Assignment 2

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. 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

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 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

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

12. Is K sensitive to outliers?

Ans. K-means can be quite sensitive to outliers. So if you think you need to remove them, I would rather remove them first, or use an algorithm that is more robust to noise. For example k medians is more robust and very similar to k-means, or you use DBSCAN.

13.Why is K means better?

Ans. Advantages of k means:

a. Relatively simple to implement.

b. Scales to large data sets.

c. Guarantees convergence.

d. Can warm-start the positions of centroids.

e. Easily adapts to new examples.

f. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

14. Is K means a deterministic algorithm?

Ans. K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in polynomial time and doesn't clearly know the next step. This is because some problems have a great degree of randomness to them. These algorithms usually have 2 steps — 1) Guessing step 2) Assignment step. On similar lines is the K-means algorithm. The K-Means algorithm divides the data space into K clusters such that the total variance of all data points with respect to the cluster mean is minimized.