Machine Learning (Assignment2)

Q1 to Q11 has 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

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

Answer: b) 1 and 2

3 Can decision trees be used for performing clustering?

a) True

b) False

Answer: b) False

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

Answer: b) 2 only

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

a) 0

b) 1

c) 2

d) 3

Answer: b) 1

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

a) Yes

b) No

Answer: b) No

7 Is it possible that the 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

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 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

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

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

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

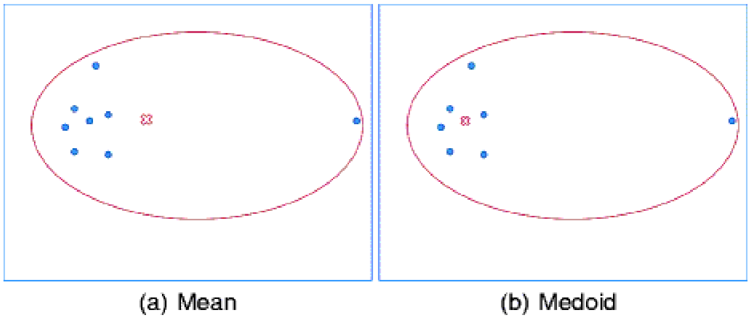
d) All of the above

Answer: d) All of the above

12) Is K sensitive to outliers?

Answer: Yes, K-means clustering is sensitive to outliers. An outlier is a data point that is significantly different from the other data points in the dataset. In K-means, the algorithm tries to minimize the sum of squared distances between the data points and their assigned cluster centers. If an outlier is present, it can increase the distance between the cluster center and the other data points, which can result in the creation of an inappropriate cluster. The presence of outliers can also affect the calculation of the mean, which is used to calculate the new cluster centers. As a result, the cluster centers can be shifted in the direction of the outlier, resulting in the creation of a suboptimal cluster.

- To address this issue, there are several techniques that can be used, such as removing the outliers from the dataset or using a clustering algorithm that is less sensitive to outliers, such as K-medoids or hierarchical clustering. Alternatively, the data points can be transformed to be less sensitive to outliers, such as using a log transformation.



13) Why is K-means better?

Answer: K-means is a popular clustering algorithm due to its simplicity, scalability, and efficiency. It is also a well-studied algorithm that has been used in a variety of applications, such as image segmentation, text clustering, and customer segmentation.

- One of the main advantages of K-means is its simplicity. The algorithm is easy to implement and understand, making it a good starting point for many clustering tasks. Additionally, K-means is computationally efficient and can handle large datasets with many features. This is because the algorithm only requires a few parameters, such as the number of clusters and the distance metric, which can be easily specified by the user.

- Another advantage of K-means is that it can handle continuous and categorical data. The algorithm can handle different types of data by using appropriate distance measures and preprocessing techniques. Additionally, K-means can take high-dimensional data by using feature selection or feature extraction techniques to reduce the dimensionality of the data.

- Finally, K-means can be easily extended and customized to suit different applications. For example, the algorithm can be modified to handle different types of constraints, such as the minimum or maximum size of the clusters. Additionally, K-means can be used in conjunction with other techniques, such as feature selection, to improve the accuracy of the clustering results.

14) Is K-means a deterministic algorithm?

Answer: Yes, K-means is a deterministic algorithm. This means that given the same input data and parameters, the algorithm will always produce the same output. The algorithm works by iteratively assigning data points to the nearest cluster center, calculating the new cluster centers based on the assigned data points, and repeating this process until convergence.

- However, there are situations where K-means can produce different results due to the presence of local minima. A local minimum occurs when the algorithm gets stuck in a suboptimal solution, rather than the global optimal solution. This can occur when the initial cluster centers are chosen poorly or when the dataset contains clusters of different sizes or shapes.
- To address this issue, K-means can be run multiple times with different initial cluster centers to find the optimal solution. Additionally, alternative versions of K-means, such as K-medoids or hierarchical clustering, can be used to reduce the impact of local minima.