MACHINE LEARNING

1. Which of the following methods do we use to find the best fit line for data in Linear Regression? A) Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

Ans : A

2. Which of the following statement is true about outliers in linear regression? A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers C) Can't say D) none of these

Ans : A

3. A line falls from left to right if a slope is ______? A) Positive B) Negative C) Zero D) Undefined

Ans : B

4. Which of the following will have symmetric relation between dependent variable and independent variable? A) Regression B) Correlation C) Both of them D) None of these

Ans : B

5. Which of the following is the reason for over fitting condition? A) High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these

Ans : C

6. If output involves label then that model is called as: A) Descriptive model B) Predictive modal C) Reinforcement learning D) All of the above

Ans : B

7. Lasso and Ridge regression techniques belong to _________? A) Cross validation B) Removing outliers C) SMOTE D) Regularization

Ans : D

8. To overcome with imbalance dataset which technique can be used? A) Cross validation B) Regularization C) Kernel D) SMOTE

Ans : D

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph? A) TPR and FPR B) Sensitivity and precision C) Sensitivity and Specificity D) Recall and precision

Ans : A

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less. A) True B) False

Ans : B

11. Pick the feature extraction from below: A) Construction bag of words from a email B) Apply PCA to project high dimensional data C) Removing stop words D) Forward selection.

Ans : A

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression? A) We don’t have to choose the learning rate. B) It becomes slow when number of features is very large. C) We need to iterate. D) It does not make use of dependent variable.

Ans : D

13. Explain the term regularization?

Ans : Regularization is a technique used in machine learning and statistics to prevent overfitting and improve the generalization of models. It involves adding a penalty term to the objective function that the model optimizes during training. The goal of regularization is to discourage the model from fitting the training data too closely, which can lead to poor performance on new, unseen data.

Purpose of Regularization:

1. **Preventing Overfitting**: Overfitting occurs when a model learns not only the underlying patterns in the training data but also noise and random fluctuations. This results in a model that performs well on training data but poorly on test data. Regularization helps to mitigate overfitting by imposing constraints on the model's parameters.
2. **Improving Generalization**: By reducing overfitting, regularization helps the model generalize better to new, unseen data. It promotes simpler models that capture the underlying patterns in the data without memorizing noise.

14. Which particular algorithms are used for regularization?

Ans : Several algorithms in machine learning incorporate regularization techniques to prevent overfitting and improve model generalization. Here are some commonly used algorithms that incorporate regularization:

1. **Ridge Regression**:
2. **Lasso Regression**:
3. **Elastic Net Regression**:
4. **Logistic Regression (with Regularization)**:
5. **Support Vector Machines (SVMs)**:
6. **Neural Networks (with Regularization Techniques)**:

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

Ans : In the context of linear regression, the term "error" refers to the difference between the predicted values (based on the regression model) and the actual observed values in the dataset. These errors are also known as residuals.

Error in Linear Regression Equation:

The linear regression equation typically takes the form:

y=β0+β1x1+β2x2+...+βpxp+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilony=β0+β1x1+β2x2+...+βpxp+ϵ

- **yyy**: This represents the dependent variable or the target variable that we are trying to predict.
- **β0,β1,...,βp\beta_0, \beta_1, ..., \beta_pβ0,β1,...,βp**: These are the coefficients (also known as parameters) estimated by the regression model that multiply the predictor variables x1,x2,...,xpx_1, x_2, ..., x_px1,x2,...,xp.
- **x1,x2,...,xpx_1, x_2, ..., x_px1,x2,...,xp**: These are the predictor variables (also known as independent variables or features) that are used to predict yyy.
- **ϵ\epsilonϵ**: This represents the error or residual term, which captures the difference between the actual observed values of yyy and the values predicted by the regression model.

Characteristics of Error Terms:

1. **Mean Zero**: The error terms have a mean (average) of zero. This means that, on average, the errors are evenly distributed around the regression line.
2. **Independence**: The errors are assumed to be independent of each other. The occurrence of one error does not influence the occurrence of another.
3. **Constant Variance (Homoscedasticity)**: The variance (or spread) of the error terms should be constant across all levels of the predictor variables. When the variance of the errors is not constant, it is referred to as heteroscedasticity.

Importance of Error Terms:

- The error terms ϵ\epsilonϵ are crucial in assessing the performance and validity of the linear regression model. By examining the residuals (differences between observed and predicted values), we can evaluate how well the model fits the data.
- The goal of linear regression is to estimate the coefficients β0,β1,...,βp\beta_0, \beta_1, ..., \beta_pβ0,β1,...,βp in such a way that the sum of squared residuals (errors) is minimized. This process is known as ordinary least squares (OLS) regression.

Conclusion:

In summary, the term "error" in the context of linear regression represents the difference between the observed values and the values predicted by the regression model. Understanding and analyzing these errors are essential for evaluating the model's performance, making predictions, and assessing the validity of the assumptions underlying linear regression.