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:
   a) 2 Only
   b) 1 and 2
   c) 1 and 3
   d) 2 and 3
   Ans.(a)2 Only

2. Sentiment Analysis is an example of:
   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement
   Options:
   a) 1 Only
   b) 1 and 2
   c) 1 and 3
   d) 1, 2 and 4
   Ans.(d) 1, 2 and 4

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
   Options:
   a) 1 only
   b) 2 only
   c) 1 and 2
   d) None of the above
   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

MACHINE LEARNING

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:
   a) 1, 3 and 4
   b) 1, 2 and 3
   c) 1, 2 and 4
   d) All of the above
   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) 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.
    Options:
    a) 1 only
    b) 2 only
    c) 3 and 4
    d) All of the above
    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-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. The k-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier.

MACHINE LEARNING

Since in k-means, you'll be taking the mean a lot, you wind up a lot of outlier-sensitive calculations. That's why we have the k-medians algorithm. It just uses the median rather than the mean and less is sensitive to outliers.

- 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.

13. Why is K means better?

Ans. k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets. K-means is the simplest. To implement and to run. All you need to do is choose "k" and run it a number of times. Most more clever algorithms (in particular the good ones) are much harder to implement efficiently and have much more parameters to set. Other clustering algorithms with better features tend to be more expensive.
Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

14. Is K means a deterministic algorithm?

Ans.k-means is a partitioning-based clustering algorithm. k-means method for clustering is an iterative process in which an initial partition of given k clusters is then improved by applying a search algorithm to the data. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.
Simplifying, given a pre-defined number (k) of clusters, the algorithm:

- begins with an initial set of k cluster centers (i.e. the centroids)
- re-assigns objects to the closest centroids
- recalculates centroids according to new memberships of the data points.
- repeats the last two steps until a consistent result is found or until the maximum number of iterations is reached.

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.
A deterministic algorithm is an algorithm which given a particular input, will always produce the same output, with the underlying machine always passing through the same sequence of states. Since K-mean clustering doesn't have this property. The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters. Initially k number of so-called centroids is chosen. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids

MACHINE LEARNING