DATE:

Assignment No: DMW-2

* Title: clustering techniques

* Problem Statement:
Consider a suitable dataset for clustering of data instances in different groups and apply different clustering techniques. (minimum two)

* Objective: To,
1) learn various clustering techniques
2) plot clusters using suitable tools.

* outcomes: Students will be able to,
1) learn K-means and hierarchial clustering.
2) plot clusters using suitable tools.

* S/W and H/W requirements:
Linux, 64-bit PC, intel i5 processor, 500 GB HDD

* Concepts related to theory:
Clustering is the process of grouping a set of data object into multiple groups or clusters so that objects within clusters have high similarity but dismissal to objects in other clusters based on distance measures.
Techniques → Partitioning method
Hierarchical
Density based clustering
Grid based clustering

DATE:

Cluster analysis - Partitioning of set of objects into subsets.
Applications : Pattern recognization, biology, security, etc.

Partitioning Methods are -

1) K-means clustering : is a method of classifying grouping items into K groups (where K is the number of pre-chosen groups). The grouping is done by minimizing the distance-squared sum between items and corresponding centroids.

Algorithms-
1) Specify number of clusters K.
2) Initialize centroids by first shuffling the dataset and then randomly selecting c data points for the centroid without replacement.
3) Keep iterating until there is no change in centroid i.e. assignment of data points to cluster isn't changing.
4) Compute the sum of the squared distance between data points and all centroids.
5) Assign each data point to the closet cluster.
6) Compute the centroids for the clustering by taking the average of all the data points that belong to each cluster.

2) Hierarchical Clustering - involves creating clusters that have predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organised in a hierarchy. There are two type of hierarchical clustering - Divisive and Agglomerative clustering.

a) Agglomerative clustering - It is a bottom up clustering method. We assign each observation to its own clusters, then compute the similarity (eg: distance) between each clusters and join two most similar clusters. Finally, repeat step 2 and 3 until only a single cluster is left.

Algorithm -

Given : A set of objects $\{x, \ldots, x_n\}$

A distance function dist (c_1, c_2)

for $i = 1$ to n

$c_i = \{x_i\}$

end for

$c = \{c_1, \ldots, c_i, \ldots c_n\}$

while $c_{size} > 1$ do

→ $(c_{min1}, c_{min2}) = \text{mindis}(c_i, c_j)$:

$c_i, c_j \in c$

→ remove c_{min1} & c_{min2} from c

→ add $\{c_{min1}, c_{min2}\}$ to c

→ $L = L + 1$

end while

* Conclusion:

Thus, we have learned clustering and successfully implemented hierarchical and K-means clustering and displayed our observations using matplotlib using python.