Flip Robo

Machine Learning

Assignment – 1

Ans 1: Option b - > 4

Ans 2: Option d -> 1,2 & 4

Ans 3: Option a -> Interpreting and profiling clusters

Ans 4: Option a -> Euclidean distance

Ans 5: Ooption b -> Divisive Clustering

Ans 6: Option d -> All Answers are correct

And 7: Option a -> Divide Data points into groups

Ans 8: Option b -> Unsupervised learning

Ans 9: Option d -> All of the above

Ans 10: Option a -> K-means clustering algorithm

Ans 11: Option d-> All of the above

Ans 12: Option a-> Labelled Data

Q 13. How is cluster analysis calculated?

Ans :

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

In short the steps are :-

Step 1: Choose the number of clusters k.

Step 2: Select k random points from the data as centroids. ...

Step 3: Assign all the points to the closest cluster centroid. ...

Step 4: Recompute the centroids of newly formed clusters. ...

Step 5: Repeat steps 3 and 4.

Q 14. How is cluster quality measured?

Ans: To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

The silhouette analysis measures how well an observation is clustered and it estimates the **average distance between clusters**. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters.

For each observation i, the silhouette width si is calculated as follows

1. For each observation i, calculate the average dissimilarity ai, between i and all other points of the cluster to which i belongs.
2. For all other clusters C, to which i does not belong, calculate the average dissimilarity d(i,C) of i to all observations of C. The smallest of these d(i,C) is defined as bi=minCd(i,C). The value of bi can be seen as the dissimilarity between i and its “neighbour” cluster, i.e., the nearest one to which it does not belong.
3. Finally the silhouette width of the observation i is defined by the formula: Si=(bi−ai)/max(ai,bi).

Silhouette width can be interpreted as follow:

- Observations with a large Si (almost 1) are very well clustered.
- A small Si (around 0) means that the observation lies between two clusters.
- Observations with a negative Si are probably placed in the wrong cluster.

Q 15. What is cluster analysis and its types?

Ans: Cluster analysis in statistics is a method to organize data by clustering data points in a particular cluster. Rightly put, cluster analysis is a way of putting data points with similar characteristics in one group so that they differ from other data points of other clusters.

When it comes to Machine learning algorithms, clustering in machine learning is a prominent statistical technique that has evolved over time.

A set of clustering algorithms, empowered by the duo of cluster analysis and machine learning, have emerged on the surface that enables organizations and industries to run through their data and cluster data in an organized manner.

Clustering is one of the most renowned unsupervised machine learning algorithms that has been known to humankind.
Broadly, there are 2 types of cluster analysis methods. On the basis of the categorization of data sets into a particular cluster, cluster analysis can be divided into 2 types - hard and soft clustering. They are as follows -

1. Hard Clustering

   In a given dataset, it is possible for a data researcher to organize clusters in a manner that a single dataset is placed in only one of the total number of given clusters.
   This implies that a hard-core classification of datasets is required in order to organize and classify data accordingly.

   For instance, a clustering algorithm classifies data points in one cluster such that they have the maximum similarity. However, there are no other grounds of similarity with data sets belonging to other clusters.

2. Soft Clustering

   The second class of cluster analysis is Soft Clustering. Unlike hard clustering that requires a given data point to belong to only a cluster at a time, soft clustering follows a different rule.
   In the case of soft clustering, a given data point can belong to more than one cluster at a time. This means that a fuzzy classification of datasets characterizes soft clustering.
   Fuzzy Clustering Algorithm in Machine learning is a renowned unsupervised algorithm for processing data into soft clusters.