MACHINE LEARNING ASSIGNMENT - 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

Answer 1. a) 2 Only

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 2. d) 1, 2 and 4

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

Answer 3. 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
Answer 4. a) 1 only

5. What is the minimum no. of variables/ features required to perform clustering?
a) 0
b) 1
c) 2
d) 3
Answer 5. b) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?
a) Yes
b) No
Answer 6. 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
Answer 7. 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
Answer 8. 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 9. 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 10. 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 11. 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 12. The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center.

13. Why is K means better?
Answer 13. Because it Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters. Relatively simple to implement.

14. Is K means a deterministic algorithm?
Answer 14. One of the significant drawbacks of K-Means is its non-deterministic nature. K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters. Besides, each run of the algorithm for the same dataset may yield a different output.