MACHINE LEARNING

Q1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

b) 4

Q2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes

d) 1, 2 and 4

Q3. The most important part of is selecting the variables on which clustering is based

d) formulating the clustering problem

Q4. The most commonly used measure of similarity is the or its square.

a) Euclidean distance

Q5. ___ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

b) Divisive clustering

Q6. Which of the following is required by K-means clustering?

d) All answers are correct

Q7. The goal of clustering is to

a) Divide the data points into groups

Q8. Clustering is a

b) Unsupervised learning

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

d) All of the above

Q10. Which version of the clustering algorithm is most sensitive to outliers?

a) K-means clustering algorithm

Q11. Which of the following is a bad characteristic of a dataset for clustering analysis?

d) All of the above

Q12. For clustering, we do not require

a) Labeled Data

Q13. How is cluster analysis calculated?

- Centroid Clustering: In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres.
- Density Clustering: In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.
- Distribution Clustering**:** It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.
- Grid Clustering: It is performed on grid cells, used for a multi-dimensional data set.
- Hierarchical Clustering**:** In this method, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters. The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

Q14. How is cluster quality measured?

Measures for Quality of Clustering:
If all the data objects in the cluster are highly similar then the cluster has high quality. The measure of quality of Clustering by using:

1. Dissimilarity/Similarity metric: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). 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, Mahalanobis distance, and Cosine 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.

Example: Consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth.

Consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster.

Then, consider the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q (C2, Cg) > Q (C1, Cg).

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. According to the rag bag method, a heterogeneous object is put into a rag bag category.

Example: Consider a clustering C1 and a cluster C ∈ C1 so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth.

Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories.

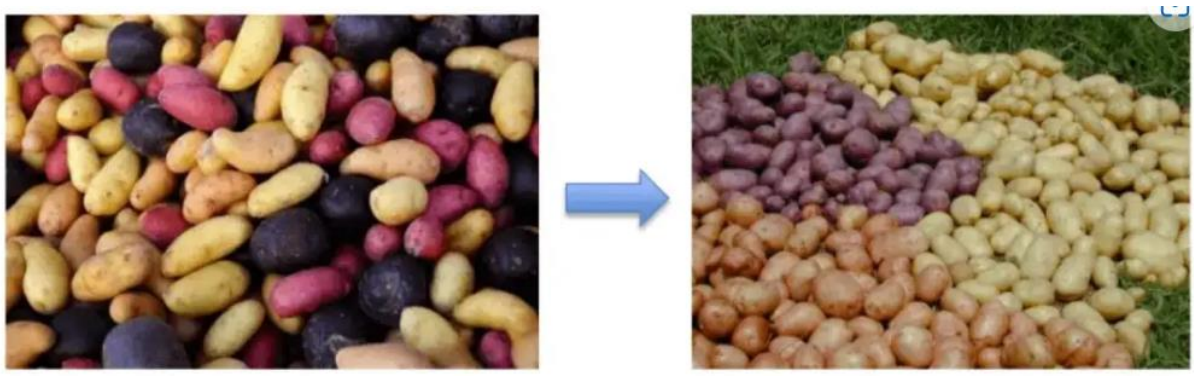
Consider, the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg)>Q(C1, Cg).

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 as the pieces of clusters are distinctive.

Suppose clustering C1 has split into three clusters, C11 = {d1, . . . , dn}, C12 = {dn+1}, and C13 = {dn+2}.
Let clustering C2 also split into three clusters, namely C1 = {d1, . . . , dn−1}, C2 = {dn}, and C3 = {dn+1,dn+2}. As C1 splits the small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q(C2, Cg) > Q(C1, Cg).

Q15. What is cluster analysis and its types?

Cluster analysis is a type of unsupervised machine learning technique, is a multivariate statistical technique that groups observations based on some of their features or variables. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).



sample Cluster/group

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

Types of Cluster Analysis

There are a number of approaches to cluster analysis, including:

- Centroid Clustering: In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres.
- Density Clustering: In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.
- Distribution Clustering**:** It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.
- Grid Clustering: It is performed on grid cells, used for a multi-dimensional data set.
- Hierarchical Clustering**:** In this method, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters. The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.