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

Ans. b) 1 and 2

2. Sentiment Analysis is an example of: i) Regression ii) Classification iii) Clustering iv) Reinforcement

Ans. c) 1 and 3

3. Can decision trees be used for performing clustering?

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:

Ans. a) 1 only

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

Ans. b) 1

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

Ans. b) No

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

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 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

Ans. d) All of the above

9. Which of the following algorithms is most sensitive to outliers?

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:

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

Ans. 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?

Ans. The k-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier.

An example, is the average of the salaries of the following people:

$50k, $20k, $35k, $65k and $1 Million

The average ends up being ($50k + $20k + $35k + $65k + $1MM) / 5 = $1170k / 5 = $234k.

If we did not have the $1MM outlier, the average would have been ($50k + $20k + $35k + $65k) / 4 = $170k / 4 = $42.5k.

Note that the two average results are wildly different from one another.

Given that k-means clustering is an unsupervised algorithm, it is up to the interpreter to determine whether this makes sense or not for a given data set.

13. Why is K means better?

Ans. K-means Clustering is a data segmentation technique. It divides the data into n parts/clusters where each cluster tries to have data points which are very close to each other. i. e. minimum variance in the cluster. That's why K means better .

14. Is K means a deterministic algorithm?

Ans. K means is not a deterministic algorithm . It is one of the most significant drawback of K means. K means starts with a random set of data points as initial centroids . This random selection influences the quality of the resulting cluster .