Sri
28/11/22

Unsupervised Machine Learning.

Machine learning

- Supervised ML
  - Classification
    - Logistic Regression
    - SVC
    - DTC
    - RFC
    - Adaboost
    - XGBoost
    - GradientBoo
    - KNNC
  - Regression
    - Linear Regression
    - Logistic "
    - SVR
    - DT Regressor
    - Random forest regr
    - Gradient Boost Regr
    - Adaboost regr
    - XG Boost regr
    - KNNR
- Unsupervised ML → we are performing clustering of the data
  - K-means - K-means++
  - Hierarchial clustering
  - DB Scan.

Target / dependent / supervised

Supervised Algorithms

Eg:

Height	weight	BMI → supervised	Country
170	60	21	India
180	65	22	UK
160	70	22	USA
165	75	18	India
140	55	19	USA

Unsupervised ML
↓
Clustering → grouping

based on above dataset, we can make based on country
3 clusters or 3 groups.
→ one group — one group
India, USA, UK → one group.

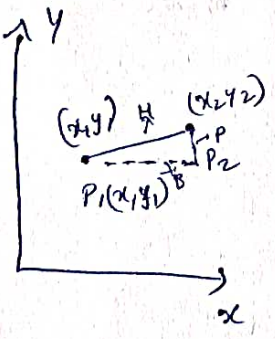
Mathematically
1. K-means
2. Hirarchial
3. DBScan

(1) K-means

Based on Similarity Measurement we can do clustering :-

- Based on distance : Euclidean distance measure
  Manhatten
  Cosine
  Tanimoto
  Squared Euclidean

Data → similarity → Distance → Euclidean distance.

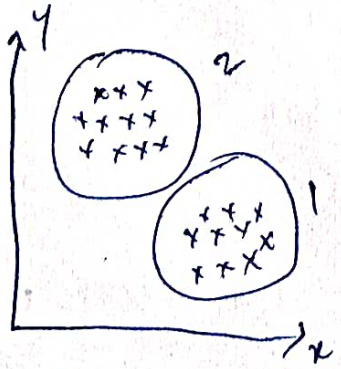


Pythagorus theorem

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

$$(P_1 P_2)\ H = \sqrt{P^2 + B^2}$$

$$D(P_1P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



y x
_ _
_ _
_ _

Height	weight	→ Clustering → based on distance we can find the similarity
185	72	c_1
170	56	c_2
168	60	
179	68	
182	72	
188	77	
180	71	
180	70	
160	84	
183	88	
180	67	

- Centroid
- Distance
- Mean

} K-means

Height	Weight
185	72
170	56
168	
179	
182	
188	

no. of Centroid

- ELBOW method → we can decide the no. of clusters using Elbow
- WCSS → we can evaluate within clusters Sum of Square

[inter clusters / Intra Cluster]

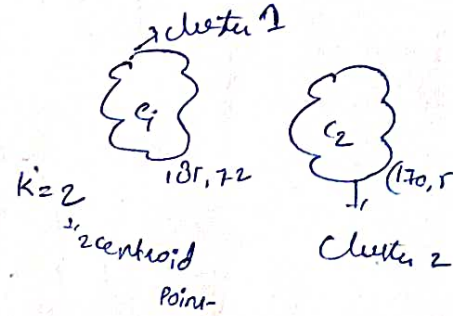
Evaluation methods:-

- Dunn Index
- Silhoutee Coff

- Centroid [randomly] → center value around that we need to build the clusters

here centroid = 2
K = 2

(185, 72) (170, 56)
C_1 C_2



Euclidean distance → d(C1,3)
d(C2,3)

C1→3 C2→3
5 8

No. of centroid = No. of clusters

C_1 3
(185, 72) (168, 60)

$$\sqrt{(168-185)^2+(60-72)^2}$$

$D(C_1,3) = \sqrt{20.80} = 4.4$

(170, 56)
C_2 (168, 60)

$D(C_2,3) = \sqrt{(170-168)^2 + (56-60)^2}$

$= \sqrt{20} = 4.4$

Update the cluster centroid

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

$(169, 58)$

3rd cluster is belong to 2nd cluster.

here we are considering min distance.

for cluster 1, = 20.08

for cluster 2, = 4.4.

since we can say that 3rd row belongs to cluster '2'.

Distance fn $D(C_1, 4) = \sqrt{(181-179)^2 + (72-68)^2} = 7.21$

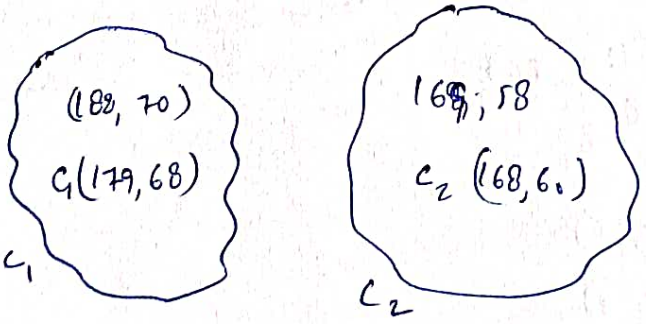
$D(C_2, 4) = \sqrt{(169-179)^2 + (58-68)^2} = 14.42.$

4th value belongs to C_1. bec. of less distance.

or belongs to C_1 because of low distance.

Update the cluster centroid

$\left(\frac{179+181}{2}\right), \left(\frac{68+72}{2}\right) = (182, 70)$



for (182, 72)

distance fn $d(C_1, 5) = \sqrt{(182-182)^2 + (72-70)^2} = 2$

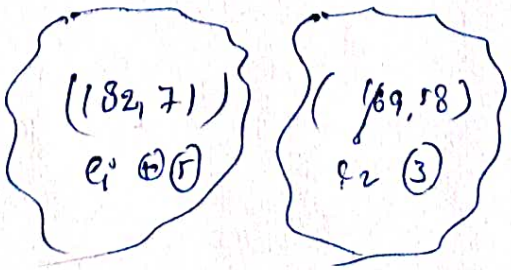
$d(C_2, 5) = (182-169)^2 + (72-58)^2 = 19.1$

(5 belongs to C_1)

it belongs to 1 cluster

$= \frac{182+182}{2}, \frac{70+72}{2}$

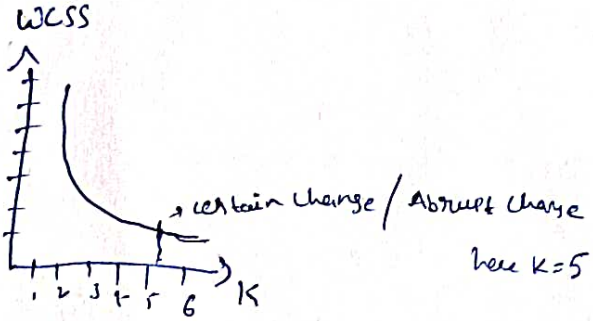
$C_1 = (182, 71)$



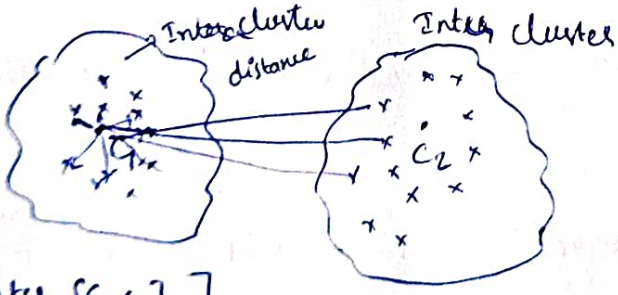
K-means

1. Centroid
2. Distance [compute, min]
3. Include point in cluster, update the centroid

How can we consider how many clusters we need & choose, based on ELBOW Method.

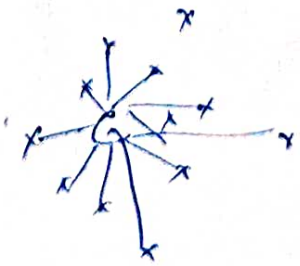


WCSS :- Within Cluster Sum of Square



[Inter $[C_1, C_2]$
Intra : within cluster]

K=1 → 1 centroid / 1 cluster



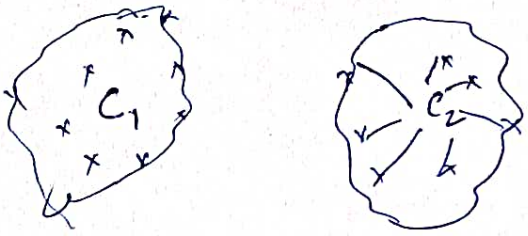
⇒ WCSS

Within cluster sum of square

WCSS,

K=2

-two glows.



$WCSS_2$ =

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

$WCSS_1 > WCSS_2$

K=3

$WSSS_1 > WCSS_2 > WCSS_3$

Difference b/n K-means/ K-means++

- The diff b/n lies in: the selection of centroids around which the clustering takes place.
- K means++ removes the drawback of K means which is it is dependent on initialization of centroid.
- Guarantee convergence :- Can warm-start the positions of centroids. Easily to new examples. Generalizes to clusters of diff. shapes and sizes, such as elliptical clusters.

How to validate clusters :- $K=5$

Eg in regression based on R^2 / adjusted R^2 $\quad R^2 = [0, 1]$ if $R^2=0$ → it's a worst
$R^2=1$ → it's best.

Classification based on Auc/RoC / confusion Matrix

1. Dunn Index
2. Silhoute score

1. Dunn Index $= \dfrac{\text{min dis}(x_i, x_j)}{\text{max dis}(y_i, y_j)}$

2. Silhoute score $= \dfrac{b_i - a_i}{\max(b_i - a_i)} \Rightarrow [-1 \text{ to } 1]$; worst -1, best $+1$

a_i = intra clusters

b_i = inter clusters