Sunday, June 2, 2024 10:25 AM

Unsupervised learning

$D \Rightarrow \{x_i, y_i\}$ → supervised learning

$D = \{x_i\}$ → unsupervised learning

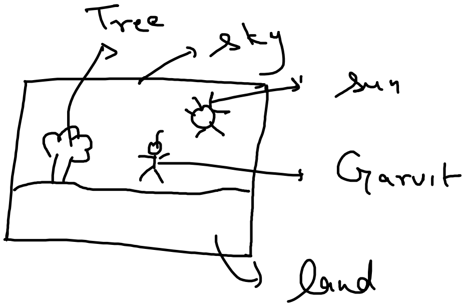
Clustering
↓
KMeans

Applications: e-commerce: group customers → location, income, gender etc.

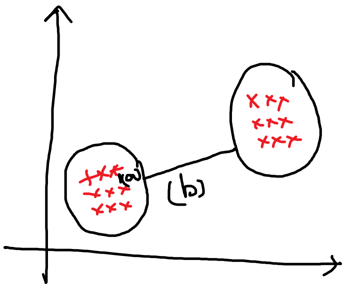
⇒ Review Analysis : Amazon Reviews

+ve —ve

→ Image Segmentation:



Metrics

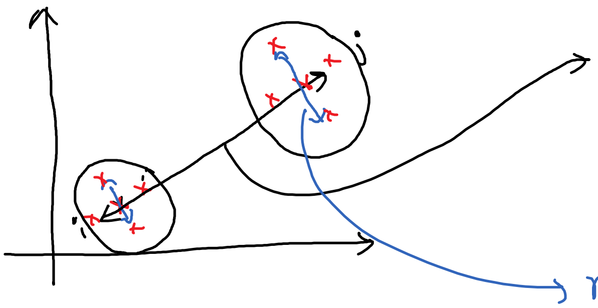


⇒ Intercluster distance (b)

⇒ Intracluster distance

Characteristics: a) → intracluster distance should be small
(good cluster) b) → intercluster distance should be large

Dunn's Index ⇒ $\dfrac{\max d(i,j) \longrightarrow \text{intercluster distance}}{\max d'(k) \longrightarrow \text{intracluster distance}}$



max d(i,j) ⇒ distance b/w farthest points in different cluster.

max d'(k) ⇒ distance b/w farthest points in same cluster

b) Silhouette's Score ⇒ $\dfrac{b-a}{\max(b,a)}$ ⟶ b ⇒ avg intercluster distance

[−1, +1]

a ⇒ avg intracluster distance

∴

case1: a ⇒ min ⇒ 0   b ⇒ b (best scenario)

$$SS = \frac{b-0}{\max(b,0)} = \frac{b}{b} = 1$$

Case 2: $b < a \,,\; b = 0 \,,\; a = a$ completely wrong

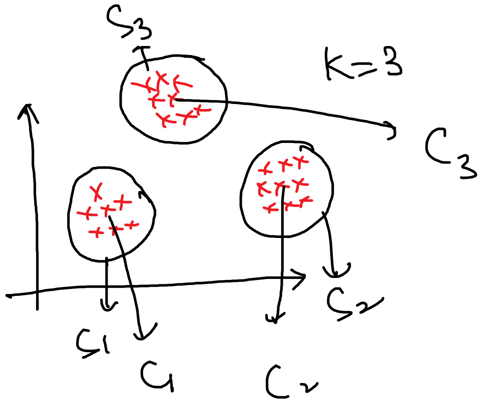
$$SS = \frac{0-a}{\max(0,a)} = \frac{-a}{a} = -1$$

Case3: $a = b$

$$SS = \frac{a-a}{\max(a,a)} = \frac{0}{a} = 0$$

$\#$clusters $\Leftarrow$ **K-Means** $\rightarrow$ Mean(μ)

↳ Centroid



$C_1 , C_2 , C_3 \Rightarrow$ centroids

$S_1 , S_2 , S_3 \Rightarrow$ sets

$S_1 \cap S_2 = \phi \quad S_3 \cap S_1 = \phi$

$S_2 \cap S_3 = \phi$

$$C = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$x_i \in S_i$

$$MOF \Rightarrow C^* = \underset{C_1, C_2, C_3 -- C_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \| x_i - C_i \|^2$$

↳ intracluster distance ↓

$x_i \in S_i$

$S_i \cap S_j = \phi$

nphard problem

Lloyd's Algorithm:

① Randomly choose k datapoints as centroids

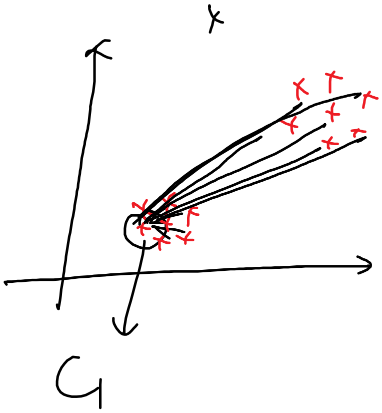
② Assignment: For each point, select the nearest centroid with the help of distance & add that point to the corresponding cluster.

③ Updation: Recalculate centroids,

$$C_i = \frac{1}{S_i} \sum_{i=1}^{n} x_i \qquad x_i \in S_i$$

④ Repeat step ② & ③ till convergence

KMeans ++



datapoints	distance
x_1	d_1
x_2	d_2
|	|
|	|
x_n	d_n

distance $\propto$ prob. of being picked as centroid

KMeans ⇒ affected by outliers