MACHINE LEARNING WORKSHEET -1

1. A) Least Square Error

2. A) Linear regression is sensitive to outliers

3. B) Negative

4. B) Correlation

5. C) Low bias and high variance

6. B) Predictive modal

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) 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 to prevent overfitting and improve the generalization performance of a model. Overfitting occurs when a model fits the training data too closely, capturing noise or random fluctuations in the data rather than the underlying patterns. Regularization introduces a penalty term to the model's objective function, discouraging overly complex models with too many features or large parameter values.

There are different types of regularization techniques, and two common ones are L1 regularization (Lasso) and L2 regularization (Ridge):

1. L1 Regularization (Lasso):

- It adds a penalty term proportional to the absolute values of the coefficients.

- The regularization term is the sum of the absolute values of the coefficients multiplied by a regularization parameter (lambda or alpha).

- L1 regularization has a tendency to shrink some coefficients to exactly zero, effectively performing feature selection by eliminating irrelevant features.

2. L2 Regularization (Ridge):

- It adds a penalty term proportional to the square of the coefficients.

- The regularization term is the sum of the squared values of the coefficients multiplied by a regularization parameter (lambda or alpha).

- L2 regularization tends to shrink the coefficients but does not usually reduce them to exactly zero. It helps to prevent large coefficients and reduce the impact of individual features.

14. Regularization techniques can be applied to various machine learning algorithms to prevent overfitting. Here are some common algorithms where regularization is often employed:

1. Linear Regression:

   - Regularization can be applied to linear regression models to avoid overfitting.

2. Logistic Regression:

   - Regularization is commonly used in logistic regression to penalize large coefficients.

3. Support Vector Machines (SVM):

   - SVM algorithms can benefit from regularization, especially when dealing with high-dimensional data.

4. Neural Networks:

   - Regularization is frequently used in neural networks to prevent overfitting. Techniques like L1 and L2 regularization can be applied to the weights in the network.

5. Decision Trees:

   - While decision trees themselves may not explicitly use regularization, ensemble methods like Random Forests and Gradient Boosted Trees can benefit from regularization techniques applied to the individual trees.

6. Elastic Net:

   - Elastic Net is a linear regression model with both L1 and L2 regularization terms. It combines features of both Lasso and Ridge regression.

7. Lasso Regression (L1 regularization):

   - Specifically used when a sparse model is desired by driving some coefficients to exactly zero.

8. Ridge Regression (L2 regularization):

- Used when preventing large coefficients and multicollinearity is important.

9. Principal Component Regression (PCR):

- In PCR, which combines principal component analysis (PCA) with linear regression, regularization techniques can be applied to control overfitting.

15. In linear regression, the error term, also known as the residual, represents the difference between the actual observed value of the dependent variable (y) and the value predicted by the regression model. It captures the inherent variability and uncertainty that cannot be explained by the linear relationship between the independent variables (x) and the dependent variable.

Here's a breakdown of its key aspects:

1. Equation: The linear regression equation is typically written as: $y = \beta 0 + \beta 1x + \varepsilon$ where:
   - y: dependent variable
   - $\beta 0$: intercept
   - $\beta 1$: slope coefficient
   - x: independent variable
   - ε: error term
2. Sources of Error: The error term arises from various factors, including:
   - Unobserved or unmeasured variables: There might be important factors influencing the dependent variable that haven't been included in the model.
   - Inherent randomness: Some events or variations are inherently unpredictable or random, leading to deviations from the predicted values.
   - Measurement errors: Imperfections in data collection or measurement can introduce errors.
   - Model misspecification: The linear model might not be the most appropriate representation of the true relationship between the variables.
3. Assumptions about the Error Term: Linear regression makes key assumptions about the error term to ensure valid model inferences:
   - Zero mean: The average of the errors across all observations should be zero.
   - Constant variance (homoscedasticity): The variance of the errors should be constant across all levels of the independent variable.
   - Normal distribution: The errors are assumed to be normally distributed.
   - Independent errors: The errors for different observations should be independent of each other.

4. Importance of Residual Analysis: Examining the residuals (the actual errors) is crucial for:
   - Assessing model fit: Large residuals indicate that the model might not be capturing the underlying relationship well.
   - Identifying outliers: Outliers can have a disproportionate effect on the model and might need to be addressed.
   - Checking assumptions: Residual plots can help visualize if the error assumptions are met.
5. Minimizing Error: The goal of linear regression is to find the model that minimizes the overall error, typically using the least squares method. This involves finding the values of β_0 and β_1 that lead to the smallest sum of squared residuals.

Python Worksheet-1

Answer

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) _abc and c)abc2

10.b) raise