Name: Syeda Saba Hussain Batch: DS2311

Internship#0523

Worksheet Set 1

STATISTICS

1. a) True
2. a) Central Limit Theorem
3. b) Modelling bounded count data
4. d) All of the above mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. Normal Distribution is a symmetric probability distribution characterized by its bell-shaped curve defined by its mean and standard deviation. It is, being the core of statistics and data analysis, is a fundamental concept since it encompasses properties and concentration of various characters in real-world phenomena.
11. Dealing with missing data demands attention to the pattern and mechanisms of missingness, imputing missing data with appropriate imputation methods, and verification that the missing has been handled in a correct manner. T
    The choice of imputation technique would depend upon the specific context, nature of data, amount of missingness, and also assumptions about the missing data mechanism. Whilst some of them are now more popular, one clear message from this analysis is that multiple imputation should be preferred when possible as being a more robust approach featuring the truth in accounts of uncertainty about the imputations made.
    Mean/Median/Mode Imputation: Replace the missing values in a data set using overall mean, median, or mode of the observed data (in case of continuous and categorical variables). This method does not change the dependent variable's actual relationship with other independent variables.
    Regression Imputation: This technique is implemented by deploying regression models to forecast missing values of any provided other variable in the dataset. It is potentially less biased than mean imputation but its limitation lies in the fact that the model's predictors are concepts that are correlated with the missing values.
    K-Nearest Neighbors (K-NN) imputation: The missing values are imputed based on the values of the k-nearest neighbors (records with most similar attribute) using distance metrics like Euclidean distance.
    Data-Driven Methods: Missing values can be deduced by machine learning algorithms, such as decision trees or random forests applied in light of the other variables present in the data set. These methods have a much better ability to capture complex relationships but are generally more sensitive to validation and computationally expensive.

12. A/B testing, also known as split testing, is a method of comparing two versions of a webpage, app, email, or other digital content against each other to determine which one performs better in terms of a specific metric or objective. The goal of A/B testing is to identify changes that increase a desired outcome, such as conversions, click-through rates, or user engagement.
13. Mean imputation is a method to handle missing data by replacing the missing values with the mean of the observed values for that variable.
    While mean imputation is a simple and accessible method to handle missing data, its limitations and potential biases make it a less preferred approach in many research and analytical contexts, especially when the missingness mechanism is non-random or related to other variables. It is essential to consider the underlying assumptions, implications, and potential impacts of mean imputation on the integrity and validity of the data and subsequent analyses. Whenever possible, more sophisticated imputation techniques that account for uncertainty, variability, and relationships between variables should be considered and applied to ensure robust and reliable results.
14. Linear regression is a statistical method used to model and analyze the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X_1, X_2 ,…, $X_{n)}$The primary goal of linear regression is to establish a linear relationship between the independent variables and the dependent variable to make predictions or infer associations based on observed data.
    The simple linear regression model can be represented as:
    $Y= \beta_0 +\beta_1 X+\epsilon$
    Where:
    Y is the dependent variable (response variable).
    X is the independent variable (predictor variable).
    β_0 is the intercept, representing the value of Y when X=0.
    β_1 is the slope coefficient, indicating the change in Y for a one-unit change in X.
    ϵ represents the error term, capturing the variability in Y that is not explained by X.
15. Statistics plays a foundational role, providing essential tools, techniques, and methodologies for analyzing, interpreting, and deriving insights from data.
    Here are some key branches of statistics are:
    a) Descriptive Statistics
    b) Inferential Statistics
    c) Probability Theory
    d) Machine Learning and Statistical Learning
    e) Time Series Analysis
    f) Spatial and Geospatial Statistics
    g) Multivariate Analysis
    h) Experimental Design and A/B Testing
    i) Statistical Computing and Data Visualization
    j) Bayesian Statistics and Probabilistic Modeling

MACHINE LEARNING

1. D) Both A and B
2. A) Linear regression is sensitive to outliers
3. B) Negative
4. C) Both of them
5. C) Low bias and high variance
6. B) Predictive model
7. D) Regularization
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. A) Construction bag of words from an email
12. A) We don’t have to choose the learning rate
    B) It becomes slow when number of features is very large.

13. Regularization is a technique used in machine learning and statistical modeling to prevent overfitting and improve the generalization performance of models. The primary goal of regularization is to add additional constraints or penalties to a model to ensure that it remains simpler and less likely to fit noise or fluctuations in the training data, thus enhancing its ability to perform well on unseen or test data.

14. Regularization is not limited to specific algorithms; instead, it is a concept that can be applied across various machine learning algorithms to prevent overfitting and improve model generalization. However, certain algorithms are commonly associated with specific regularization techniques. Here are some algorithms and their associated regularization methods:

Linear Regression:

Ridge Regression: Adds an L2 penalty term to the linear regression objective function to shrink the coefficients toward zero.

Lasso Regression: Incorporates an L1 penalty term, encouraging sparsity by driving some coefficients to exactly zero, thus performing feature selection.

Elastic Net: Combines L1 and L2 penalties to leverage the benefits of both regularization techniques.

Logistic Regression:

Similar to linear regression, logistic regression can also use L1 or L2 regularization to prevent overfitting, control model complexity, and improve classification performance on unseen data.

Neural Networks:

Dropout: A regularization technique specific to neural networks, where a fraction of neurons is randomly deactivated during training to prevent co-adaptation of neurons, improve convergence, and enhance generalization performance.

L1 and L2 Regularization: Neural networks can also incorporate L1 and L2 regularization techniques to control overfitting, reduce model complexity, and improve learning stability.

Support Vector Machines (SVM): It can utilize L1 or L2 regularization techniques, such as soft margin SVMs, to incorporate regularization, control the margin width, and handle non-linearly separable data using kernel methods like the radial basis function (RBF) kernel.

Tree-Based Algorithms (Random Forest, Gradient Boosting): It incorporate built-in mechanisms to prevent overfitting, techniques like pruning, limiting tree depth, and ensemble methods (bagging, boosting) can indirectly act as regularization techniques by constraining model complexity and improving generalization.

Regularized Regression Models: Algorithms like Ridge Regression, Lasso Regression, and Elastic Net are explicitly designed with built-in regularization mechanisms to prevent overfitting, control model complexity, and enhance predictive performance by incorporating penalty terms into the optimization process.

15. The term **"error"** refers to the difference between the observed values (actual values) and the predicted values generated by the linear regression model. These errors, also known as residuals, quantify the discrepancy between the model's predictions and the actual outcomes in the dataset.