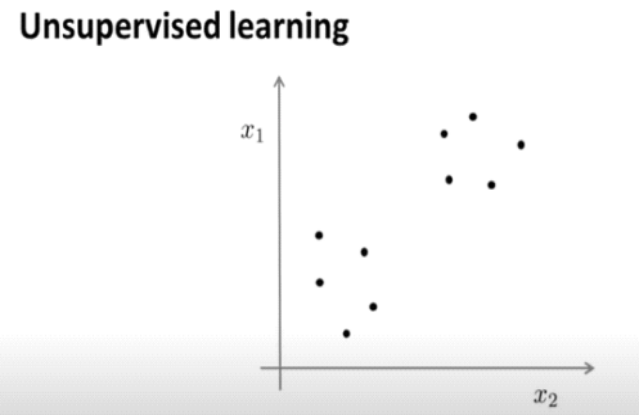
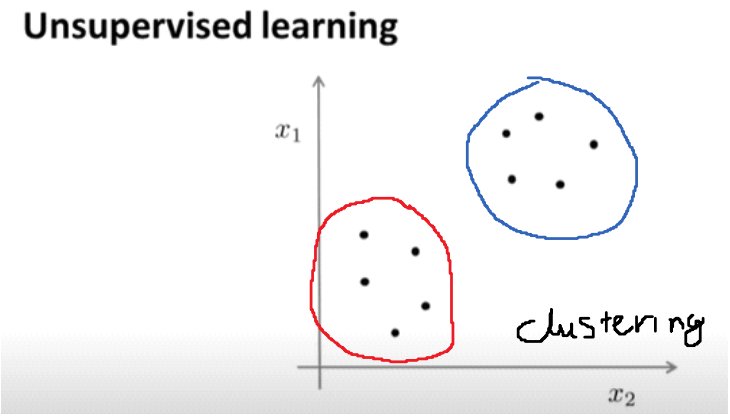
Clustering

Un supervised Learning

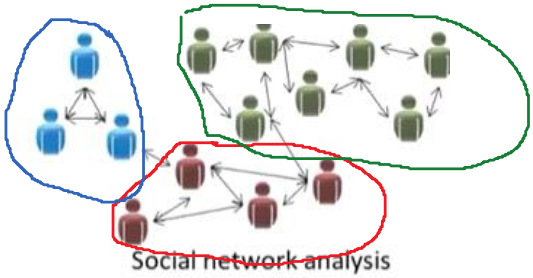




Phân cụm các dữ liệu có tính chất giống nhau

Ứng dụng

→ Market segmentation

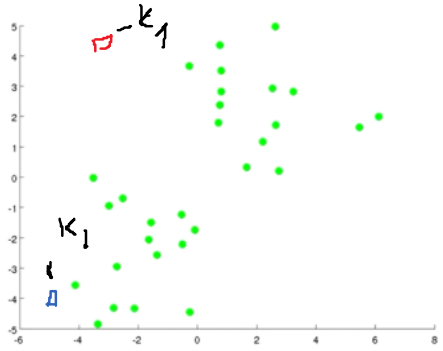


Social network analysis

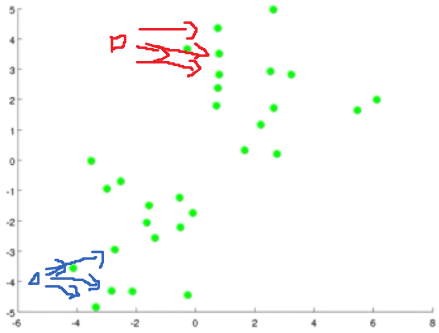
thuật toán cơ bản trong Unsupervised Learning

các bước của thuật toán K-mean

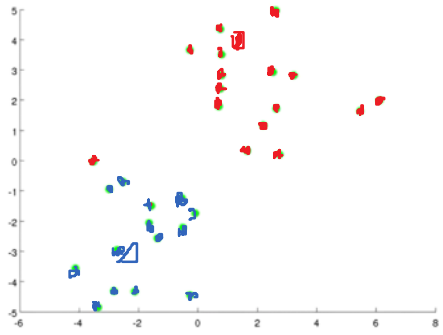
B_1: Random K center



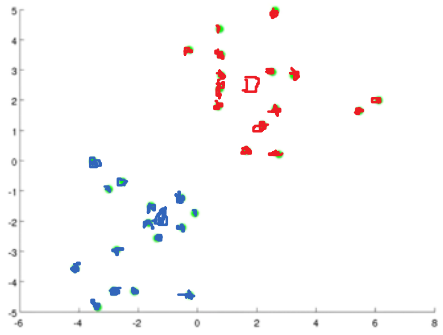
B_2: tính KC center tới những điểm gần nhất



B_3: di chuyển center đến trung tâm các điểm



B_4: tiếp tục tính KC và dịch chuyển



thuật toán

- Input: K

+ training Set

mô tả

sinh ngẫu nhiên K center $u_1, u_2, u_3, \dots, u_K$

{

gán nhãn

for i = 1 to m (m là số example data)

$c^{(i)}$ = chỉ số của center gần $x^{(i)}$

for j = 1 to K

u_j = trung tâm (trung bình) của tất cả các điểm có nhãn j

}

Optimization

$c^{(i)}$ = index của center mà $x^{(i)}$ nằm trong

μ_k = trung tâm (center) của nhóm K

$\mu_{c^{(i)}}$ = center mà $x^{(i)}$ đã được dán nhãn

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$$

$$\min_{c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_K} J$$

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$ } giữ nguyên μ_K và cập nhật

for k = 1 to K

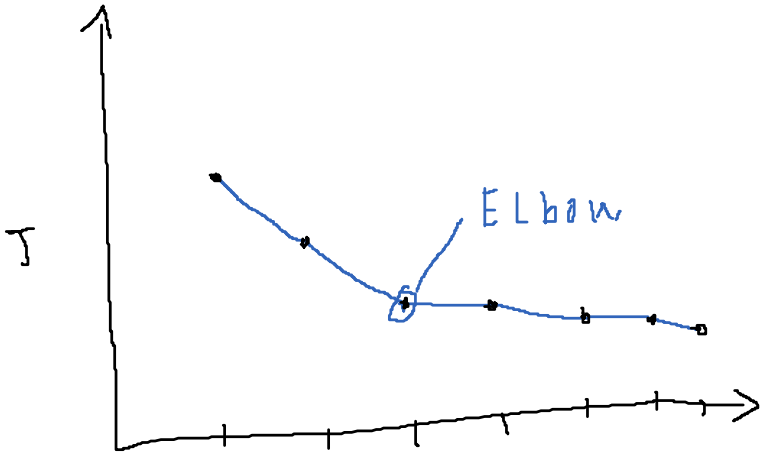
μ_k := average (mean) of points assigned to cluster k } cập nhật μ_K

}

Choosing value of Key

Elbow method hoặc là theo mục đích

VD: cần 6 size cho $K = 6$



K

Cho tập dữ liệu $X = [x_1, x_2, \dots, x_N] \in R^{d \times N}$ và $K < N$

và center $m_1, m_2, \dots, m_K \in R^{1 \times K}$

Với mỗi điểm x_i đặt 1 Vecto $y_i = [y_{1K}, y_{2i}, y_{3i}, \dots, y_{Ki}]$

y là vector one-hot $[0, 1, 0, 0, \dots 0]$

để khoảng cách tới center nhỏ nhất

$$\|x_1 - m_K\|_2^2 \qquad (\text{norm } 2)$$

vì $x_i \in$ cluster (center K) $\Rightarrow \begin{cases} y_{iK} = 1 & \forall j \neq K \\ y_{ij} = 0 \end{cases}$

$$\Rightarrow y_i \|x_i - m_K\|_2^2 = \sum_{j=1}^{K} y_{ij} \cdot \|x_i - m_j\|$$

- K = số cluster
- tính kc từ x_i -> center
- one-hot nên chỉ có K c ngắn nhất

áp dụng cho toàn bộ data set

$$J(M, Y) = \sum_{i=1}^{m} \sum_{j=1}^{K} y_{ij} \|x_i - m_j\|_2^2$$

$$\Rightarrow \min_{M,Y} J$$

$$M, Y = \underset{M,Y}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{K} y_{ij} \|x_i - m_j\|_2^2$$

thỏa mãn $y_{ij} \in \{0, 1\}$ và $\sum_{j=1}^{K} y_{ij} = 1 \Leftrightarrow y_i$ one hot

* khái niệm argmin là giá trị của biến để hàm số min

VD $f(x) = x^2 - 2x + 1 = (x-1)^2$

$\min f(x) = 0$

$\operatorname{argmin} f(x) = 1$ vì $\min f(x)$ tại $x = 1$

$\min f(x) = 0$

$\arg\min_x f(x) = 1$ vì $\min f(x)$ tại $x = 1$

giải bài toán: xem Kế giải Y và M khi cái còn lại cố định

Cố định M tìm Y: Sau khi tìm được hoàn di chuyển các center

$$x_i = \arg\min_{y_i} \sum_{j=1}^{K} y_{ij} \|x_i - m_j\|_2^2$$

$$\Leftrightarrow i = \arg\min_{j} \sum_{j=0}^{r} \|x_i - m_j\|_2^2$$

Cố định Y tìm M: Sau khi đánh nhãn cho tất data

$$m_j = \arg\min_{m_j} \sum_{i=1}^{N} y_{ij} \|x_i - m_j\|_2^2$$ N tập dữ liệu

Đặt $l(m_j) = \sum_{i=1}^{N} y_{ij} \|x_i - m_j\|_2^2$

$$\Rightarrow \frac{\partial l(m_j)}{\partial m_j} = \sum_{i=1}^{N} 2 y_{ij} (m_j - x_i) \quad (I)$$

giả (I) = 0

$$\Rightarrow m_j \sum_{i=1}^{N} 2 y_{ij} = \sum_{i=1}^{N} 2 y_{ij} \cdot x_i$$

$$\Rightarrow m_j = \frac{\sum_{i=1}^{N} y_{ij} x_i}{\sum_{i=1}^{N} y_{ij}}$$

tổng các $x_i \in$ cluster j

phép đếm số lượng $x_i \in$ cluster j

$\Rightarrow m_j = \frac{\text{tổng}}{\text{SL}} \Leftrightarrow$ trung bình cộng của cluster j