Group - B

Assignment No : 06

Aim: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

Objective:

- To understand Clustering.
- To solve K-Means clustering/ hierarchical clustering.

Hardware Requirement:

- PC with RAM 256 MB or above
- PC with RAM i3 or above

Software Requirement:

- Python
- Jupiter Notebook
- Chrome

Theory:

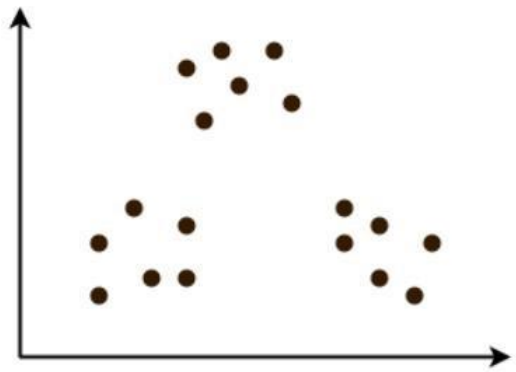
K-MEAN Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

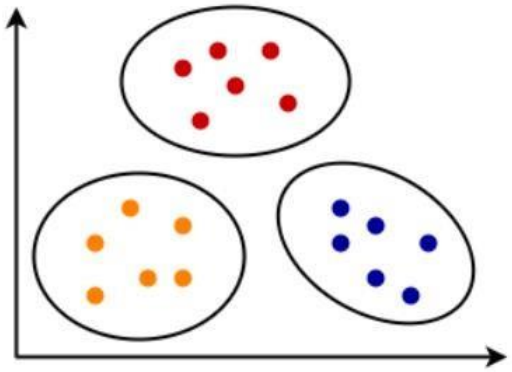
it is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.



Before K-Means



After K-Means

Grouping below data using k-mean Cluttering

SRNO	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	177	76

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Clusters	Height	Weight
K_1	185	72
K_2	170	56

Step-3: Calculate Euclidean distance

$$\sqrt{(Xo - Xc)^2 + (Yo - Yc)^2}$$

Xo: Height Observed value
Xc : Height Cluster value
Yo: Weight Observed value
Yc : Weight Cluster value

Euclidean distance for K_1	Euclidean distance for K_2
$\sqrt{(168 - 185)^2 + (60 - 72)^2}$ =20.80	$\sqrt{(168 - 170)^2 + (60 - 56)^2}$ =4.48

Assign each data point to their closest centroid, which will form the predefined K clusters, which would be K_2
Before Changing Value of K_2 mean value should be calculated of previous and new K_2

Updated centroid after first iteration

Clusters	Height	Weight
K_1	185	72
K_2	(170+168) / 2 =169	(56+60) / 2 =58

Step-4: Repeat Step 3 until it reaches last set

Algorithm :

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

Hierarchical Clustering in Machine Learning

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

1. **Agglomerative:** Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive**: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

Conclusion:

In this way, we learnt about K-Means clustering/ hierarchical clustering

Oral Question:

1. **What is Clustering?**
2. **Explain KMEAN Clustering**
3. **What is Hierarchical clustering**
4. **Solve the following using k-mean cluster for 2 centroids**

SRNO	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72