MACHINE LEARNING

D) Both A and B.

A) Linear regression is sensitive to outliers.

B) Negative.

C) Both of them.

C) Low bias and high variance.

B) Predictive model.

D) Regularization.

D) SMOTE.

A) TPR and FPR.

B) False.

A) Construction bag of words from an email.

A) We don’t have to choose the learning rate and B) It becomes slow when the number of features is very large.

Regularization is a technique used in machine learning to prevent overfitting by adding a penalty term to the model's objective function. It helps in controlling the complexity of the model by discouraging large coefficient values.

Some algorithms used for regularization include Ridge Regression, Lasso Regression, and Elastic Net.

Error in the linear regression equation refers to the difference between the actual values and the predicted values generated by the linear regression model. The goal of linear regression is to minimize this error, typically measured using metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

PYTHON – WORKSHEET 1

C) %

B) 0

C) 24

D) 0

D) 6

C) the finally block will be executed no matter if the try block raises an error or not.

A) It is used to raise an exception.

C) in defining a generator

A) abc & C) abc2

A) yield & B) raise

For rest of the Programming Questions Jupyter file submitted.

STATISTICS WORKSHEET-1

a) True

a) Central Limit Theorem

b) Modeling bounded count data

c) The square of a standard normal random variable follows what is called chi-squared distribution

c) Poisson

b) False

b) Hypothesis

a) 0

c) Outliers cannot conform to the regression relationship

Subjective Questions:

What do you understand by the term Normal Distribution?

Ans: The Normal Distribution, also known as the Gaussian Distribution, is a symmetric probability distribution that is characterized by its bell-shaped curve. It is widely used in statistics and probability theory to model many real-world phenomena. In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution. The shape of the distribution is completely determined by its mean and standard deviation.

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

Ans: Handling missing data is crucial in data analysis. Some common techniques for handling missing data include:

Mean/Median imputation: Filling missing values with the mean or median of the available data.

Forward/Backward fill: Carrying the last observed value forward or the next observed value backward.

Interpolation: Estimating missing values based on the trend or pattern of the existing data points.

Multiple Imputation: Creating multiple plausible values for missing data and analyzing the results.

Deletion: Removing rows or columns with missing data (carefully, as this can lead to biased results).

What is A/B testing?

Ans: A/B testing, also known as split testing or bucket testing, is a method used to compare two versions of a webpage or application against each other to determine which one performs better. It involves dividing users into two or more groups and exposing each group to a different version of the content. Statistical analysis is then performed to determine if there is a significant difference in user behavior or outcomes between the groups. A/B testing is commonly used in marketing, product development, and user experience optimization.

Is mean imputation of missing data an acceptable practice?

Ans: Mean imputation is a simple technique for handling missing data, but it has limitations. While it can help preserve the sample size and maintain simple statistical properties, it may introduce bias and underestimate variability. Mean imputation assumes that the missing values are missing completely at random (MCAR) and that the missingness is unrelated to the underlying data. In cases where these assumptions are not met, mean imputation may lead to inaccurate results. More advanced imputation techniques, like multiple imputation or regression imputation, are often recommended.

What is linear regression in statistics?

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables or features). The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the variables. This equation allows us to make predictions or understand how changes in the independent variables affect the dependent variable. In simple linear regression, there is one independent variable, while in multiple linear regression, there are multiple independent variables.

What are the various branches of statistics?

Ans: Statistics is a broad field that encompasses various branches, including:

Descriptive Statistics: Involves summarizing and describing data using measures like mean, median, and standard deviation.

Inferential Statistics: Focuses on making inferences and predictions about populations based on sample data.

Probability Theory: Deals with uncertainty and randomness, providing the foundation for statistical analysis.

Biostatistics: Applies statistical methods to analyze and interpret biological and health-related data.

Econometrics: Applies statistical methods to economic data and models.

Social Statistics: Analyzes social phenomena and trends using statistical methods.

Machine Learning and Data Mining: Involves using statistical techniques to develop algorithms for predictive modeling and pattern recognition.

Bayesian Statistics: Focuses on updating beliefs and making decisions based on prior knowledge and new data.

Spatial Statistics: Analyzes data with spatial relationships, commonly used in geography and environmental studies.