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

ANS: 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

ANS: d) 1, 2 and 4

3.Can decision trees be used for performing clustering?

a) True
b) False
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:
a) 1 only
b) 2 only
c) 1 and 2
d) None of the above
ANS: a) 1 only

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

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

ANS: b) 1

MACHINE LEARNING

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

a) Yes
b) No

ANS: 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

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 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

ANS: 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

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:

a) 1 only
b) 2 only
c) 3 and 4
d) All of the above

ANS: d) All of the above

MACHINE LEARNING

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

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?

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. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center.

Because the mean, as a statistic, is generally sensitive to outliers.

The mean of 2 , 2 , 2 , 3 , 3 , 3, 4 , 4 , 4 is 3 .

If we add a single 23 to that, the mean becomes 5, which is larger than *any* of the other values.

Since in k-means, you'll be taking the mean a lot, you wind up with a lot of outlier-sensitive calculations.That's why we have the k-medians algorithm. It just uses the median rather than the mean and is less sensitive to outliers.

K-means is one of ten popular clustering algorithms. However, k-means performs poorly due to the presence of outliers in real datasets. Besides, a different distance metric makes a variation in data clustering accuracy. Improve the clustering accuracy of k-means is still an active topic among researchers of the data clustering community from outliers removal and distance metrics perspectives. Herein, a novel modification of the k-means algorithm is proposed based on Tukey’s rule in conjunction with a new distance metric. The standard Tukey rule is modified to remove the outliers adaptively by considering whether the data is distributed to the left, right or even to the input data's mean value. The elimination of outliers is applied in the proposed modification of the k-means before calculating the centroids to minimize the outliers' influences. Meanwhile, a new distance metric is proposed to assign each data point to the nearest cluster. In this research, the modified k-means significantly improves the clustering accuracy and centroids convergence. Moreover, the proposed distance metric's overall performance outperforms most of the literature distance metrics. This manuscript's presented work demonstrates the significance of the proposed technique to improve the overall clustering accuracy up to 80.57% on nine standard multivariate datasets.

13.Why is K means better?

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). According to a tutorial, for any word/value/key that needs to be ‘vector quantized’, it is by calculating the distance from all the code vectors and assign the index of the code vector with the minimum distance to this value. For example, clustering can be applied to MP3 files, cellular phones are the general areas that use this technique.

K-means has been around since the 1970s and fares better than other clustering algorithms like density-based, expectation-maximisation. It is one of the most robust methods, especially for image segmentation and image annotation projects. According to some users, K-means is very simple and easy to implement.

Unsupervised learning has emerged as the most effective technique for discovering patterns in data. K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

It is also being used to develop labels on top of the supervised models. This is one of the most widely used techniques for market or customer segmentation wherein the company’s data can be segregated into clusters and used to identify certain patterns which leads to a more customised approach. This technique comprises machine learning algorithms through which data analysts can draw inferences from datasets without labelled responses. Cluster analysis is also widely used for exploratory data analysis to find hidden patterns or grouping in data.

MACHINE LEARNING

Unsupervised Learning Algorithms Can Be Divided Into Two Wide Categories:

Clustering: A clustering problem is where one can find the inherent groupings in the data, such as grouping customers by purchasing behaviour.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y. Some of the common clustering algorithms are hierarchical clustering, Gaussian mixture models and K-means clustering. The last one is considered one of the simplest unsupervised learning algorithms, wherein data is split into k distinct clusters based on distance to the centroid of a cluster.

According to some users, K-means is very simple and easy to implement. However, it is unlikely to be the state-of-the-art, but for straightforward clustering, it is also a part of a larger data-processing pipeline, K-means is a reasonable default choice, at least until you figure out that the clustering step is your bottleneck in terms of overall performance.

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 of cluster centres; repeat this several times to form a “deep” network of feature representations) gives image classification results that are competitive with much more complex / intimidating deep neural network models. In fact, a lot of k-means applications are now done using support vector machines.

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

Weakness

- The results given are usually dependent on the initial values for the means.
- And the way to initialise the means is not specified, one can start by randomly choosing K of the samples.

The algorithm clusters into k groups and here k is the input parameter. In this procedure, a dataset is classified through a certain number of clusters, commonly known as k clusters and the main idea is to define k centres, one for each cluster.

These centers should be placed in a way since different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. However, the main disadvantage is one has to specify the number of clusters as an input in the algorithm.

14.Is K means a deterministic algorithm?

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. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.

A non-deterministic algorithm can provide different outputs for the same input on different executions. Unlike a deterministic algorithm which produces only a single output for the same input even on different runs, a non-deterministic algorithm travels in various routes to arrive at the different outcomes.

Non-deterministic algorithms are useful for finding approximate solutions, when an exact solution is difficult or expensive to derive using a deterministic algorithm.

The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

Method: We propose an improved, density based version of K-Means, which involves a novel and systematic method for selecting initial centroids.

K-Means has many drawbacks too. One of the significant drawbacks of K-Means is its non-deterministic nature. K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters.

MACHINE LEARNING

A deterministic signal gives only a single output for the same input even when it is on different runs. On the contrary, a non-deterministic algorithm gives different outputs, as it travels through different routes.

In a situation where finding an exact solution is expensive and difficult using deterministic algorithms, non-deterministic algorithms are used. They are helpful when it comes to finding approximate solutions.

A non-deterministic algorithm can run on a deterministic computer with multiple parallel processors, and usually takes two phases and output steps. The first phase is the guessing phase, and the second is the verifying phase. In the first phase, we make use of arbitrary characters to run the problem, and in verifying phase, it returns true or false values for the chosen string. An example that follows the concept of a non-deterministic algorithm is the problem of P vs NP in computing theory. They are used in solving problems that allow multiple outcomes. Every output returned is valid, irrespective of their difference in choices during the running process. Clustering has been widely applied in interpreting the underlying patterns in microarray gene expression profiles, and many clustering algorithms have been devised for the same. K-means is one of the popular algorithms for gene data clustering due to its simplicity and computational efficiency. But, K-means algorithm is highly sensitive to the choice of initial cluster centers. Thus, the algorithm easily gets trapped with local optimum if the initial centers are chosen randomly. This paper proposes a deterministic initialization algorithm for K-means (DK-means) by exploring a set of probable centers through a constrained bi-partitioning approach. The proposed algorithm is compared with classical K-means with random initialization and improved K-means variants such as K-means++ and MinMax algorithms. It is also compared with three deterministic initialization methods. Experimental analysis on gene expression datasets demonstrates that DK-means achieves improved results in terms of faster and stable convergence, and better cluster quality as compared to other algorithms.