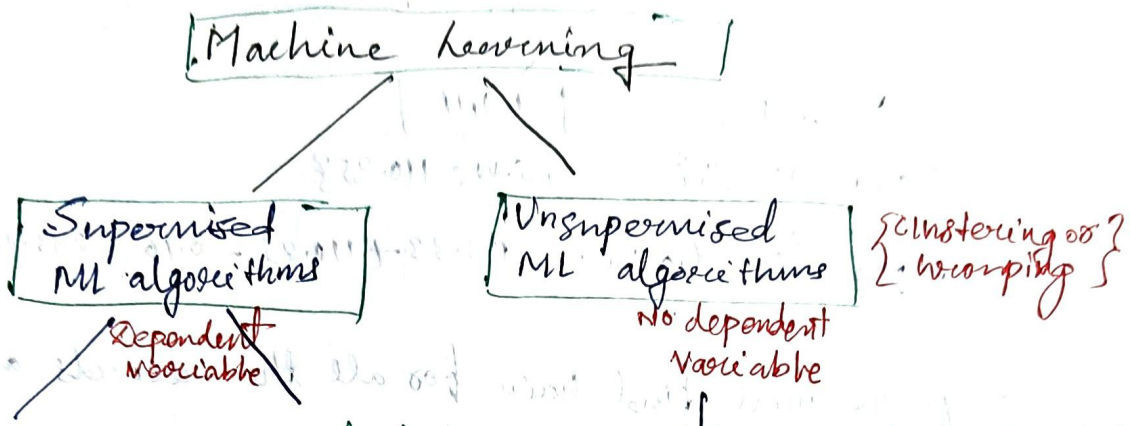
28-NOV-2022

Unsupervised learning



Regression
1. Linear Regression
2. SVR
3. DTR
4. RFR
5. GBR
6. XBR
7. L1 & L2 Reg.

Classification
1. Logistic Regression
2. SVC
3. DTC
4. RFC
5. XBC
6. GBC / ABC

Unsupervised
1. K-Means and K-Means++ (updated version)
2. Hierarchical
3. DBScan

Eg: Market Segmentation

8. KNN (can be used for both supervised and unsupervised problem).

Dataset

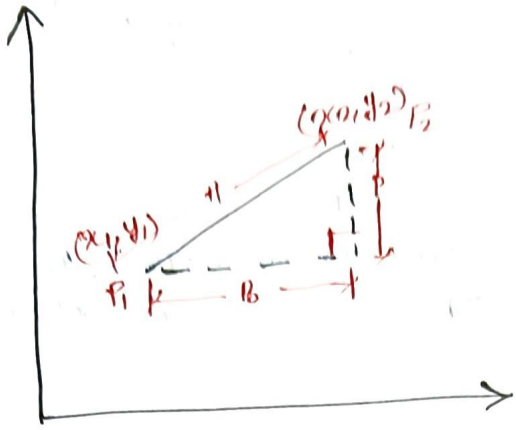
Height	Weight	BMI	Country
170	60	21	IND
180	65	22	UK
160	70	20	USA
165	75	18	IND
145	85	19	USA

3 groups {IND, USA, UK}

- Based on similarity, grouping happened.
(by correlation, by distance)
  - Euclidean distance
  - ~~Manhattan distance~~
  - Manhatten distance
  - Cosine distance
  - Tanimoto distance
  - Squared Euclidean distance

K-Means

⇒ K-means → Data → Similarity → Distance → Euclidian distance



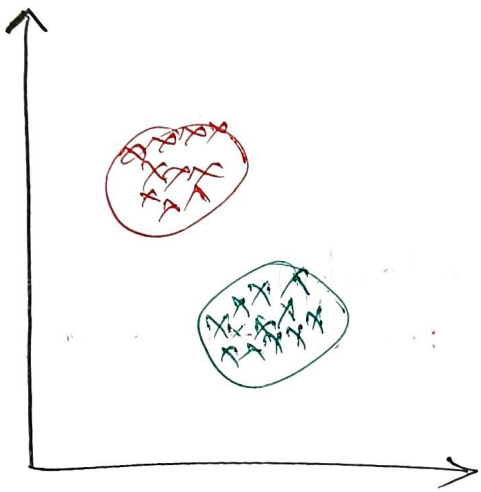
$$H^2 = P^2 + B^2$$

$$H = \sqrt{P^2 + B^2}$$

Distance b/w P_1 & P_2,

$$H = \sqrt{P^2 + B^2}$$

$$D(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Two Group

	Height	Weight
①	185	72
②	170	56
③	168	60
④	179	68
⑤	182	72
⑥	188	77
⑦	180	71
⑧	180	70
⑨	183	84
⑩	180	88
⑪	180	67
⑫	167	76

① Centroid
② Distance
③ Mean

} connect these 3 to understand K-Means

> K-Means
↓
Number of centroid → {Centroid value (random if, we have to build cluster)}

→ How many centroid should we take?
To choose K, we use elbow method. We use WCSS (within cluster sum of square) in elbow method.

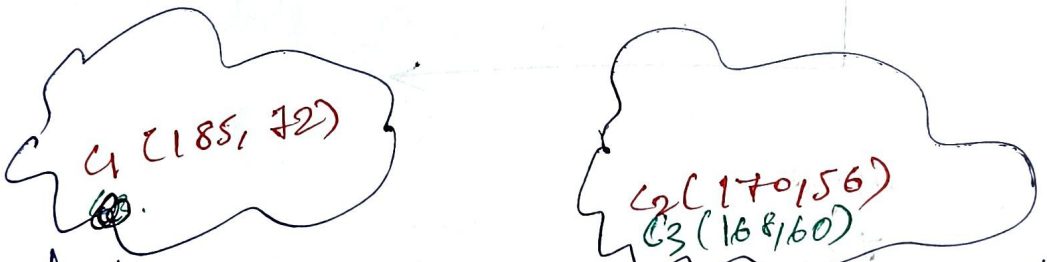
WCSS
- Inter cluster
- Intra cluster

Evaluation of clustering model

① Dunn Index
② Silhouette score

Steps for K-Means

① Initialize centroid (Randomly)
Lets take K = 2 {Number of centroid = no. of cluster}



→ find euclidian distance, b/w C1 and C3, C2 and C3
Lets assume, C3 (168, 60)

d(C1, C3) = 3 | (less distance) ↓ 3rd point will be assigned to C2.
d(C2, C3) = 8

Actual distance

$$d(C_1, C_3) = \sqrt{(168-185)^2 + (60-72)^2} = 20.80$$

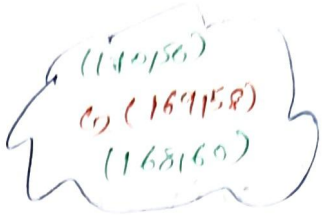
$$d(C_2, C_3) = \sqrt{170 \times (168-170)^2 + (60-56)^2} = 4.14$$ (less distance) ↓ C3 belong to 2nd cluster (C2)

Now, update the cluster 2 centroid.

updated centroid = $\left(\frac{170+168}{2}, \frac{56+60}{2}\right)$

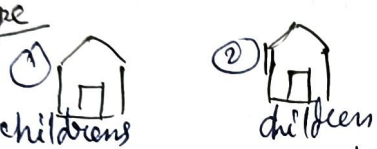
New centroid = (169, 58)

> Every time centroid will get updated after adding the addition of new points



> Similarly, find the distance to each point and keep updating centroid till the end.

Example



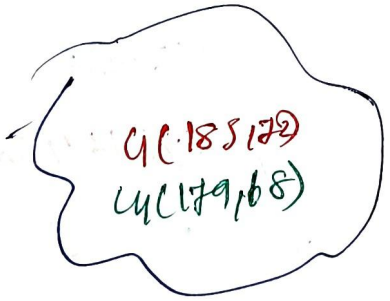
→ House 1 children has more similarity to house 1 people as compared to house 2.

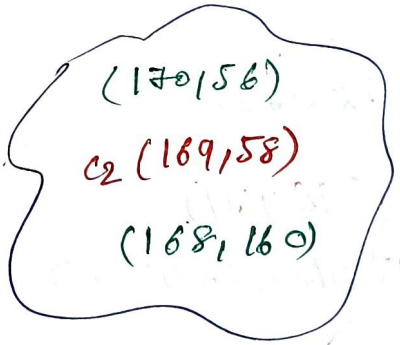
4th point $c_4(179, 68)$

find actual distance wrt c_1 and c_2

$d_1(c_1, c_4) = \sqrt{(179-185)^2 + (68-72)^2} = 7.21$ → (less distance as compared to d_2. ↓ Belong to c_1.)

$d_2(c_2, c_4) = \sqrt{(179-169)^2 + (68-58)^2} = 14.14$





Now, update the cluster 1 centroid.

New centroid = $\left(\frac{185+179}{2}, \frac{72+68}{2}\right)$

= (182, 70)

(185,72)
C1 (182,70)
(179,68)

(170,56)
C2 (169,58)
(168,60)

5th point, C5 (182,72),
Find actual distance w.r.t C5,

$d_1(C_1, C_5) = \sqrt{(182-182)^2 + (72-70)^2} = 2$ (less distance ∴ Belong to C1)

$d_2(C_1, C_5) = \sqrt{(182-169)^2 + (72-58)^2} = 19.1$

(185,72)
C1 (182,70)
(179,68), (182,72)

(170,56)
C2 (169,58)
(168,60)

Now, update the centroid of cluster 1,

New centroid $= \left(\frac{182+182}{2}, \frac{70+72}{2}\right)$

$= (182, 71)$

(185,72)
C1 (182,71)
(179,68) (182,72)

(170,56)
C2 (169,58)
(168,60)

→ Similarly find the distance w.r.t each point and keep updating centroid. Like this we can get final 2 clusters.

> Keep iterating till we reach all the data points.

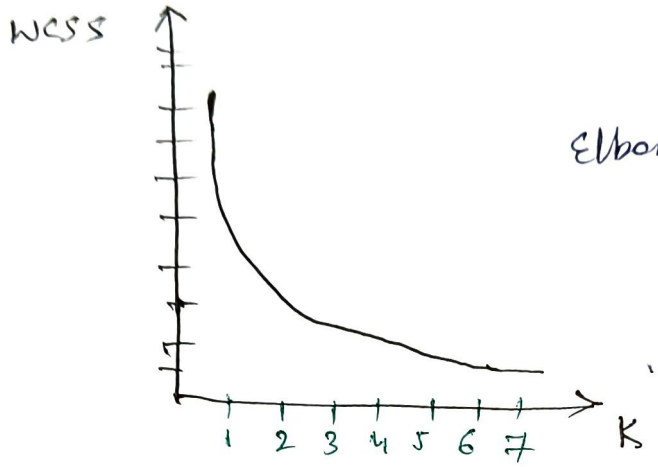
> If both distance is same, any cluster can be choosen.

Elbow method and WCSS

Ans

K=2 (How can we decide?)

→ Using elbow method, we can decide.

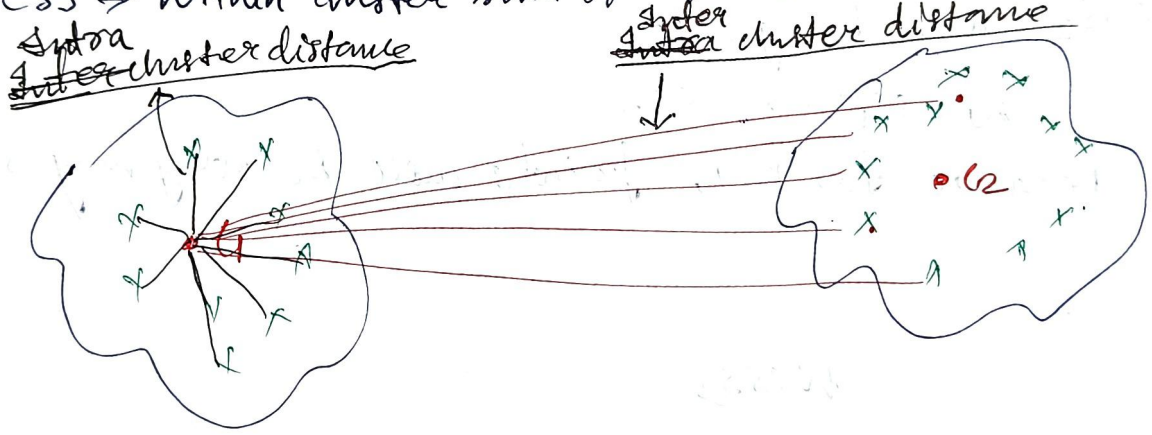


Elbow curve.

WCSS → within cluster sum of squares

Intra cluster distance

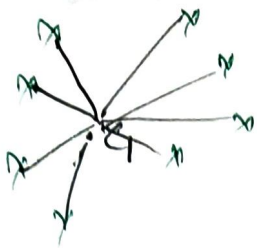
Inter cluster distance



{ Intra cluster distance – within cluster
Inter cluster distance – B/w the cluster }

WCSS

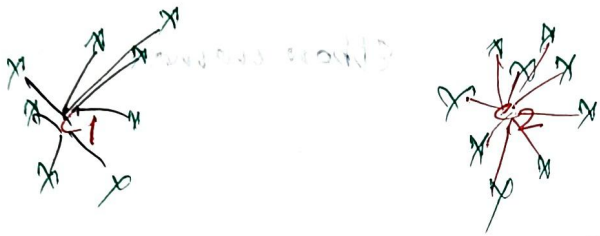
For K=1,



$$WCSS_1 = \sum_{i=1}^{n} d(C_1, x_i)^2$$

For K=2,

Two clusters,



$WCSS_2$

which WCSS will be greater?

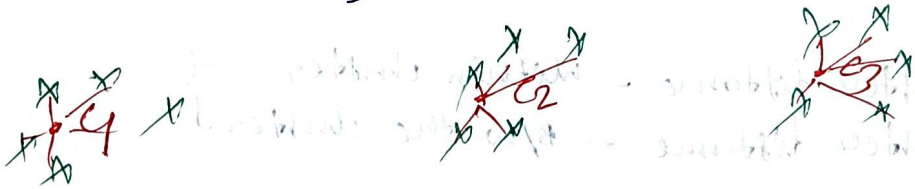
→ $WCSS_1$ will be greater.

$WCSS_1 > WCSS_2$

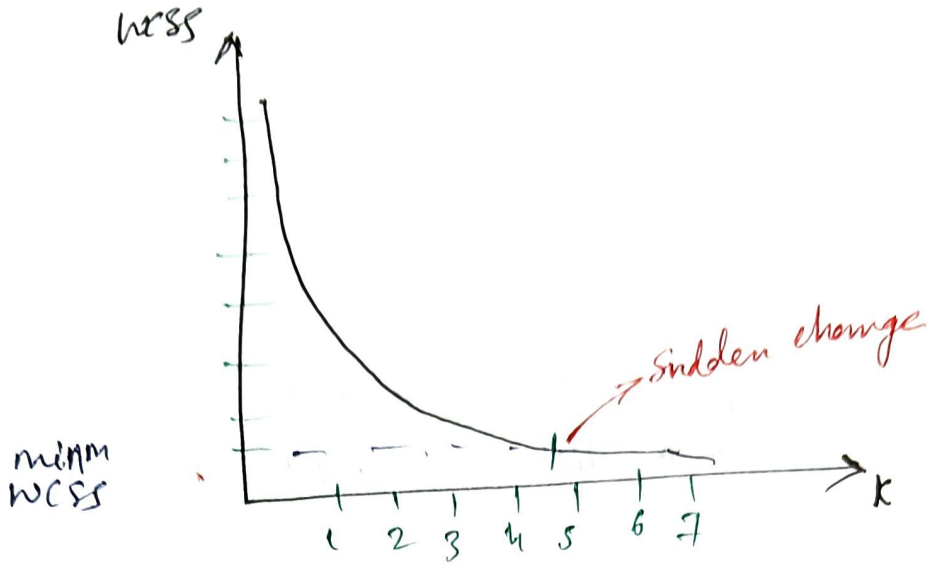
because for K=1, points will be very very scattered.

For K=3,

$WCSS_3$



$$WCSS_1 > WCSS_2 > WCSS_3$$



⇒ choose K=5, because elt has abrupt changes.

> Take + average value of WCSS if no. of cluster is greater than 1.

Q what is the diff b/w K-means and K-means++?

Ans

Validation of the clustering

1. Dunn Index
2. Silhouette Score

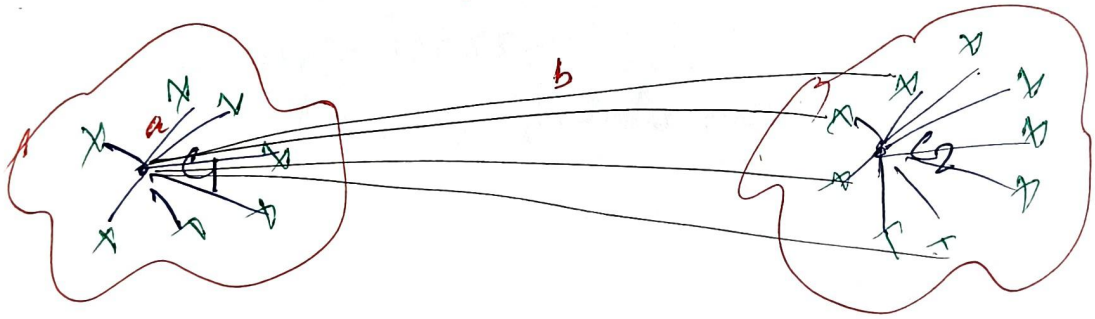
1. Dunn Index

$$\text{Dunn Index} = \frac{\max \text{distance}(x_i, x_j)}{\max \text{distance}(y_i, y_j)}$$

2. Silhouette Score $[-1 \leq S.S \leq +1]$

$$\text{Silhouette score} = \frac{b_i - a_i}{\max(b_i - a_i)} \quad \{-1 \leq S.S \leq 1\}$$

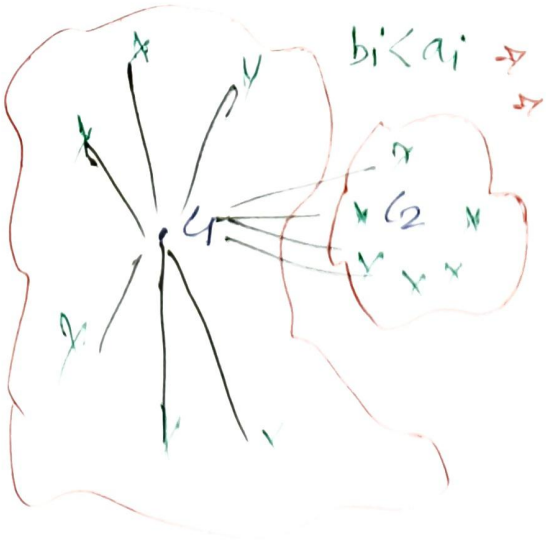
a_i = Intra cluster (within cluster)
b_i = Inter cluster (B/w the cluster)



$\{-1$, (worst model)
$+1$, (Best model) $\}$

Example: let's assume $b = 50$, $a = 10$
$50 - 10 = 40$ (less scattered data, good Model)

$b = 10$, $a = 50$
$10 - 50 = -40$ (more scattered data, not good)



$b_i < a_i$ → Worst model.

→ In same cluster, values are more scattered.

Q How to make a best model or optimise solution?

Ans Custom learning / Custom model / Semi supervised learning

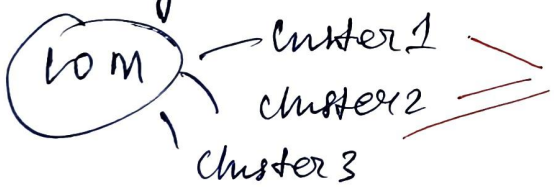
> combination of supervised and unsupervised learning.

Dataset

weight	Height	Gender
170	55	M
180	60	F
165	70	M
180	80	F
155	50	F
160	100	F

Q Build classification model and optimise result?

Ans We can perform clustering 1st and then classification. Basically we are segregating the data.



⇒ then any supervised ML algorithm.

> For categorical features, we can find hamming distance.