Chapter Clustering

K-means based method

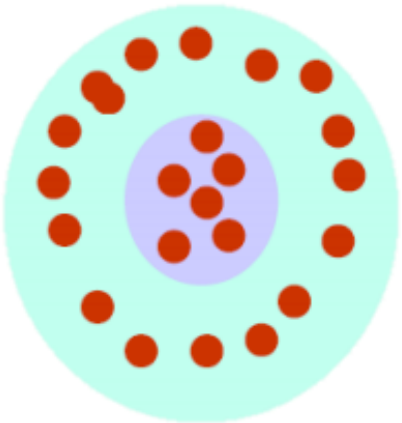
Nguyễn Giáp Nguyên Sinh
Nguyễn Quốc Long
Khoa Khoa học và Kỹ thuật Máy tính
Đại học Bách Khoa TP.HCM

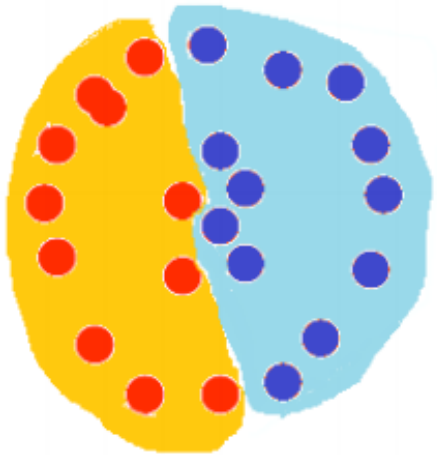
Nội dung

Phân cụm

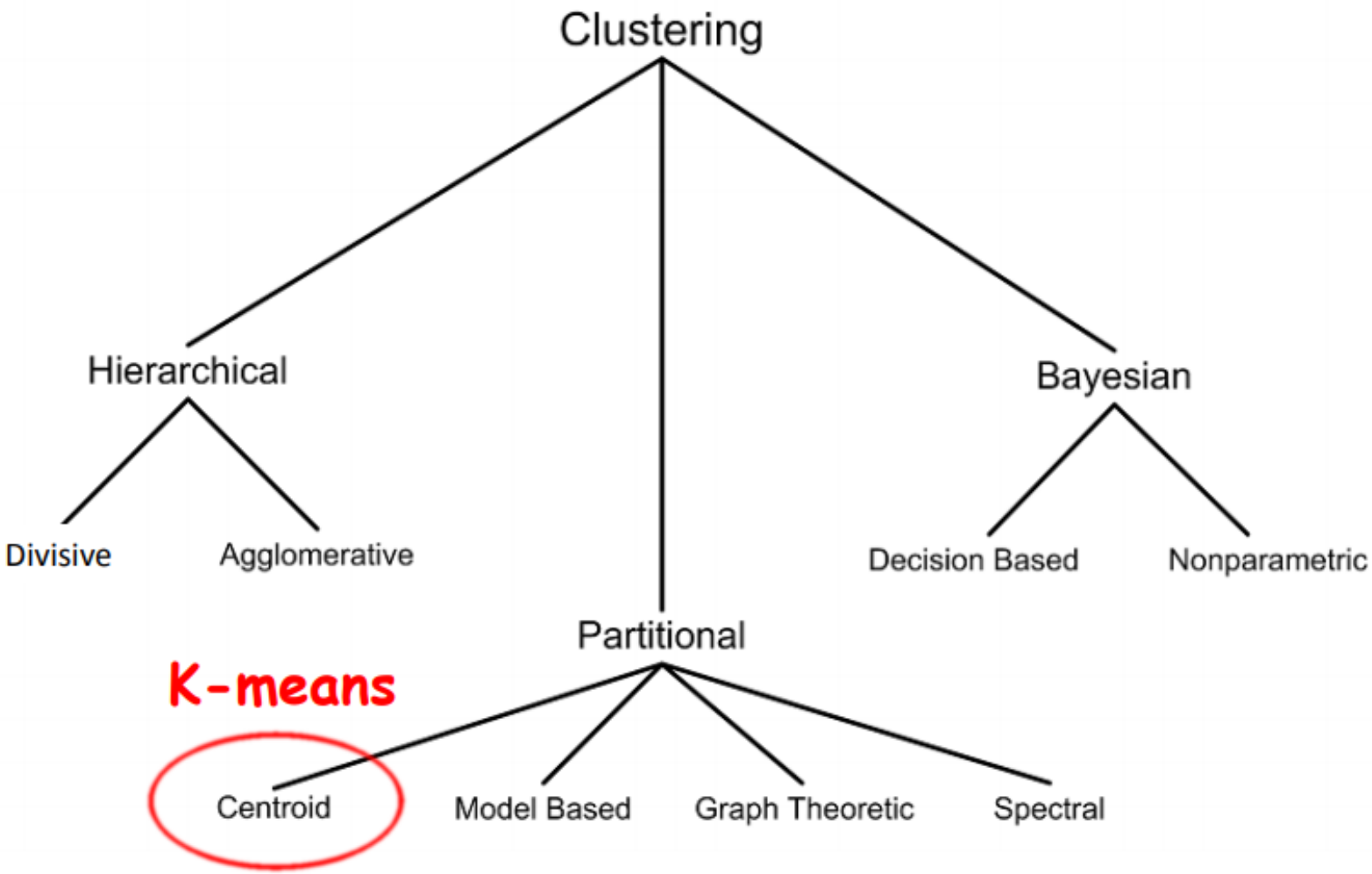
Phân cụm là gì ?

- Việc tổ chức dữ liệu chưa được gán nhãn (label) vào các nhóm tương tự nhau được gọi là phân cụm
- Một cụm là một tập hợp các phần tử dữ liệu có sự giống nhau về mặt dữ liệu và sẽ khác với các phần tử dữ liệu ở các cụm khác.





Kỹ thuật phân cụm



Vấn đề

Khi nào nghĩ đến K-means

- Không biết nhãn(label) của điểm dữ liệu
- Mục đích : Phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong một cụm có tính chất giống nhau.

K-means cluster

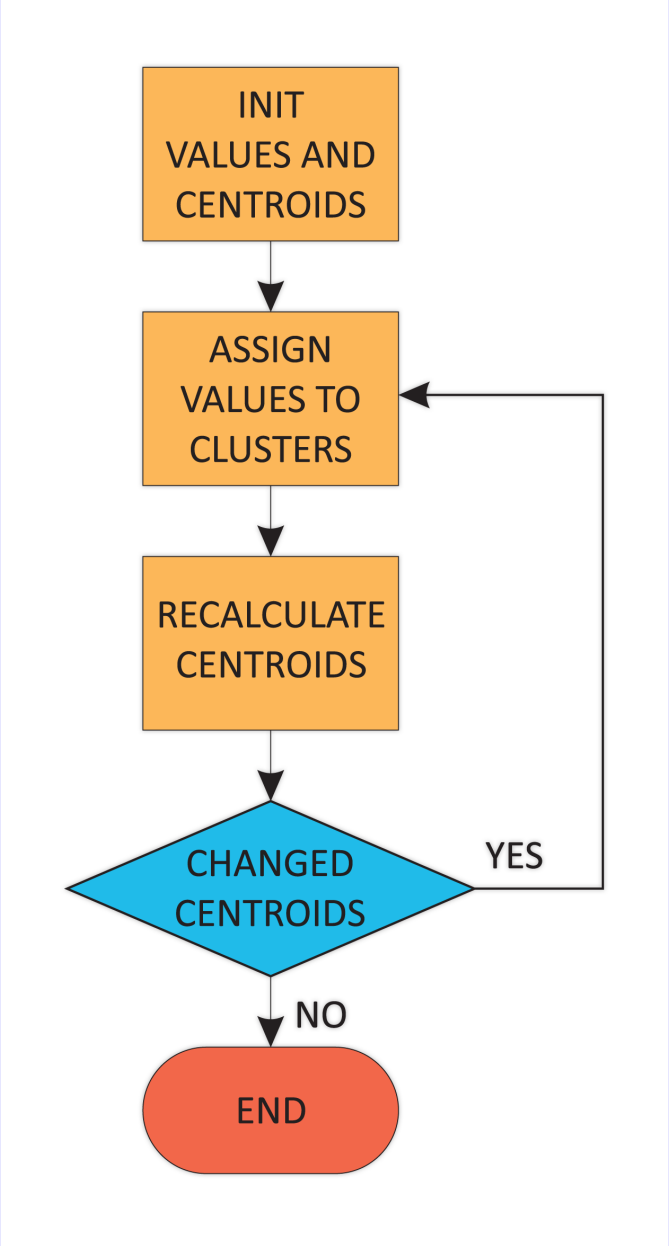
- K-means được đề xuất bởi MacQueen năm 1967
- Giải thuật k-means chia tập dữ liệu thành k cụm (cluster)
  - Mỗi cluster có một điểm trung tâm , gọi là *centroid*
  - K được chỉ định bởi nhân viên phân tích dữ liệu (Data analytics)

Giải thật k-means

Nhập giá trị k, giải thuật k-means sẽ thực thi các bước như sau:

1. Chọn ngẫu nhiên k điểm dữ liệu làm *centroids*, điểm trung tâm của cụm dữ liệu.
2. Phân mỗi điểm dữ liệu vào cluster có điểm trung tâm (center) gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Cơ chế của giải thuật K-means có thể tổng quát bằng sơ đồ dưới đây:



Cách tính khoảng cách và điểm trung tâm

Cách tính khoảng cách trong giải thuật K-means

Để tìm được điểm dữ liệu gần với điểm trung tâm nào nhất thì ta dựa vào giá trị nhỏ nhất của hàm tính khoảng cách Euclidean.

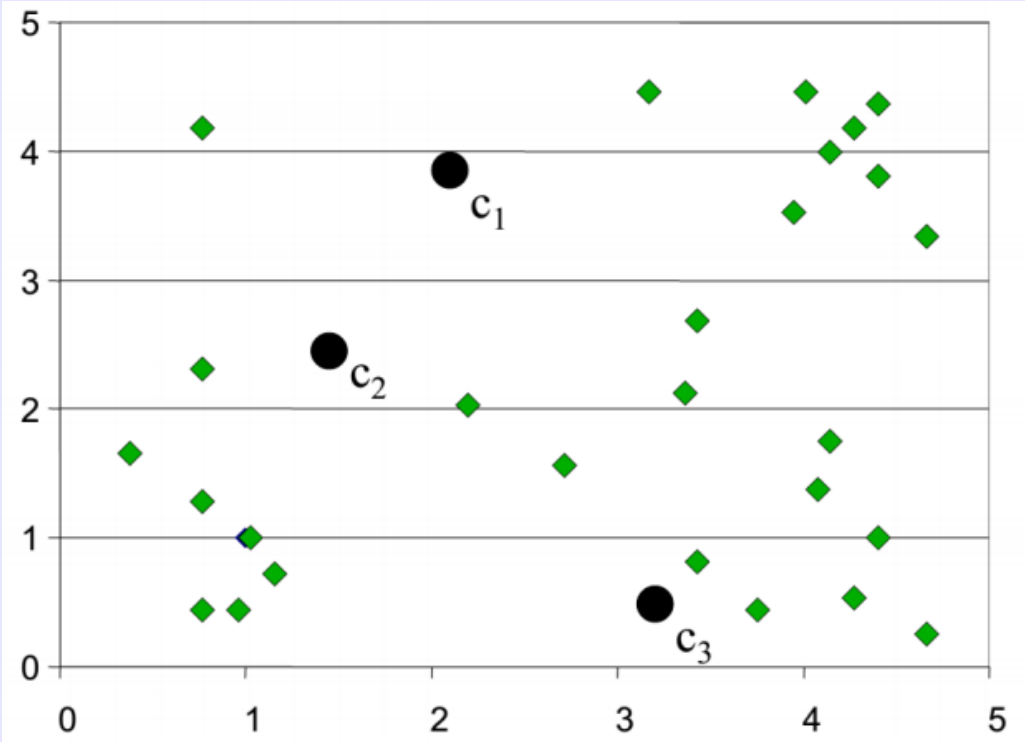
- Với dữ liệu 1 chiều, với 2 điểm dữ liệu p và q
  $\sqrt{p-q} = |p-q|$
- Với dữ liệu 2 chiều, với 2 điểm dữ liệu p (p1,p2) và q (q1, q2)
  $\sqrt{(p1-q1)^2 + (p2-q2)^2}$
- Với dữ liệu 3 chiều, với 2 dữ liệu p(p1,p2,p3) và q(q1,q2,q3)
  $\sqrt{(p1-q1)^2 + (p2-q2)^2 + (p3-q3)^2}$
- Với dữ liệu n chiều, với 2 dữ liệu p(p1,p2,...,pn) và q(q1,q2,...,qn)
  $\sqrt{(p1-q1)^2 + (p2-q2)^2 + ... + (pn-qn)^2}$

Cách tính điểm trung tâm

Điểm trung tâm *centroid* của cluster là trung bình cộng của các điểm trong cluster đó.

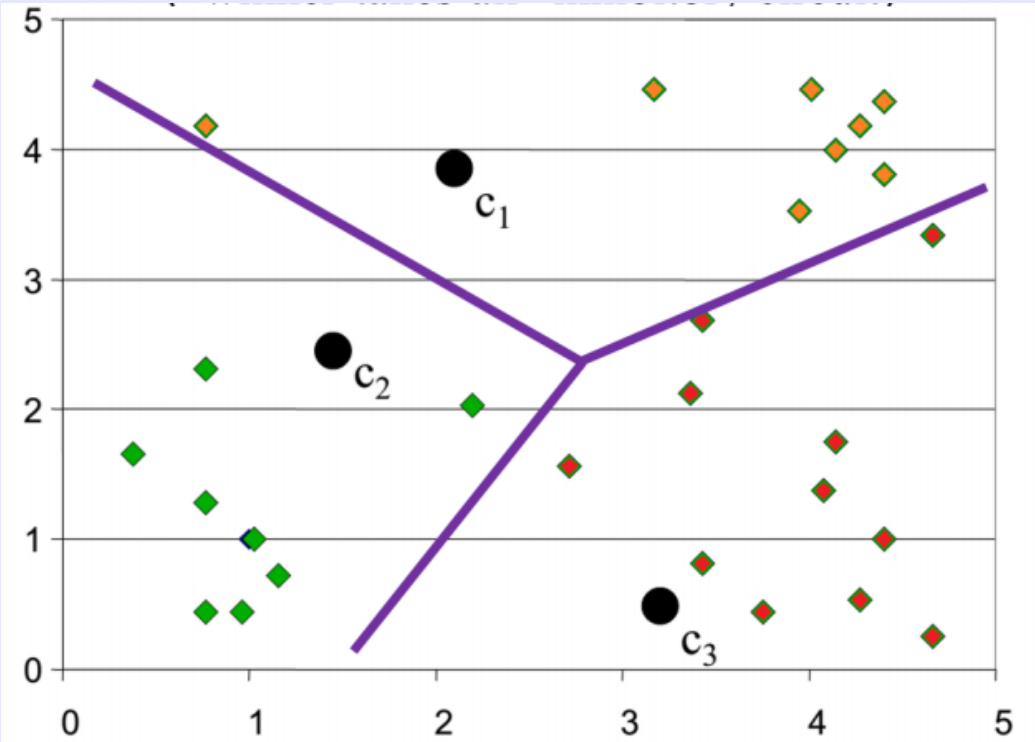
Mô phỏng K-means

Chọn ngẫu nhiên các centroid



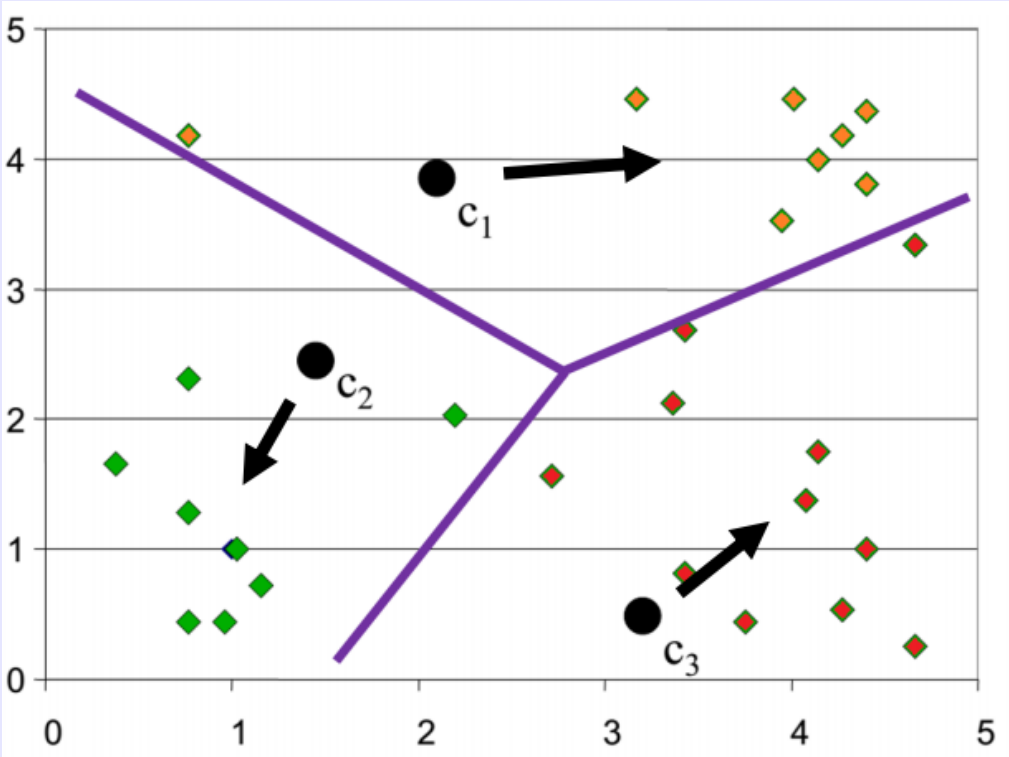
Mô phỏng K-means

Xác định cluster cho từng điểm dữ liệu



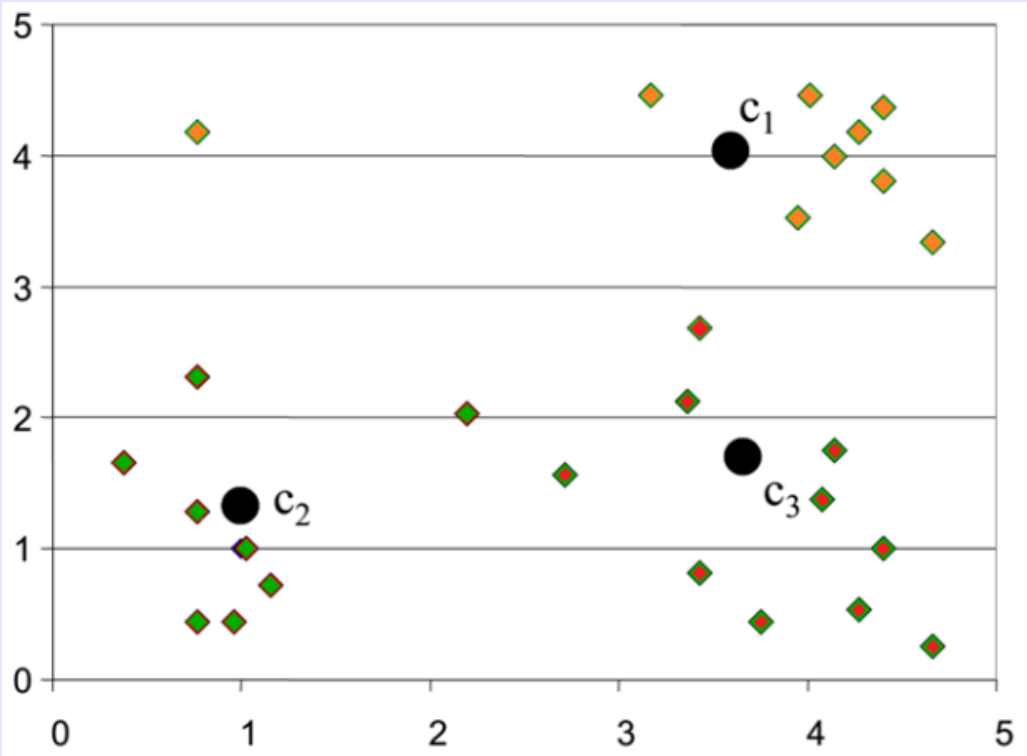
Mô phỏng K-means

Xác định lại centroid cho các cluster



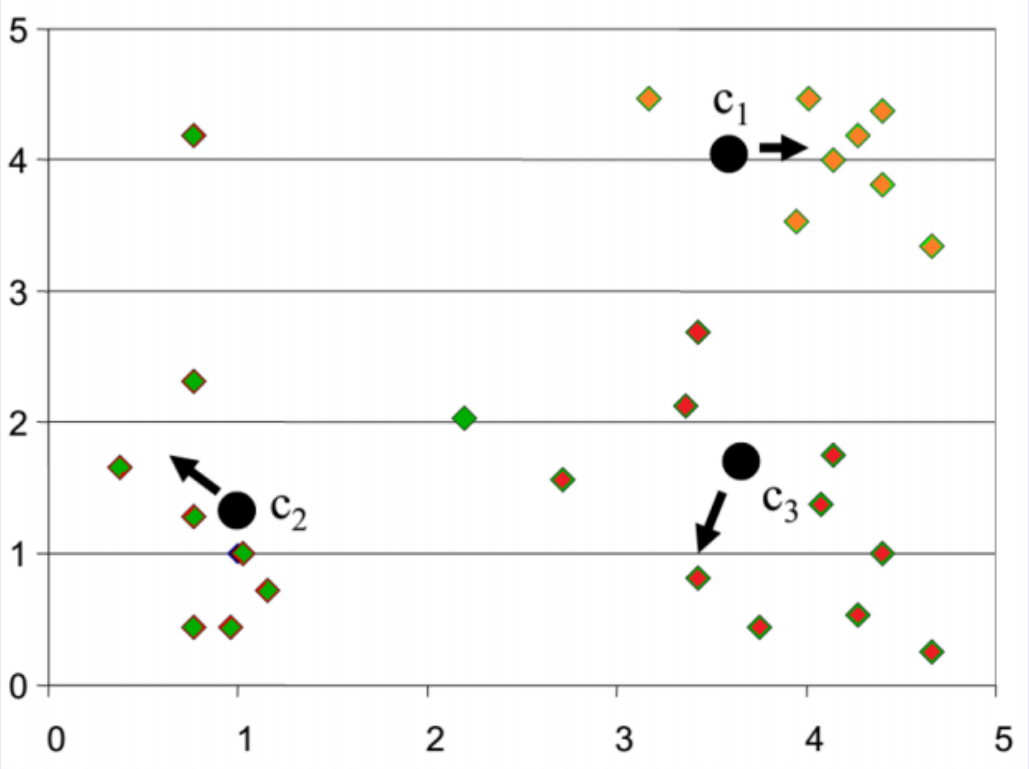
Mô phỏng K-means

Kết quả của vòng lặp thứ nhất



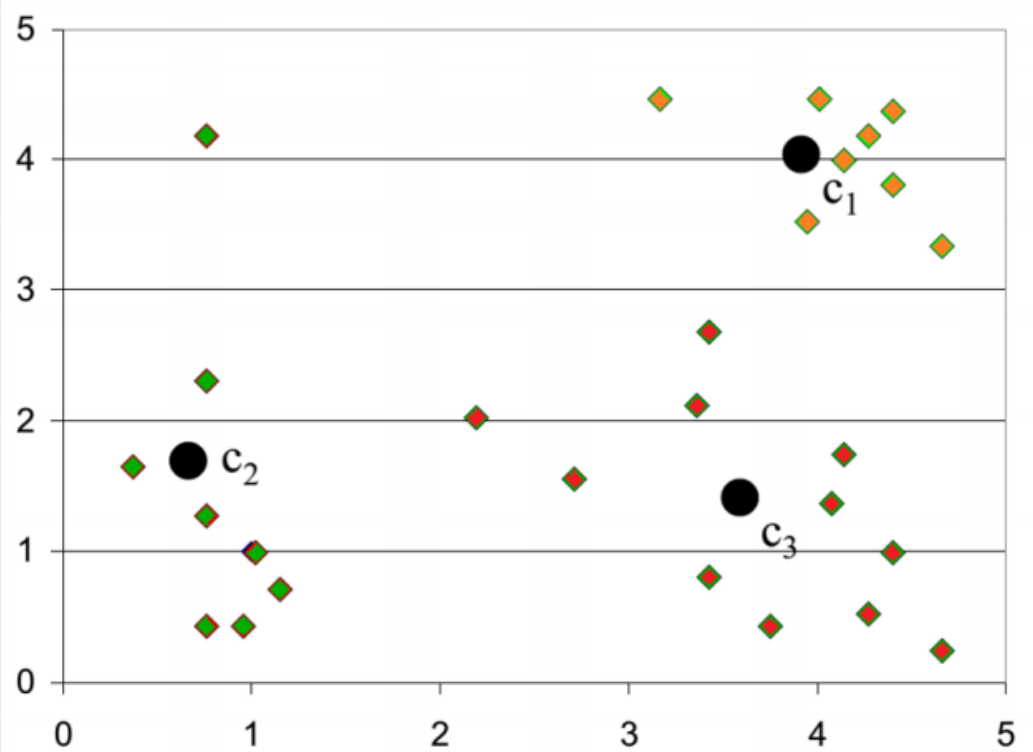
Mô phỏng K-means

Vòng lặp thứ 2, xác định lại các điểm centroid



Mô phỏng K-means

Kết quả của vòng lặp thứ hai



Điểm mạnh và điểm yếu

Điểm mạnh

- Đơn giản: Dễ dàng hiểu và thực thi
- Rất hiệu quả: Độ phức tạp chỉ là O(tkn), trong đó:
  - n là số lượng điểm dữ liệu (data point)
  - k là số cụm (cluster)
  - t là số lần lặp cho đến khi hội tụ

  Vì k và t đều nhỏ nên giải thuật K-mean là thuật toán có độ phức tạp tuyến tính

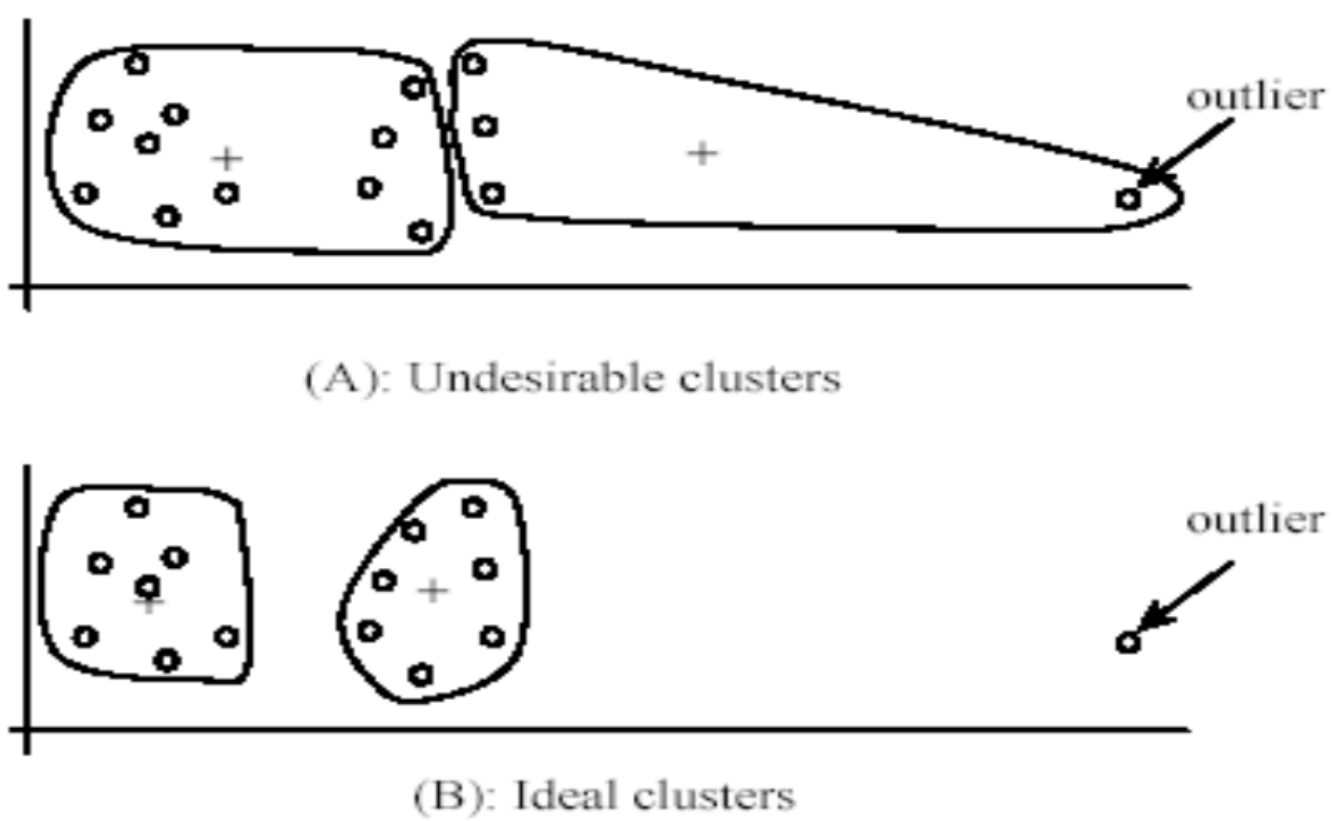
k-means là giải thuật phổ biến.

Điểm yếu

- Thuật toán chỉ áp dụng được khi các giá trị trung bình được xác định.
- Phải xác định số lượng cụm (cluster)
- Giải thuật bị ảnh hưởng bởi các điểm *outliers*
  - Outliers là những điểm dữ liệu ở quá xa so với các điểm dữ liệu khác.
  - Outliers tồn tại do lỗi trong quá trình ghi dữ liệu hoặc một số điểm dữ liệu đặc biệt có các giá trị rất khác nhau.

Điểm yếu

Outliers

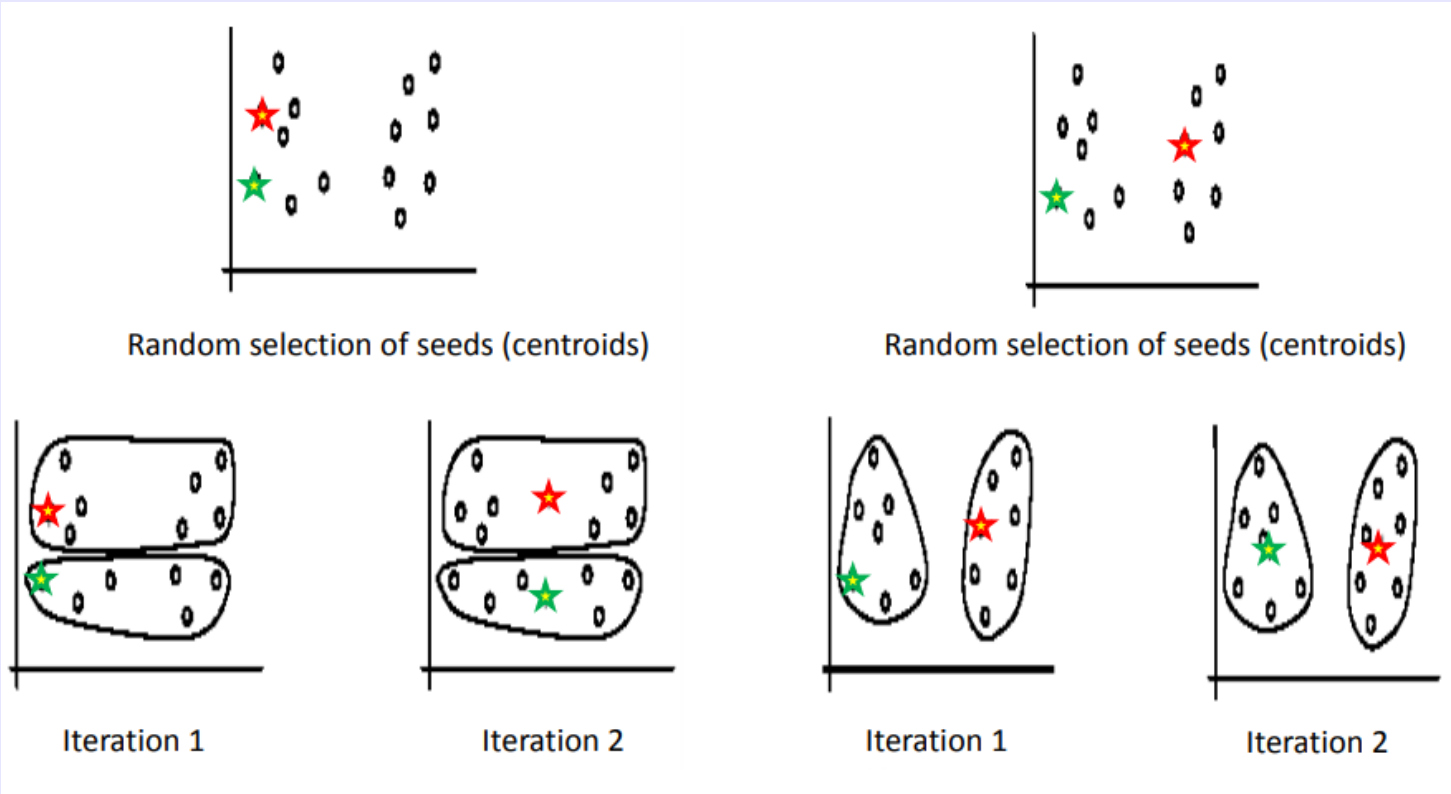


Giải quyết Outlier

- Xóa các điểm dữ liệu ở xa centroid so với các điểm dữ liệu khác.
  - Để đảm bảo an toàn, chúng ta có thể theo dõi các điểm này qua vài lần lặp trước khi quyết loại bỏ.
- Một cách khác, bằng cách lấy mẫu ngẫu nhiên: Vì trong khi lấy mẫu, chúng ta chỉ chọn tập hợp con của các điểm dữ liệu, nên rất ít khả năng chọn trúng điểm outlier.
  - Gán những điểm dữ liệu còn lại cho các clusters theo khoảng cách hoặc phân lớp.

Điểm yếu

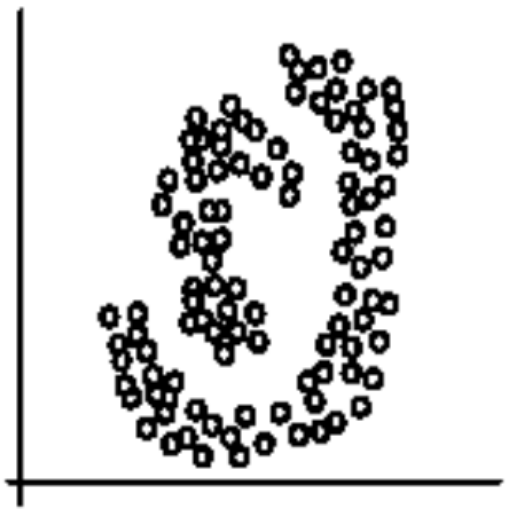
Sensitivity to initial seeds



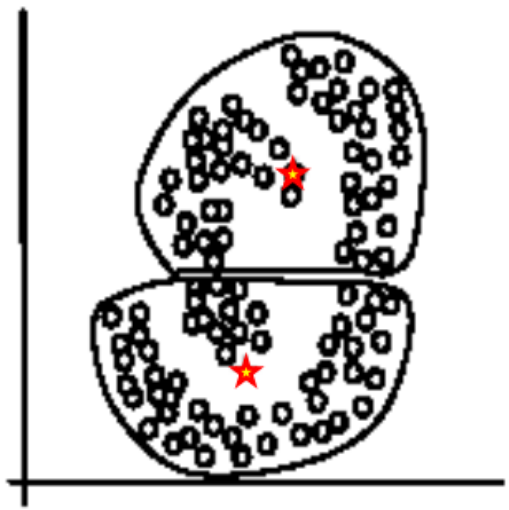
Điểm yếu

Cấu trúc dữ liệu đặc biệt

Thuật toán k-means không phù hợp để khai thác các cluster không phải là hyper - ellipsoids (hyper-spheres)



(A): Two natural clusters



(B): k-means clusters

Tổng kết

Tổng kết giải thuật K-means

- Mặc dù có điểm yếu nhưng k-means vẫn là giải thuật phổ biến do tính đơn giản và tính hiệu quả của nó.
- Chưa có bằng chứng nào chứng tỏ giải thuật phân cụm nào là tốt nhất.

Tham khảo

1. http://www.mit.edu/ 9.54/fall14/slides/Class13.pdf
2. https://www.saedsayad.com/clustering_kmeans.htm
3. Bing Liu, Web data Mining, Springer, Second Edition, 2007