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

It is a scaling technique method in which data points are shifted and rescaled so that they end up in a range of 0 to 1. It is also known as min-max.

1NF:

It is a level of normalization in DBMS. A relation is said to be in 1 normal form in DBMS (or 1NF) when it consists of an atomic value. In simpler words, 1NF states that a table's attribute would not be able to hold various values- it will only be able to hold an attribute of a single value.

2NF:

A relation is said to be in the 2nd Normal Form in DBMS (or 2NF) when it is in the First Normal Form but has no non-prime attribute functionally dependent on any candidate key's proper subset in a relation. A relation's non-prime attribute refers to that attribute that isn't a part of a relation's candidate key.

3NF:

It is a normalization level in DBMS. A relation is said to be in 3rd normal form in DBMS (or 3NF) when it is in the second normal form, but no transitive dependency exists for a non-prime attribute.

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

Distribution: A statistical distribution, or probability distribution, describes how values are distributed for a field. In other words, the statistical distribution shows which values are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution.

Frequency distribution is a representation of statistical data. It shows how often a possible outcome occurs in a sample group. In this article, we take a closer look at frequency distribution, including how to construct one from sample data, plus provide examples of the ways different sectors may apply this type of visualization.

A Probability distribution can be use in either of the following ways. By using a discrete distribution, you can define the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has only countable values, for example, a list of non-negative integers.

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

Decision theory, in statistics is a set of quantitative methods for reaching optimal decisions. A solvable decision problem must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their consequences.

Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypothesis and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

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

The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. It favours larger partitions. It quantifies the amount of inequality that exists in a population. The Gini coefficient is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality. Sometimes these statistics are reported in terms of percentages, with numbers between 0 and 100.

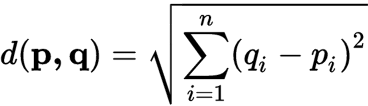
The Gini index is calculated as the ratio of the area between the perfect equality line and the Lorenz curve (A) divided by the total area under the perfect equality line (A + B).

1. **What is entropy?**

Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning. Further, in other words, we can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.

1. **What is Euclidean distance?**

 Euclidean distance is defined as **the distance between two points**. In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.

****

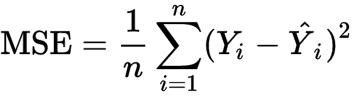
|  |  |  |
| --- | --- | --- |
|  | p, q = | two points in Euclidean n-space |
|  | qi, pi = | Euclidean vectors, starting from the origin of the space (initial point) |
|  | n= | n-space |

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

Both the quantities are used to measure the linear relation between the two variables. But the value of covariance lies between -∞ and ∞ which does not indicate the degree or the magnitude of the relation, covariance is given by the formula:

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

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.



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

variance is used in calculating standard deviation, which is a measure of how spread out a set of data is. Covariance is used in calculating correlation, which is a measure of how two variables relate to each other.  whereas the MSE measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated.