Học Máy

(Machine Learning)

Thân Quang Khoát

khoattq@soict.hust.edu.vn

Viện Công nghệ thông tin và Truyền thông
Trường Đại học Bách Khoa Hà Nội

Năm học 2013-2014

Nội dung môn học:

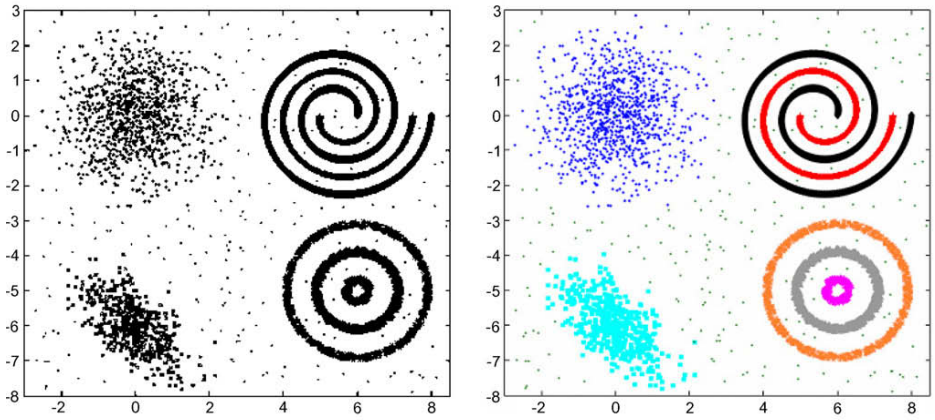
- Giới thiệu chung
- Các phương pháp học dựa trên xác suất
- Các phương pháp học có giám sát
- Chuẩn hoá
- Đánh giá hiệu năng hệ thống học máy
- **Các phương pháp học không giám sát**
  - **Giới thiệu về phân cụm**
  - **Phương pháp k-Means**

Hai vấn đề học

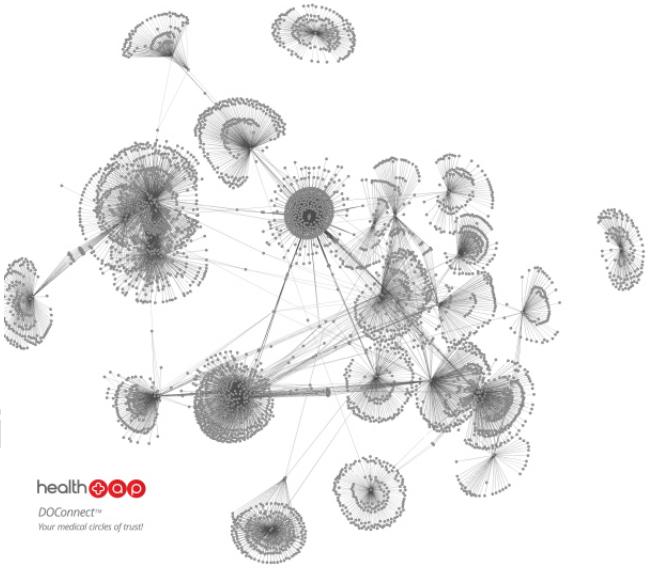
- **Học có giám sát (Supervised learning)**
  - Tập dữ liệu học (training data) bao gồm các ví dụ, mà mỗi ví dụ được *gắn kèm với một nhãn lớp hoặc giá trị đầu ra mong muốn.*
  - Mục đích là học (xấp xỉ) một hàm (vd: một phân lớp, một hàm hồi quy,...) phù hợp với tập dữ liệu hiện có.
  - Hàm học được sau đó sẽ được dùng để phân lớp hoặc dự đoán đối với các ví dụ mới.
- **Học không giám sát (Unsupervised learning)**
  - Tập học (training data) bao gồm các ví dụ, mà mỗi ví dụ *không có thông tin về nhãn lớp hoặc giá trị đầu ra mong muốn.*
  - Mục đích là tìm ra (học) các cụm, các cấu trúc, các quan hệ tồn tại ẩn trong tập dữ liệu hiện có.

Ví dụ về học không giám sát (1)

- Phân cụm (clustering)
  - Phát hiện các cụm dữ liệu, cụm tính chất,…

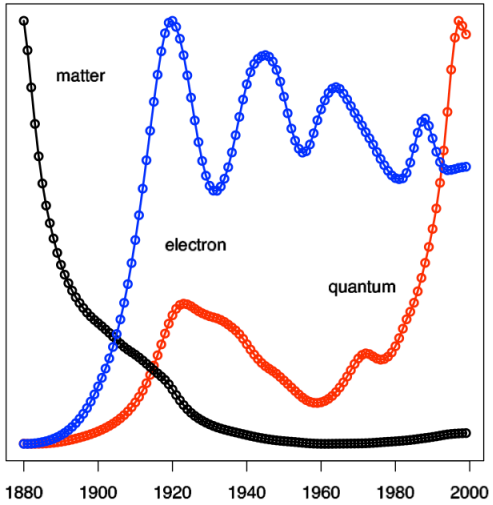


- Community detection
  - Phát hiện các cộng đồng trong mạng xã hội

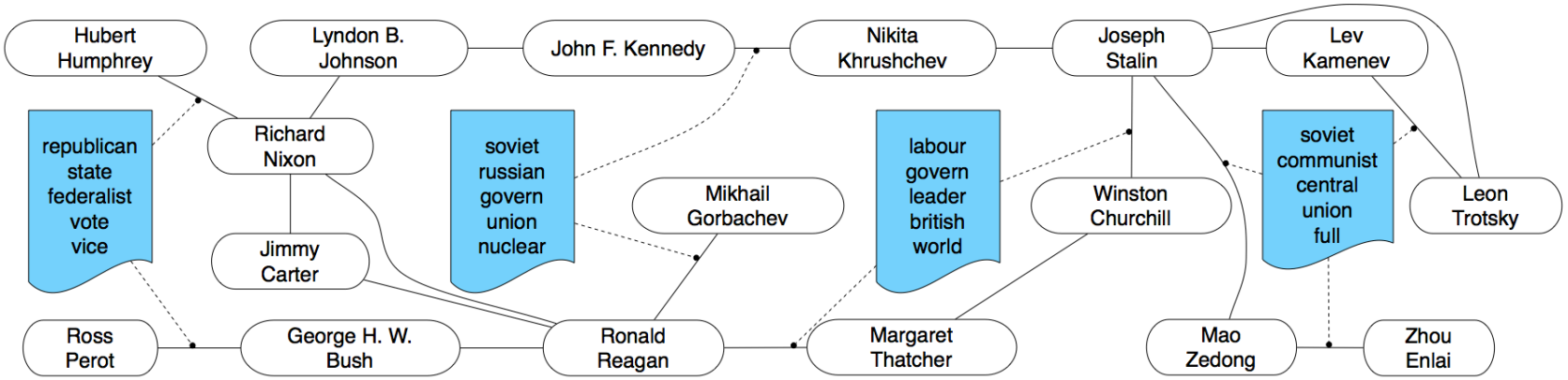


Ví dụ về học không giám sát (2)

- Trends detection
  - Phát hiện xu hướng, thị yếu,…



- Entity-interaction analysis



Phân cụm

- Học phân cụm
  - Đầu vào: một tập dữ liệu không có nhãn (các ví dụ không có nhãn lớp hoặc giá trị đầu ra mong muốn)
  - Đầu ra: các cụm (nhóm) của các ví dụ
- Một **cụm (cluster)** là một tập các ví dụ
  - Tương tự với nhau (theo một ý nghĩa, đánh giá nào đó)
  - Khác biệt với các ví dụ thuộc các cụm khác

Phân cụm (2)

- Giải thuật phân cụm
  - **Dựa trên phân hoạch (Partition-based clustering)**
  - **Dựa trên tích tụ phân cấp (Hierarchical clustering)**
  - Bản đồ tự tổ thức (Self-organizing map – SOM)
  - Các mô hình hỗn hợp (Mixture models)
  - …
- Đánh giá chất lượng phân cụm (Clustering quality)
  - Khoảng cách/sự khác biệt *giữa các cụm* → Cần được *cực đại hóa*
  - Khoảng cách/sự khác biệt *bên trong một cụm* → Cần được *cực tiểu hóa*

Phương pháp k-Means

- Là phương pháp phân cụm phổ biến nhất trong các phương pháp dựa trên phân hoạch (partition-based clustering)
- Tập dữ liệu $D=\{x_1,x_2,\ldots,x_r\}$
  - x_i là một ví dụ (một vectơ trong một không gian *n* chiều)
- Giải thuật k-means phân chia tập dữ liệu thành k cụm
  - Mỗi cụm (cluster) có một điểm trung tâm, được gọi là **centroid**
  - k (tổng số các cụm thu được) là một giá trị được cho trước (vd: được chỉ định bởi người thiết kế hệ thống phân cụm)

k-Means: Các bước chính

Với một giá trị k được xác định trước

- Bước 1. Chọn ngẫu nhiên k ví dụ (được gọi là **các hạt nhân – seeds**) để sử dụng làm *các điểm trung tâm ban đầu* (*initial centroids*) của *k* cụm.
- Bước 2. Lặp liên tục hai bước sau cho đến khi *gặp điều kiện hội tụ* (convergence criterion):
  - Bước 2.1. Đối với mỗi ví dụ, *gán nó vào cụm* (trong số k cụm) mà có tâm (centroid) gần ví dụ đó nhất.
  - Bước 2.2. Đối với mỗi cụm, *tính toán lại điểm trung tâm (centroid) của nó* dựa trên tất cả các ví dụ thuộc vào cụm đó.

***k*-means**(D, k)

- D: Tập ví dụ học
- k: Số lượng cụm kết quả (thu được)

Lựa chọn ngẫu nhiên k ví dụ trong tập D để làm các điểm trung tâm ban đầu (initial centroids)

while not CONVERGENCE

  for each ví dụ x∈D

    Tính các khoảng cách từ x đến các điểm trung tâm (centroid)

    Gán x vào cụm có điểm trung tâm (centroid) gần x nhất

  end for

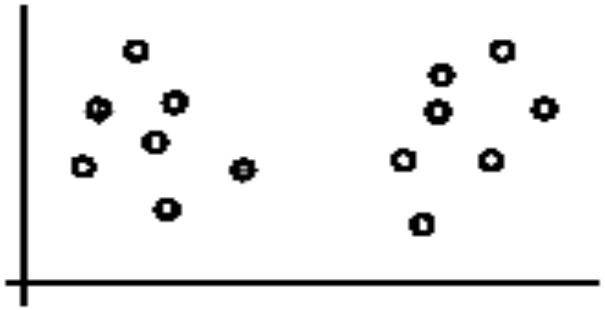
  for each cụm

    Tính (xác định) lại điểm trung tâm (centroid) dựa trên các ví dụ hiện thời đang thuộc vào cụm này

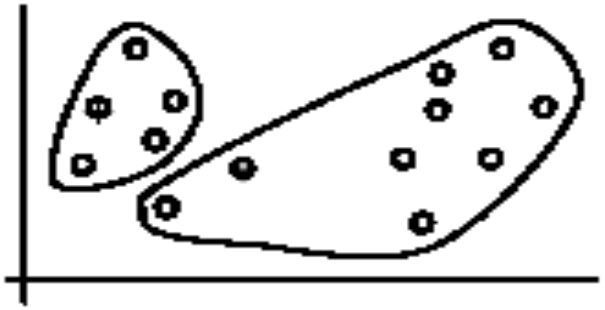
end while

return {k cụm kết quả}

k-Means: Minh họa (1)



(A). Random selection of k centers

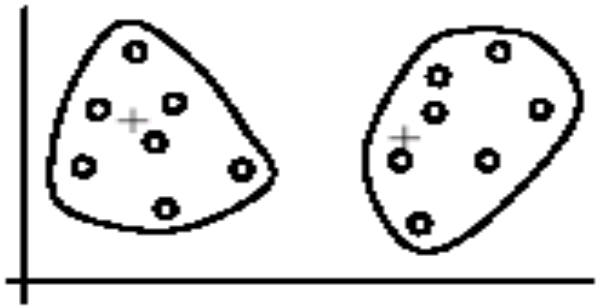


Iteration 1: (B). Cluster assignment

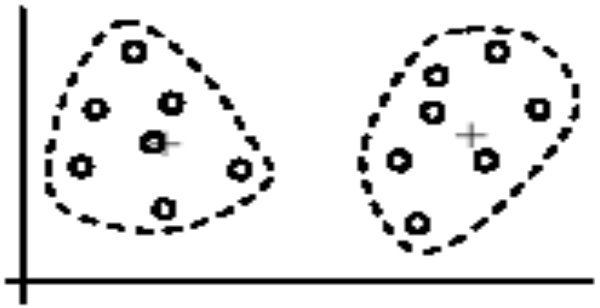
(C). Re-compute centroids

[Liu, 2006]

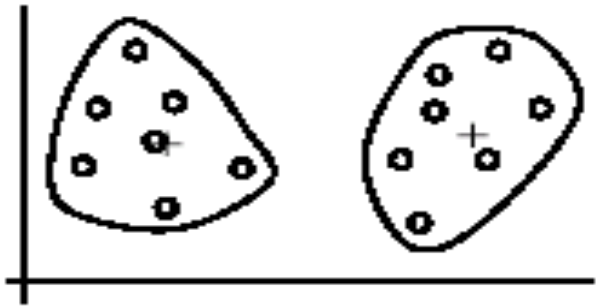
k-Means: Minh họa (2)



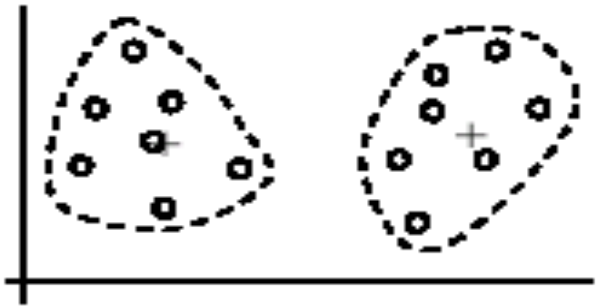
Iteration 2: (D). Cluster assignment



(E). Re-compute centroids



Iteration 3: (F). Cluster assignment



(G). Re-compute centroids

[Liu, 2006]

k-Means: Điều kiện hội tụ

Quá trình phân cụm kết thúc, nếu:

- Không có (hoặc có không đáng kể) việc gán lại các ví dụ vào các cụm khác, *hoặc*
- Không có (hoặc có không đáng kể) thay đổi về các điểm trung tâm (centroids) của các cụm, *hoặc*
- Giảm không đáng kể về tổng lỗi phân cụm:

$$Error = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mathbf{m_i})^2$$

  - C_i: Cụm thứ i
  - $\mathbf{m_i}$: Điểm trung tâm (centroid) của cụm C_i
  - $d(\mathbf{x}, \mathbf{m}_i)$: Khoảng cách (khác biệt) giữa ví dụ $\mathbf{x}$ và điểm trung tâm $\mathbf{m}_i$

k-Means: Điểm trung tâm, Hàm khoảng cách

- Xác định điểm trung tâm: Điểm trung bình (*Mean centroid*)

$$\mathbf{m_i} = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

  - (vectơ) $\boldsymbol{m_i}$ là điểm trung tâm (centroid) của cụm C_i
  - $|C_i|$ kích thước của cụm C_i (tổng số ví dụ trong C_i)

- Hàm khoảng cách: *Euclidean distance*

$$d(\mathbf{x}, \mathbf{m_i}) = \|\mathbf{x} - \mathbf{m_i}\| = \sqrt{(x_1 - m_{i1})^2 + (x_2 - m_{i2})^2 + \ldots + (x_n - m_{in})^2}$$

  - (vectơ) $\boldsymbol{m_i}$ là điểm trung tâm (centroid) của cụm C_i
  - $d(\boldsymbol{x}, \boldsymbol{m_i})$ là khoảng cách giữa ví dụ $\boldsymbol{x}$ và điểm trung tâm $\boldsymbol{m_i}$

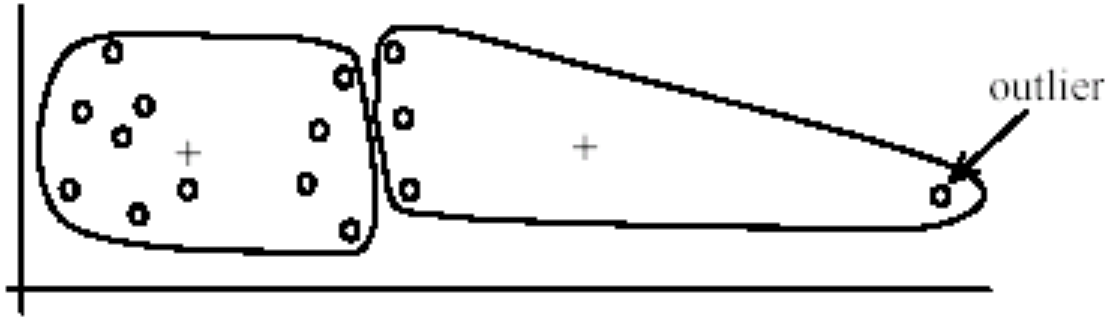
k-Means: Các ưu điểm

- Đơn giản: dễ cài đặt, rất dễ hiểu
- Rất linh động: cho phép dùng nhiều độ đo khoảng cách khác nhau → phù hợp với các loại dữ liệu khác nhau.
- Hiệu quả
  - Độ phức tạp về thời gian ~ $O(r.k.t)$
    - r: Tổng số các ví dụ (kích thước của tập dữ liệu)
    - k: Tổng số cụm thu được
    - t: Tổng số bước lặp (của quá trình phân cụm)
  - Nếu cả 2 giá trị k và t đều nhỏ, thì giải thuật k-means được xem như là có độ phức tạp ở mức tuyến tính
- k-means là giải thuật phân cụm được dùng phổ biến nhất

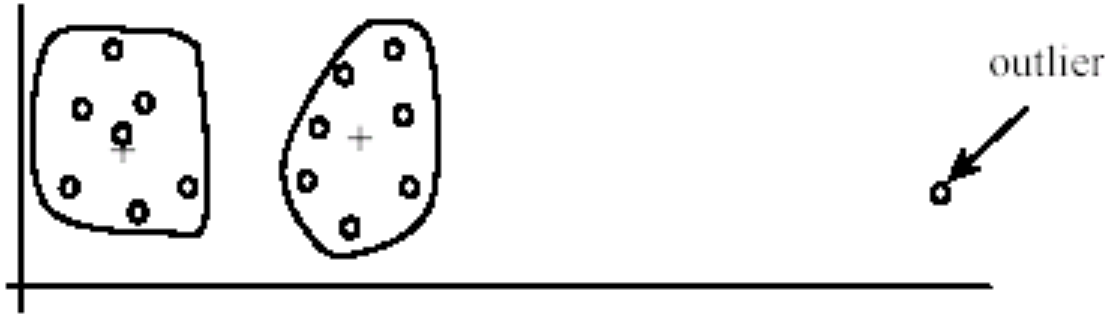
K-means: Các nhược điểm (1)

- Số cụm k phải được xác định trước
  - Thường ta không biết chính xác!
- Giải thuật k-means nhạy cảm (gặp lỗi) với ***các ví dụ ngoại lai (outliers)***
  - Các ví dụ ngoại lai là các ví dụ (rất) khác biệt với tất các ví dụ khác
  - Các ví dụ ngoại lai có thể do lỗi trong quá trình thu thập/lưu dữ liệu
  - Các ví dụ ngoại lai có các giá trị thuộc tính (rất) khác biệt với các giá trị thuộc tính của các ví dụ khác

k-Means: Các ví dụ ngoại lai



(A): Undesirable clusters



(B): Ideal clusters

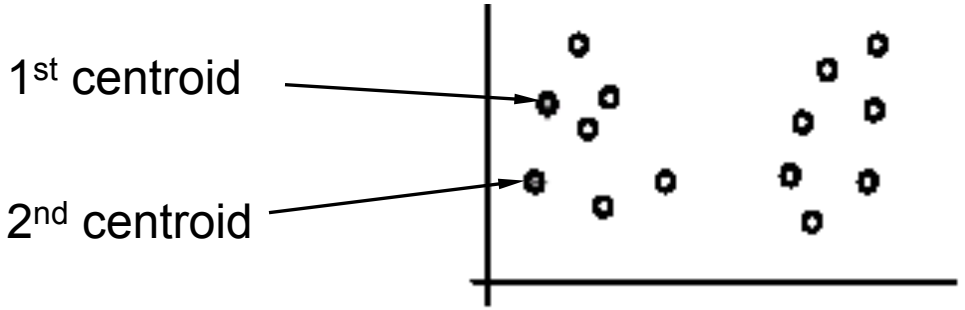
[Liu, 2006]

Giải quyết vấn đề ngoại lai

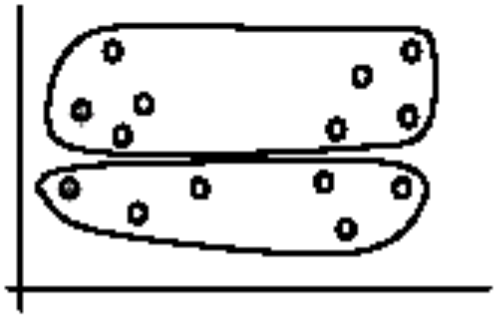
- Giải pháp 1: Trong quá trình phân cụm, cần loại bỏ một số các ví dụ quá khác biệt với (cách xa) các điểm trung tâm (centroids) so với các ví dụ khác
  - Để chắc chắn (không loại nhầm), theo dõi các ví dụ ngoại lai (outliers) qua một vài (thay vì chỉ 1) bước lặp phân cụm, trước khi quyết định loại bỏ
- Giải pháp 2: Thực hiện việc lấy mẫu ngẫu nhiên (a random sampling)
  - Do quá trình lấy mẫu chỉ lựa chọn một tập con nhỏ của tập dữ liệu ban đầu, nên khả năng một ngoại lai (outlier) được chọn là rất nhỏ
  - Gán các ví dụ còn lại của tập dữ liệu vào các cụm tùy theo đánh giá về khoảng cách (hoặc độ tương tự)

k-Means: Các nhược điểm (2)

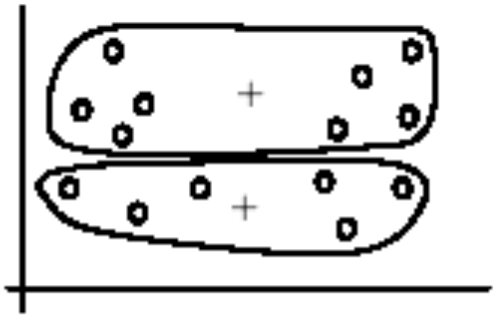
- Giải thuật k-means phụ thuộc vào việc chọn các điểm trung tâm ban đầu (initial centroids)



(A). Random selection of seeds (centroids)



(B). Iteration 1

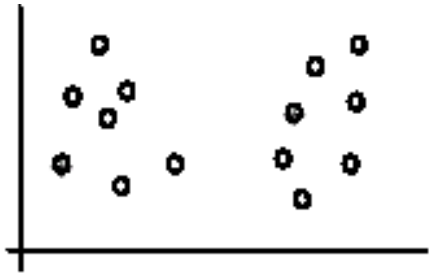


(C). Iteration 2

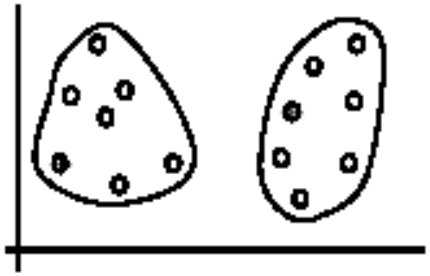
[Liu, 2006]

k-Means: Các hạt nhân ban đầu (1)

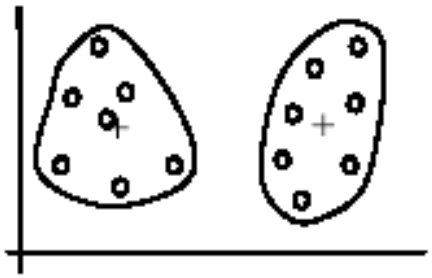
- Sử dụng các hạt nhân (seeds) khác nhau → Kết quả tốt hơn!
  - Thực hiện giải thuật k-means nhiều lần, mỗi lần bắt đầu với một tập (khác lần trước) các hạt nhân được chọn ngẫu nhiên



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



(B). Iteration 1



(C). Iteration 2

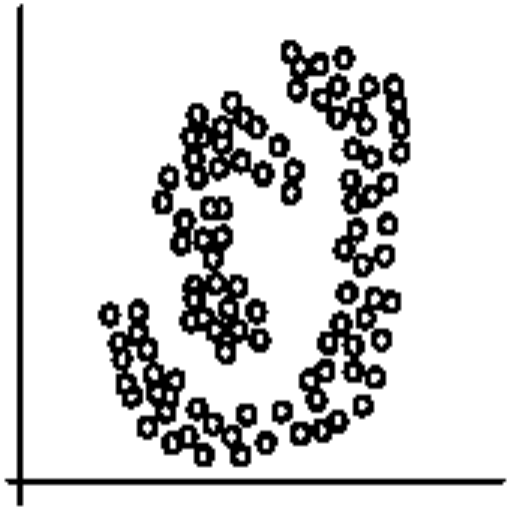
[Liu, 2006]

k-Means: Các hạt nhân ban đầu (2)

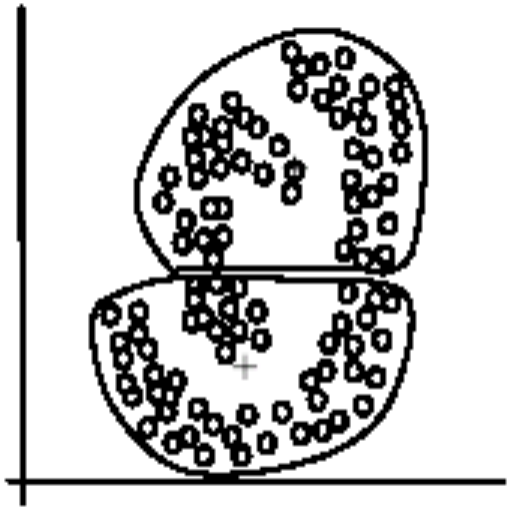
- Một cách chọn hạt nhân nên dùng:
  - Lựa chọn ngẫu nhiên hạt nhân thứ 1 ($\boldsymbol{m_1}$)
  - Lựa chọn hạt nhân thứ 2 ($\boldsymbol{m_2}$) càng xa càng tốt so với hạt nhân thứ 1
  - …
  - Lựa chọn hạt nhân thứ *i* ($\boldsymbol{m_i}$) càng xa càng tốt so với hạt nhân gần nhất trong số $\{\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_{i-1}}\}$
  - ...

k-Means: Các nhược điểm (3)

- Giải thuật k-means không phù hợp để phát hiện các cụm (nhóm) không có dạng hình elip hoặc hình cầu.
- Cải thiện??



(A): Two natural clusters



(B): k-means clusters

[Liu, 2006]

k-Means: Tổng kết

- Mặc dù có những nhược điểm như trên, k-means vẫn là giải thuật phổ biến nhất được dùng để giải quyết các bài toán phân cụm – do tính đơn giản và hiệu quả.
  - Các giải thuật phân cụm khác cũng có các nhược điểm riêng.
- Về tổng quát, không có lý thuyết nào chứng minh rằng một giải thuật phân cụm khác hiệu quả hơn k-means.
  - Một số giải thuật phân cụm có thể phù hợp hơn một số giải thuật khác đối với một số kiểu tập dữ liệu nhất định, hoặc đối với một số bài toán ứng dụng nhất định.
- So sánh hiệu năng của các giải thuật phân cụm là một nhiệm vụ khó khăn (thách thức).
  - Làm sao để biết được các cụm kết quả thu được là chính xác?

Tài liệu tham khảo

- B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.

Câu hỏi ôn tập

- Làm thế nào để phân cụm tốt trong trường hợp các cụm không phân bố theo hình cầu?
- Làm sao để phân một ví dụ mới vào các cụm đã học?