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**

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**

3. Can decision trees be used for performing clustering?
   a) **True**
   b) False

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

5. What is the minimum no. of variables/ features required to perform clustering?
   a) 0
   b) 1
   c) 2
   d) 3

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

MACHINE LEARNING

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
   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

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**

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**

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

12. Is K sensitive to outliers?
13. Why is K means better?
14. Is K means a deterministic algorithm?

12. Is K sensitive to outliers?

The K-Means algorithm is sensitive to the outliers. K-Means is a well-studied clustering problem that finds applications in many fields related to unsupervised learning. It is known that k-means clustering is highly sensitive to the isolated points. Such outliers can significantly influence the final cluster configuration and should be removed to obtain quality solutions.

Introduction:
Clustering is an important research branch of data mining. The k-means algorithm is one of the most popular clustering methods [1]. When performing k-means clustering, we usually use a local search to find the solution [2, 3], i.e., selecting k points as the initial cluster centers and then optimizing them by an iterative process to minimize the following objective function (see, for example, [4, 5]):

where is the j-th data point belonging to the i-th cluster . It is well known that the solution of equation is affected by the initial values of . In order to choose properly, the k-means++ algorithm picks out a set of points as the initial center points whose distances between each other are as large as possible. However, this method for choosing the initial center points is sensitive to outliers. Some methods use the subsets of the original data set to determine . For instance, the CLARA and CLARANS algorithms use PAM to calculate the initial cluster centers from the random subsets of the original data set. The sampling-based methods weaken the sensitivity because the sampling process can discard some outliers in the original data set, but it cannot guarantee all outliers to be ignored in the sampling process. Therefore, the remaining outliers in subsets still affect the clustering results.

The automatic clustering algorithms are attracting more and more attention from the academic community, e.g., the density-based spatial clustering of applications with noise (DBSCAN) algorithm, depth difference-based clustering algorithm, and Tanir’s method. Recently, a new automatic clustering algorithm named I-nice was proposed in. Inspired by the observation point mechanism of I-nice algorithm, we propose a two-stage k-means clustering algorithm in this paper to find the cluster centers from a subset of the original data set with all outliers removed. In the first stage, we select a small subset of original data set based on a set of nondegenerate observation points. The subset contains only all the higher density points of the original data set and does not have the outliers. Therefore, it is a good representation of the original data set for finding the proper cluster centers. In the second stage, we perform the k-means algorithm on the subset to obtain a set of cluster centers and then the other points in the original data set can be clustered accordingly.

Selecting the subset in the first stage is based on a set of d+1 nondegenerate observation points that are assigned to the data space Rd , where *d* is the dimension of data points. For each observation point, we compute a set of distances between it and all data points in the original data set. The set of distances generates a distance distribution with respect to the observation point. From the distance distribution, we identify the dense areas and extract the subset of data points in the dense areas. Then, we take the intersection of all d+1 subsets of data points in all dense areas from those d+1 distance distributions. After refining this intersection subset of data points, we obtain a subset without outliers of the original data set. Therefore, it can be used to find the proper cluster centers. Finally, we conduct some convictive experiments to validate the effectiveness of our proposed algorithm and the experimental results demonstrate that our proposed algorithm is robust to outliers.

13. Why is K means better?

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

What is K-Means Clustering:

You can use cluster analysis to group data points according to the similarities between them. This practice has a widespread application in business analytics and can help you to achieve your business goals. You can use the k-means algorithm to maximize the similarity of data points within clusters and minimize the similarity of points in different clusters.

It is an unsupervised algorithm that does not make use of labelled data or a training dataset. This type of algorithm is suitable for use when you have categorical data (e.g. grouping based on category, subcategory and brand).

As far as clustering algorithms go, it is simple and flexible to use in your retail business. With it you are also able to cluster large data sets in a short amount of time which is necessary when you work with retail data.

How to use a K-Means Clustering Algorithm:

1. Collect And Clean Your Data:

For a clustering Algorithm to be used, you will need to ensure that your data is in a standardised format. Each row acts as an observation and each column acts as a clustering variable or parameter. You will need to remove any missing data or add in an estimated value. Clustering variables must be standardised so that they can be compared and used to create groupings in the data set.

2. **Select the number of clusters you would like to use:**
When you use a k-means clustering algorithm, you will need to select the number of clusters you would like to work with. Selecting the optimal number of clusters is important because this will fall somewhere between full localization or standardization.

   Working with the optimal number of clusters for your retail data and market environment will facilitate the use of resources in a more efficient and effective manner. You can select the number of clusters using industry-related knowledge or three different statistical methods when you use the k-means algorithm.

   The Elbow method: To determine the optimal number of clusters, you will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, you will then need to calculate the total within-cluster sum of squares (wss). You can then plot the values of wss on the y-axis and the number of clusters (k) on the x-axis. The optimal number of clusters can be read off the graph at the x-axis.

   The Silhouette coefficient: To determine the optimal number of clusters, you will need to measure the quality of the clusters that were created. This value determines how closely each data point is to the centroid of its cluster. A high average silhouette coefficient indicates successful clusters. This method checks the silhouette coefficient for different values of k. The optimal number of clusters is, therefore, the maximized silhouette value for the data set.

   The Gap Statistic: To determine the optimal number of clusters, you will need to know the variation between clusters for different values of k with their expected values of distribution with no clusters.

3. **Run the clustering algorithm:**

The k-means algorithm identifies mean points called centroids in the data. It then assigns each data point to a centroid to form the initial clusters. The algorithm will measure the distances between each point and the centroids and assign each point where this distance is minimized.

You can retrieve the following information from a cluster analysis which can be used to profile, analyse and target each cluster effectively:

- The vector of integers (this identifies which cluster each point is allocated to);
- A matrix of cluster centres (the clusters can also be represented on an axis);
- The total sum of squares within each cluster;
- The inter-cluster sum of squares; and
- The number of data points within each cluster.

4. **Evaluate your k-means clusters:**
You can evaluate your clusters using inertia or the silhouette coefficient. Inertia measures how far apart the data points in a cluster are from each other. Inertia is measured from 0 upwards and a small inertia value indicates successful clustering.

15. Is K means a deterministic algorithm?

K-means Clustering is undoubtedly one of the most popular unsupervised learning algorithm. The reason behind it being used so frequently is the strong yet simple statistical backbone. This story would first explain the logical approach behind K-Means clustering, then it would bring forth a practical drawback and a few suggestions to avoid it.

K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in

MACHINE LEARNING

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.

However, the approach that compiler takes does not involve Multivariate Calculus as it seems. Rather, the approach taken is iterative. Now, like any deterministic algorithm it has 2 phases. Guessing phase: Randomly initializing k means in the data space(Mu(k)s). Now, all the data points X(i)s (1,m) are assigned to clusters in accordance to which cluster mean they are closer to. Mathematically, this step tries to minimize the within cluster variance. Hence, every point is now assigned a cluster. Next is the assignment step. All the cluster means (Mu(k)s)are now assigned to the mean of the data points in the cluster.

Now similar to the most of non-deterministic algorithms, K-Means has a bad habit. Which is that every time you run a K-Means clustering it would give you different results. The situation gets even worsened when you are unsure if the any modification to the K-Means would improve the results.

1. How to choose the value of K?

One should rely on the problem statement for this. For example in a tree specie classification problem, if one know the number of possible specie and given that all of them appear in significant numbers in the data-set, one can assign the number of species to K.

2. There is no versatile approach for this issue. Rather one should focus on initializing the cluster means with the best possible estimate. There could be any statistical approach for this.