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

Answer: b) 1 and 2

2. Sentiment Analysis is an example of:

Answer: a) 1 Only

3. Can decision trees be used for performing clustering?

Answer: 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

Answer: a) 1 only

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

Answer: b) 1

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

Answer: a) Yes

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

Answer: 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.

Answer: d) All of the above

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

Answer: 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.

Answer: b) 2 only

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

Answer: 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?

Answer: The answer to you question is yes. K-means can be used as outlier detection. BUT, more attention needs to be given for the definition of outliers. In K-means, using the symmetric distance measure is the key component to define the samples that belonging to the same cluster. symmetric distance measurement gives similar weight to each dimension (feature) this may not always be the case for defining outliers.

13. Why is K means better?

Answer: K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code-vectors (the centroids of different clusters). According to a tutorial, for any word/value/key that needs to be 'vector quantized', it is by calculating the distance from all the code vectors and assign the index of the code vector with the minimum distance to this value. For example, clustering can be applied to MP3 files, cellular phones are the general areas that use this technique**.**

14. Is K means a deterministic algorithm?

Answer: Clustering algorithm with steps involving randomness usually give different results on different executions for the same dataset. This non-deterministic nature of algorithms such as the K-Means clustering algorithm limits their applicability in areas such as cancer subtype prediction using gene expression data. It is hard to sensibly compare the results of such algorithms with those of other algorithms. The non-deterministic nature of K-Means is due to its random selection of data points as *initial centroids*.