03-03-2021 KANAV BANSAL

LECTURE-58

UNSUPERVISED LEARNING:

Unsupervised learning is where we only have input data(X) and no corresponding output variables.

→ The goal is to model the underlying structure (or) distribution in the data in order to learn more about the data.

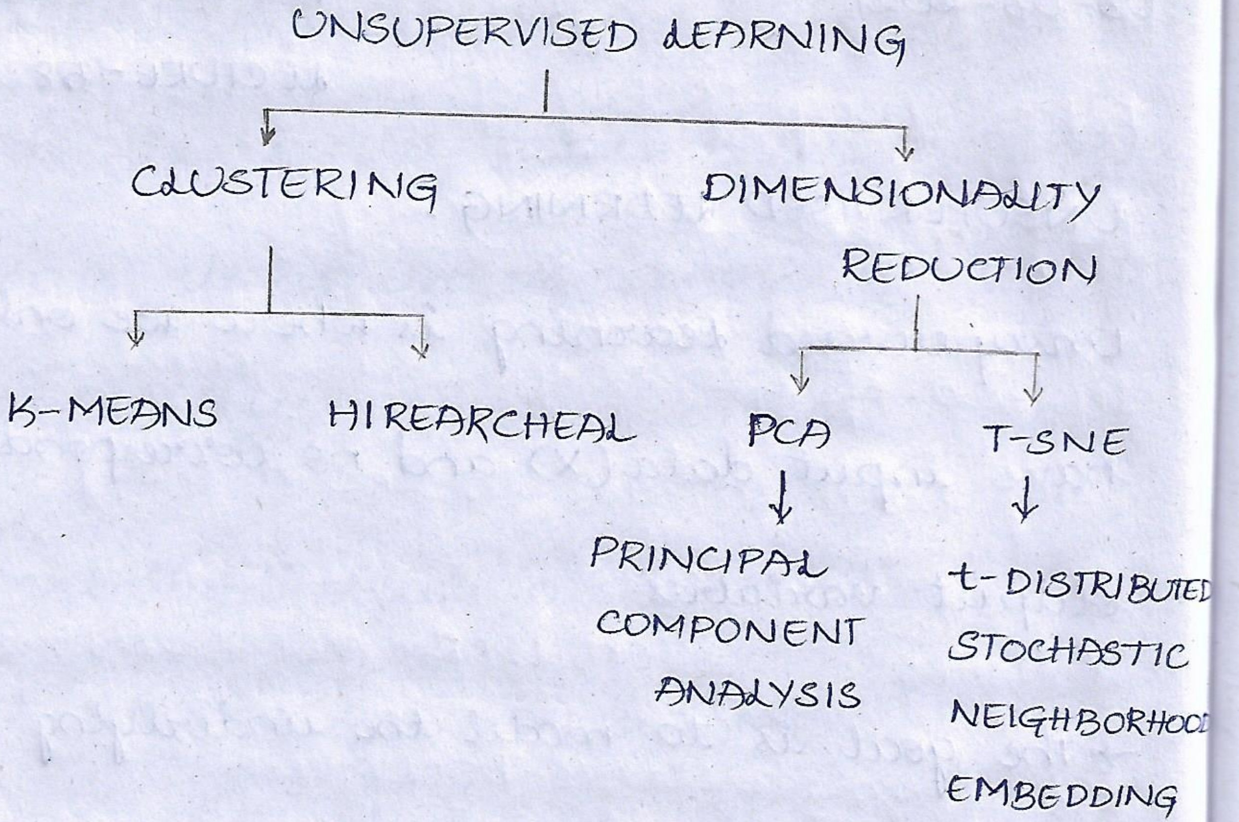
TECHNIQUES OF UNSUPERVISED LEARNING:

Common Algorithms used in unsupervised learning include

→ clustering

→ Anomaly Detection

→ Neural Networks & approaches for learning latent variable models.



GUASSIAN MIXTURE MODEL (GMM):

The traditional GMM for pattern recognition is an unsupervised learning method.

→ GMM ~~model~~ are a probabilistic model for representing normally distributed subpopulations within a over population.

→ Estimating the parameters of the individual normal distribution components is a canonical problem in modelling data with GMM's.

CLUSTERING: NOT A PREDICTION

clustering is a machine learning technique that involves the grouping of data points.

→ In theory, Data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features.

WHY CLUSTERING IS USED?

→ Useful for explorating/ exploring data.

K-MEANS CLUSTERING:

An unsupervised learning algorithm, which groups unlabeled dataset into different clusters.

→ Its 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.

The Hyperparameter in K-means clustering is 'K'.

*K-means ~~is~~ & Hirearcheal clustering are HARD CLUSTERING & GMM is a SOFT CLUSTERING.

INTER CLUSTER DISTANCE: AS MAX. AS POSSIBLE

It shows the distance between data point with cluster center.

INTRA CLUSTER DISTANCE: AS MIN. AS POSSIBLE

It shows the distance between the data point of one cluster with the other data point in other cluster.

STEPS INVOLVED IN K-MEANS CLUSTERING:

→ Initialise 'K':

- Randomly pick 'K' points from data & assign them as centroids $c_1, c_2, \ldots, c_k$

°We generally take 'K' value as '3'.

→ Assignment step:

- Iterate i.e., for each data point (x) in data
  - * Select nearest centroid C_J where $J \in [1,2,3]$
  - * Add 'x' in S_J → list of S_J [ ]

→ Recompute the centroids

$$\Rightarrow c_i = \frac{1}{n} \sum x_j \text{ where } x_j \in S_i$$

→ Repeat step 2 & 3, until it reaches convergence

NOTE:

So, In step-3:

- The minimum distance in the data is considered

$c_1 \rightarrow S_1[\quad], \; c_2 \rightarrow S_2[\quad], \; c_3 \rightarrow S_3[\quad]$

From c_1, c_2, c_3 pick the middle values.

*→ Centroids doesn't move if they are at the center.

WHAT DOES K-SCORE MEAN?

The K-means objective is to reduce the sum of squares of the distances of points from their respective cluster centroids.

→ It has other names like J-squared error function, J-score (or) within-cluster sum of squares.

→ This value tells how internally coherent the clusters are.

$$\Rightarrow \min \sum_{i=1}^{K} \underset{\substack{\downarrow \\ \text{forall}(\forall) \\ \text{data points } (x)}}{\sum} \| x - \underset{\substack{\downarrow \\ \text{Nearest centroid}}}{c_i} \|^2 \text{ where } x \in s_i$$

ADVANTAGES:

↳ If variables are huge, then K-means most of the times computationally faster than hierarchical clustering, if we keep 'K' small.

↳ K-means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

DISADVANTAGES :

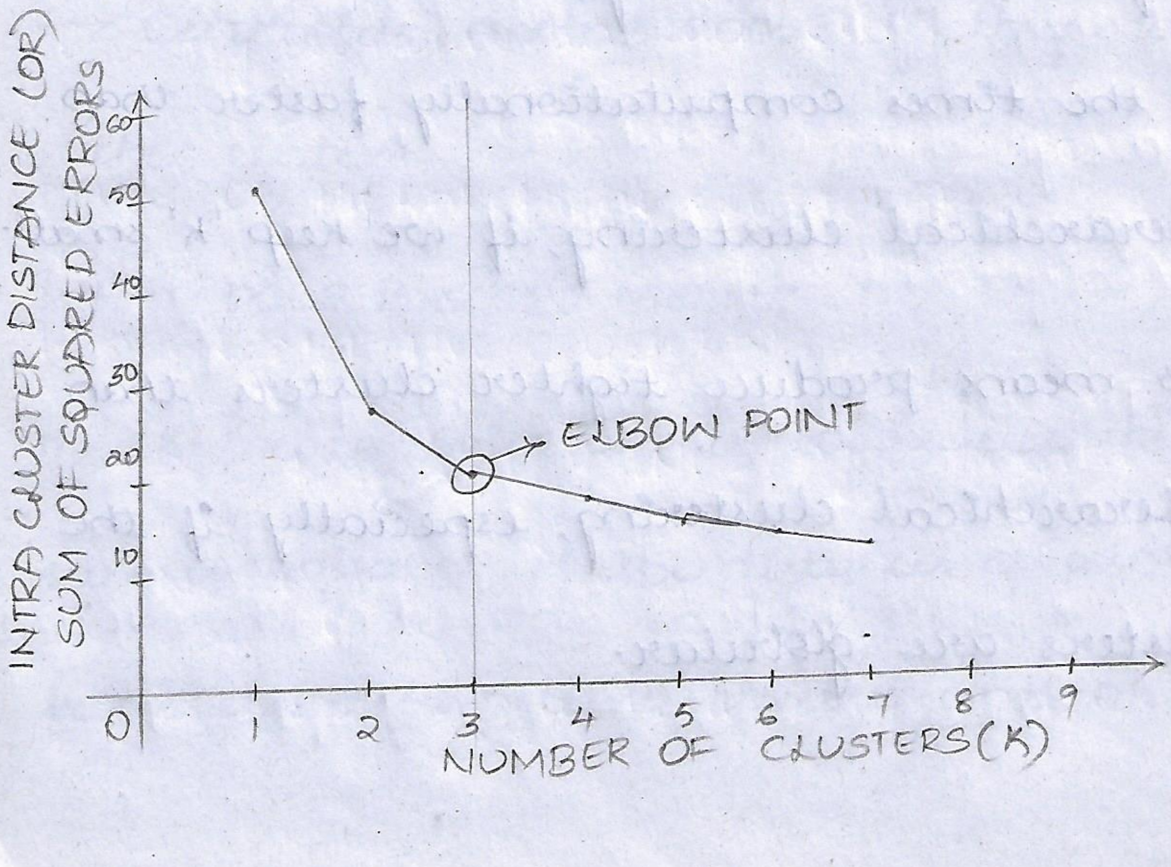
→ Difficult to predict K-value.

HOW TO FIND THE BEST VALUE OF 'K' in K-MEANS?

1. Compute K-means clustering algorithm for different values of K.

2. For each K, calculate the total within-cluster sum of square (WSS).

3. Plot the curve of 'WSS' according to the number of clusters K.



ELBOW POINT:

the point which defines the optimal number of clusters is known as "Elbow point".

→ It can be used as a visual measure to find out the best pick for the value of K.

NOTE:

As the K value increases, the intra cluster distance decreases.