ASSIGNMENT 5

Title : K Means Clustering.

Problem Statement :
(a) Apply Data Pre processing
(b) Perform data preparation
(c) Apply M.L algorithm
(d) Evaluate model
(e) Apply Cross Validation and evaluate Model

Objective :
This assignment will help the students to realize how to do clustering using K-means clustering algorithm.

Theory :

Introduction to K-Means Clustering Algo :
K-Means is a type of unsupervised learning which is used when you have an unlabeled data. The goal of this algorithm is to find groups in the data, with the no. of groups represented by the variable K.
The results of the K-means clustering algorithm :
① The centroids of the K-clusters, which can be used to label new data.
② Labels for the training data.

③ Rather than defining groups before looking at the data, clustering allows you to find the groups that have formed organically.

This introduction to the K-means algo. covers:
- Common business cases where K-means is used.
- The steps involved in running the algorithm.

Some examples of use cases are:

// Behavioral segmentation
- Segment by purchase history
- Segment by activities on application / platform.
- Create profiles based on activity monitoring.
- Define personas based on interests.

// Inventory categorization
- Group inventory by manufacturing metrics.
- Group inventory by sales activity.

// Sorting sensor measurements
- Detect activity types in motion sensor's
- Identify groups in health monitoring

// Detecting bots or anomalies
- Separate valid activity group from
- Group valid activity to clean up outlier detection. In addition, monitoring if a tracked data point switches between groups over time can be used to detect meaningful changes in the data.

Algorithm:

- The K means clustering algo. uses iterative refinement to produce a final result. The algo. inputs are the no. of clusters K and the data set.
- The data set is a collection of features for each data point. The algo. starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the dataset. The algo. then iterates between two steps:

1. Data assignment step:
Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared euclidean distance.

$$\underset{c_i \in C}{\operatorname{argmin}} \; dist(c_i, x)^2$$

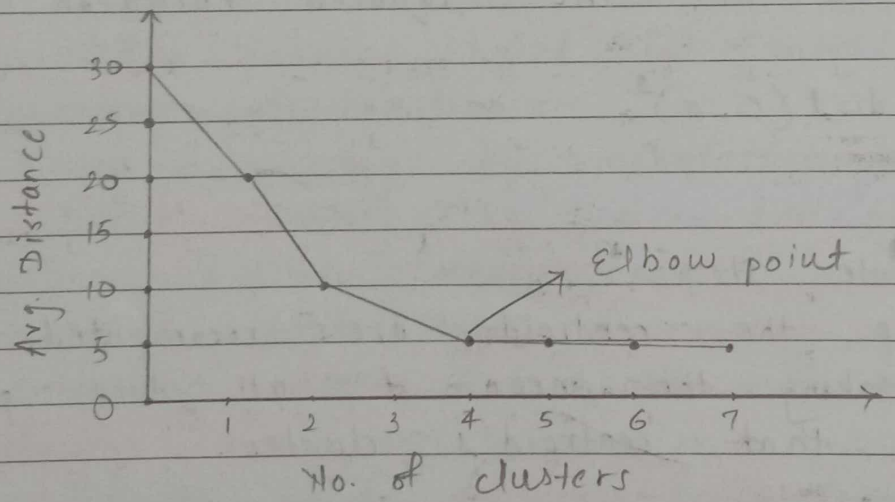
2. Centroid update step:
In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

The algo. iterates btn. steps one & two until a stopping criteria is met. This algo. is guaranteed to converge to a result. The result may be a local optimum.

Choosing K :

The algo. described above finds the clusters and data set labels for a particular prechosen K. To find the no. of clusters in data, user needs to run the K-means algo. for a range of K values and compare the results. Since increasing the no. of clusters will always reduce the dist. to data points, increasing K will always decrease this metric. Thus, this metric can't be used as the sole target. A no. of other techniques exist for validating K, i.e. cross validation, info. criteria, the silhouette method & the G-means algorithm.



Conclusion :

Successfully implemented K-means clustering algorithm for given dataset and problem statement.