MACHINE LEARNING

In Q1 to Q11, only one option is correct, choose the correct option:

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

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

3. A line falls from left to right if a slope is _______?

A) Positive

B) Negative

C) Zero

D) Undefined

4. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression

C) Both of them

B) Correlation

D) None of these

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

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

7. Lasso and Ridge regression techniques belong to _________?

A) Cross validation

C) SMOTE

B) Removing outliers

D) Regularization

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation

B) Regularization

C) Kernel

D) SMOTE

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

C) Sensitivity and Specificity

B) Sensitivity and precision

D) Recall and precision

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

A) True

B) False

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

In Q12, more than one options are correct, choose all the correct options:

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.

MACHINE LEARNING

ASSIGNMENT – 39

Answer them briefly.

13. Explain the term regularization?

Answer: Regularization is a technique used in machine learning and statistics to prevent overfitting by adding a penalty to the model's complexity. It helps in improving the generalization of the model on unseen data by discouraging overly complex models that might fit the training data too closely.

How Regularization Works

Regularization works by modifying the loss function used to train a model. The standard loss function (such as Mean Squared Error for linear regression) is augmented with an additional term that penalizes large coefficients or complex models. The penalty term helps to constrain the magnitude of the coefficients, which in turn can lead to simpler models that are less likely to overfit.

Types of Regularization

1. **L1 Regularization (Lasso):**
   - **Effect:** Adds the absolute values of the coefficients to the loss function.
   - **Outcome:** Can shrink some coefficients to zero, effectively performing feature selection.
2. **L2 Regularization (Ridge):**
   - **Effect:** Adds the squared values of the coefficients to the loss function.
   - **Outcome:** Reduces the size of the coefficients but does not set them to zero.
3. **Elastic Net Regularization:**
   - **Effect:** Combines both L1 and L2 regularization penalties.
   - **Outcome:** Provides a balance between feature selection (L1) and coefficient shrinkage (L2).

Benefits of Regularization

- **Prevents Overfitting:** By penalizing large coefficients, regularization discourages the model from fitting noise in the training data.
- **Improves Generalization:** Helps the model to perform better on unseen data by simplifying the model.
- **Feature Selection:** L1 regularization (Lasso) can help in identifying and selecting relevant features by shrinking some coefficients to zero.

When applying regularization, the penalty term is controlled by a hyperparameter, which determines the strength of the regularization. The optimal value for this hyperparameter is typically found through techniques such as cross-validation. Regularization is a crucial tool for building robust models that generalize well to new data.

14. Which particular algorithms are used for regularization?

Answer: Algorithms incorporate regularization to improve model performance and prevent overfitting. Here are some of the most commonly used algorithms that involve regularization:

1. Linear Regression with Regularization

- **Lasso Regression (L1 Regularization):** Adds a penalty equal to the absolute value of the magnitude of coefficients. It can shrink some coefficients to zero, performing feature selection.
- **Ridge Regression (L2 Regularization):** Adds a penalty equal to the square of the magnitude of coefficients. It helps to reduce the size of coefficients but does not eliminate any features.
- **Elastic Net Regression:** Combines both L1 and L2 penalties. It balances between feature selection and coefficient shrinkage.

2. Logistic Regression with Regularization

- Similar to linear regression, logistic regression can use L1 (Lasso), L2 (Ridge), or a combination of both (Elastic Net) regularization to prevent overfitting and improve generalization.

3. Support Vector Machines (SVM)

- **SVM with L2 Regularization:** SVMs can be regularized using an L2 penalty on the weights, which helps in controlling the margin and avoiding overfitting.

4. Generalized Linear Models (GLM)

- **Regularized GLMs:** Generalized Linear Models can include L1 or L2 regularization to handle various types of distributions and link functions while controlling model complexity.

5. Neural Networks

- **Dropout:** A form of regularization used in neural networks where randomly selected neurons are ignored during training. This helps prevent overfitting by ensuring that the network does not rely too heavily on any particular set of neurons.
- **Weight Regularization:** L1 and L2 regularization can be applied to the weights of neural networks to prevent overfitting.

6. Decision Trees and Ensemble Methods

- **Pruning:** Regularization in decision trees can be achieved by pruning, which involves removing parts of the tree that do not provide significant power in predicting target variables.
- **Random Forests and Gradient Boosting Machines:** Regularization can be applied in these ensemble methods through techniques such as limiting the depth of trees, controlling the number of trees, and applying shrinkage (in the case of gradient boosting).

7. Regularized Principal Component Analysis (PCA)

- **Sparse PCA:** Applies L1 regularization to the principal components to encourage sparsity, which can help in feature selection and improve interpretability.

8. Regularized Logistic Regression in Classification

- **Ridge Logistic Regression:** Uses L2 regularization to control the size of coefficients in logistic regression.
- **Lasso Logistic Regression:** Uses L1 regularization to promote sparsity and feature selection in logistic regression.

Regularization is a versatile concept applied across various algorithms to enhance model performance, especially in the context of high-dimensional data and to ensure models generalize well to unseen data.

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

Answer: In linear regression, the term "error" typically refers to the difference between the observed values and the values predicted by the model. It is crucial for understanding how well the linear regression model fits the data and for assessing its performance.

Here’s a detailed explanation of the term "error" in linear regression:

1. Definition of Error

- **Error (or Residual):** The error for a given data point is defined as the difference between the observed value and the predicted for that data point.

2. Role of Error in Linear Regression

- **Model Fitting:** The goal of linear regression is to minimize the errors across all data points by adjusting the model parameters (slope and intercept) so that the predicted values are as close as possible to the actual values.

- **Error Terms in the Linear Regression Equation:** In a linear regression model, the relationship between the independent variable x and the dependent variable y is expressed as:

3. Error Metrics

To evaluate the performance of a linear regression model, several metrics are used that involve errors:

- **Mean Absolute Error (MAE):** The average of the absolute errors across all observations.
- **Mean Squared Error (MSE):** The average of the squared errors, which penalizes larger errors more heavily.
- **Root Mean Squared Error (RMSE):** The square root of the MSE, providing error in the same units as the dependent variable.
- **Residual Sum of Squares (RSS):** The sum of the squared errors, used in the context of the least squares method.

4. Importance of Error in Model Evaluation

- **Goodness of Fit:** Errors help in assessing how well the model fits the data. Lower errors indicate a better fit.
- **Diagnostic Tool:** Analysing the pattern of errors can reveal if the model assumptions are violated or if there are outliers in the data.
- **Model Improvement:** Regularly monitoring errors helps in refining the model by adjusting parameters or choosing different features.

In summary, the error in a linear regression model is a measure of how well the model's predictions match the actual values, and it plays a critical role in evaluating and improving the model's performance.