Machine learning

1. A) least square Error

2. A) Linear regression is sensitive to outliers

3. B) Negative

4. B) Correlation

5. C) Low blas and high variance

6. B) predictive model

7. D) regularization

8. D) smote

9. A) TPR and FPR

10. B) false

11. B) Apply PCA to project high dimensional data

12. A) B)

13. Explain the term regularization?

Ans. Regularization is a technique used in machine learning to prevent overfitting by adding a penalty to the model's complexity. This helps the model generalize better to new data.

There are two main types of regularization:

1. **Lasso (L1) Regularization**: Adds the absolute values of the coefficients as a penalty term. It can shrink some coefficients to zero, effectively performing feature selection.
2. **Ridge (L2) Regularization**: Adds the squared values of the coefficients as a penalty term. It shrinks the coefficients but does not set them to zero.

14. Which particular algorithms are used for regularization?

Ans. Regularization techniques are commonly used in the following algorithms to prevent overfitting

1. Linear Regression:
   - Ridge Regression (L2 Regularization): Adds the squared magnitude of the coefficients as a penalty term.
   - Lasso Regression (L1 Regularization): Adds the absolute value of the coefficients as a penalty term.

2. Logistic Regression:

   - Both L1 and L2 regularization can be applied to logistic regression to improve generalization.

3. Support Vector Machines (SVM):

   - Regularization helps control the trade-off between maximizing the margin and minimizing classification error.

4. Neural Networks:

   - Dropout: Randomly drops units during training to prevent overfitting.

   - L2 Regularization: Adds a penalty term to the loss function based on the squared magnitude of the weights.

15. Explain the term error present in linear regression equation?

Ans. In a linear regression equation, the **error term** (also known as the residual) represents the difference between the observed values and the values predicted by the model. Mathematically, it is expressed as:

Error=y−y^

where (y) is the actual observed value and (\hat{y}) is the predicted value from the regression line.

PYTHON – WORKSHEET 1

1. C) %

2. B) 0

3. C) 24

4. A) 2

5. D) 6

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

7. A) it is used to raise an exception

8. C) in defining a generator

9. A) C)

10. A) B)

STATISTICS WORKSHEET-1

1. A) true

2. A) central limit theorem

3. B) modeling bounded count data

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

5. C) Poisson

6. B) false

7. B) hypothesis

8. A) 0

9. C) Outliers cannot conform to the regression relationship

10 . What do you understand by the term Normal Distribution?

Ans. The **normal distribution**, or Gaussian distribution, is a continuous probability distribution symmetric around its mean.

Key characteristics of a normal distribution include:

- **Symmetry**: The left and right sides of the curve are mirror images.
- **Mean, Median, and Mode**: All three measures of central tendency are equal and located at the center of the distribution.

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

Ans. Handling missing data is crucial for maintaining the integrity of your dataset. Here are some common imputation techniques:

1. Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the column. This is simple but can distort the data distribution.
2. Forward/Backward Fill: Use the previous or next value to fill in missing data. This is useful for time series data.
3. Interpolation: Estimate missing values based on other data points. Linear interpolation is a common method.
4. K-Nearest Neighbors (KNN) Imputation: Use the values from the nearest neighbors to impute missing data. This method considers the similarity between data points.
5. Multiple Imputation: Create multiple imputed datasets and combine the results. This accounts for the uncertainty in the imputation process.

6. Model-Based Imputation: Use machine learning models to predict and fill in missing values based on other features in the dataset.

12. What is A/B testing?

Ans. A/B testing, also known as split testing, is a method used to compare two versions of a variable to determine which one performs better. This involves randomly dividing a sample into two groups: one group sees version A (the control), and the other sees version B (the variation)

13. Is the imputation of missing data acceptable practice?

Ans. Mean imputation is a common technique for handling missing data, where missing values are replaced with the mean of the available data. While it is simple and easy to implement, it has some drawbacks:

- Distorts Variability: It reduces the variability in the data, which can lead to biased statistical estimates.
- Ignores Relationships: It does not consider the relationships between variables, potentially leading to inaccurate results.

14. What is linear regression in statistics?

Ans. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data The simplest form, simple linear regression, involves one independent variable and one dependent variable, and the relationship is represented by a straight line

15. . What are the various branches of statistics?

Ans. Statistics is broadly divided into two main branches:

1. **Descriptive Statistics**: This branch deals with the collection, organization, summarization, and presentation of data. It includes measures like mean, median, mode, and standard deviation, and uses tools like charts, graphs, and tables to describe the data
2. **Inferential Statistics:** This branch involves making predictions or inferences about a population based on a sample of data. It includes hypothesis testing, confidence intervals, and regression analysis, allowing statisticians to draw conclusions and make decisions based on data