**MACHINE LEARNING**

**ASSIGNMENT – 1**

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

Ans : d) 2

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

Ans : 1. Data points with outliers

2. Data points with different densities

4. Data points with non-convex shapes

3. The most important part of is\_\_\_\_\_ selecting the variables on which clustering is based.

Ans : d) formulating the clustering problem

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

a) Euclidean distance

5. \_\_\_\_\_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.

Ans: b) Divisive clustering

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

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

Ans d) all ans are correct

7. The goal of clustering is to-

a) Divide the data points into groups

b) Classify the data point into different classes

c) Predict the output values of input data points

d) All of the above

Ans : d) All of the above

8. Clustering is a-

Ans : b) Unsupervised learning

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

Ans : Hierarchical Clustering

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

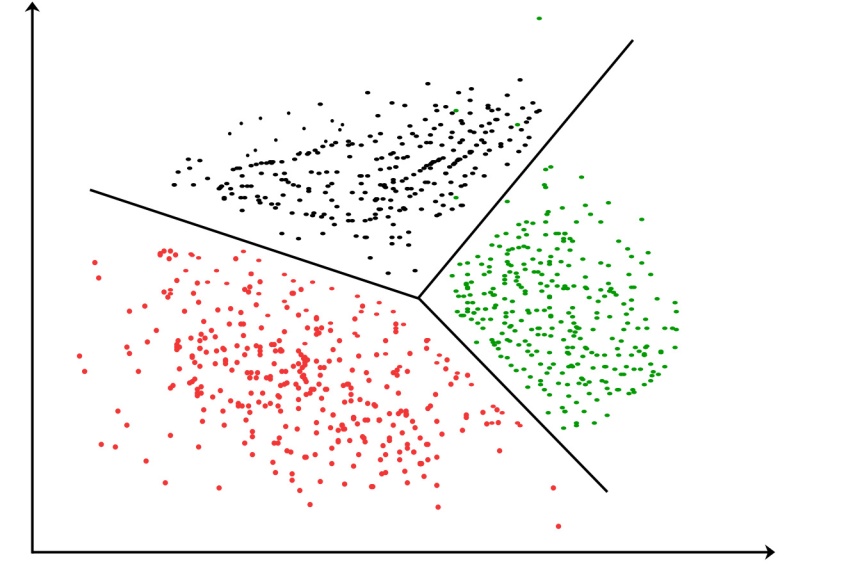
Ans : a) K-means clustering algorithm

12. For clustering, we do not require-

Ans : a) Labeled data

13. How is cluster analysis calculated?

Ans : A cluster is the collection of data objects which are similar to each other within the same group. The data objects of a  cluster are dissimilar to data objects of other groups or clusters.



Clustering Approaches:

1. **Partitioning approach**: The partitioning approach constructs various partitions and then evaluates them by some criterion, e.g., minimizing the sum of square errors. It adopts exclusive cluster separation(each object belongs to exactly one group) and uses iterative relocation techniques to improve the partitioning by moving objects from one group to another. It uses a greedy approach and approach at a local optimum. It finds clusters with spherical shapes in small to medium size databases.

Partitioning approach methods:

k-means

k-medoids

CLARINS

2. **Density-based approach**: This approach is based on connectivity and density functions. It divides the set of objects into multiple exclusive clusters or a hierarchy of clusters. Density-based methods:

DBSACN

OPTICS

3**. Grid-based approach**: This approach quantizes objects into a finite number of cells that form a grid structure. Fast processing time and independent of a number of data objects. Grid-based Clustering method is the efficient approach for spatial data mining problems.

Grid-based approach methods:

STING

WaveCluster

CLIQUE

4. **Hierarchical approach**: This creates a hierarchical decomposition of the data objects by using some measures. Hierarchical approach methods:

Diana

Agnes

BIRCH

CAMELEON

14. How is cluster quality measured

Ans : Measures for Quality of Clustering:

If all the data objects in the cluster are highly similar then the cluster has high quality. 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, 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.

Let us 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. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define 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, we should put the heterogeneous object into a rag bag category.

Let us 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. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define 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 )

**WORKSHEET 1 SQL**

1. Which of the following is/are DDL commands in SQL?

Ans : A) Create and D) ALTER

2. Which of the following is/are DML commands in SQL?

A) Update B) Delete

3. Full form of SQL is:

Ans : B) Structured Query Language

4. Full form of DDL is:

Ans : B) Data Definition Language

5. DML is:

Ans : A) Data Manipulation Language

6. Which of the following statements can be used to create a table with column B int type and C float type?

Ans : C) Create Table A (B int,C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

Ans : B) Alter Table A ADD COLUMN D float

8. Which of the following statements can be used to drop the column added in the above question?

Ans : B) Alter Table A Drop Column D

9. Which of the following statements can be used to change the data type (from float to int ) of the column D of table A created in above questions?

Ans : B) Alter Table A Alter Column D int

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

Ans : A) Alter Table A Add Constraint Primary Key B

11. What is data-warehouse

Ans : Data Warehousing integrates data and information collected from various sources into one comprehensive database.

12. What is the difference between OLTP VS OLAP?

OLTP is a transactional processing while OLAP is an analytical processing system. OLTP is a system that manages transaction-oriented applications on the internet for example, ATM. OLAP is an online system that reports to multidimensional analytical queries like financial reporting, forecasting, etc.

13. What are the various characteristics of data-warehouse?

Ans : **Subject-oriented**: A data warehouse typically provides information on a topic (such as a sales inventory or supply chain) rather than company operations.

**Time-variant**: Time variant keys (e.g., for the date, month, time) are typically present.

**Integrated**: A data warehouse combines data from various sources. These may include a cloud, relational databases, flat files, structured and semi-structured data, metadata, and master data. The sources are combined in a manner that’s consistent, relatable, and ideally certifiable, providing a business with confidence in the data’s quality.

**Persistent and non-volatile**: Prior data isn’t deleted when new data is added. Historical data is preserved for comparisons, trends, and analytics

14. What is Star-Schema

Ans : A star schema is a database organizational structure optimized for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data, and one or more smaller dimensional tables that store attributes about the data. It is called a star schema because the fact table sits at the center of the logical diagram, and the small dimensional tables branch off to form the points of the star.

15.What do you mean by SETL

Ans. :SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets.

SETL provides two basic aggregate data types: unordered sets, and sequences (the latter also called tuples). The elements of sets and tuples can be of any arbitrary type, including sets and tuples themselves. Maps are provided as sets of pairs (i.e., tuples of length 2) and can have arbitrary domain and range types. Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

**STATISTICS WORKSHEET-1**

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

10. What do you understand by the term Normal Distribution?

Ans : The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.

11. How do you handle missing data? What imputation techniques do you recommend?

Ans :