ASSIGNMENT –2

MACHINE LEARNING

1. Movie Recommendation systems are an example of:

i) Classification

ii) Clustering

iii) Regression

Ans = ii) clustering, iii) Regression

2. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcemen

Ans= i) Regression ii) Classification iv) Reinforcemen

3. Can decision trees be used for performing clustering?

a) True b) False

Ans = a) True

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

i) Capping and flooring of variables

ii) Removal of outliers

Ans = i) Capping and flooring of variables

5. What is the minimum no. of variables/ features required to perform clustering?

a) 0

b) 1

c) 2

d) 3

Ans = b) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?

a) Yes

b) No

Ans = b) No

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

a) Yes

b) No

c) Can't say

d) None of these

Ans = a) Yes

8. Which of the following can act as possible termination conditions in K-Means?

i) For a fixed number of iterations.

ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum.

iii) Centroids do not change between successive iterations.

iv) Terminate when RSS falls below a threshold.

Ans = All of the above

9. Which of the following algorithms is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm

Ans= a) K- means clustering algorithm

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

i) Creating different models for different cluster groups.

ii) Creating an input feature for cluster ids as an ordinal variable.

iii) Creating an input feature for cluster centroids as a continuous variable.

iv) Creating an input feature for cluster size as a continuous variable.

Ans= All of the above

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

a) Proximity function used

b) of data points used

c) of variables used

d) All of the above

Ans= d) All of the above

12. Is K sensitive to outliers?

Ans= The *K*-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. *K*-medoids clustering is a variant of *K*-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, *K*-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center.

13. Why is K means better?

Ans= k-Means Advantages and Disadvantages

Advantages of k-means

Relatively simple to implement.

Scales to large data sets.

Guarantees convergence.

Can warm-start the positions of centroids.

Easily adapts to new examples.

Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Disadvantages of k-means

Choosing k **manually.**

Use the “Loss vs. Clusters” plot to find the optimal (k), as discussed in Interpret Results.

Being dependent on initial values.

For a low k, you can mitigate this dependence by running k-means several times with different initial values and picking the best result. Ask increases, you need advanced versions of k-means to pick better values of the initial centroids (called **k-means seeding**). For a full discussion of k- means seeding see, A Comparative study of Efficient Initialization Methods for the K-Means Clustering Algorithm by M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela.

Clustering data of varying sizes and density.

k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means as described in the Advantages section.

14. Is K means a deterministic algorithm?

Ans= K-Means is one of the most used algorithms for data clustering and the usual clustering method for benchmarking. Despite its wide application it is well-known that it suffers from a series of disadvantages; it is only able to find local minima and the positions of the initial clustering centres (centroids) can greatly affect the clustering solution. Over the years many K-Means variations and initialisation techniques have been proposed with different degrees of complexity. In this study we focus on common K-Means variations along with a range of deterministic and stochastic initialisation techniques. We show that, on average, more sophisticated initialisation techniques alleviate the need for complex clustering methods. Furthermore, deterministic methods perform better than stochastic methods. However, there is a trade-off: less sophisticated stochastic methods, executed multiple times, can result in better clustering. Factoring in execution time, deterministic methods can be competitive and result in a good clustering solution. These conclusions are obtained through extensive benchmarking using a range of synthetic model generators and real-world data sets.