Assignment DMW-2

Title :- Clustering Techniques.

Problem Statement :
Consider a suitable dataset for clustering of data instances in different groups apply different clustering techniques visualize the clusters using suitable tools.

Objectives :-
- Understand the working of k means & ~~DBSCAN~~ Hierarchical clustering techniques.
- Implement clustering models using python functions & libraries.

Outcomes :
Students will be able to :
- Understand the working of K Means and ~~DBSCAN~~ Hierarchical clustering techniques.
- Implement clustering models using python functions & libraries.

s/w & H/w Requirements :
fedora 20 / windows 10 , R studio.

Theory :-
1) k means clustering :-
It is one of the most commonly used unsupervised Machine learning clustering techniques. It is a centroid based clustering techniques. that needs you to decide

the number of cluster (centroids) & randomly placer the cluster centroids to begin the clustering process. The goal is to divide N observations into K clusters repeatedly until more groups can be formed.

Advantages of K means:
1. Easy to understand & implement
2. Can handle large ~~documents~~ datasets well

Disadvantages:
1. Sensitive to number of cluster / centroids
2. Does not work well with outliers.
3. Get difficult in high diamentional spaces as distance between points increases & Euclidean distance diverges.
4. Gets slower as number of diamentions increases.

K means Algorithm:
1. Decide the number of cluster. This number is called K & number of cluster is equal to the number of centroider, Based on the value of K generate the coordinates for K random centroides
2. for every point calculate the Euclidean distance between the point & each of the centroides
3. Assign the point to its ~~nearest~~ nearest centroides The points assigned to same centroid form a cluster.
4. Once clusters are formed, calculate new

Centroid for each cluster by taking the cluster mean. cluster mean is the mean of the x & y coordinates of all pts belonging to the cluster

5. Repeate ~~the~~ these steps until convergence

Elbow Method to find optimal number of cluster for K means:

1. for different values of k execute the following steps:
2. for each cluster calculate the sum squared distance of every pt.
3. Add the sum squared distances of each cluster to get total sum
4. Keep adding the total sum for each k to a list.
5. Plot the sum of squared distances & K ~~from~~ from the list.
6. Select the K at which a sharp change occurs (looks like an elbow).

2) Hierarchical clustering

Also called as Hierarchical cluster analysis or HCA or an unsupervised clustering algorithm ~~with~~ which involves creating cluster that have predominant ordering from top to bottom

for eg:- all files and folders on our hard disk are organized hierarchy.

The algorithm groups similar objects in groups called 'clusters'. The endpoint is set of clusters or groups where each cluster is distinct from each other cluster and the object within each cluster are broadly similar to each other

This clustering technique is divided into two types:-
1. Agglomerative Hierarchical clustering
2. Divisive Hierarchical clustering.

1. Agglomerative Hierarchical clustering.
It a bottom up approch: each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.
Some common linkage methods are:
1) Complete linkage
2) Single linkage
3) Average linkage
4) Centroid linkage.

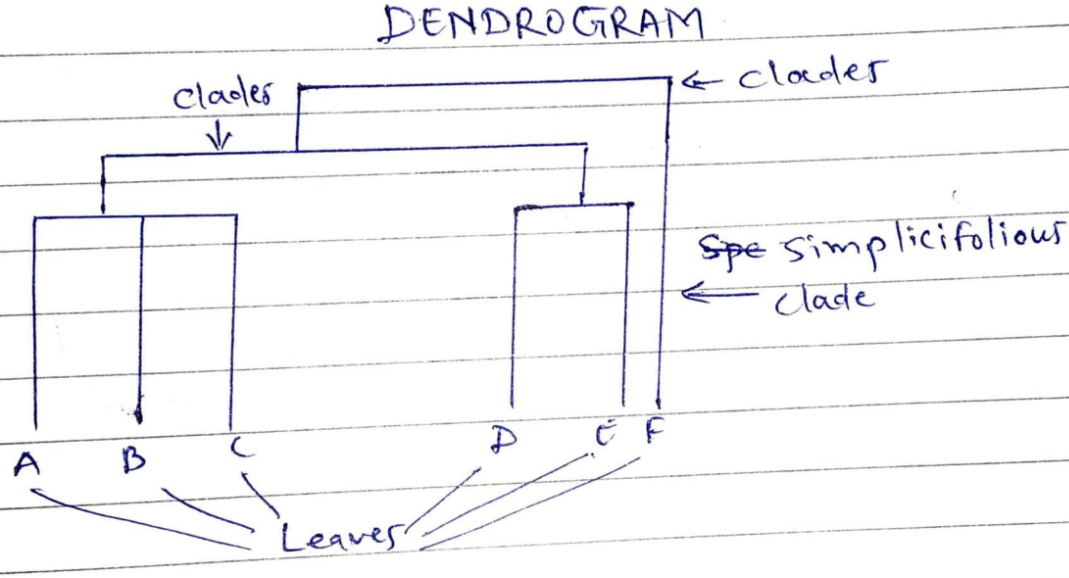
2. Divisive Hierarchical clustering.
In divisive or DIANA is top-down clustering method where we assign all of the observations to single cluster and then partition the cluster to two least similar cluster. Finally we proceed recursively on each cluster until there is one cluster for each observation So this clustering approch is exactly oppusite to Agglomerative clustering.

What is ~~dendo~~ Dendrogram?

1) A dendrogram is a type of tree diagram showing hierarchical relationships between sets of data

2) Dendrogram contains the memory of hierarchical clustering algorithm so just by looking at dendrogram you can tell how cluster are formed.

Parts of dendrogram



1) Clades :- clades are branch and are arranged according to how similar they are.
2) Each clade has one or more leaves

Conclusion:- Hence we understood and implemented kmeans and Hierarchical clustering techniques on iris dataset and visualized the clusters using suitable tools.