Machine learning worksheet 2

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

Answere:- (B)

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

Answere:- (D)

3. Can decision trees be used for performing clustering?

a) True

b) False

Answere:- (B)

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

Answere:- (B)

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

a) 0

b) 1

c) 2

d) 3

Answere:- (B)

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

a) Yes

b) No

Answere:- (B)

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

Answer:- (A)

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

Answere:- (D)

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

Answere:- (A)

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

Answere:- (D)

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.

Answere:- (D)

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

1. Is K sensitive to outliers?

K-means is one of the most sensitive algorithms to outliers. K-means is a centroid-based algorithm which means that it finds the cluster center by taking the mean of the data points in a cluster. Outliers can have a large impact on the mean and can cause the cluster center to be shifted away from the majority of the data points. As a result, the cluster assignments of the data points may be affected, leading to suboptimal clusters. Additionally, K-means algorithm is sensitive to the scale of the variables.

2. Why is K means better?

K-means is a popular and widely used clustering algorithm for several reasons:

Simplicity: K-means is a simple and easy-to-understand algorithm that can be implemented in a relatively short amount of code.

Speed: K-means is computationally efficient and scales well to large datasets, making it suitable for handling large amounts of data.

Flexibility: K-means can be used to cluster data with any number of dimensions and can work with different types of data (e.g. continuous, categorical, etc.).

Good results with a large number of clusters: K-means is able to partition data into any number of clusters, which makes it suitable for data with a large number of clusters.

Generalization: K-means can generalize well to new data and can be used as a preprocessing step to improve the performance of other algorithms.

Robustness: K-means is relatively robust to initial conditions, meaning that it can produce consistent results even if the initial starting points for the clusters are not ideal

3. Is K means a deterministic algorithm?

K-means is a deterministic algorithm, meaning that it will always produce the same output for a given set of input data and parameters, if it is provided with the same initial centroids. However, it is sensitive to the initial starting points of the clusters, also known as centroids, and can produce different results based on the chosen initial centroids