22 May 2023 21:54

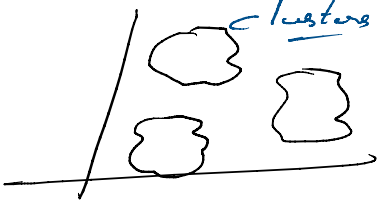
Unsupervised Learning

Training machine learning model with only input variables

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

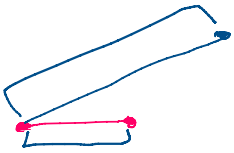
Eg. ① Market Basket Analysis

② Mall Customer Segmentation



→ K-Means Algorithm

* K-Mean algorithm comes under unsupervised Learning and also called as clustering algorithm.
* K-Mean is a clustering algorithm which is used to classify unlabelled data [x] into groups/clusters based on similarity



Eg:- Mall Customer Segmentation.

K → No of clusters

Similarity → Nearest distance b/w observations

Distance Measures.

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

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

How K-Mean Algorithm works?

① Plot data

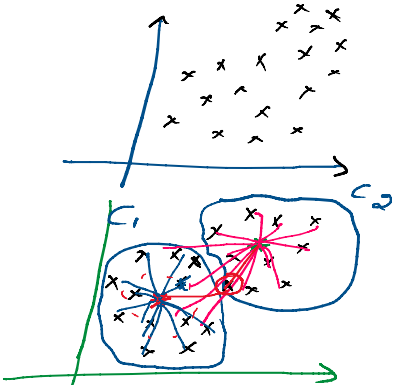
② Define no of clusters

K=2 → Create two clusters

③ It will randomly create two clusters.

④ Initialise Centroids in each Cluster.

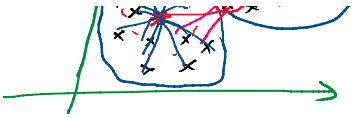
Centroids → Cluster centers



④ Initialize centroids for the clusters.

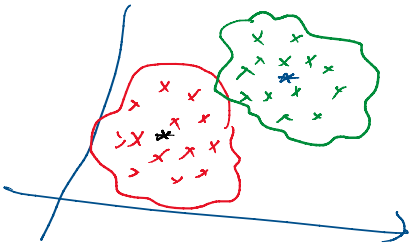
Centroids → Cluster centers

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



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

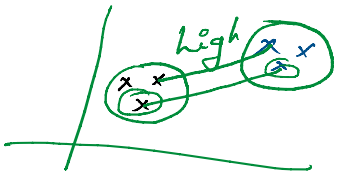
→ find distance between data and centroid, if it is near to first cluster then it belongs to first cluster.



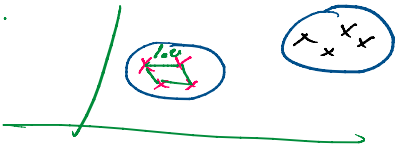
⑥ Reinitialize the centroids

⑦ Repeat 5th step until you get clearer clusters.

Aim of K-Mean Algorithm



* Intercluster distance should be high.
  - Distance blw observation in two clusters should be high.
* Intracluster distance should be less.
  - Distance of observations within the cluster should be very less



How to Evaluate K-Mean Model?

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

a → distance within the cluster / intracluster

b → distance between the clusters / intercluster

Range of Silhouette = $[-1, 1]$.

value is near to 1 → Clearer clusters

value is near to -1 → Clusters are not clear / overlapping.