**TECHNICAL ASSIGNMENTS**

**IMPORTANT DEFINITIONS**

1. **What is data normalization? How is it different from database normalization (1st/2nd/3rd)?**

The process of reducing data redundancy in a table and improving data integrity is called data normalization. It is a way of organizing data in a database. Normalization, also known as database normalization or data normalization, is an important part of relational database design because it helps to improve the speed, accuracy, and efficiency of the database.

* **1NF**: A relation is in first normal form if every attribute in that relation is singled valued attribute.

A table is in 1 NF if and only if:

* There are only Single Valued Attributes.
* Attribute Domain does not change.
* There is a unique name for every Attribute/Column.
* The order in which data is stored does not matter.
* **2NF**: Second Normal Form (2NF) is based on the concept of full functional dependency. Second Normal Form applies to relations with composite keys, that is, relations with a primary key composed of two or more attributes.

A relation is in 2NF if

* It has No Partial Dependency, i.e., no non-prime attribute is dependent on any proper subset of any candidate key of the table.
* **3NF**: A relation will be in 3NF if it is in 2NF and not contain any transitive partial dependency.3NF is used to reduce the data duplication. It is also used to achieve the data integrity.
* **BCNF**: BCNF is the advance version of 3NF. **It is stricter than 3NF**. A table is in BCNF if every functional dependency X → Y, X is the super key of the table.

1. **What is a distribution? What are the uses for frequency and probability distribution?**

A distribution of data represents the number of times each value occurs in a dataset. A statistical distribution can be a frequency distribution or a probability distribution. The data distribution helps in knowing the mean, median, mode and helps in knowing more about the data.

* **Frequency distribution**: A frequency distribution is a graph or dataset organized to show the frequency(number of times) of occurrence of each possible outcome of a repeated event observed many times. It is usually represented using a histogram.
* **A probability distribution:** The probability distribution gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences.

1. **What is a decision? How's it different from inference?**

* In computational theory, **a decision problem** is a computational problem that can be posed as a yes–no question of the input values. For example, deciding by means of an algorithm whether a given natural number is even.
* **Statistical inference** refers to the theory, methods, and practice of forming judgements about the parameters of a population and the reliability of statistical relationships, typically on the basis of random samples.

1. **Google- what is Gini in probability, and explain in your own terms**

As much I understand the Gini Coefficient is used to measure income inequality for a given population. It ranges between 0 and 1, where 0 indicates lowest inequality and 1 being the highest level of inequality.

1. **What is entropy?**

Entropy is a scientific concept as well as a measurable physical property that is mostly associated with a state of disorder, randomness, or uncertainty.

Information Entropy or Shannon's entropy quantifies the amount of uncertainty (or surprise) involved in the value of a random variable or the outcome of a random process. Its significance in the decision tree is that it allows us to estimate the impurity or heterogeneity of the target variable.

1. **What is Euclidean distance?**

In mathematics, Euclidean distance between two points in the Euclidean space is the length of a line segment between the two points.

1. **What's the difference between correlation and covariance?**

Covariance and correlation both primarily assess the relationship between variables. Correlation measures the association between two variables while covariance describes the extent to which one variable is related to another.

1. **What is mean squared error?**

**The mean squared error(MSE)** is the sum of the squared difference of each value of the dataset and the mean value of the variables in the dataset.

The MSE is used to see the accuracy of the model that is how well the model fits the data.

1. **What is the difference between covariance, standard deviation and mean squared error?**

Covariance measures the direction of the relationship between two variables. A positive covariance implies that both the variables have high or low values together. A negative covariance implies that if one variable has high value then the other has low values. This is given by:

Mean squared error measures the amount of error in a statistical model and is given as:

Standard Deviation is a measure of the dispersion of the actual values in the data from the central tendency. It conveys the spread of the data. It is given as the square root of variance.