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

Least Square Error: This method involves minimizing the sum of squared differences between the predicted values by the linear regression model and the actual observed values in the dataset. The goal is to find the line that minimizes the overall distance between the predicted and observed values.

Maximum Likelihood: This method involves finding the line that maximizes the likelihood of observing the given data under the assumption that the data follows a linear regression model. It uses statistical techniques to estimate the parameters of the model that maximize the likelihood function.

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

Outliers can have a significant impact on linear regression models. Linear regression works by minimizing the sum of the squared residuals between the predicted values and the actual values. Outliers, which are extreme values that deviate significantly from the majority of the data, can pull the regression line towards them, resulting in a poor fit for the rest of the data.

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
B) Correlation
C) Both of them
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

The reason for overfitting is when a model has low bias and high variance. Bias refers to the error introduced by approximating a real-world problem with a simplified model. A low bias means that the model is very flexible and can capture complex patterns and relationships in the data. However, high variance means that the model is highly sensitive to variations in the training data and can produce very different results for different training sets.

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

f the output of a model involves labels, it is typically referred to as a predictive model. Predictive models are designed to predict or classify a label or outcome based on input features or variables. These models use historical data to learn patterns and relationships and make predictions or classifications on new, unseen data.

7. Lasso and Ridge regression techniques belong to_________?

A) Cross validation
B) Removing outliers
C) SMOTE
D) Regularization

D) Regularization

Lasso and Ridge regression techniques belong to the category of regularization methods. Regularization is a technique used to prevent overfitting in regression models by adding a penalty term to the objective function.

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

A) Cross validation B) Regularization

C) Kernel D) SMOTE

SMOTE addresses this issue by creating synthetic samples for the minority class by interpolating between existing samples. It randomly selects a sample from the minority class and identifies its nearest neighbors. Then, it generates synthetic samples along the line segments connecting the selected sample and its neighbors

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

The AUC-ROC (Area Under the Receiver Operating Characteristic) curve is a widely used evaluation metric for binary classification problems. It is used to assess the performance of a classifier by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds.

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

A) True B) False

In the AUC-ROC curve, a higher area under the curve (AUC) is generally indicative of a better model performance. The AUC ranges from 0 to 1, where a value of 0.5 represents a random classifier, and a value of 1 represents a perfect classifier.

A higher AUC value implies that the model has better discrimination ability in distinguishing between the positive and negative classes.

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

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

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?\

Regularization is a method for “constraining” or “regularizing” the size of the coefficients, thus “shrinking” them towards zero. It reduces model variance and thus minimizes overfitting.

14. Which particular algorithms are used for regularization?

L1 Regularization (Lasso): In L1 regularization, a penalty is added to the model's objective function proportional to the absolute values of the model's coefficients. This encourages sparsity in the parameter values, effectively driving some of them to zero.

L2 Regularization (Ridge): L2 regularization adds a penalty term proportional to the squared values of the model's coefficients. This penalty term encourages smaller and more evenly distributed parameter values, reducing the impact of individual features.

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

In linear regression, the term "error" refers to the difference between the actual observed values and the predicted values generated by the linear regression model. It represents the discrepancy or deviation between the predicted values and the true values of the dependent variable.

The goal of linear regression is to minimize the error, also known as the residual or the residual error. This error is typically represented as the vertical distance between the observed data points and the regression line.