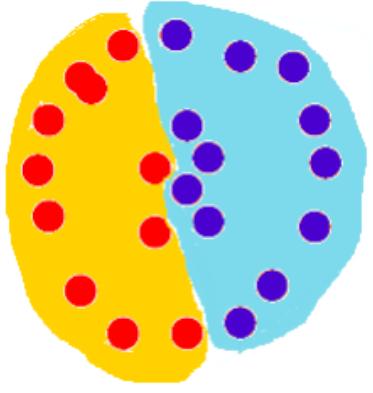
Bài 11:
Học không giám sát - Gom cụm

Nội dung bài học

Giới thiệu

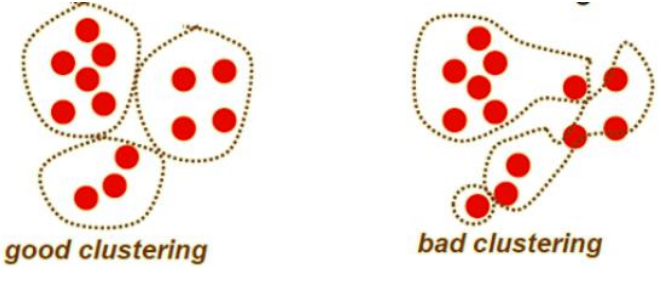
- Tổ chức dữ liệu không được gắn nhãn thành các nhóm tương tự gọi là cụm.
- Một cụm là một tập hợp các mục dữ liệu "tương tự" và "khác nhau" tới các mục dữ liệu trong các cụm khác.



Cần gì cho học không giám sát?

1. Độ đo khoảng cách:
   a. Độ đo tương đồng $s(x_i, x_k)$: lớn nếu x_i, x_k là tương đồng
   b. Độ đo bất đồng (khoảng cách) $d(x_i, x_k)$: nhỏ nếu x_i, x_k là tương đồng

1. Hàm tiêu chí để đánh giá cụm



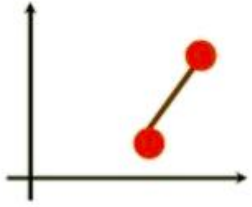
1. Thuật toán để tính cụm (tối ưu hàm tiêu chí)

1. Độ đo khoảng cách (bất đồng)

- Euclidean distance

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} \left(x_i^{(k)} - x_j^{(k)}\right)^2}$$

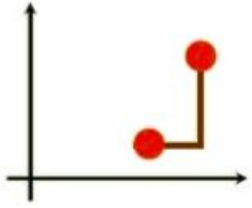
  - translation invariant



- Manhattan (city block) distance

$$d(x_i, x_j) = \sum_{k=1}^{d} \left|x_i^{(k)} - x_j^{(k)}\right|$$

  - approximation to Euclidean distance, cheaper to compute



- They are special cases of **Minkowski distance**:

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{m} \left|x_{ik} - x_{jk}\right|^p\right)^{\frac{1}{p}}$$

(p is a positive integer)

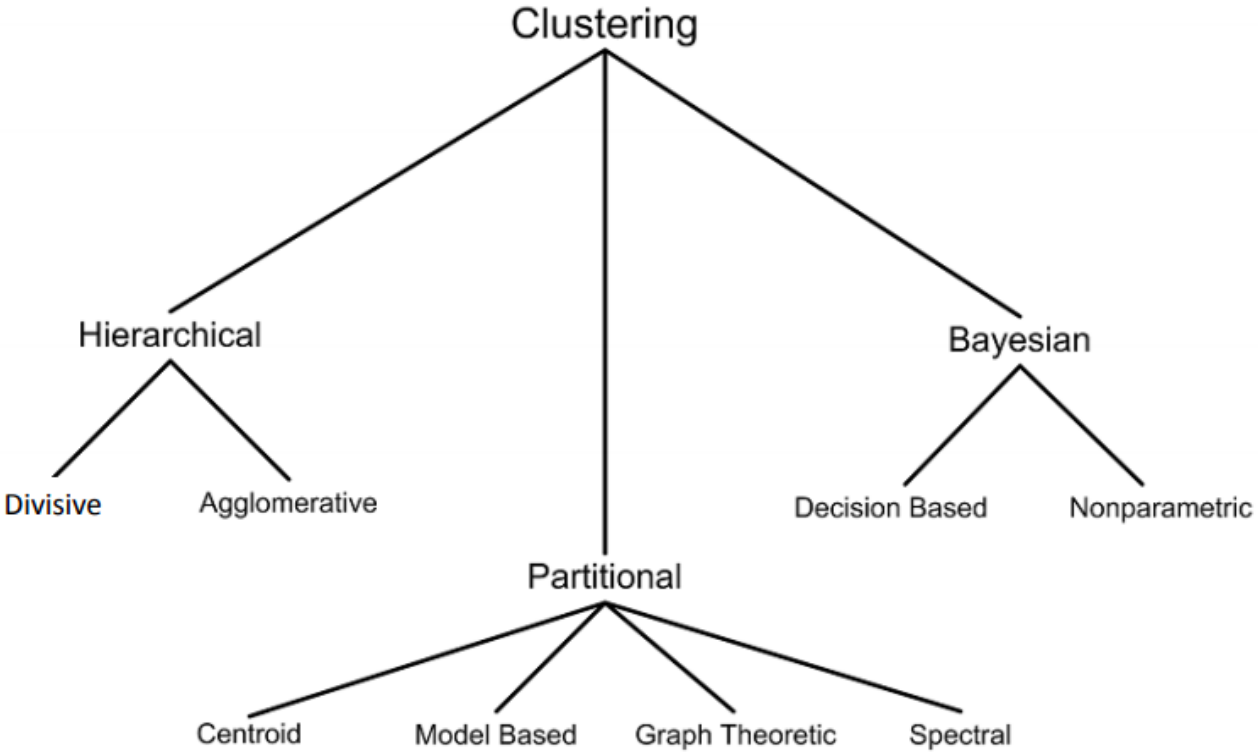
2. Đánh giá cụm

- Sự gắn kết nội bộ (Intra-cluster cohesion):
  - Độ gắn kết chỉ ra khoảng cách giữa các điểm dữ liệu trong 1 cụm đến tâm của cụm
  - Tổng sai số bình phương (Sum of squared error) là đơn vị đo thường được sử dụng
- Sự phân tách cụm (Inter-cluster separation):
  - Sự phân tách nghĩa là các tâm của các cụm khác nhau phải cách xa nhau

Cần bao nhiêu cụm?

- Hướng tiếp cận:
  - Cố định số cụm tới một số ***k***
  - Tìm ra cụm tốt nhất theo hàm tiêu chí (số lượng cụm có thể khác)

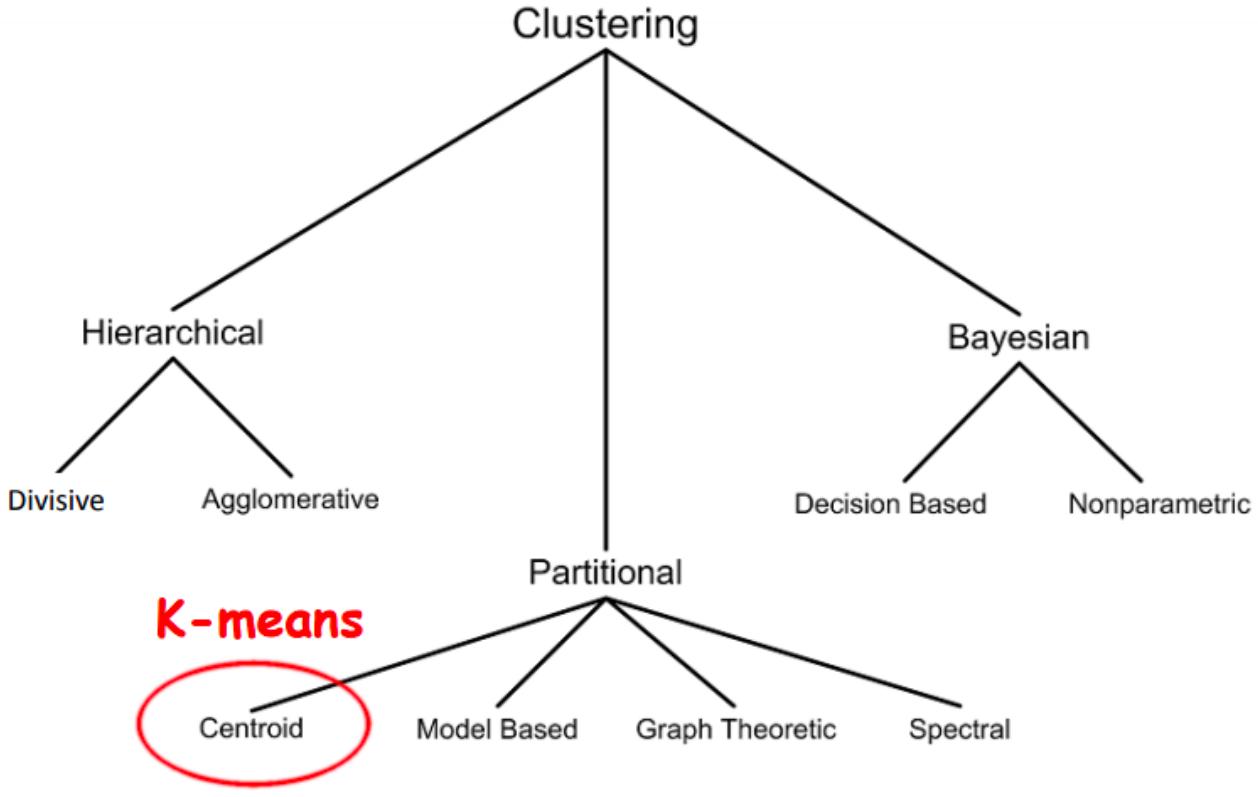
Các kỹ thuật gom cụm



Các kỹ thuật gom cụm

- Các thuật toán phân cấp (hierarchical) tìm các cụm liên tiếp bằng cách sử dụng các cụm đã được thiết lập trước đó. Các thuật toán này có thể là cộng gộp (từ dưới lên) hoặc chia (từ trên xuống)
  - Thuật toán cộng gộp (agglomerative) bắt đầu với mỗi phần tử là một cụm riêng biệt và hợp nhất chúng thành các cụm lớn hơn liên tiếp
  - Các thuật toán chia (divisive) bắt đầu với toàn bộ tập hợp và tiến hành chia nó thành các cụm nhỏ hơn liên tiếp
- Các thuật toán phân chia (partition) thường xác định tất cả các cụm cùng một lúc, nhưng cũng có thể được sử dụng như các thuật toán chia trong phân nhóm phân cấp
- Các thuật toán Bayes phát sinh phân phối hậu kỳ (posteriori distribution) trên tập hợp các phân chia (partition) của dữ liệu

Các kỹ thuật gom cụm



K-means

- K-means (MacQueen, 1967) is một thuật toán gom cụm phân chia
  - Giả sử có tập các điểm D = $\{x_1, x_2, ..., x_n\}$
    - trong đó, $x_i = (x_{i1}, x_{i2}, ..., x_{ir})$ là 1 vector trong $X \subseteq \mathbf{R}^r$, và r là số chiều của vector
  - Thuật toán k-means phân chia dữ liệu thành k cụm:
    - Một cụm có một trung tâm gọi là **centroid**
    - k được xác định bởi người dùng

K-means

- Cho k, thuật toán k-means hoạt động như sau:
  1. Chọn k (ngẫu nhiên) điểm (seed) để khởi tạo các tâm (centroid)
  2. Gán các điểm dữ liệu đến các tâm gần nhất
  3. Tính toán lại tâm sử dụng các thành viên (membership) của cụm
  4. Nếu tiêu chí hội tụ chưa thỏa, lặp lại bước 2 và 3

K-means - Tiêu chí hội tụ

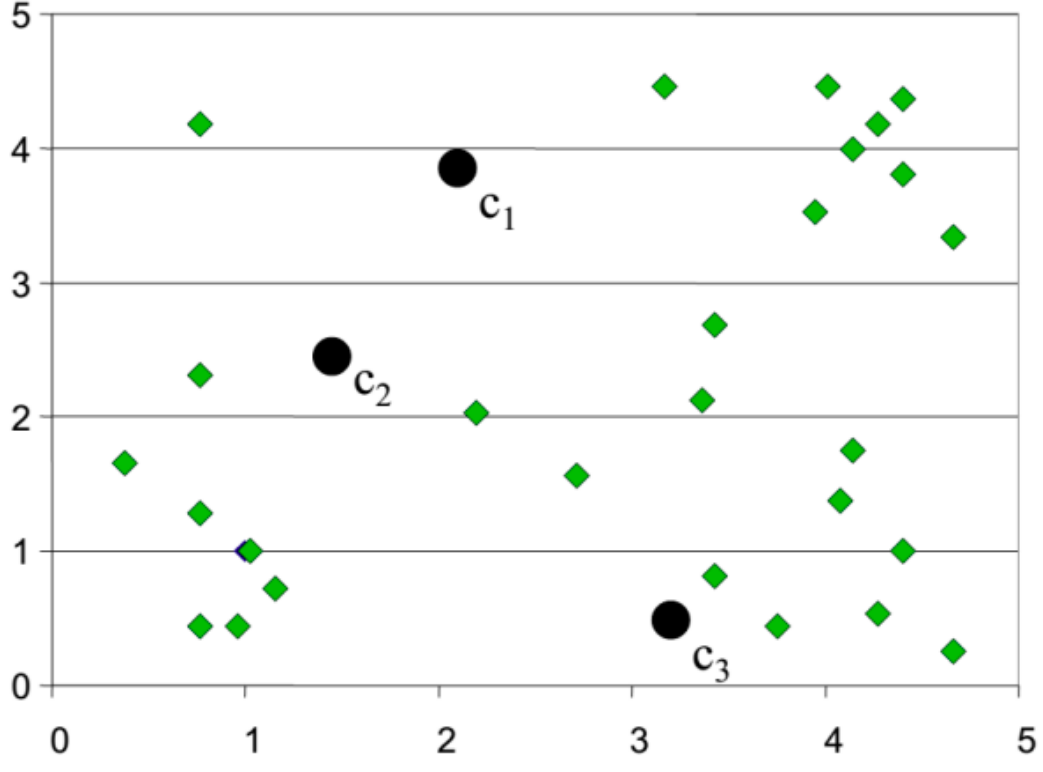
- K-means hội tụ khi thỏa 1 trong những điều kiện:
  - Không có (hoặc tối thiểu) khả năng gán các điểm dữ liệu đến các cụm khác
  - Không có (hoặc tối thiểu) thay đổi tâm
  - Giảm tối thiểu đối với độ đo sai số bình phương (SSE):

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m}_j)^2$$

- C_j is the jth cluster,
- $\mathbf{m}_j$ is the centroid of cluster C_j (the mean vector of all the data points in C_j),
- $d(\mathbf{x}, \mathbf{m}_j)$ is the (Eucledian) distance between data point $\mathbf{x}$ and centroid $\mathbf{m}_j$.

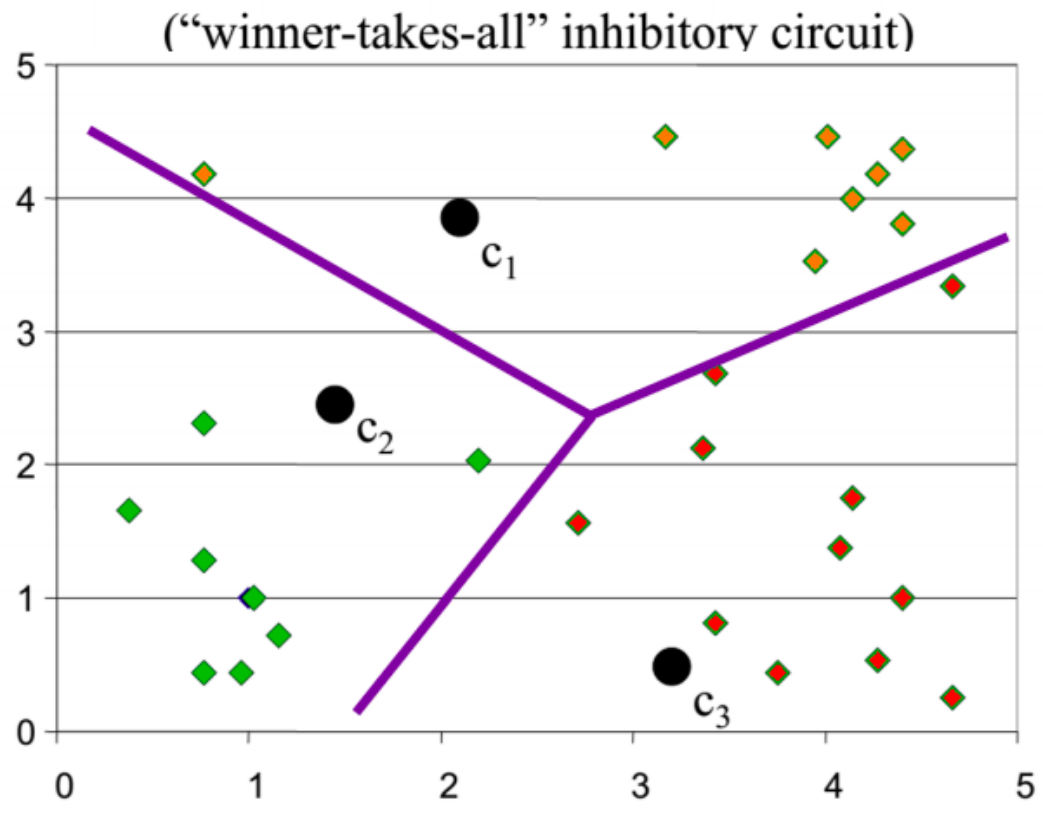
Minh họa k-means

- Khởi tạo ngẫu nhiên các tâm cụm



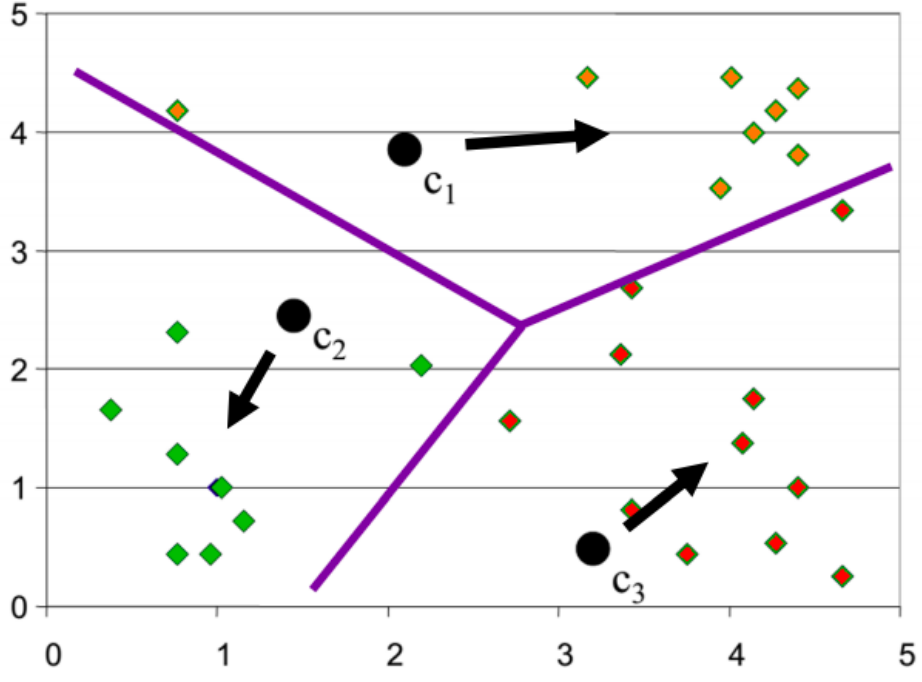
Minh họa k-means

- Xác đinh thành viên cho từng cum



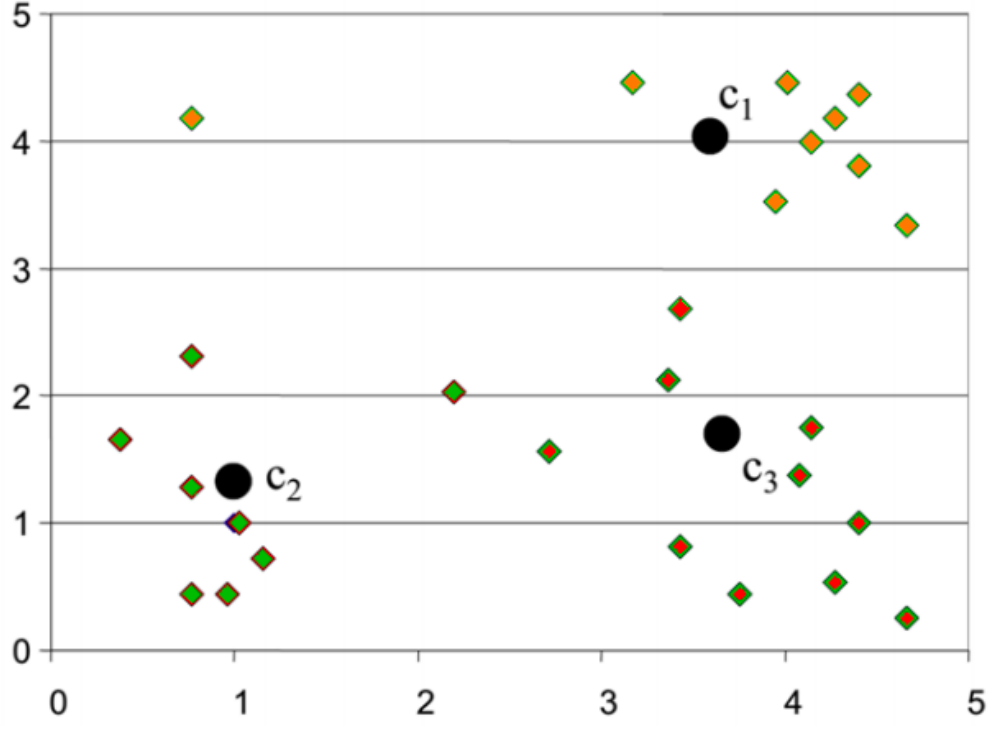
Minh họa k-means

- Tính toán lại các tâm cụm



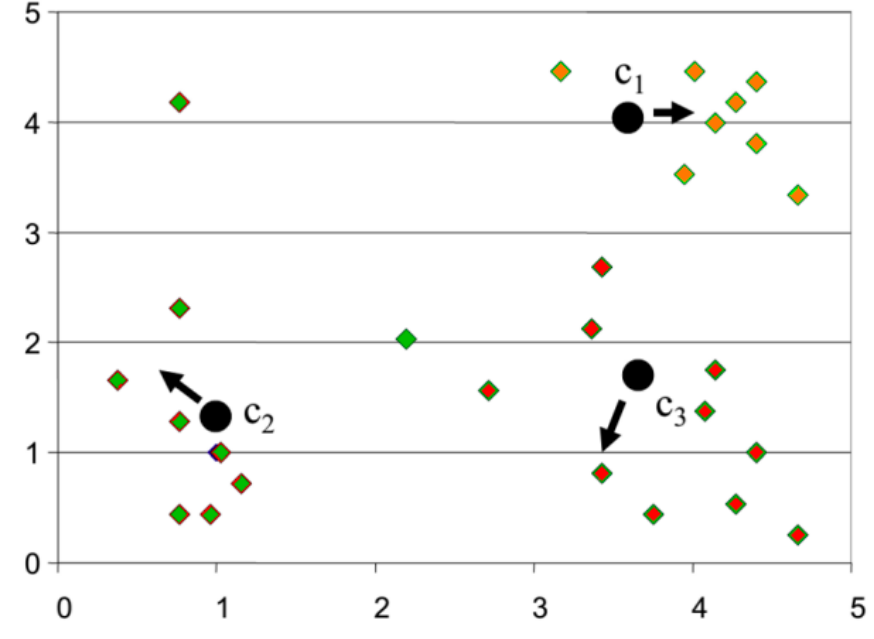
Minh họa k-means

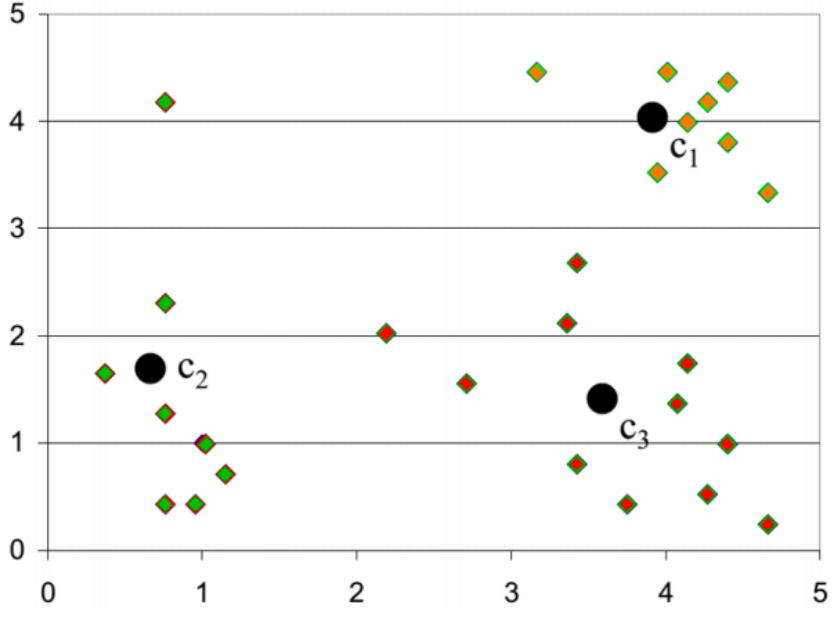
- Kết quả của lần lặp thứ nhất:



Minh họa k-means

- Lần lặp thứ hai:





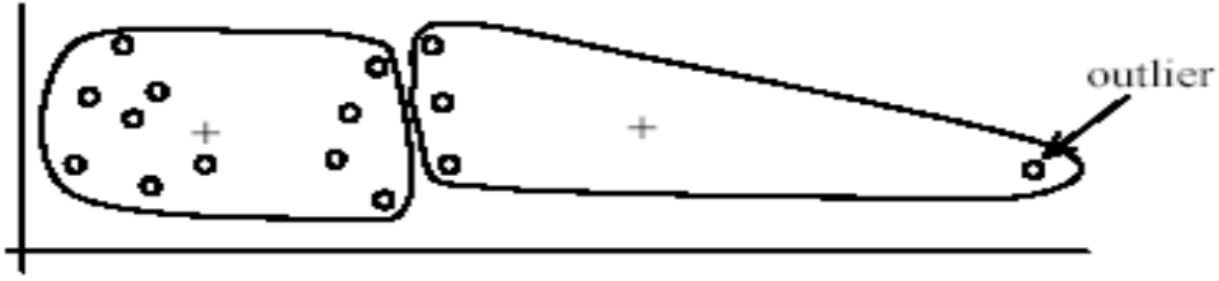
Tại sao dùng k-means?

- Ưu điểm:
  - Đơn giản: dễ hiểu, dễ cài đặt
  - Hiệu quả: độ phức tạp O(tkn)
    - n: số điểm
    - k: số cụm
    - t: số lần lặp
  - K-means là thuật toán gom cụm phổ biến nhất

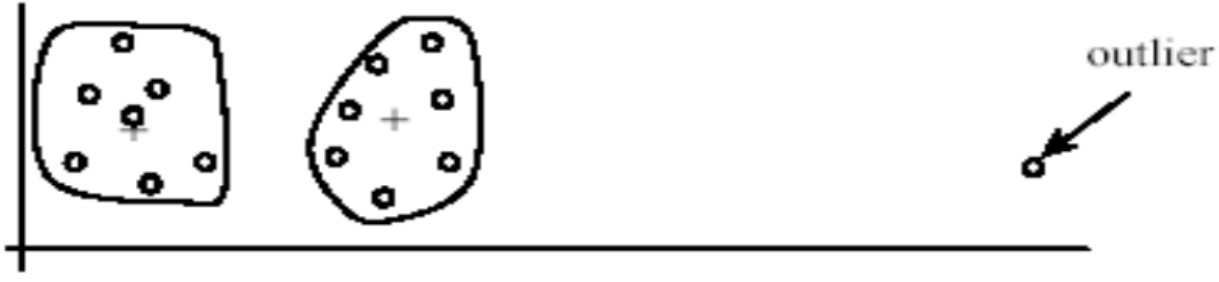
Tại sao dùng k-means?

- Khuyết điểm:
  - Thuật toán chỉ hoạt động với dữ liệu số
    - Đối với dữ liệu phân loại (category) -> k-mode (tâm được biểu diễn bằng giá trị xuất hiện nhiều nhất)
  - Người dùng phải xác định k
  - Nhạy cảm với các điểm ngoại lệ (outlier):
    - Điểm ngoại lệ là các điểm nằm xa các điểm dữ liệu
    - Điểm ngoại lệ có thể do lỗi trong quá trình thu thập dữ liệu hoặc các điểm đặc biệt có giá trị khác

Điểm ngoại lệ



(A): Undesirable clusters

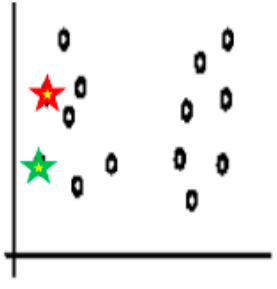


(B): Ideal clusters

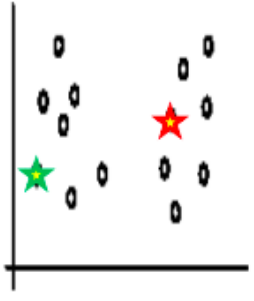
Điểm ngoại lệ

- Xóa bỏ những điểm nằm quá xa tâm hơn các điểm khác
- Thực hiện lấy mẫu ngẫu nhiên: chọn một tập nhỏ các điểm dữ liệu, khả năng chọn trúng các điểm ngoại lệ sẽ nhỏ hơn

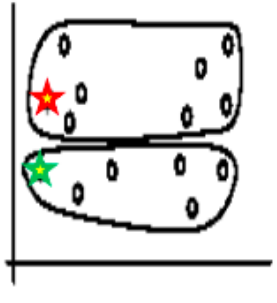
Nhạy cảm với khởi tạo ban đầu



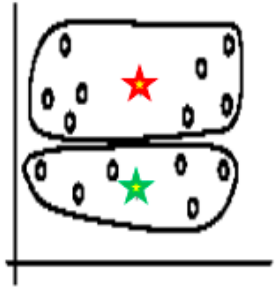
Random selection of seeds (centroids)



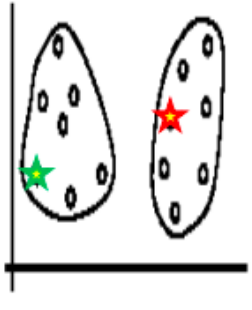
Random selection of seeds (centroids)



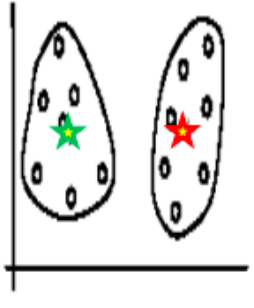
Iteration 1



Iteration 2



Iteration 1



Iteration 2

Tổng kết

- Mặc dù có các điểm yếu như trên, k-means vẫn là thuật toán phổ biến do tính đơn giản và hiệu quả
- Không có bằng chứng rõ ràng rằng các thuật toán gom cụm khác thực hiện tốt hơn
- Thực hiện việc so sánh các thuật toán gom cụm là một tác vụ khó (không ai biết được cụm đúng)

Demo

Thank you and happy learning !!!