

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

Answer: 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

Answer: d) 1, 2 and 4

3. Can decision trees be used for performing clustering?

- a) True
- b) False

Answer: 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: a) 1 only

ASSIGNMENT – 2 MACHINE LEARNING

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

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

Answer: b) 1

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

- a) Yes
- b) No

Answer: 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: 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

Answer: 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: a) K-means clustering algorithm

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

ASSIGNMENT – 2 MACHINE 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

- 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

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

12. Is K sensitive to outliers?

Answer:

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.

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.

13. Why is K means better?

Answer:

K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code-vectors (the centroids of different clusters)

Strengths of K-Means Clustering Algorithm

ASSIGNMENT – 2 MACHINE LEARNING

K-means is used to learn feature representations for images (use k-means to cluster small patches of pixels from natural images, then represent images in the basis cluster centres; repeat this several times to form a “deep” network of feature representations)

Lot of k-means applications are now done using support vector machines.

- 1) It gives good results
- 2) It is already implemented in the software
- 3) Number of clusters has to be fixed before
- 4) Dependent of the initialisation parameters and the chosen distance

14. Is K means a deterministic algorithm?

Answer:

The basic k-means clustering is based on a **non-deterministic algorithm**. This means that running the algorithm several times on the same data, could give different results.

This non-deterministic nature of algorithms such as the K-Means clustering algorithm limits their applicability in areas such as cancer subtype prediction. It is hard to sensibly compare the results of such algorithms with those of other algorithms. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.

K-Means

Clustering

Cancer subtype prediction

Centroid initialization