Dữ Liệu Lớn

Clustering with K-means

Thân Quang Khoát

khoattq@soict.hust.edu.vn

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

Năm 2019

Nội dung khoá học

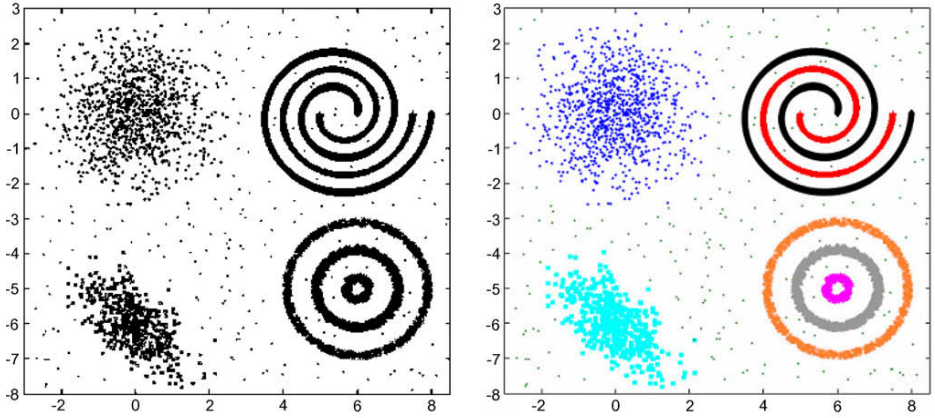
- Overview of data analytics/science
- Basic statistics
- Python and programming tools
- Exploratory data analysis
- Data integration and preprocessing
- **Prediction with machine learning**
- Data visualization
- Evaluation of analysis results
- Basics of natural language processing
- Anomaly detection
- Big data analysis
- Capstone project

1. Hai bài toán học

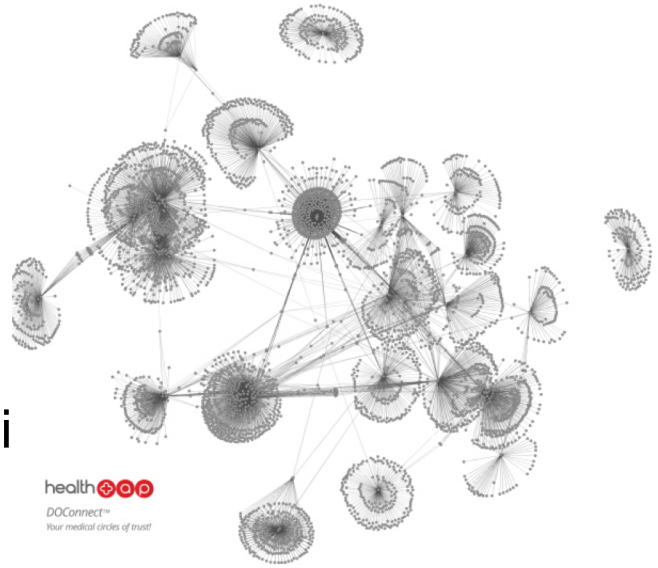
- **Học có giám sát (Supervised learning)**
  - Tập dữ liệu học (*training data*) bao gồm các quan sát (*examples, observations*), mà mỗi quan sát được *gắn kèm với một giá trị đầu ra mong muốn.*
  - Ta cần học một hàm (vd: một phân lớp, một hàm hồi quy,...) phù hợp với tập dữ liệu hiện có.
  - Hàm học được sau đó sẽ được dùng để dự đoán cho các quan sát mới.
- **Học không giám sát (Unsupervised learning)**
  - Tập học (training data) bao gồm các quan sát, mà mỗi quan sát *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,…

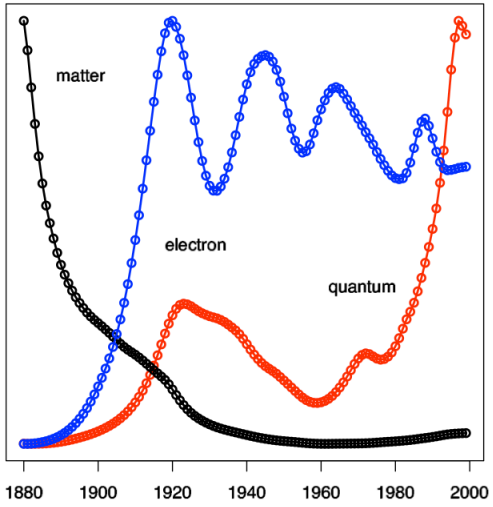


- Community detection
  - Phát hiện các cộng đồng trong mạng xã hội

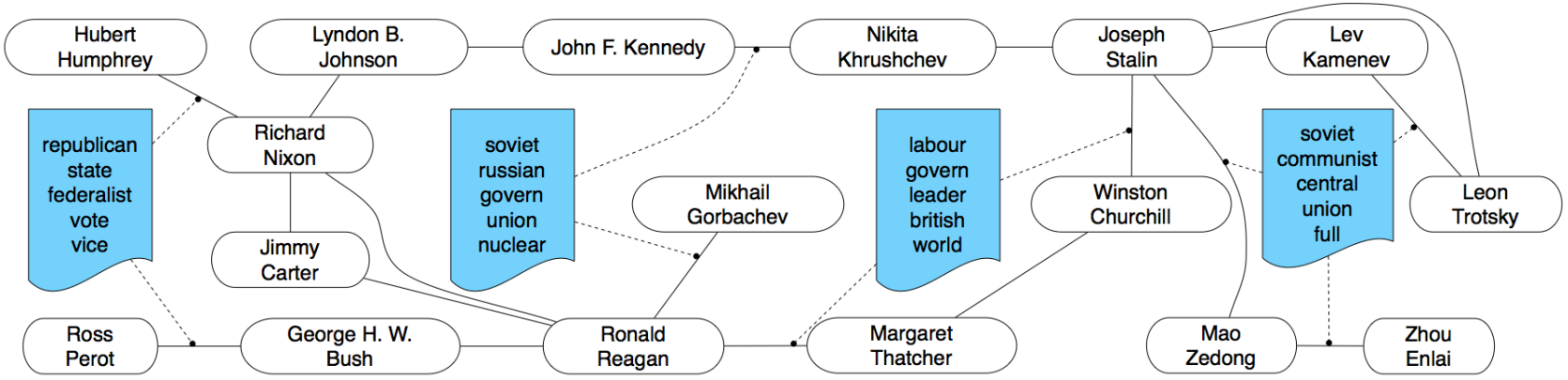


Ví dụ về học không giám sát (2)

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



- Entity-interaction analysis



2. Phân cụm

- Phân cụm (clustering)
  - Đầu vào: một tập dữ liệu $\{x_1, \ldots, x_M\}$ không có nhãn (hoặc giá trị đầu ra mong muốn)
  - Đầu ra: các cụm (nhóm) của các quan sát
- Một **cụm (cluster)** là một tập các quan sát
  - Tương tự với nhau (theo một ý nghĩa, đánh giá nào đó)
  - Khác biệt với các quan sát thuộc các cụm khác

Phân cụm

- 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*

3. Phương pháp K-means

- K-means được giới thiệu đầu tiên bởi Lloyd năm 1957.
- Là phương pháp phân cụm phổ biến nhất trong các phương pháp dựa trên phân hoạch (partition-based clustering)
- Biểu diễn dữ liệu: $D=\{x_1,x_2,\ldots,x_r\}$
  - x_i là một quan sát (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

Đầu vào: tập học **D**, số lượng cụm k, khoảng cách d(x,y)

- Bước 1. Chọn ngẫu nhiên k quan sát (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 quan sát, *gán nó vào cụm* (trong số k cụm) mà có tâm (centroid) gần nó nhất.
  - Bước 2.2. Đối với mỗi cụm, *tính toán lại điểm trung tâm của nó* dựa trên tất cả các quan sát thuộc vào cụm đó.

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

D: Tập học

k: Số lượng cụm kết quả (thu được)

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

while not CONVERGENCE

for each 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

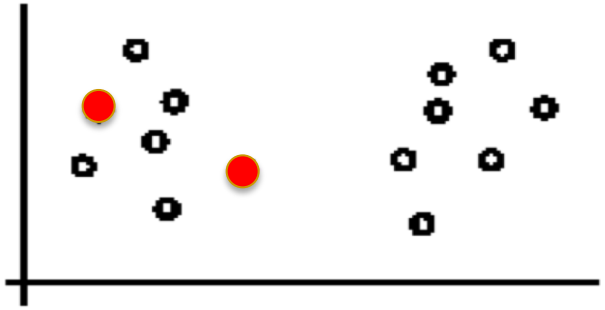
for each cụm

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

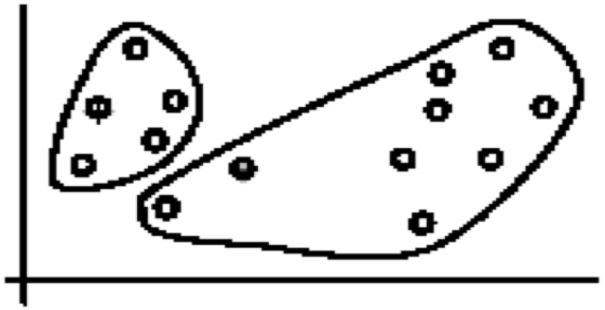
end while

return {k cụm kết quả}

K-means: Minh họa (1)



(A). Random selection of k centers

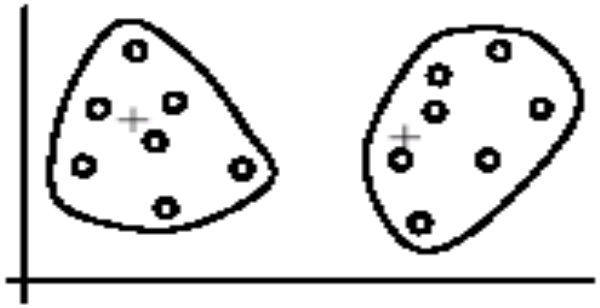


Iteration 1: (B). Cluster assignment

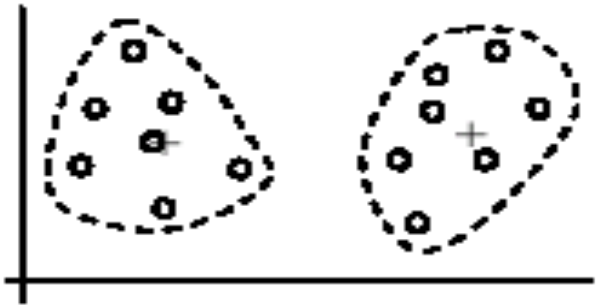
(C). Re-compute centroids

[Liu, 2006]

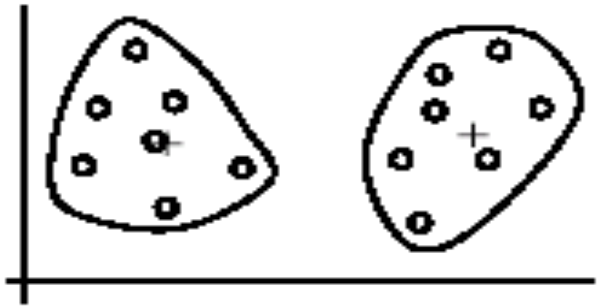
K-means: Minh họa (2)



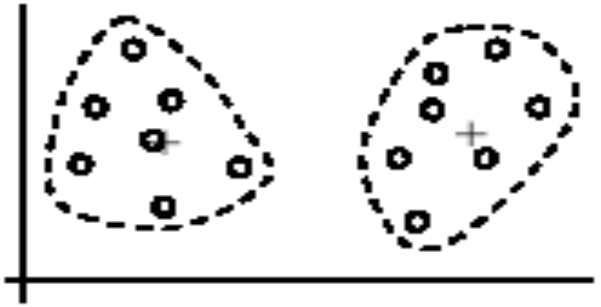
Iteration 2: (D). Cluster assignment



(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 quan sát 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 quan sát $\mathbf{x}$ 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 quan sát trong C_i)

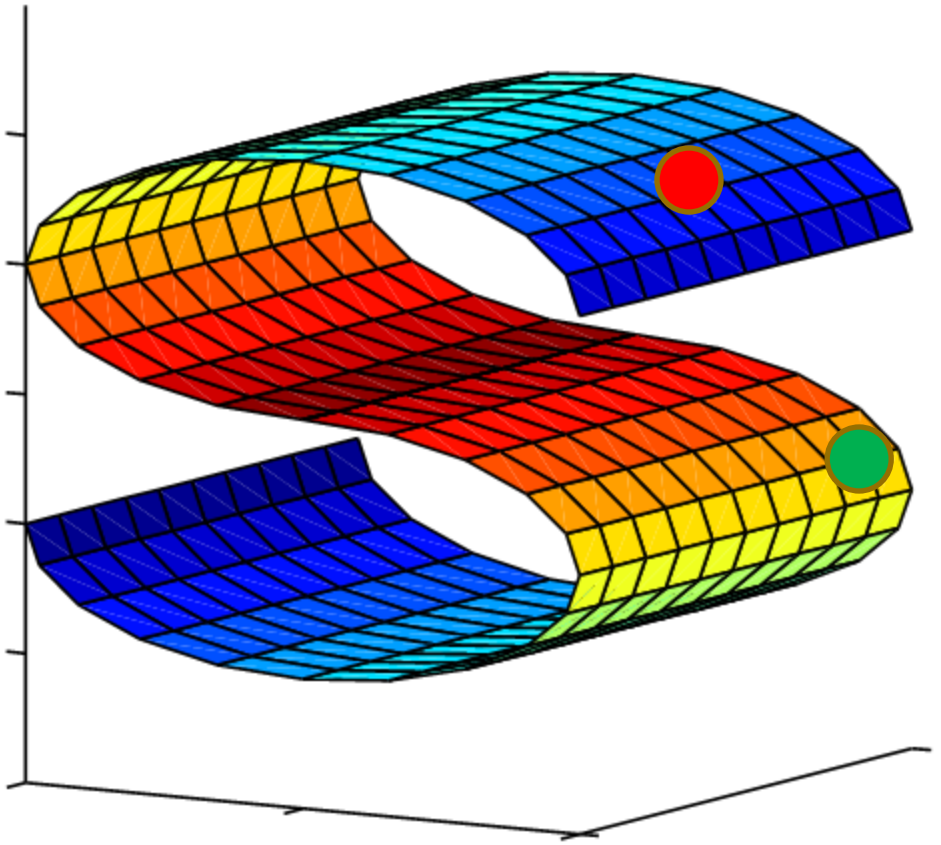
- Hàm khoảng cách: *Euclidean distance*

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

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

K-means: hàm khoảng cách

- Hàm khoảng cách
  - Mỗi hàm sẽ tương ứng với một cách nhìn về dữ liệu.
  - Vô hạn hàm!!!
  - Chọn hàm nào?
- Có thể thay bằng độ đo tương đồng (similarity measure)



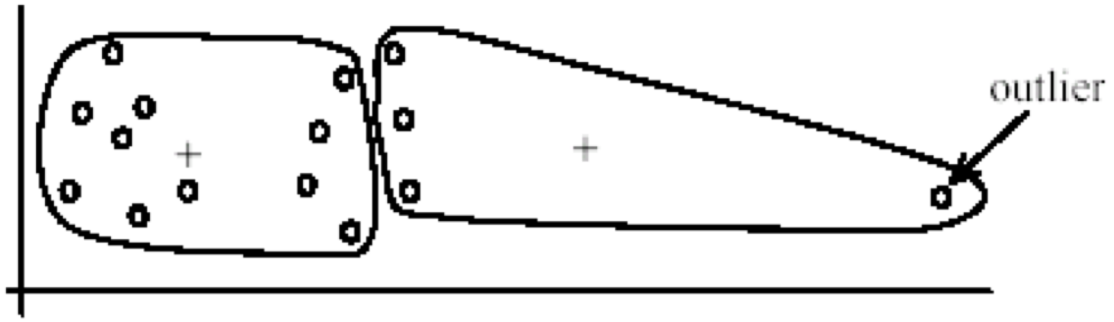
K-means: Các ưu điểm

- 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 (khi dùng độ đo Euclide)
  - Độ phức tạp tính toán tại mỗi bước ~ $O(r.k)$
    - r: Tổng số các quan sát (kích thước của tập dữ liệu)
    - k: Tổng số cụm thu được
  - Thuật toán có độ phức tạp trung bình là đa thức.
- K-means là giải thuật phân cụm được dùng phổ biến nhất

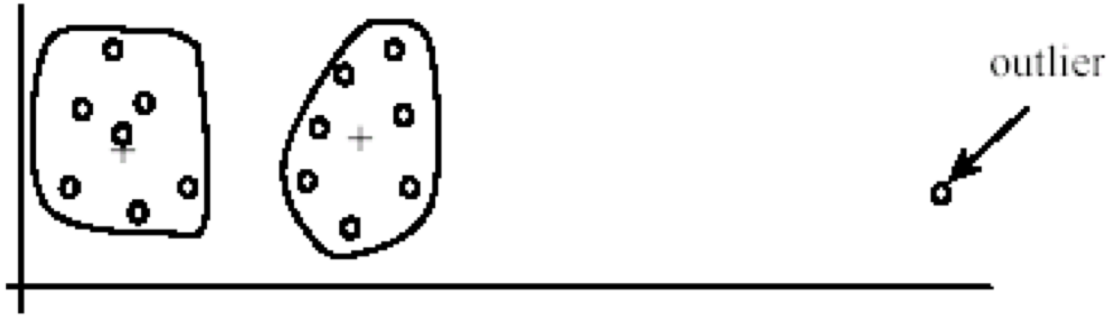
K-means: Các nhược điểm (1)

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

K-means: ngoại lai



(A): Undesirable clusters



(B): Ideal clusters

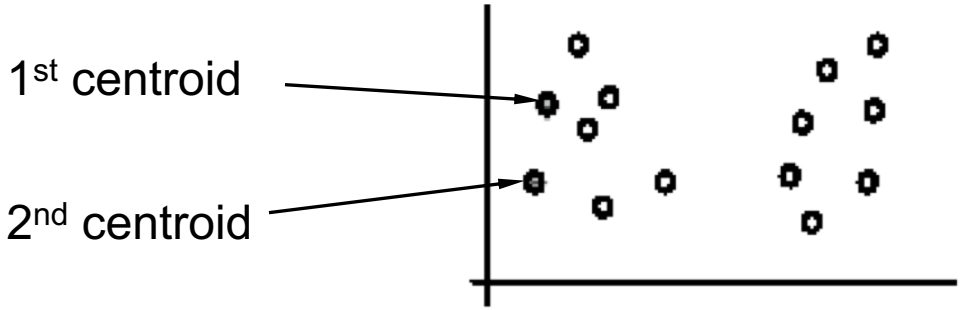
[Liu, 2006]

Giải quyết vấn đề ngoại lai

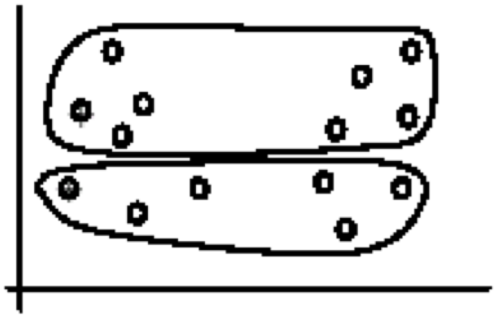
- Giải pháp 1: Trong quá trình phân cụm, cần loại bỏ một số các quan sát quá khác biệt với (cách xa) các điểm trung tâm (centroids) so với các quan sát khác
  - Để chắc chắn (không loại nhầm), theo dõi các quan sát ngoại lai (outliers) qua một vài (thay vì chỉ 1) bước lặp phân cụm, trước khi quyết định loại bỏ
- Giải pháp 2: Thực hiện việc lấy ngẫu nhiên (random sampling) một tập nhỏ từ **D** để học K cụm
  - Do đây là tập con nhỏ của tập dữ liệu ban đầu, nên khả năng một ngoại lai (outlier) được chọn là nhỏ
  - Gán các quan sát 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)

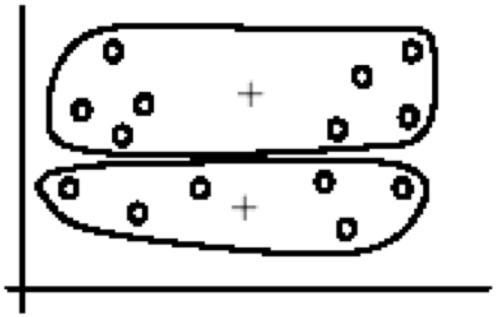
- Giải thuật *K*-means phụ thuộc vào việc chọn các điểm trung tâm ban đầu (initial centroids)



(A). Random selection of seeds (centroids)



(B). Iteration 1

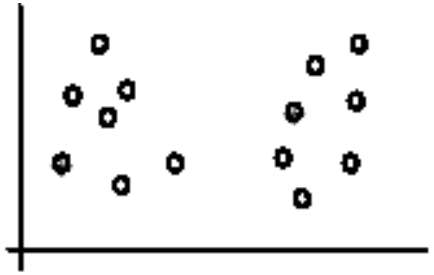


(C). Iteration 2

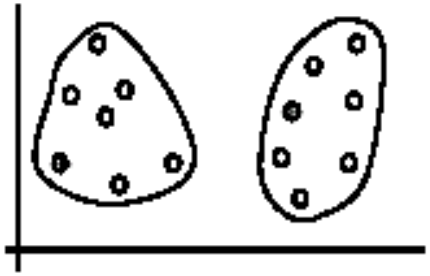
[Liu, 2006]

K-means: Các hạt nhân ban đầu (1)

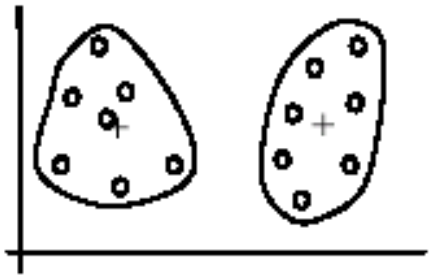
- Kết hợp nhiều kết quả phân cụm với nhau → Kết quả tốt hơn!
  - Thực hiện giải thuật *K*-means nhiều lần, mỗi lần bắt đầu với một tập các hạt nhân được chọn ngẫu nhiên



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



(B). Iteration 1



(C). Iteration 2

[Liu, 2006]

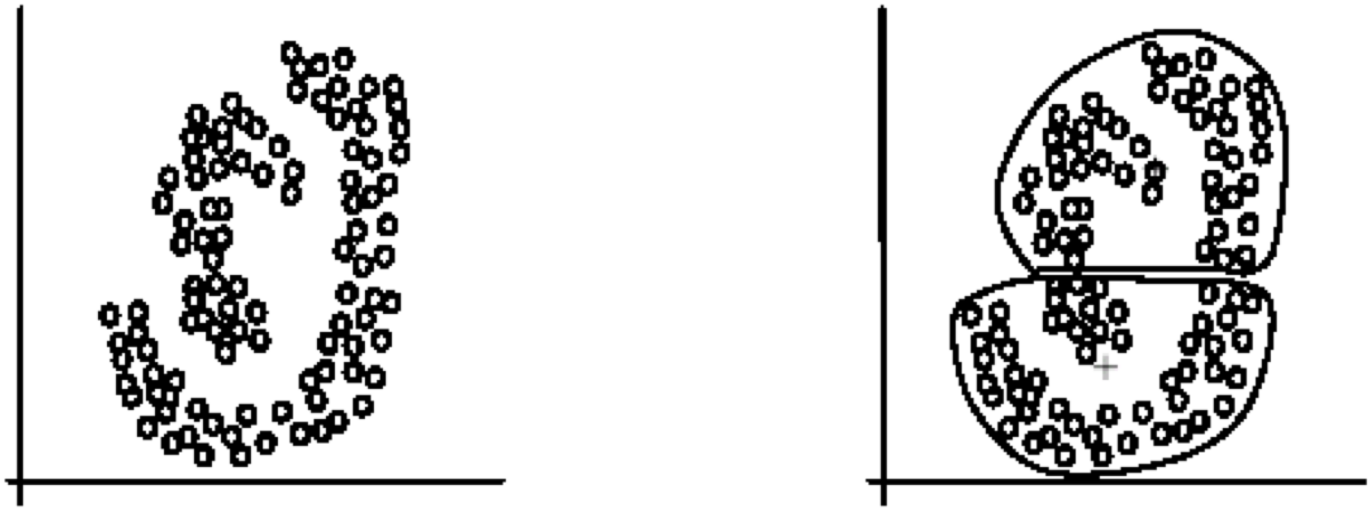
K-means: Các hạt nhân ban đầu (2)

- 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 $\{\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_{i-1}}\}$
  - ...
- Đây được gọi là phương pháp **K-means++**

[Arthur, D.; Vassilvitskii, 2007]

K-means: Các nhược điểm (3)

- K-means (với khoảng cách Euclid) phù hợp với các cụm hình cầu.
- *K-means không phù hợp để phát hiện các cụm (nhóm) không có dạng hình cầu.*
- Cải thiện??



(A): Two natural clusters [Liu, 2006] (B): *k*-means clusters

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.
- 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?

4. Online K-means

- K-means:
  - Cần dùng toàn bộ dữ liệu tại mỗi bước lặp
  - Do đó không thể làm việc khi dữ liệu quá lớn (big data)
  - Không phù hợp với luồng dữ liệu (stream data, dữ liệu đến liên tục)
- *Online K-means* cải thiện nhược điểm của K-means, cho phép ta phân cụm dữ liệu rất lớn, hoặc phân cụm luồng dữ liệu.
  - Được phát triển từ K-means [Bottou, 1998].
  - Sử dụng tư tưởng học trực tuyến (online learning) và gradient ngẫu nhiên (stochastic gradient)

Online K-means: ý tưởng

- K-means tìm K tâm cụm và gán các quan sát $\{x_1, \ldots, x_M\}$ vào các cụm đó bằng cách cực tiểu hoá hàm lỗi sau

$$Q(w) = \sum_{i=1}^{M} \| x_i - w(x_i) \|_2^2$$

  - Trong đó $w(x_i)$ là tâm gần nhất với x_i.

- Online K-means cực tiểu hàm Q theo phương pháp leo đồi và dùng thông tin đạo hàm (gradient) của Q.
  - Tuy nhiên tại mỗi bước lặp *t* ta chỉ lấy một phần thông tin gradient,
  - Phần gradient này thu được từ các quan sát tại bước *t*. Ví dụ:

$$x_t - w_t(x_t)$$

Online K-means: thuật toán

- Khởi tạo K tâm ban đầu.
- Cập nhật các tâm mỗi khi một điểm dữ liệu mới đến:
  - *Tại bước t, lấy một quan sát x_t .*
  - *Tìm tâm w_t gần nhất với x_t. Sau đó cập nhật lại w_t như sau:*

$$w_{t+1} = w_t + \gamma_t (x_t - w_t)$$

- Chú ý: tốc độ học $\{\gamma_1, \gamma_2, ...\}$ là dãy hệ số dương nên được chọn thoả mãn các điều kiện sau

$$\sum_{t=1}^{\infty} \gamma_t = \infty; \sum_{t=1}^{\infty} \gamma_t^2 < \infty$$

Online K-means: tốc độ học

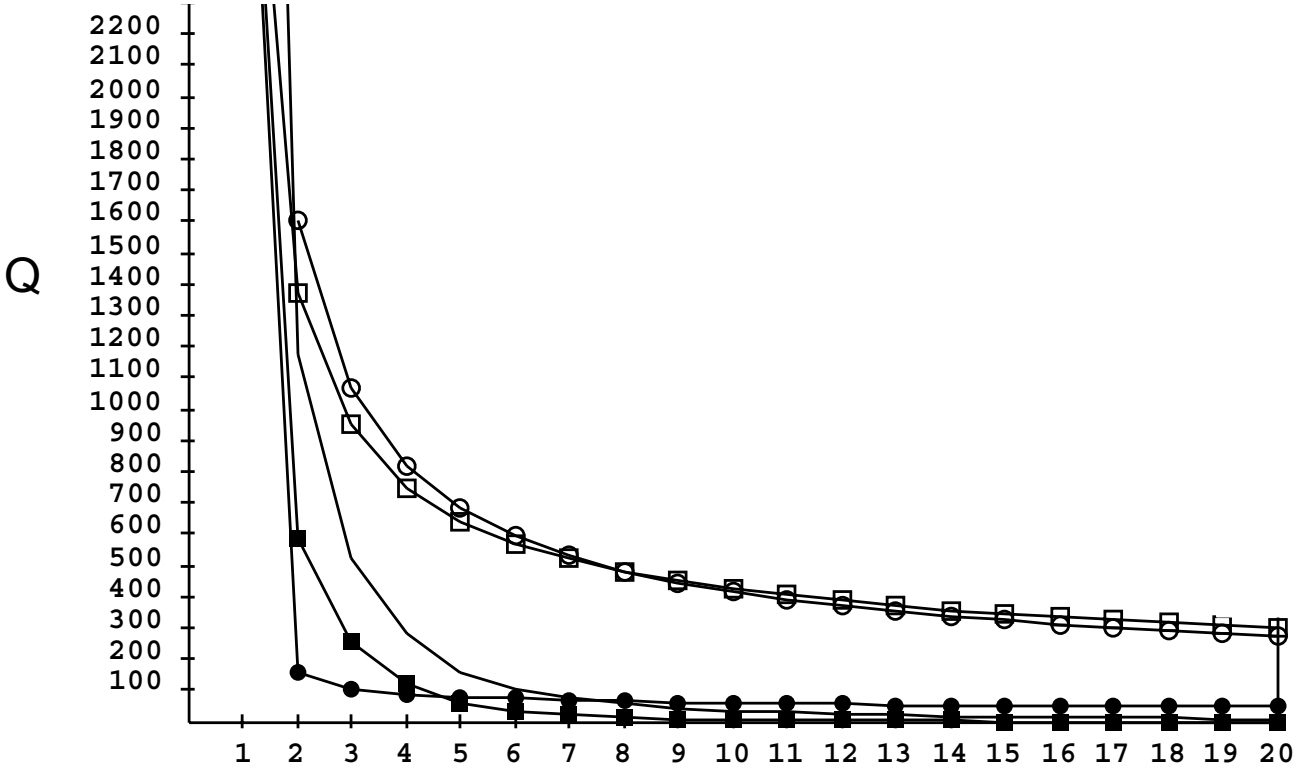
- Một cách lựa chọn tốc độ học hay dùng:

$$\gamma_t = \left(t + \tau\right)^{-\kappa}$$

- τ, κ là các hằng số dương.
- $\kappa \in$ (0.5, 1] là tốc độ lãng quên. κ càng lớn thì sẽ nhớ quá khứ càng lâu; các quan sát mới càng ít đóng góp vào mô hình hơn.

Online K-means: tốc độ hội tụ

- Hàm Q giảm khi số lần lặp tăng lên.
  (so sánh các phương pháp khác nhau)



Online K-means (hình tròn đen),

K-means (hình vuông đen)

Dùng một phần Q' để tối ưu hàm Q (hình tròn trắng),

Dùng hết Q' để tối ưu hàm Q (hình vuông trắng)

Tài liệu tham khảo

- Arthur, D., Manthey, B., & Röglin, H. (2011). Smoothed analysis of the k-means method. *Journal of the ACM (JACM)*, 58(5), 19.
- Bottou, Léon. Online learning and stochastic approximations. *On-line learning in neural networks* 17 (1998).
- B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.
- Lloyd, S., 1982. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* 28, 129–137. Originally as an unpublished Bell laboratories Technical Note (1957).
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.
- Arthur, D.; Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035.

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 quan sát mới vào các cụm đã học?