05 October 2024 11:33

Clustering → grouping

↓ X_i

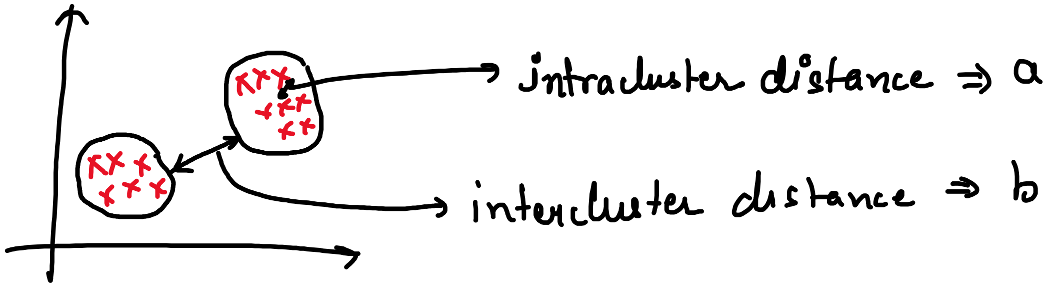
↳ Unsupervised learning

Applications: ① ecommerce ⇒ customer segmentation

② Review Analysis

③ Image Segmentations

Metrics



Characteristic of a good cluster:

1) intra cluster distance must be small

2) intercluster " " " large

Silhoutte's Score:

b ⇒ avg intercluster

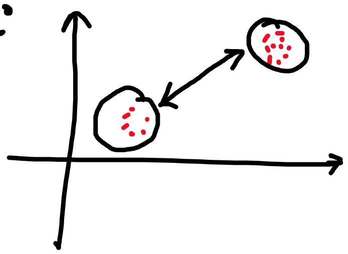
Silhouette Score:

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

b ⇒ avg intercluster distance

a ⇒ avg intracluster distance.

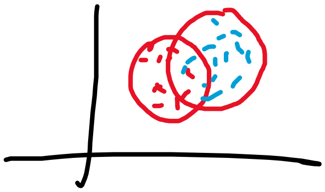
Case 1:



$a \Rightarrow 0 \Rightarrow \min$ $b \Rightarrow b$

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

Case 2:



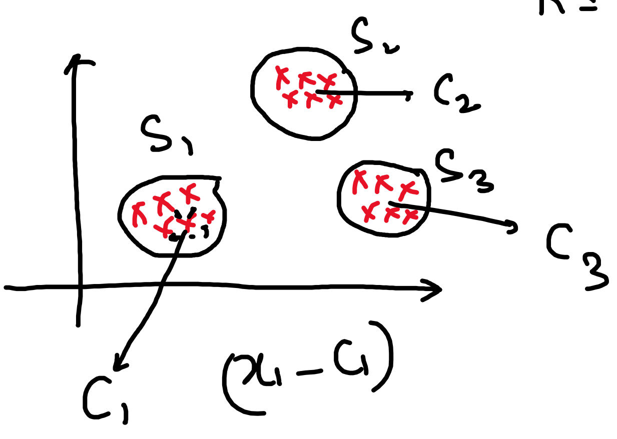
$a = b$

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

KMeans → Mean (Centroid)

KMeans ⇒ # cluster

K = 3



$C_1, C_2, C_3 \Rightarrow$ centroids

$S_1, S_2, S_3 \Rightarrow$ sets

$$S_1 \cap S_2 = \emptyset$$

$$S_2 \cap S_3 = \emptyset$$

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

$$x_i \in S_i$$

$$S_3 \cap S_1 = \emptyset$$

<u>MOF</u>: $C^* = \text{argmin} \sum_{1=1}^{k} \overbrace{\sum_{x_1 \in S_1} \|X - C_1\|^2}$

↳ intracluster distance

↓

np hard problem

<u>Lloyd's Algorithm</u>

① Randomly choose pts as centroids.

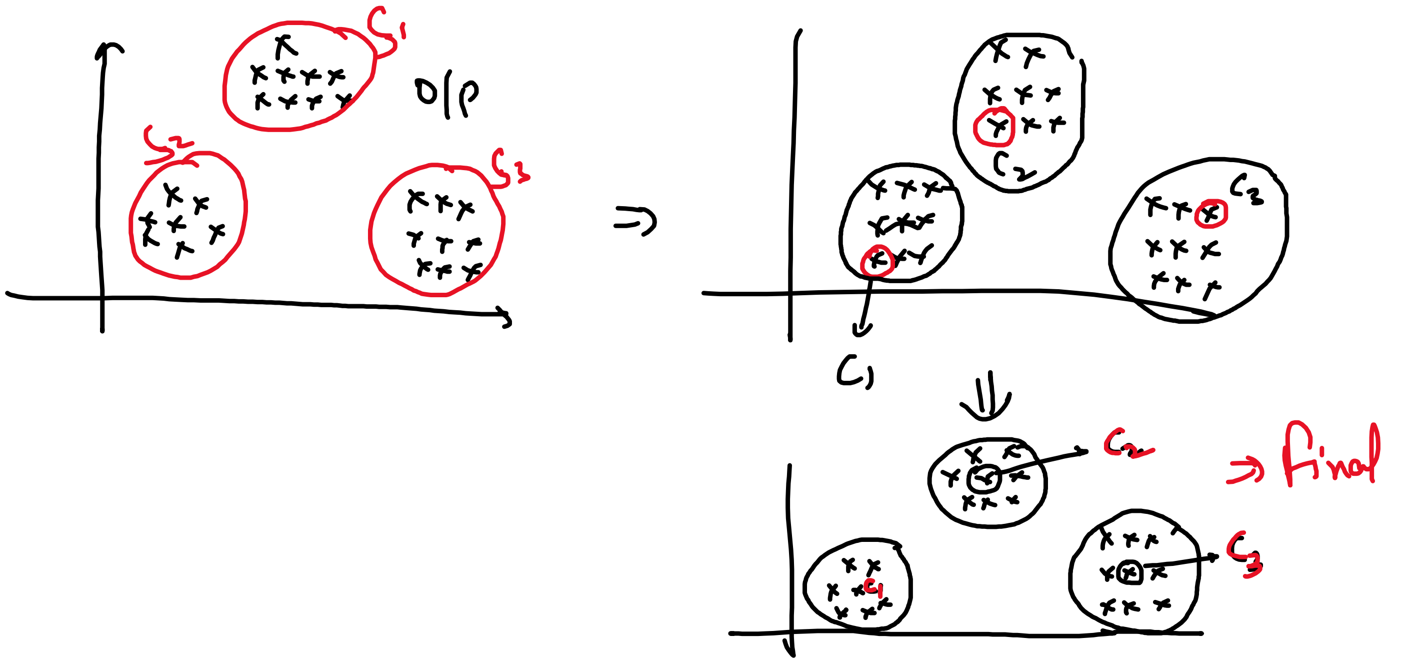
② Assignment: for each pt, select the nearest Centroid (with help of distance) & add that pt to Corresponding cluster

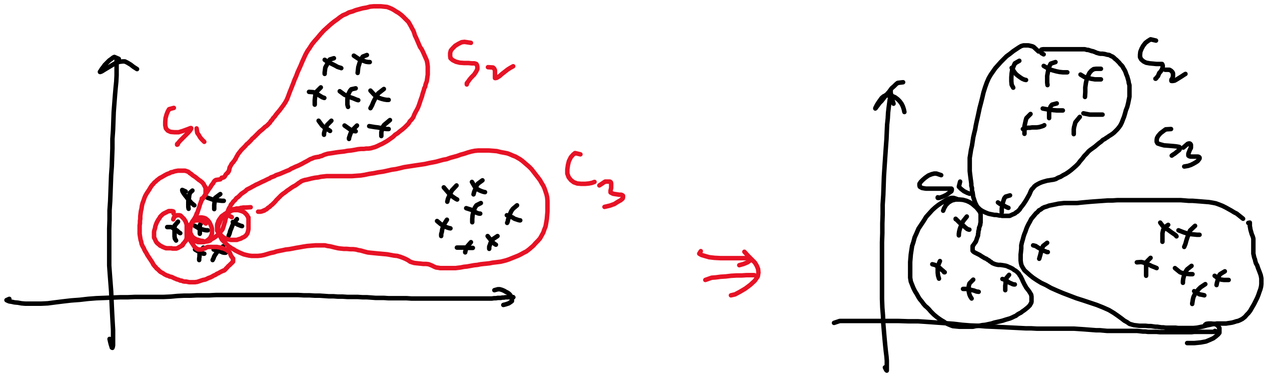
③ Update. Recalulate Centroid

$$\underbrace{C_1 = \frac{1}{S_1}\sum_{1=1}^{n} x_i}_{x_1 \in S_1}$$

④ T[illegible]

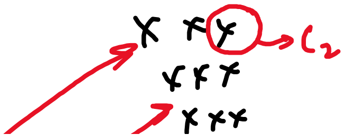
④ There ② & ③ steps will be repeated till centroid Stop changing



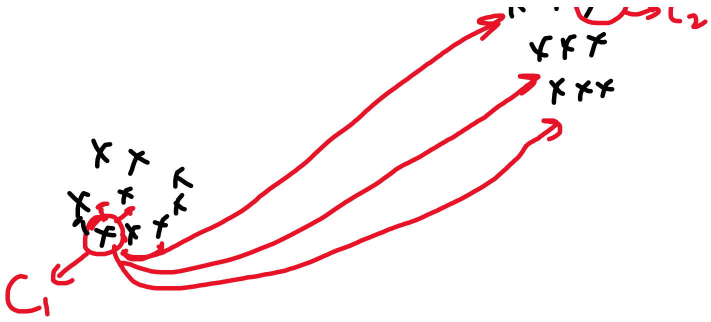


KMeans++ :

① Choose only one random centroid



datapoint	distance from C1
x_1	d_1
x_2	d_2



x_1	d_1
x_2	d_2
x_3	d_3
|	|
x_n	d_n

distance $\propto$ probability

KMeans is sensitive to outliers.