MACHINE LEARNING

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

Correct Answer – (b) 1 and 2

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

Correct Answer – (b) 1, 2 and 4

3. Can decision trees be used for performing clustering?
   a) True
   b) False

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

Correct Answer – (a) 1 only

5. What is the minimum no. of variables/ features required to perform clustering?
   a) 0
   b) 1

c) 2
d) 3

Correct Answer – (b) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?
a) Yes
b) No

Correct Answer – (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

Correct 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 with a 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

Correct Answer – (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

Correct 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.
Options:
a) 1 only
b) 2 only
c) 3 and 4
d) All of the above

Correct Answer – (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

Correct Answer – (d) All of the above

12. Is K sensitive to outliers?

Answer – Yes, the k-means clustering algorithm is sensitive to outliers.

K-means is a centroid-based clustering algorithm where each cluster is represented by a centroid (the mean of the points in the cluster). The algorithm works by iteratively assigning each data point to the nearest centroid and then updating the centroid based on the points in the cluster.

However, if there are outliers (data points that are significantly different from the rest of the data), they can pull the centroid away from the center of the cluster, leading to suboptimal clustering results. This is because the distance metric used in k-means (usually Euclidean distance) is influenced by outliers and can distort the overall clustering structure.

Therefore, it is important to preprocess the data and remove any outliers before applying the k-means algorithm, or to use alternative clustering algorithms that are less sensitive to outliers.

13. Why is K means better?

Answer – K-means is a popular clustering algorithm that partitions data into K clusters based on their similarity. There are several reasons why K-means can be considered better compared to other clustering algorithms:

1. Simple and easy to implement: K-means is a simple and easy-to-implement algorithm, which makes it accessible to users with varying levels of technical

knowledge. The algorithm is fast and scalable, making it suitable for large datasets.

2. Interpretable: The algorithm is also interpretable, meaning that the results can be easily understood and visualized. K-means generates clusters that are mutually exclusive and collectively exhaustive, which simplifies interpretation.
3. Flexibility: K-means is a versatile algorithm that can be applied to a wide range of data types, including numeric, categorical, and binary data.
4. Efficiency: The algorithm is efficient in terms of computational resources required to cluster the data. It uses an iterative process to converge on the final solution, which helps to reduce the computational cost.
5. Performance: K-means has been shown to perform well on many datasets, particularly those with well-separated clusters. It is also robust to noise and outliers in the data.

14. Is K means a deterministic algorithm?

Answer – Yes, K-means is a deterministic algorithm. This means that if you run the algorithm multiple times on the same dataset with the same parameters, you will always get the same result. The algorithm follows a fixed set of rules and procedures to partition the data into K clusters based on their similarity, so the output is deterministic and predictable. However, the initial placement of the centroids can affect the final outcome, so it is common practice to run the algorithm multiple times with different initializations to ensure that the optimal solution is found.