**Machine Learning**

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

Ans: 4

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

Ans: 1,2 and 4

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

Ans: formulating the clustering problem

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

Ans: 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: Divisive Clustering

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

Ans: All of the above

7. The goal of clustering is to-

Ans: All of the above

8. Clustering is a -

Ans: Unsupervised Learning

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

Ans: K-Means Clustering

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

Ans: K-Means Clustering

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

Ans: All of the above

12. For clustering, we do not require-

Ans: Labeled Data

13. How is clustering analysis calculated?

Ans: Initially the given number of clusters are created. Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated. If an object is detected whose distance to the centre of gravity of the own cluster is greater than the distance to the center of gravity (centroid) of another cluster, this object is shifted to the other cluster. Finally, the centroids of the two changed clusters are calculated again. These steps are repeated until each object is located in a cluster with the smallest distance to its centroid.

14. How is cluster quality measured?

Ans: Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance.

There are majorly two types of measures to assess the clustering performance.

(i) Extrinsic Measures which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) Intrinsic Measures that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

15. What is cluster analysis and its types?

Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

Types of Clustering :-

Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. The algorithm will start by randomly selecting centroids (cluster centers) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iteratations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters.

Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points, the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data. Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to a cluster.

Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.

**SQL**

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

Ans: Create and Alter

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

Ans: Update

3. Full form of SQL is:

Ans: Structured Query Language

4. Full form of DDL is:

Ans: Data Definition Language

5. DML is:

Ans: 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: 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 : 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: 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: 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: Alter Table A Add Primary key B

11. What is data-warehouse?

Ans: A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting. It is a blend of technologies and components which aids the strategic use of data. It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

12. What is the difference between OLTP VS OLAP?

Ans: OLTP and OLAP: The two terms look similar but refer to different kinds of systems. Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

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

Ans: Data warehouse can be controlled when the user has a shared way of explaining the trends that are introduced as specific subject. Below are major characteristics of data warehouse:

Subject-oriented –  
A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc. A data warehouse never put emphasis only current operations, instead it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

Integrated –  
Integration means founding a shared entity to scale all the similar data from the different databases. A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc.

Time-Variant –  
In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated.

Non-Volatile –  
As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment.

14. What is Star-Schema??

Ans: Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary case of the snowflake schema. It is also efficient for handling basic queries. It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points.

15. What do you mean by SETL?

Ans: Short for Set Theory as a Language (or Set Language), SETL is a [high-level programming language](https://www.webopedia.com/definitions/high-level-language/) that’s based on the mathematical theory of sets. It was developed in the early 1970’s by mathematician Professor J. Schwartz. SETL is an interpreted language with a [syntax](https://www.webopedia.com/definitions/syntax/) that is resembles [C](https://www.webopedia.com/definitions/c-language/) and in many cases similar to [Perl](https://www.webopedia.com/definitions/perl/). In SETL every statement is terminated by a semicolon. [Variable](https://www.webopedia.com/definitions/variable/) names are case-insensitive and are automatically determined by their last assignment.

**Statistics**

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

Ans: True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans: Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans: All of the mentioned

4. Point out the correct statement.

Ans: All of the mentioned

5. \_\_\_\_\_\_ random variables are used to model rates

Ans: Poisson

6.Usually replacing the standard error by its estimated value does change the CLT.

Ans: True

7. Which of the following testing is concerned with making decisions using data?

Ans: Hypothesis

8. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

Ans: 0

9. Which of the following statement is incorrect with respect to outliers?

Ans: Outliers cannot conform to the regression relationship

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

Ans: Normal distribution, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp). A normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

Below are some of the imputation techniques:-

Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

Pros:

Complete removal of data with missing values results in robust and highly accurate model Deleting a particular row or a column with no specific information is better, since it does not have a high weightage.

Cons:

Loss of information and data. Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset

Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

Pros:

This is a better approach when the data size is small . It can prevent data loss which results in removal of the rows and columns

Cons:

Imputing the approximations add variance and bias. Works poorly compared to other multiple-imputations method

Assigning An Unique Category

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. This strategy will add more information into the dataset which will result in the change of variance. Since they are categorical, we need to find one hot encoding to convert it to a numeric form for the algorithm to understand it.

I will prefer to use mean median or mode in order to handle the missing values.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training.

Pros:

This is a better approach when the data size is small. It can prevent data loss which results in removal of the rows and columns

Cons:

Imputing the approximations add variance and bias. Works poorly compared to other multiple-imputations method

14. What is linear regression in statistics?

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable

15. What are the various branches of statistics?

Descriptive Statistics

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

Inferential Statistics

It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.