UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

NHẬN DẠNG
PATTERN RECOGNITION

Gom cụm- Clustering

1

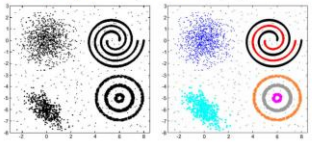
Phân loại dựa vào phương thức học

- **Supervised Learning** (Học có giám sát)
  - Classification (Phân loại, phân lớp)
  - Regression (Hồi quy)
- **Unsupervised Learning** (Học không giám sát)
  - **Clustering** (phân nhóm, cụm)
  - Association (luật)
- **Semi-Supervised Learning** (Học bán giám sát)
- **Reinforcement Learning** (Học Củng Cố)

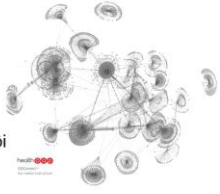
2

Clustering

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



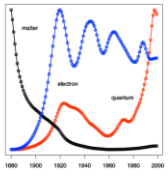
- Community detection
  - Phát hiện các cộng đồng trong mạng xã hội



3

Clustering

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



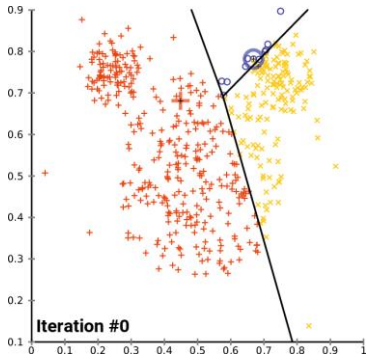
4

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: $D=\{x_1,x_2,\ldots,x_t\}$
  - x_i là một quan sát (một vecto 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)

5

Phương pháp K-means



6

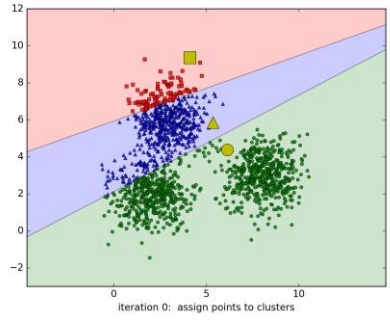
K-Means – Các bước thực hiện

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.

7

K-Means – Các bước thực hiện



8

K-Means – Các bước thực hiện

***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

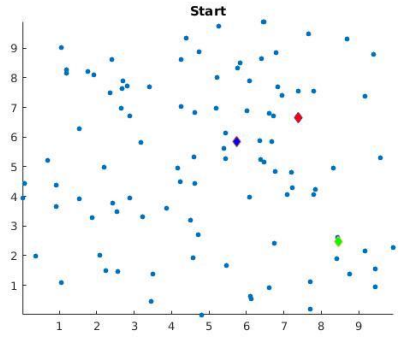
Tính (xác định) lại điểm trung tâm (centroid) dựa trên các quan sát hiện thời đang thuộc vào cụm này

end while

return {k cụm kết quả}

9

K-Means – Các bước thực hiện



10

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 $\mathbf{m}_i$

11

K-Means – Điểm trung tâm và 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*

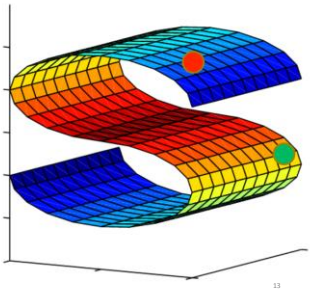
$$d(\mathbf{x}, \mathbf{m_i}) = \|\mathbf{x} - \mathbf{m_i}\| = \sqrt{(x_1 - m_{i1})^2 + (x_2 - m_{i2})^2 + \ldots + (x_n - m_{in})^2}$$

  - (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 $\boldsymbol{m_i}$

12

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)



13

K-Means –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 Euclidean)
  - 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

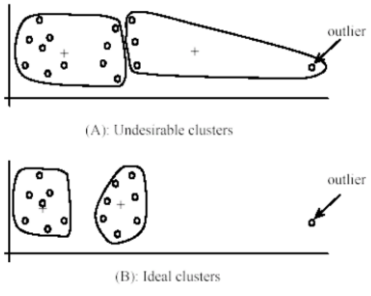
14

K-Means –nhược điểm

- 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

15

K-Means –trường hợp outlier



[Liu, 2006]

16

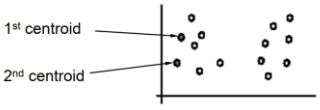
K-Means – loại bỏ Outlier

- 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)

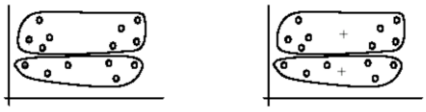
17

K-Means –nhược điểm

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



(A). Random selection of seeds (centroids)



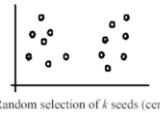
(B). Iteration 1 (C). Iteration 2

[Liu, 2006]

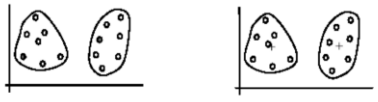
18

K-Means –Khởi tạo hạt nhân ban đầu

- 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]

19

K-Means –Khởi tạo hạt nhân ban đầu

- 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 $\{\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_{i-1}}\}$
  - …

20

K-Means

- 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?

21

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.

22

K-Means ví dụ

https://machinelearningcoban.com/2017/01/01/kmeans/

23

Tài liệu tham khảo

- Slide Machine learning TQ Khoat , ĐHBK HN
- https://www.coursera.org/learn/machine-learning?
- https://machinelearningcoban.com/
- https://machinelearningmastery.com/
- Slide Machine learning Thầy Tùng, ĐH Thủy Lợi
- Lý thuyết nhận dạng – Ngô Hữu Phúc

24