Machine Learning Day 11: Clustering

13/12/2022

Machine Learning

Supervised Machine Learning | Unsupervised Machine Learning

★ **Unsupervised Machine Learning:**

Unsupervised Machine Learning uses machine learning algorithm to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groups without the need of human intervention.

Difference Between Supervised and Unsupervised Learning

⇒ In supervised machine Learning, input data is provided to the model along with the output. The goal of supervised learning is to train the model so that it can predict the output when it is given new data.

⇒ In unsupervised learning, only input data is provided to the model. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning does not need any supervision. Instead, it finds patterns from the data by its own.

Unsupervised Machine Learning Algorithm

1. K- means → K-mean ++
2. Hierarchical Clustering
3. Dbscan clustering

UnSupervised machine Learning

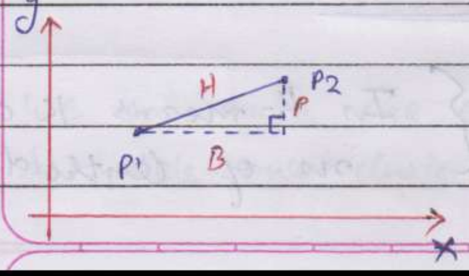
Height	weight	BMI	Country
170	60	21	IND
180	65	22	UK
160	70	20	USA
165	75	18	IND
140	55	19	USA

In Supervised machine Learning we predict the target feature (BMI) using independent features (Height, weight).

In Unsupervised machine Learning we make clusters based on country column. Clustering means grouping of data.

1. K- means :-

Data → Similarity → Distance → Euclidean Distance



Phythagoras Theorem.

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

SAHIL JOSAN

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

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

Distance betn P_1, P_2

Euclidean Distance

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

DATASET : Perform Clustering on this dataset

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
160	70
183	84
180	88
180	67
167	76

Important point of K-means :-

1. Centroid
2. Distance
3. mean

{ In K-means K denotes the no. of Centroid }

Another important points:-
→ ELBOW method.
→ WCSS: within cluster sum of square.
→ Intercluster
→ Intracluster.

For evaluation of clustering method.
→ Dunn Index.
→ Silhoute Score / Silhoute Coefficient.

★ Now when we get the dataset 1st we need to find out the Centroid and that we will find randomly Intially we take two centroids.

Here 1st record 185 72 } These points we are
and 2nd record 170 56 } considering as Centroid
and around these centroid we are creating our cluster.

First Centroid C_1 (185, 72) C_2 2nd Centroid. (170, 56)

Centroid is the center value around which we are creating our clusters.

Here we have 2 centroid that means our K value in K-means is 2.

We can take one cluster also, but everything will be inside one cluster only

Now we have our Two Centroids C_1 and C_2. Now we will calculate the Euclidean distance between each centroid and all other points. means we take one point and calculate its distance from C_1 and C_2 both. whose distance is minimum we will put that point in that respective centroid.

For example. Suppose for 3rd point.

Distance $(C_1, 3) = 5$
Distance $(C_2, 3) = 8$

Here 3rd point is near to Centroid C_1, so it will go in cluster C_1

C_1, 3	C_2
(185, 72)	(170, 56)

Comming back to our dataset Now we calculate our actual distance.

	Height	Weight	
①	185	72	C_1
②	170	56	C_2
③	168	60	
④	179	68	
⑤	182	72	
⑥	188	77	
⑦	180	71	
⑧	160	70	

C_1	③
(185, 72)	(168, 60)

3rd point Euclidean distance.

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

$$= \sqrt{433}$$

$$(C_1, 3) = 20.8$$

Now Distance between C_2 and point 3
(170, 56) (168, 60)

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

$$\text{Distance}((2,3) = 4.472$$

So Distance of point 3 is lower to C_2 then C_1, therefore point 3 will belong to C_2 cluster.

⇒ Similarly calculate distance of point 4 from C_1 and C_2

$C_1 = (185, 72)$ $C_2 = (170, 56)$ point 4 = (179, 68)

$$\text{Distance}(C_1, 4) = \sqrt{(185-179)^2 + (72-68)^2}$$

$$= 7.211$$

$$\text{Distance}(C_2, 4) = \sqrt{(170-179)^2 + (56-68)^2}$$

$$= 15$$

∴ Point 4 will belong to cluster C_1

Now when we add point 3 to centroid C_2 the centroid C_2 will be updated.

$$\text{New } C_2 = \frac{170+168}{2}, \frac{56+60}{2}$$

$$\text{New } C_2 = (169, 58)$$

Similarly when we add point 4 to centroid C_1

$$\text{New } C_1 = \frac{185+179}{2}, \frac{72+68}{2}$$

$$\text{New } C_1 = (182, 70)$$

=> Now calculate distance of point 5 from updated C_1 and update C_2

point 5 = (182, 72) $C_1 = (182, 70)$ $C_2 = (169, 58)$

$$\text{Distance}(C_1, 5) = \sqrt{(182-182)^2 + (70-72)^2}$$

$$= 2$$

$$\text{Distance}(C_2, 5) = \sqrt{(182-169)^2 + (72-58)^2}$$

$$= 19.10$$

point 5 will add to centroid C_1

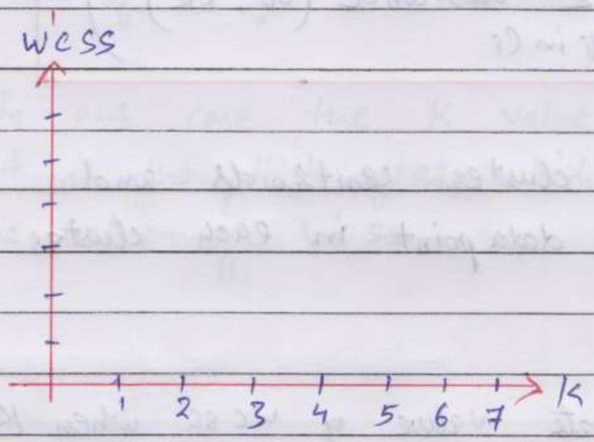
Now C_1 will be more updated

$$\text{Update } C_1 = \left(\frac{182+182}{2}, \frac{70+72}{2}\right)$$

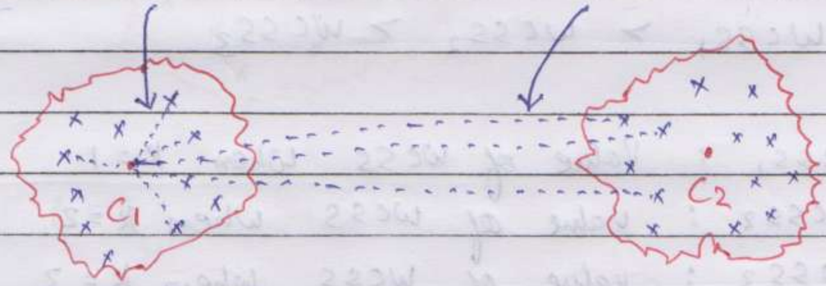
$$C_1 = (182, 71)$$

C_2 will remain same : (169, 58)

Elbow Method



WCSS: within cluster sum of Square



Intracluster distance is the distance between a data item and the cluster centroid with a cluster.

Intercluster distance is the distance between the data items in distinct clusters.

For K=1 means one centroid

$$\boxed{WCSS = \sum_{d_i \text{ in } C_i}^{d_n} distance(d_i, C_k)^2}$$

where C is the cluster centroids.
d_i is the data point in each cluster.

For $K = n$ means 2, 3, 4 etc centroid

$$\boxed{WCSS = \sum_{c_K}^{c_n} \left(\sum_{d_i \text{ in } C_i}^{d_m} \text{distance}\,(d_i, C_K)^2 \right)}$$

where C is the cluster centroids and
d is the data point in each cluster.

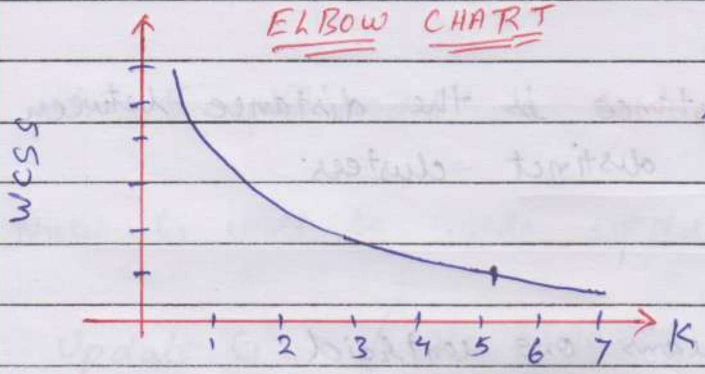
So If we calculate value of WCSS when $K=1$ then that value will be greater then WCSS value with $K=2$

$$WCSS_1 > WCSS_2 > WCSS_3$$

$WCSS_1$: Value of WCSS when $K=1$
$WCSS_2$: Value of WCSS when $K=2$
$WCSS_3$: value of WCSS when $K=3$

So. out WCSS vs K Graph will become.

ELBOW CHART



At some point, you will find sundden change and then there will be no change.
Means after 5 the value of WCSS when

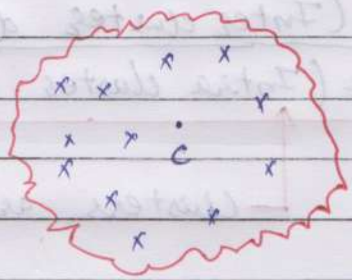
$K = 5, 6, 7$ will be same.

Ques Now comming to the Question what should be the value of K ?

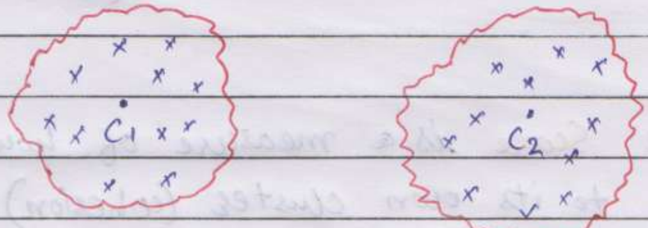
Ans: In our case the K value should be 5, because it is the point after which there is no change in the value of WCSS.

when $K=1$

Cluster will be Big and so the value of WCSS

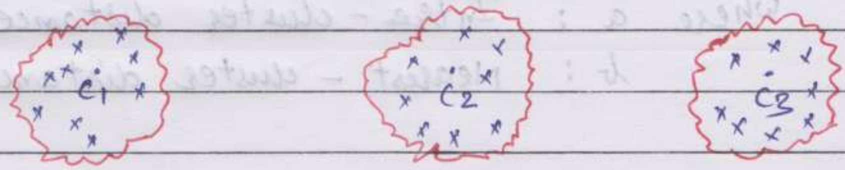


when $K=2$



Cluster will be smaller and so the value of WCSS will be smaller then $K=1$

when $K=3$



Now clusters will be much smaller because same points will be divided into 3 clusters.

How to validate cluster ?

1. Dunn Index (DI)
2. Silhoute Score

1. Dunn Index is calculated as a ratio of the smallest inter - cluster distance to the largest intra - cluster distance.

Clusters are far apart

$$\boxed{\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}}$$

Clusters are compact

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

2. Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

$$\boxed{\text{Silhouette Score} = \frac{b_i - a_i}{\max(a, b)}}$$

where a : Intra - cluster distance
b : Nearest - cluster distance

Silhouette score value ranges from -1 to 1.

1: means clusters are well apart from each other and clearly distinguished.