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: d) 2 and 3

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

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: Yes, K-means can be used as outlier detection. 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?

ANS: K-means is like the Exchange Sort algorithm. Easy to understand, helps one get into the topic, but should never be used for anything real, ever. In the case of Exchange Sort, even Bubble Sort is better because it can stop early if the array is partially sorted. In the case of K-means, the EM algorithm is the same algorithm but assumes Gaussian distributions for clusters instead of the uniform distribution assumption of K-means. K-means is an edge case of E-M when all clusters have diagonal covariance matrices.

14. 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. 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. K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data.

THE END