***Machine Learning Questions***

**13. How is cluster analysis calculated?**

Cluster analysis means grouping a set of objects in such a way that objects in the same group are more similar (in some sense) to each other than to those in other groups. It is a main task of exploratory [data mining](https://en.wikipedia.org/wiki/Data_mining), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

**The methods used to get the optimal number of clusters:**

**The elbow method:** The sum of squares at each number of clusters is calculated and graphed a change of slope from steep to shallow is observed to get the optimal number of clusters.

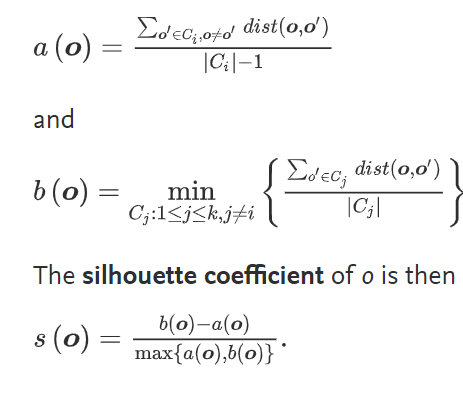
**The Gap Statistic**: The [gap statistic](http://www.web.stanford.edu/~hastie/Papers/gap.pdf) compares the total within intra-cluster variation for different values of **k** with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic.

**14. How is cluster quality measured?**

Clustering-quality measures may also be used to help in clustering model-selection by comparing different clusterings over the same data set.

Intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are.

The **silhouette coefficient** is such a measure. For a data set, D, of n objects, suppose D is partitioned into k clusters, C1, …, Ck. For each object **o** ∈ D, we calculate a (**o**) as the average distance between **o** and all other objects in the cluster to which **o** belongs. Similarly, b(**o**) is the minimum average distance from **o** to all clusters to which **o** does not belong. Formally, suppose **o** ∈Ci (1 ≤ i ≤ k); then



* The value of the silhouette coefficient is between −1 and 1
* The smaller the value of **a(o)**, the more compact the cluster.

**15. What is cluster analysis and its types?**

Cluster analysis means grouping a set of objects in such a way that objects in the same group are more similar (in some sense) to each other than to those in other groups.

## 1. Centroid Clustering

k-means is the most widely-used centroid-based clustering algorithm.

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.

### 2. Connectivity-based clustering (hierarchical clustering):

It is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance.

## 3. 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 .

## 4. Density Clustering

Density clustering groups data points by how densely populated they are.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**SQL\_Worksheet question**

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

**12. What is the difference between OLTP VS OLAP?**

**OLTP:(**Online transaction processing)

* It is characterized by a large number of short on-line transactions.
* The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments
* In OLTP database there is detailed and current data,

**OLAP(**Online analytical processing )

* It is characterized by relatively low volume of transactions.
* Queries are often very complex and involve aggregations.
* OLAP applications are widely used by Data Mining techniques.

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

Major characteristics of data warehouse are:-

1**.Subject oriented**: A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations. It focuses on demonstrating and analysis of data to make various decision.

2. **Integrated**:A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. This integration helps in effective analysis of data.

### 3.Time-Variant :The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly.

**4.** **Non-volatile**: Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it.Data is read-only and periodically refreshed. This also helps to analyze historical data and understand what & when happened

**14. What is Star-Schema?**

A star schema is composed of one or more central fact tables, a set of dimension tables, and the joins that relate the dimension tables to the fact tables.

* A **fact table** contains data columns for the numeric measurements of a business.
* **Dimension tables** store descriptions of the characteristics of a business.
* The **join constraints** in a star schema define the relationships between a fact table and its dimension tables.

**15. What do you mean by SETL?**

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

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Statistics worksheet question**

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

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean.it is a bell shaped symmetrical graph with an infinitely long base.Mean ,median & mode all are located at center

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

Missing data can be handled by 2 ways:

* Deleting Rows with **missing values**.

If there is very less null values present in data set,we can directly drop them.

* Impute **missing values** for continuous & categorical variable.
* Mean imputation for continuous data.
* Mode imputation for categorical data.

**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.

An A/B test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

**13. Is mean imputation of missing data acceptable practice?**

Mean imputation can be risky incase there are many null values as the mean may be biased

It distorts the covariance with the remaining dataset variables. The higher the percentage of missing values, the higher the distortions

**14. What is linear regression in statistics?**

Linear Regression is the simplest & most extensively used statistical technique for predictive modelling analysis.It is a way to expalin the relationship between a dependent & independent variable.

**15. What are the various branches of statistics?**

Main branch of statistics are :-

1.Descriptive statistics:

It refers to method for summarizing & organizing the information in a dataset. there are two general types of descriptive statistic that are used to describe data:

* **Measures of central tendency:**
* Spred of data

2.Inferential statistics**:**

This involves drawing right conclusion from the statistical analysis that has been performed using descriptive analysis.