Worksheet - 2

Machine learning

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 D) 2 and 3

Question 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– D) 1, 2 and 4

Question 3 Can decision trees be used for performing clustering?

a) True b) False

Answer– A) True

Question 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 – A) 1 only

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

a) 0 b) 1 c) 2 d) 3

Answer– B) 1

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

a) Yes b) No

Answer– B) No

Question 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) Yes

Question 8 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 – D) None of these

Question 9 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) Yes

Question 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

Question 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

Question 12- Is K sensitive to outliers?

Answer K-Means can be quite sensitive to outliers. The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. 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 centre closer to the outlier.

Question 13- Why is K means better?

Answer - Advantages of k-means

- Relatively simple to implement

- Scales to large dataset
- Guarantees convergence
- Can warm up start the position of centroids
- Easily adapts to new examples
- Generalises to clusters of different shapes and sizes, such as elliptical clusters.

Question 14- Why is K means better?

Answer - 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.