LP2 - Assignment A2

Title:- Clustering techniques.

Problem statement:- Consider a suitable dataset for clustering of data instances in different groups visualize clusters using suitable tool.

Objective:- Understand the working of K-means & K-medoids clustering techniques.
Implement clustering models.

Outcomes:- Students will be able to learn and implement clustering algo. & apply them on a suitable dataset.

S/w & H/w :-
1) Python 3
2) 64 bit OS.

Theory:-

K-means:-

- It is a centroid based partitioning technique.
- It uses centroid of a cluster C_i to represent that cluster.
- K-means defines the centroid of cluster as the mean value of points within cluster.
- First it randomly selects k of the objects in D, each of which initially represents a cluster mean for each of the remaining obj.

- An object is assigned to the cluster to which it is most similar based on Euclidean distance between the object & cluster mean.
- The algorithm then iteratively improves within cluster variation.

Advantages:-
1) It is easy to understand & implement.
2) It can handle large datasets.

Disadvantages:-
1) It is sensitive to no. of clusters chosen.
2) Does not work with outliers.
3) Gets slower as no. of dimensions increases

K-medoids:-

- Instead of taking mean value of objects in a cluster as a reference point, we can pick actual object to represent the cluster using one representative objects per cluster.
- Each remaining object is assigned to the cluster of which the representative object is most similar.

- The partitioning method is then performed based on the principle of minimizing sum of dissimilarities between each object p & its corr. representative object.
- That is an absolute error criteria :-

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, O_i)$$

- The partitioning around methods is a polarization of k-medoids clustering.
- Tackles the problem in an iterative, greedy way.
- Like the k-means algo, the initial representative object are chosen arbitrarily.

Advantages :-
- More robust than k-means in the presence of noise & outliers because mediod is less influenced by outliers or other extreme values than a mean.

Disadvantages :-
- Complexity of each iteration in k-mediod algo. is $O(k(n-k)^2)$. For large value of n & k, computation very costly & much more costly than k-means.

Algorithms:-

k-means:-

1) Arbitrary choose k objects from D as the initial cluster centers.

2) Repeat:

3) (re) assign each object to the cluster is most similar, based on the mean value of the objects in the cluster.

4) Update the cluster means, that is calculate the mean value of the objects for each cluster.

5) Until no change.

k-medoids:-

1) Arbitrary choose k-objects in D as the initial representative objects or seeds.

2) Repeat.

3) Assign each remaining object to the cluster with the nearest representative obj

4) Randomly select a non-representative ob

5) Compute the cost S, of swapping rep. object O_j with O_{random}.

6) Repeat until no change.

Test case :-

We apply k-means on k-medoids clustering algorithms on Iris datasets with k=3 in each case & verify the number of items in each cluster.

Conclusion :-

We have successfully applied k-means & k-medoids clustering techniques & visualized the clusters.