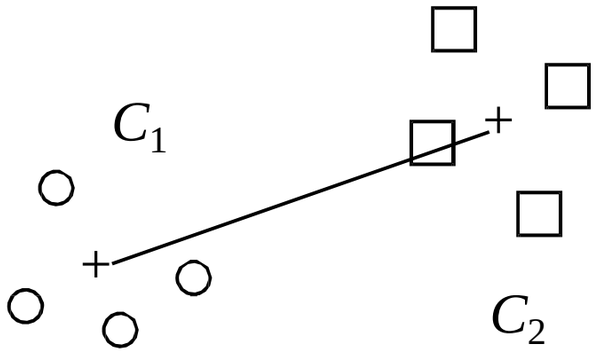
Liên kết trung bình: Khoảng cách trong liên kết trung bình (Average-link) là sự thỏa hiệp giữa các khoảng cách trong liên kết hoàn toàn (Complete-link) và liên kết đơn (Single-link)

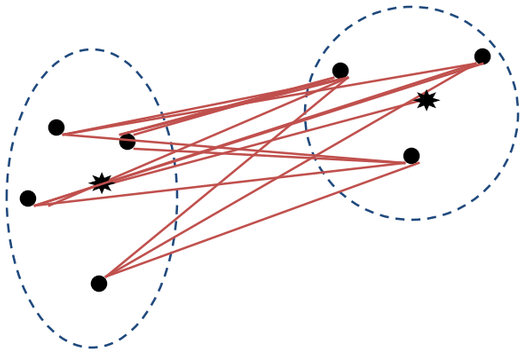
- Để giảm mức độ nhạy cảm (khả năng lỗi) của phương pháp phân cụm dựa trên liên kết đầy đối với các ngoại lai (outliers)
- Để giảm xu hướng sinh ra các cụm có dạng "chuỗi dài" của phương pháp phân cụm dựa trên liên kết đơn (dạng "chuỗi dài" không phù hợp với khái niệm tự nhiên của một cụm)

- Khoảng cách giữa 2 cụm là khoảng cách trung bình của tất cả các cặp mẫu (mỗi mẫu thuộc về một cụm)

Các phương pháp liên kết

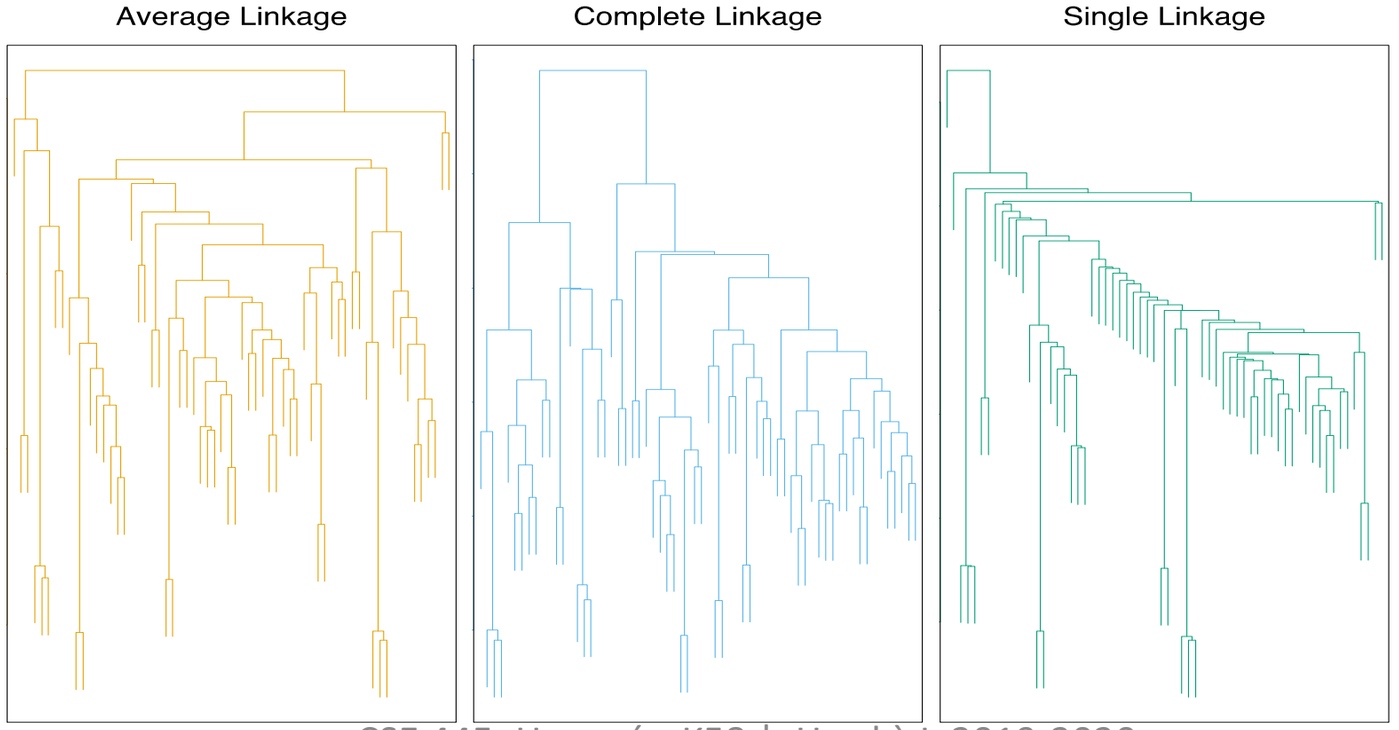
<u>Liên kết tâm:</u> Khoảng cách giữa các tâm của các mẫu tương ứng





Mối liên kết rất quan trọng

Dưới đây ta có 3 kết quả phân cụm trên cùng 1 bộ dữ liệu
Phương pháp tính mối liên kết khác nhau nhưng kết quả đem lại rất khác xa nhau
Phương pháp liên kết đầy và liên kết trung bình dường như có cỡ cụm như nhau, tuy nhiên liên kết đơn lại cho số cụm nhiều hơn vì mỗi lá của cây được hợp nhất từng lần một



Ví dụ

Cho dữ liệu gồm 6 đối tượng (các điểm trong không gian 2 chiều). Thực hiện phân cụm theo phân hoạch.

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Ví dụ

Tính ma trận khoảng cách và thực hiện theo các bước của phân cụm phân cấp

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Bộ dữ liệu Iris

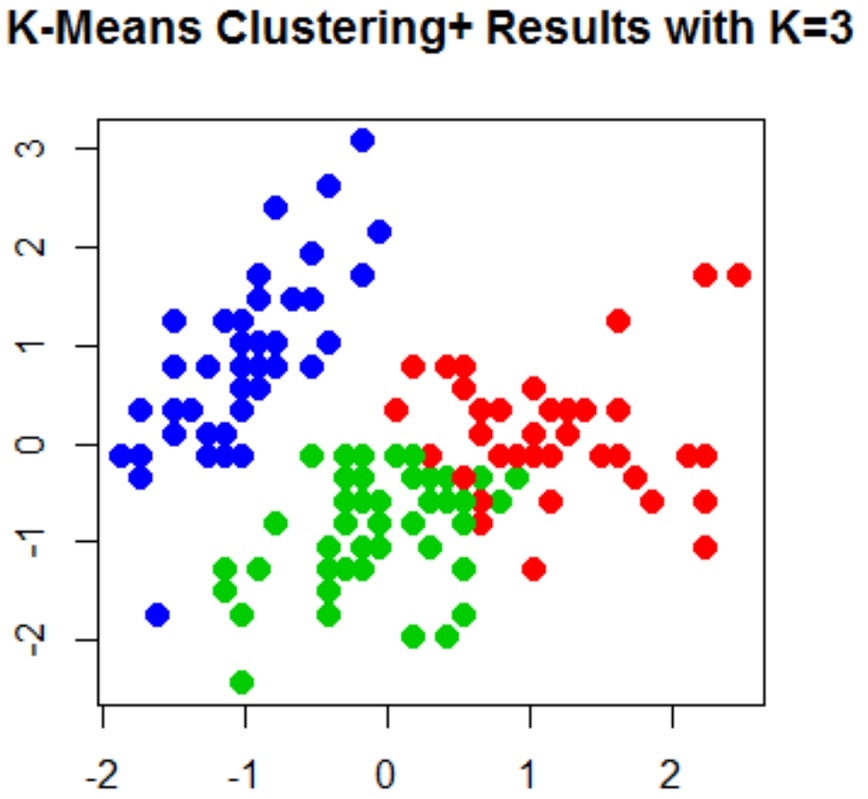
Fisher's iris data (famous)

- Có 3 loài hoa: setosa, versicolor, virginica
- 4 tiêu chí: S.Length S.Width P.Length, P.Width
- Mục tiêu: dựa vào tiêu chí để phân loại các loài hoa

```
data(iris)
head(iris)
  S.Length S.Width P.Length P.Width  Species
1      5.1     3.5      1.4     0.2 I.setosa
2      4.9     3.0      1.4     0.2 I.setosa
3      4.7     3.2      1.3     0.2 I.setosa
4      4.6     3.1      1.5     0.2 I.setosa
5      5.0     3.6      1.4     0.2 I.setosa
6      5.4     3.9      1.7     0.4 I.setosa
```

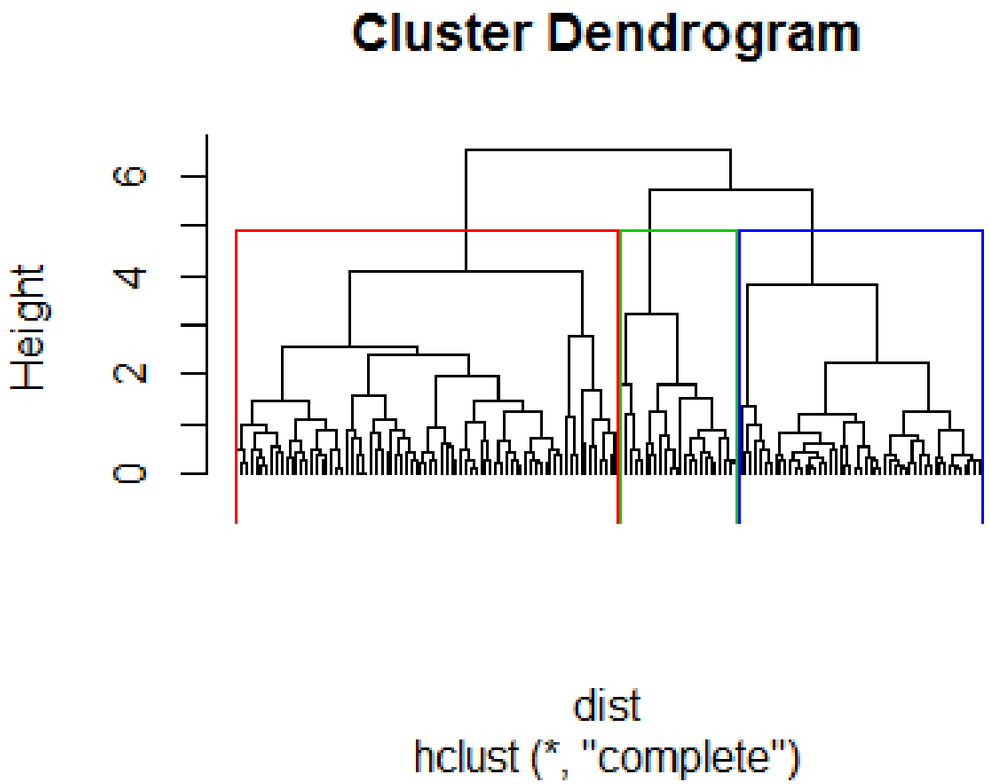
Clustering in R - Kmeans

```
data(iris)
s.iris <- scale(iris[1:4], center = TRUE,
scale = TRUE)
km.res = kmeans(s.iris, 3, nstart = 25)
km.res$cluster
plot(s.iris, col=(km.res$cluster +1),
main="K-Means Clustering+ Results with
K=3", xlab="", ylab="", pch=20, cex=2)
km.res
```



Clustering in R – Hierarchical Clustering

```
dist = dist(s.iris, method = "euclidean")
hc = hclust(dist, method = "complete")
plot(hc, labels = FALSE, hang = -1)
rect.hclust(hc, k = 3, border = 2:4)
```



Tóm lược

- Phân tích cụm: Phương pháp rất hữu hiệu để "nhận dạng" nhóm dựa vào các biến quan sát
- Chủ yếu dựa vào khoảng cách giữa các nhóm dựa trên dữ liệu thực tế
- Một phương pháp rất có ích trong thời đại "Big Data"

Câu hỏi?