Unsupervised Learning

→ Training machine learning model with only input variables.

→ Unsupervised Learning is clustering algm. which clusters entire data analysis into different groups based on similarity.

eg:- ① Market Basket Analysis.

② Mall customer Segmentation

Date ....

→ k-Means Algorithm

→ k-Mean algm comes under unsupervised Learning and also called as clustering algm.

→ k-Mean is a clustering algm which is used to clasify unlabelled data [x] into groups / clusters based on similarity.

eg: Mall customer Segmentation.

k → No. of. clusters.

Similarity → Nearest distance.

Distance Measures:-

→ Euclidean → $d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

→ Manhattan → $d = |x_2-x_1| + |y_2-y_1|$

Date ....

How K-Mean Algorithm works?

→ ① Plot data

② Define no. of. clusters. (K)

③ It will randomly create no. of. clusters.

④ Initialise Centroids in each cluster

centroids → Cluster centers.

Centroids are found by taking average of all the points in each clusters.

⑤ Assign each observation to the nearest cluster based on distance.

→ find distance between data and centroid, if its near to first ~~class~~ cluster then it belongs to first cluster.

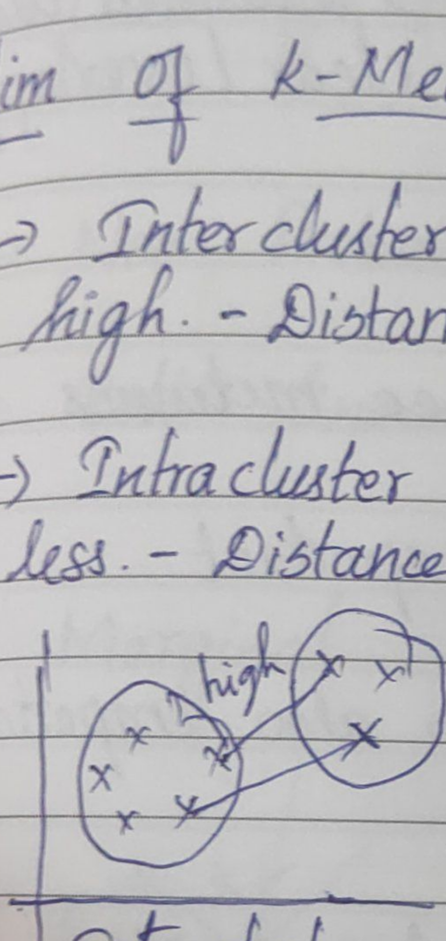
(6) Reinitialize the centroids.

(7) Repeat 5th step untill you get clearer clusters.

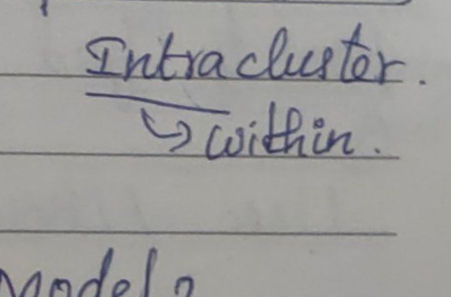
Aim of k-Mean Algorithm:-

→ Intercluster distance should be high. - Distance b/w observations in two clusters should be high.

→ Intracluster distance should be less. - Distance of observations within the cluster should be very less



Inter cluster
↳ b/w



Intracluster.
↳ within.

How to Evaluate k-Mean Model?

$$\text{Silhouette score} = \frac{b-a}{\max(a,b)}$$

a → distance within the cluster / intracluster

b → distance b/w the cluster / intercluster

Range of silhouette = $[-1, 1]$

Value is near to 1 → clearer cluster

Value is near to -1 → clusters are not clear / overlapping

Note:-

→ It will use distance metrsues.

→ Scaling is very important

→ Handling outlier is also important

How to find optimal value for k?

Use ELBOW METHOD.