MACHINE LEARNING

Q1 to Q11 have only one correct answer. Choose the correct option to answer your question.

1. Movie Recommendation systems are an example of: i) Classification ii) Clustering iii) Regression Options:

Ans = d) 2 and 3

2. Sentiment Analysis is an example of: i) Regression ii) Classification iii) Clustering iv) Reinforcement Options:

Ans = d) 1, 2 and 4

3. Can decision trees be used for performing clustering?

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 Options:

Ans = a) 1 only

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. Options:

Ans = d) 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) Kmodes 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. Options:

Ans = d) 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

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?

Ans = The K-Mean clustering algorithm is sensitive to outliers because a mean is easily influenced by extreme values. k-medoids clustering is a variant is a variant of k -means that is more robust to noises and outliers. Instead of using the mean point as the centre 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 different between mean and medoid in a 2-D example the group of point in the right form a cluster while the rightmost point is an outliers mean is greatly influenced by the outliers and thus cannot represent the correct cluster centre while medoids robust to the outliers and correctly represents the cluster center. The k-means clustering algorithm is sensitive to outliers because a mean is easily influenced by extreme values k-medoids clustering is a variant kmean that is more Robust to noises and outliers k-mean is a well stidied clustering problem that finds applications in many fields related to unsupervised learning it is knows that k-mean clustering is highly sensitive to the isolated point (called outliers) .

12. Why is K means better?

Ans = K mean algorithm is an iterative algorithm that tries to partition the dataset into defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group . It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

13. Is K means a deterministic algorithm?

Ans = the basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. In such cases, some supervision is needed to partition objects which have the same label into one cluster. This paper demonstrates how the popular k-means clustering algorithm can be profitably modified to be used as a classifier algorithm. clustering algorithm This blog post on the K-Means algorithm is part of the article series Understanding AI Algorithms. K-Means is a clustering algorithm. K-Means is an algorithm that segments data into clusters to study similarities. This includes information on customer behaviors which can be used for targeted marketing. K-Means Clustering Advantages and Disadvantages. K-Means Advantages : 1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. 2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.