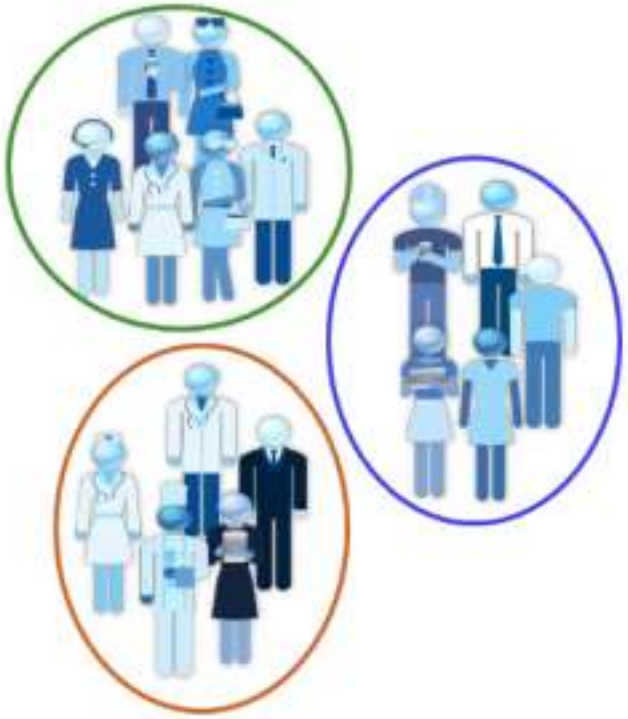
CSE: **Faculty of Computer Science and Engineering**
Thuyloi University

Phân cụm
(Clustering)

TS. Nguyễn Thị Kim Ngân

Giới thiệu



Giới thiệu

Có bao nhiêu clusters

Có bao nhiêu clusters?

2 clusters

4 clusters

6 clusters

Tuỳ thuộc vào "resolution" !

Giới thiệu

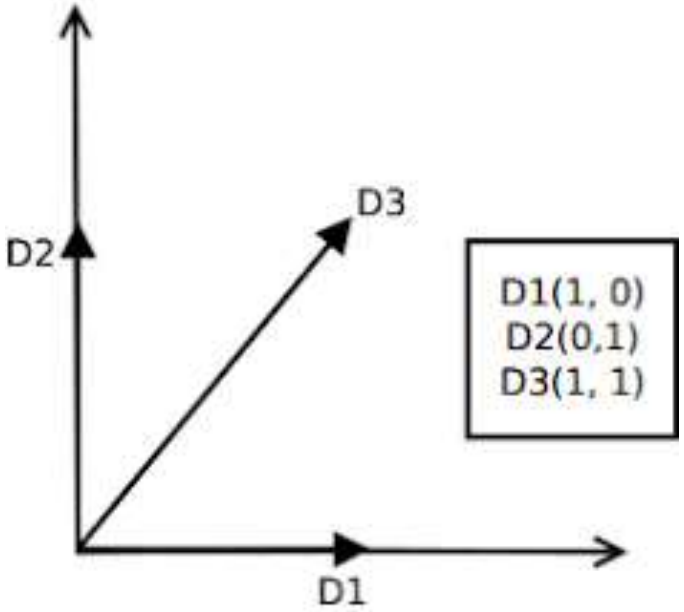
- Thuật toán phân cụm K-means thuộc loại học không giám sát
  - Không biết nhãn (label) của từng điểm dữ liệu
  - Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho *dữ liệu trong cùng một cụm có tính chất giống nhau*.
- **Ví dụ:** Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau nhưng chưa có cách phân thành một số nhóm/cụm:
  - Phân nhóm dựa trên sự tương tác giữa mỗi khách hàng với công ty: số năm là khách hàng; số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thành phố; nghề nghiệp; ….
  - Giả sử công ty có nhiều dữ liệu của rất nhiều khách hàng

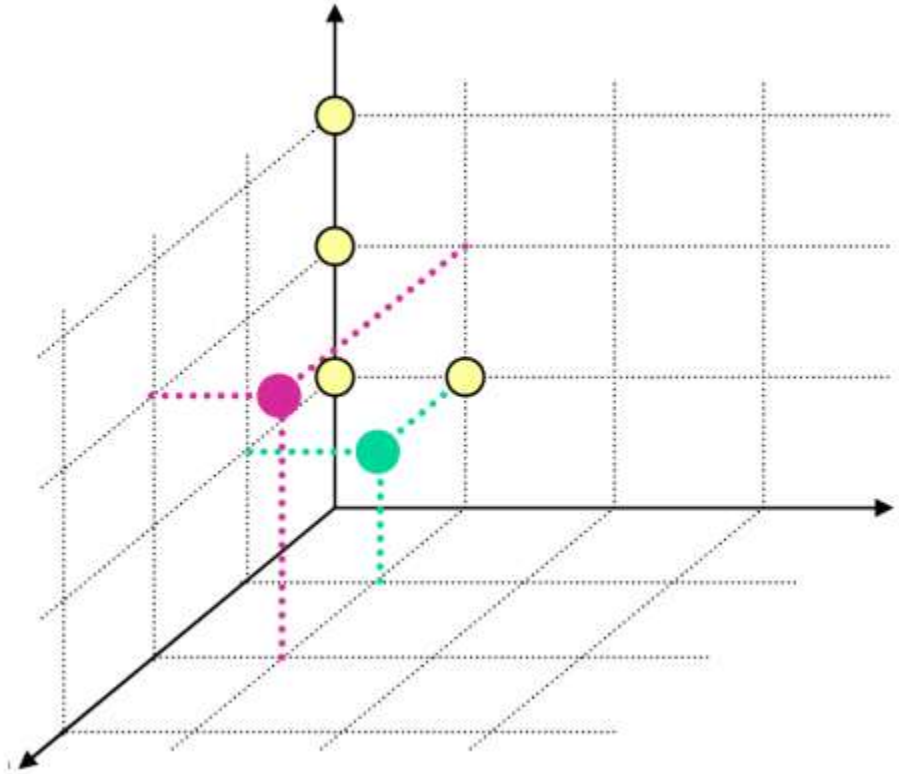
Giới thiệu

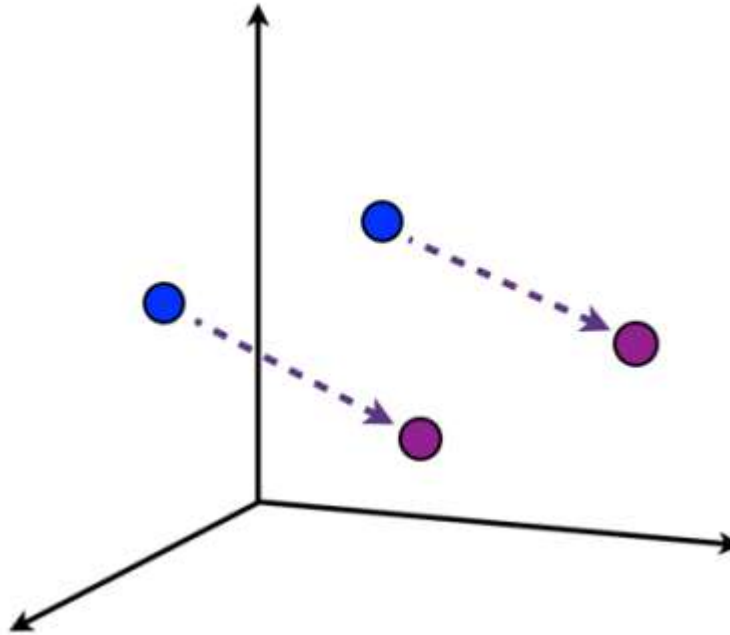
- Sau khi đã có các nhóm, nhân viên công ty lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào
- Ý tưởng đơn giản nhất về phân cụm (clustering) là tập hợp các điểm ở gần nhau trong một không gian nào đó

Giới thiệu

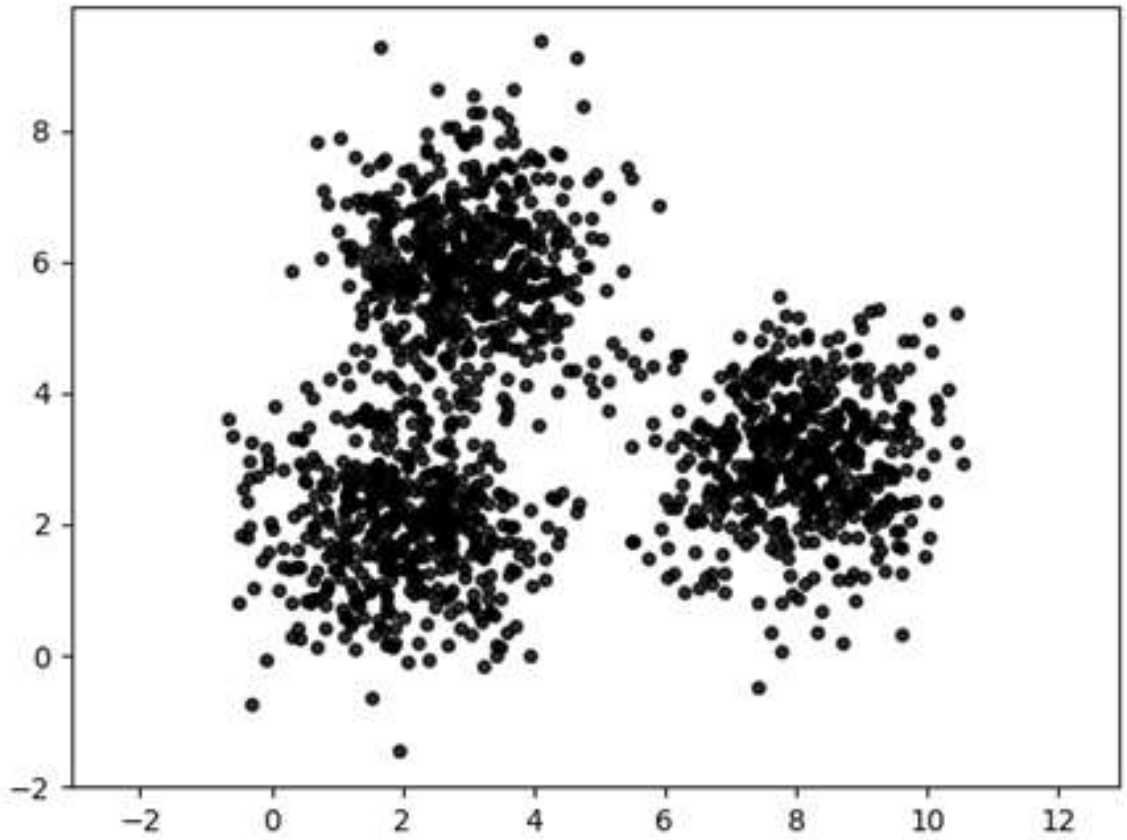
- Điểm trong không gian véc tơ



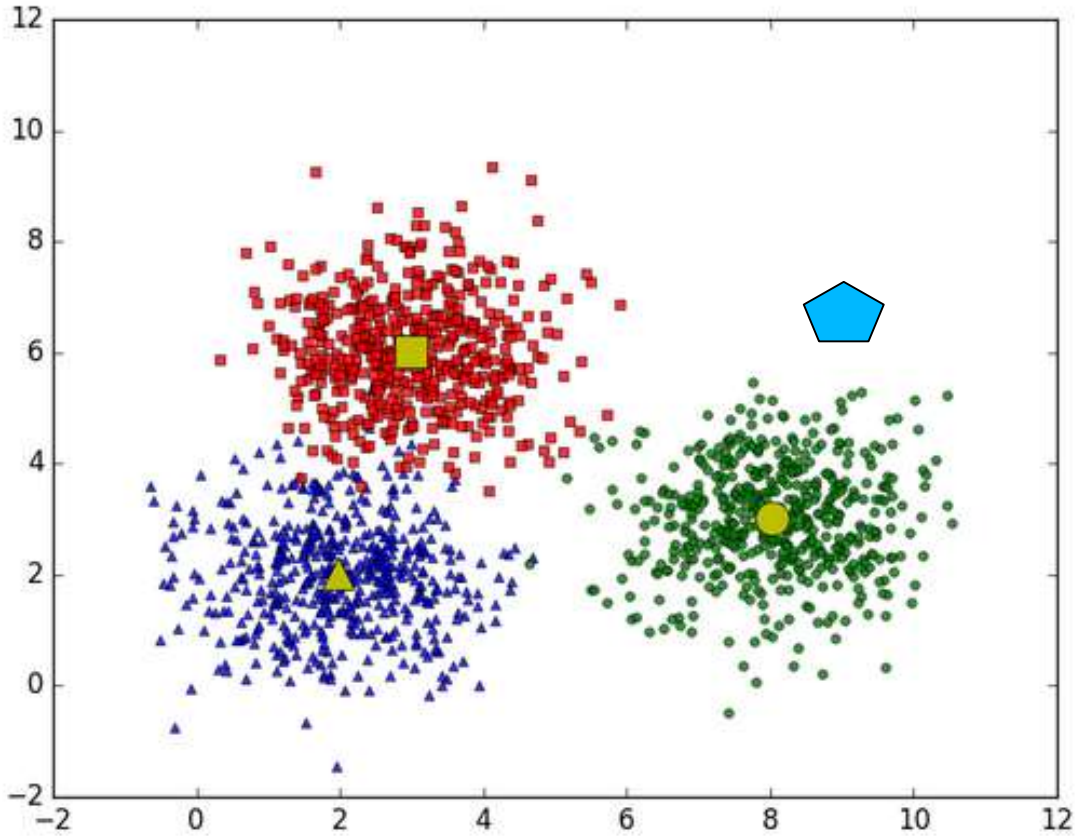




Giới thiệu



Giới thiệu



Bài toán phân nhóm

Input:

- Dữ liệu đầu vào
- Số lượng nhóm muốn tìm

Ouput:

- Chỉ ra tâm (center) của mỗi nhóm
- Phân các điểm dữ liệu vào các nhóm

Một số ký hiệu toán học

- Có N điểm dữ liệu $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_N] \in \mathbb{R}^{d \times N}$ và K là số cụm cần phân chia
- Cần tìm tâm của mỗi cụm $\mathrm{m}_1, \mathrm{m}_2, \dots, \mathrm{m}_K \in \mathbb{R}^{d \times 1}$ và nhãn của mỗi điểm dữ liệu
- Với mỗi x_i, đặt $\mathrm{y}_i = [y_{i1}, y_{i2}, \dots, y_{iK}]$ là véc tơ nhãn
- Nếu x_i được phân vào nhóm k thì $y_{ik} = 1$ và $y_{ij} = 0, \forall j \neq k$
- Ví dụ:
  - Một điểm dữ liệu có véc tơ nhãn là $[1,0, \dots, 0]$ thì điểm đó thuộc về lớp 1

Một số ký hiệu toán học

- Với mỗi x_i, đặt $y_i = [y_{i1}, y_{i2}, \dots, y_{iK}]$ là véc tơ nhãn

Nếu x_i được phân vào nhóm k thì $y_{ik} = 1$ và $y_{ij} = 0, \forall j \neq k$

- Ràng buộc của y_i có thể viết dưới dạng toán học như sau:

$$y_{ik} \in \{0, 1\}, \quad \sum_{k=1}^{K} y_{ik} = 1$$

K-means

- G/S m_k là center (hoặc representative) của mỗi cluster
  - Ước lượng tất cả các điểm được phân vào cluster này bởi m_k
  - Một điểm dữ liệu x_i được phân vào cluster k sẽ có sai số là $|x_i - m_k|$
- Chúng ta mong muốn sai số $|x_i - m_k|$ nhỏ nhất nên đại lượng dưới đây cần có giá trị nhỏ nhất:

$$\|x_i - m_k\|_2^2$$

- Vì x_i được phân vào cụm k nên $y_{ik} = 1, y_{ij} = 0, \forall j \neq k$, do đó chúng ta có:

$$y_{ik}\|x_i - m_k\|_2^2 = \sum_{j=1}^{K} y_{ij}\|x_i - m_j\|_2^2$$

Khoảng cách Euclidian

x=(x1, x2, …, xd)

y=(y1, y2, …, yd)

Khoảng cách giữa x và y:

$$\|x-y\|_2^2 = \sqrt[2]{\sum_{i=1}^{d}(xi-yi)^2} = \sqrt{(x1-y1)^2+(x2-y2)^2+\cdots+(xd-yd)^2}$$

K-means

- Sai số cho toàn bộ dữ liệu sẽ là:

$$\mathcal{L}(\mathrm{Y},\mathrm{M}) = \sum_{i=1}^{N}\sum_{j=1}^{K} \mathrm{y}_{ij}\left\|\mathrm{x}_i - \mathrm{m}_j\right\|_2^2$$

- Trong đó
  - $Y = [\mathrm{y}_1; \mathrm{y}_2; \ldots; \mathrm{y}_N]$ là ma trận véc tơ nhãn
  - $\mathrm{M} = [\mathrm{m}_1, \mathrm{m}_2, \ldots, \mathrm{m}_K]$ là ma trận tâm của mỗi cụm

K-means

- Chúng ta cần tối ưu bài toán để tìm Y và M

$$\mathbf{Y}, \mathbf{M} = \arg\min_{\mathbf{Y},\mathbf{M}} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \qquad (*)$$

$$\text{subject to: } y_{ij} \in \{0, 1\} \ \forall i, j; \quad \sum_{j=1}^{K} y_{ij} = 1 \ \forall i$$

- Bài toán (*) là một bài toán khó tìm *điểm tối ưu* vì nó có thêm các điều kiện ràng buộc.
- *Bài toán này thuộc loại rất khó tìm nghiệm tối ưu toàn cục (global optimal point, tức nghiệm làm cho hàm mất mát đạt giá trị nhỏ nhất có thể).*

Ý tưởng

- Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu.
  - Lưu ý: *điểm cực tiểu chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất*
- Giải bài toán (*) là:
  - Xen kẽ giải Y và M khi biến còn lại được cố định
  - Đây là thuật toán lặp, cũng là kỹ thuật phổ biến khi giải bài toán tối ưu
- Sau khi giải, chúng ta có được Y và M

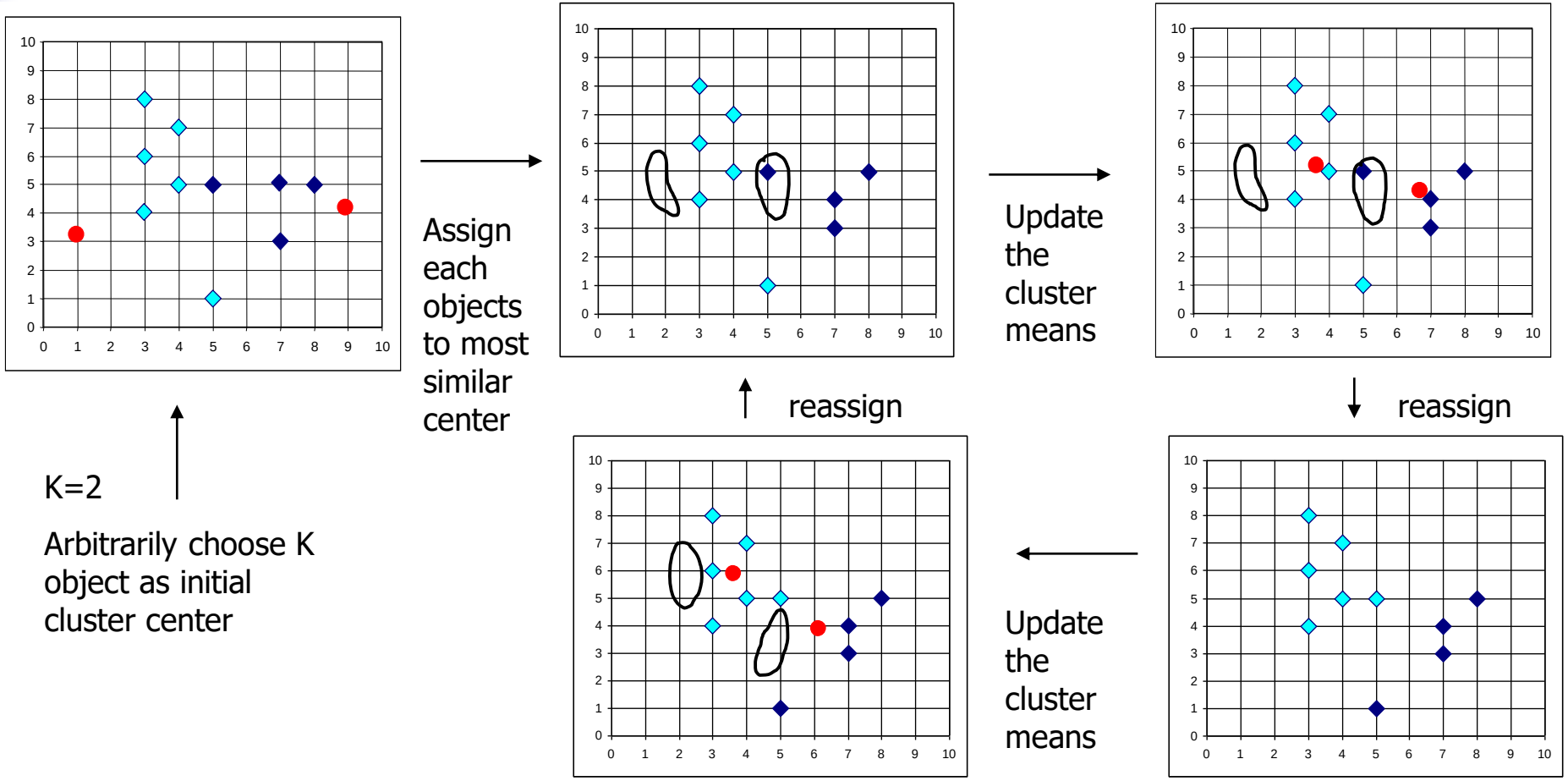
Phương pháp

Đầu vào: Dữ liệu X và số lượng cluster cần tìm K

Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y

1. Chọn K điểm bất kỳ làm các center ban đầu
2. Phân mỗi điểm dữ liệu vào cluster có 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 thuật toán dừng
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2
5. Quay lại bước 2

Ví dụ


Assign each objects to most similar center
Update the cluster means
reassign
Update the cluster means
reassign
K=2
Arbitrarily choose K object as initial cluster center

Ví dụ

Sử dụng thuật toán K-mean, khoảng cách Euclidian để phân các mẫu sau thành 2 cụm. Giả sử, tại bước khởi tạo ban đầu 2 tâm được chọn là x1, x3.

x1 = (1, 4)

x2 = (1, 6)

x3 = (2, 6)

x4 = (3, 8)

x5 = (4, 3)

x6 = (5, 2)

Ví dụ

x1 = (1, 4), x2 = (1, 6), x3 = (2, 6),

x4 = (3, 8), x5 = (4, 3), x6 = (5, 2)

c1=x1(1,4), c2=x3(2,6)

d(x1,c1)= sqrt((1-1)^2+(4-4)^2)=0

d(x1,c2)=sqrt((1-2)^2 +(4-6)^2)=sqrt(5)

d(x2,c1)= sqrt((1-1)^2 +(6-4)^2)=sqrt(4)=2

d(x2,c2)= sqrt((1-2)^2 +(6-6)^2)=sqrt(1)

d(x3,c1)= sqrt((2-1)^2 +(6-4)^2)=sqrt(5)

d(x3,c2)= sqrt((2-2)^2 +(6-6)^2)=sqrt(0)

d(x4,c1)= sqrt((3-1)^2 +(8-4)^2)=sqrt(20)

d(x4,c2)= sqrt((3-2)^2 +(8-6)^2)=sqrt(5)

d(x5,c1)= sqrt((4-1)^2 +(3-4)^2)=sqrt(10)

d(x5,c2)= sqrt((4-2)^2 +(3-6)^2)=sqrt(13)

d(x6,c1)= sqrt((5-1)^2 +(2-4)^2)=sqrt(20)

d(x6,c2)= sqrt((5-2)^2 +(2-6)^2)=sqrt(25)

Nhóm 1: x1, x5, x6

Nhóm 2: x2, x3, x4

Ví dụ

Nhóm 1: x1(1,4), x5(4,3), x6(5,2)
Nhóm 2: x2(1,6), x3(2,6), x4(3,8)

Bước 2
c1=((1+4+5)/3, (4+3+2)/3)=(10/3,3)
c2=((1+2+3)/3, (6+6+8)/3)=(2,20/3)

d(x1,c1)= sqrt((1-10/3)^2+(4-3)^2)=sqrt(58/9)
d(x1,c2)=sqrt((1-2)^2 +(4-20/3)^2)=sqrt(73/9)

d(x2,c1)= sqrt((1-10/3)^2 +(6-3)^2)=sqrt(130/9)
d(x2,c2)= sqrt((1-2)^2 +(6-20/3)^2)=sqrt(13/9)

d(x3,c1)= sqrt((2-10/3)^2 +(6-3)^2)=sqrt(97/9)
d(x3,c2)= sqrt((2-2)^2 +(6-20/3)^2)=sqrt(4/9)

d(x4,c1)= sqrt((3-10/3)^2 +(8-3)^2)=sqrt(226/9)
d(x4,c2)= sqrt((3-2)^2 +(8-20/3)^2)=sqrt(25/9)

d(x5,c1)= sqrt((4-10/3)^2 +(3-3)^2)=sqrt(4/9)
d(x5,c2)= sqrt((4-2)^2 +(3-20/3)^2)=sqrt(157/9)

d(x6,c1)= sqrt((5-10/3)^2 +(2-3)^2)=sqrt(34/9)
d(x6,c2)= sqrt((5-2)^2 +(2-20/3)^2)=sqrt(277/9)

Nhóm 1: x1, x5, x6
Nhóm 2: x2, x3, x4