Technical session – AI &ML

Pavithra Jose

2211585

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

Normalization is the process of organizing data in a database. This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency. Data normalization consists of remodeling numeric columns to a standard scale. Data normalization is considered as the development of clean data.

Database normalization is the process of organizing the attributes of the database to reduce or eliminate data redundancy. Database normalization is the process of structuring a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.

First Normal Form(1NF) - A relation is in first normal form if every attribute in that relation is singled valued attribute. If a relation contains composite or multi-valued attribute, it violates first normal form or a relation is in first normal form if it does not contain any composite or multi-valued attribute.

Second Normal Form(2NF) - To be in second normal form, a relation must be in first normal form and relation must not contain any partial dependency. A relation is in 2NF if it has No Partial Dependency, i.e., no non-prime attribute dependent on any proper subset of any candidate key of the table. If the proper subset of candidate key determines non-prime attribute, it is called partial dependency.

Third Normal Form(3NF) - A relation is in third normal form, if there is no transitive dependency for non-prime attributes as well as it is in second normal form. A relation is in 3NF if at least one of the following condition holds in every non-trivial function dependency X –> Y:

- X is a super key.
- Y is a prime attribute

If A->B and B->C are two function dependencies, then A->C is called transitive dependency.

2) What is a distribution? What are the uses for frequency and probability distribution?

A distribution is a function that shows the possible values for a variable and how often they occur. Frequency distributions are charts which show the frequency with which data values in a certain situation occur. A common example is the histogram.

While a frequency distribution gives the exact frequency or the number of times a data point occurs, a probability distribution gives the probability of occurrence of the given data point. When the number of test cases are large, the frequency distribution and the probability distributions are similar in shape.

Frequency distribution present raw data in an organized, easy-to-read format. The most frequently occurring scores can be easily identified, as are score ranges, lower and upper limits, outliers, and total number of observations between any given scores.

3) What is a decision? How's it different from inference?

Statistical decisions are decisions made based on observations of a phenomenon that obeys probabilistic laws that are not completely known. Inference is arriving at a decision or opinion by reasoning facts from known evidence or facts. Statistical inference refers to the process of drawing conclusions from the model estimation. Decision analysis involves identifying and assessing all aspects of a decision, and taking actions based on the decision that produces the most favorable outcome.

4) What is Gini in probability, and explain in your own terms

Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. Gini coefficient is derived from the Lorenz curve. The Gini index is the ratio of the area below the ‘equality line’ (an area which is exactly 0.5) minus the area below the Lorenz curve to the area below the ‘equality line’.

The Gini coefficient is a number between 0 and 1. 0 represents perfect equality and 1 represents perfect inequality. It is typically used to quantify income inequality in human populations, and in that case a gini index of 0 means that everyone earns the same income, while 1 means that one person earns all the money.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

where p_i is the probability of an object being classified to a particular class. While building the decision tree, we prefer to choose the attribute with the least Gini index as the root node.

5) What is entropy?

Entropy measures the expected amount of information conveyed by identifying the outcome of a random trial. Calculating information and entropy is a useful tool in machine learning and is used as the basis for techniques such as feature selection, building decision trees, and, more generally, fitting classification models.

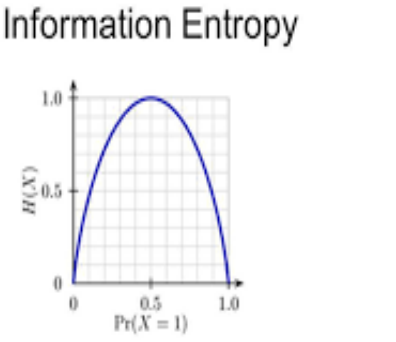
In information theory, the entropy of a random variable is the average level of "information ", "surprise", or "uncertainty" inherent in the variable's possible outcomes. That is, the more certain or the more deterministic an event is, the less information it will contain. In a nutshell, the information is an increase in uncertainty or entropy.

Claude E. Shannon had expressed this relationship between the probability and the heterogeneity or impurity in the mathematical form with the help of the following equation:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

The uncertainty or the impurity is represented as the log to base 2 of the probability of a category (pi). The index (i) refers to the number of possible categories. Here, i = 2 as our problem is a binary classification.

This equation is graphically depicted by a symmetric curve as shown below. x - axis is the probability of the event and the y-axis indicates the heterogeneity, or the impurity denoted by H(X).



- For a Bernoulli trial (X = {0,1}) the graph of entropy vs. Pr(X = 1). The highest H(X) = 1 = log(2)

6) What is euclidean distance?

Euclidean space is a two- or three-dimensional space to which the axioms and postulates of Euclidean geometry apply. Euclidean distance refers to the distance between two points in Euclidean space. In machine learning, it is commonly used to understand and measure how similar observations are to each other.

According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by:

$d((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$ where,

- (x, y) are the coordinates of one point.
- (a, b) are the coordinates of the other point.
- d is the distance between (x, y) and (a, b).

The source of this formula is the Pythagoras theorem.

7) What's the difference between correlation and covariance?

Covariance and correlation are two mathematical concepts used in statistics. Both terms are used to describe how two variables relate to each other.

Covariance can be positive, negative, or zero. A positive covariance means that the two variables tend to increase or decrease together. A negative covariance means that the two variables tend to move in opposite directions. A zero covariance means that the two variables are not related.

Correlation can only be between -1 and 1. A correlation of -1 means that the two variables are perfectly negatively correlated, which means that as one variable increases, the other decreases. A correlation of 1 means that the two variables are perfectly positively correlated, which means that as one variable increases, the other also increases. A correlation of 0 means that the two variables are not related.

<u>Covariance</u>

Covariance signifies the direction of the linear relationship between the two variables. The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Correlation

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

The main result of a correlation is called the correlation coefficient.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

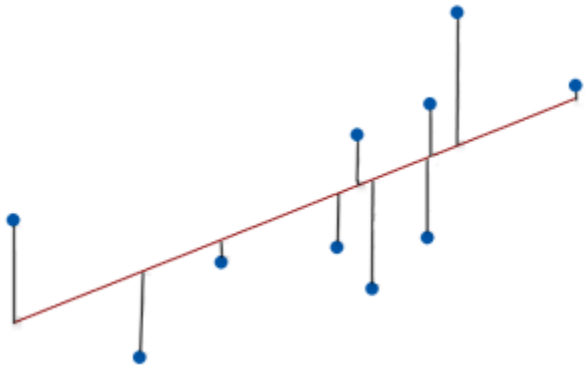
where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

8) What is mean squared error?

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

For example, in regression, the mean squared error represents the average squared residual.



This image depicts the relationship between the residuals and the mean squared error. As the data points fall closer to the regression line, the model has less error,

decreasing the MSE. A model with less error produces more precise predictions. The formula for MSE is given below:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where:

- yi is the ith observed value.
- ŷi is the corresponding predicted value.
- n = the number of observations.

The calculations for the mean squared error are similar to the variance. To find the MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all those squared values and divide by the number of observations.