WORKSHEET-2 MACHINE LEARNING

Q1 to Q12 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 – 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 – 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

FLIP ROBO

Answer – 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 – 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):
   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 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 the above

FLIP ROBO

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?

Answer – An outlier is a point which is different from the rest of data points/set. The K-means algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-means updates the cluster centres by taking the average of all the data points that are closer to each cluster centres. 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 centres closer to the outlier.

For Example:

The mean of 2, 3, 2, 3, 4, 3, 4, 4, 50, 2, 2, 3, 5, 5, 3, 4, 2 is ***5.82***

If we add only one more number i.e. 500 to above numbers then, the mean becomes ***74.875***, which is much larger than any of the other values.

We can observe that the outlier can increase the mean of the data by almost 10 times. So we can clearly say that the mean is influenced by the outliers.

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. There are other clustering algorithms out there that are less sensitive to outliers. Depending on your application it may be worth using a different approach than the k-means algorithm.

13. Why is K-means better?

Answer – K-means clustering is an unsupervised algorithm which we can use to organise the large amounts of data to generate competitive insights about the business. There are many use cases which can help you implement this practice in our business and help us compete strategically in the market.

Other clustering algorithms with better features tend to be more expensive. In this case, k- means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. K-means is the simplest to implement and to run. All we need to do is choose "k" and run it a number of times.

More clever algorithms are much harder to implement efficiently and have much more parameters to set. Most people don't need quality clusters. They actually are happy with anything remotely working for them. Plus, they don't really know what to do when they had more complex clusters. K-means, which models clusters using the simplest model ever - a centroid - is exactly what they need: massive data reduction to centroids.

Some advantages of K-means clustering:

- Relatively simple to implement.
- Scales to large data sets.

- 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.

14. Is K means a deterministic algorithm?

Answer – A Deterministic algorithm is that in which output does not change when the algorithm is run several time whereas 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, hence we can say that K-means is not a deterministic algorithm. The non-deterministic nature of K-means is due to its random selection of data points as initial centroids.

FLIP ROBO