Internship assignments- **Assignment no-1**

1.Machine learning-

Q1) The correct answer is option-b

four clusters will be appropriate answers as the dendogram horizontal line will be maximum in this.

Q2) The correct answer is option-d

K-means doesnt works well when outliers are present, datasets are spreaded with high entropy or different density & data points with non-convex shape or banana like shape.

Q3) The correct answer is option- d

in formulating the cluster problem,the selection of variables is very important and crucial on which clustering is based.

Q4) The correct answer is option-d

Euclidean distance or its square is most commanly used measure of similarity.

Q5) The correct answer is option-b

Divisive clustering is a procedure where clusters are formed by dividing this cluster into smaller and smaller clusters.

Q6) The correct answer is option-d

K-means clustering requires distance metric between centroid and datapoints,Number of cluster to be made,and initially defining the centroid equals to no of clusters by guessing.

Q7) The correct answer is option-a

The goal of clustering is to divide the data points into groups.

Q8) The correct answer is option-b

Clustering is an unsupervised machine learning technique.which only have features no label is present in such techniques.

Q9)) The correct answer is option-d

All of the above clustering methods or algorithms suffers from the problem of convergence at local optima.

Q10) The correct answer is option-a

K-means clustering technique is most sensitive to outliers.

Q11) The correct answer is option-d

As quoted in question no-2

Q12) The correct answer is option-a

As clustering is unsupervised ML technique so we dont required label in this.

Q13) In K-means clustering method firstly we have to find out the optimum no of clusters that can be made in the given data set, which we can figure out with the help of Elbow method,which is a graphical representation bw WCSS(within cluster squared sum of distance) and no of clusters.

In this graph the point where the graph line is bending like an elbow is the best point from where we can draw a line to no of clusters and figure out the correct no of clusters can be made on the given data set.

Now we figure out the ompimum no of clusters to be made, now we will allocate the equal no of centroids to the data set and will connect these centroids and will divide half, with this we have parted a certain set of data.

Now the squared sum of distance will be calculated with the help of euclian distance bw centroids and data points parted in that particular segment.and the average of the squared sum of distance will be ploted as a point on the graph.

Now the centroid will move towards that particular average point and this step will keep continuing till there is no further movement in the centroid with respect to average distance point.

in this way we divide the data into the optimum no of clusters.

Q14) We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. Dissimilarity/Similarity metric: The similarity between the clusters can be expressed in terms of a distance function. Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance for different types of data.

2. Cluster completeness: Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.

3. Ragbag: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method.

4. Small cluster preservation: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation

criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters.

Q15) Cluster analysis is grouping of datasets or objects which have similar kind of characteristics based on user selected characteristics.most important technique in data mining clustering is.

Types of clustering-

1. K-Means Clustering: –

K-means clustering is a type of unsupervised learning used when we have unlabeled data. This algorithm aims to find groups in the data.

with the help of elbow method we figure out how many clusters required for the given data set.

The data point which is closest to the centroid of the cluster gets assigned to that cluster. After an iteration, it computes the centroids of those clusters again and the process continues until a pre-defined number of iterations are completed or when the centroids of the clusters do not change after an iteration.

It is a very computationally expensive algorithm as it computes the distance of every data point with the centroids of all the clusters at each iteration. This makes it difficult for implementing the same for huge data sets.

2.Hierarchical Clustering-

Hierarchical Clustering groups (Agglomerative or also called as Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics.

In agglomerative clustering, initially, each data point acts as a cluster, and then it groups the clusters one by one. This comes under in one of the most sought-after clustering methods.

Divisive is the opposite of Agglomerative, it starts off with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram. There are different types of linkages: –

- Single Linkage: – In single linkage the distance between the two clusters is the shortest distance between points in those two clusters.

- Complete Linkage: – In complete linkage, the distance between the two clusters is the farthest distance between points in those two clusters.

- Average Linkage: – In average linkage the distance between the two clusters is the average distance of every point in the cluster with every point in another cluster.

3.DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN groups data points together based on the distance metric. It follows the criterion for a minimum number of data points. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.It takes two parameters – eps and minimum points. Eps indicates how close the data points should be to be considered as neighbors. The criterion for minimum points should be completed to consider that region as a dense region.