MACHINE LEARNING:

1. B
2. D
3. D
4. A
5. B
6. D
7. A
8. B
9. D
10. A
11. D
12. B
13. How is cluster analysis calculated?

Only three steps are necessary:

* Copy your data into the table
* Select more than one variable
* Select the number of clusters you want to calculate

Clusters can be calculated using various grouping methods. These can be divided into

* graph-theoretical
* hierarchically
* partitioning
* optimizing

1. How is cluster quality measured?

To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

1. What is cluster analysis and its types?

* Centroid Clustering
* Density Clustering
* Distribution Clustering
* Connectivity Clustering

**SQL:**

* 1. A, d
  2. A, B, C
  3. B
  4. B
  5. A
  6. C
  7. B
  8. A
  9. C
  10. A
  11. What is data-warehouse?

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 analyse business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

* 1. What is the difference between OLTP vs OLAP?

OLTP and OLAP both are the online processing systems. 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.

* 1. What are the various characteristics of data-warehouse?

Data consolidation

Data Cleaning

Data Integration

* 1. What is Star-Schema??

A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

* 1. What do you mean by SETL?

A Programmable Semantic Extract-Transform-Load Framework for Semantic Data Warehouses

**STATISTICS:**

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. What do you understand by the term Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution 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.

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

One of the most common problems I have faced in Data Cleaning/Exploratory Analysis is handling the missing values. Firstly, understand that there is NO good way to deal with missing data. I have come across different solutions for data imputation depending on the kind of problem — Time series Analysis, ML, Regression etc. and it is difficult to provide a general solution.

Remove rows with missing values – This works well if 1) the values are missing randomly2) if you don’t lose too much of the dataset after doing so.

Build another predictive model to predict the missing values – This could be a whole project in itself, so simple techniques are usually used here.

Use a model that can incorporate missing data – Like a random forest, or any tree-based method.

1. 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.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

1. Is mean imputation of missing data acceptable practice?

It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms

1. 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. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

1. What are the various branches of statistics?

Descriptive Statistics and Inferential Statistics