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 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

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 B) Removing outliers
   C) SMOTE 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 B) Sensitivity and precision
   C) Sensitivity and Specificity D) Recall and precision

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

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

13. Explain the term regularization?

Ans:- Regularization is a technique used in machine learning to prevent overfitting and improve the generalization ability of a model. The goal of regularization is to add a penalty term to the cost function, discouraging the model from becoming too complex or fitting the training data too closely. This is particularly useful when dealing with models with a large number of features or parameters.

There are two common types of regularization: L1 regularization (Lasso) and L2 regularization (Ridge).

1. L1 Regularization (Lasso): In L1 regularization, the penalty term is the absolute value of the coefficients. It adds the sum of the absolute values of the coefficients to the cost function. L1 regularization is useful for feature selection because it tends to force some of the coefficients to be exactly zero, effectively eliminating certain features from the model.

2. L2 Regularization (Ridge): In L2 regularization, the penalty term is the square of the coefficients. It adds the sum of the squared values of the coefficients to the cost function. L2 regularization is effective in preventing large weights and, thus, reduces the impact of any single feature on the model.

The regularization term is typically controlled by a hyperparameter (lambda or alpha), allowing the user to balance the trade-off between fitting the training data well and keeping the model simple. Regularization is a powerful tool for improving the robustness of machine learning models, especially in situations where the amount of training data is limited or when dealing with high-dimensional datasets.

14. Which particular algorithms are used for regularization?

Ans:- Regularization techniques can be applied to various machine learning algorithms to prevent overfitting. Some common algorithms where regularization is often used include:

1. Linear Regression: Lasso regression (L1 regularization) and Ridge regression (L2 regularization) are commonly applied to linear regression models.

2. Logistic Regression: Regularization can be applied to logistic regression models using L1 or L2 regularization.

3. Support Vector Machines (SVM): SVM models can benefit from regularization, and techniques like L2 regularization are sometimes used.

4. Neural Networks: In neural networks, weight regularization (L1 or L2) is often applied to prevent overfitting, especially when dealing with deep learning models.

5. Decision Trees: While decision trees themselves do not directly use regularization, ensemble methods like Random Forests or Gradient Boosted Trees can be regularized by controlling the number of trees or their depth.

6. Elastic Net: Elastic Net is a combination of L1 and L2 regularization, and it is used when both types of regularization are desired.

MACHINE LEARNING

It's important to note that not all algorithms require regularization, and the choice of regularization method depends on the specific characteristics of the problem and data at hand. Regularization is typically used when there is a risk of overfitting, and it helps improve the generalization performance of models.

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

Ans:- In the context of a linear regression equation, the term "error" refers to the difference between the predicted values of the dependent variable (the variable we are trying to predict) and the actual observed values. These differences are also known as residuals.

The linear regression equation aims to model the relationship between the independent variable(s) and the dependent variable by finding the line (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differences between the predicted and observed values. Mathematically, the linear regression equation can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

where:

- Y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the model.
- ε represents the error term.

The error term captures the unobserved factors or random variation that affects the dependent variable but is not accounted for by the independent variables in the model. In an ideal situation, the goal is to have the error term as small as possible, indicating that the model is doing a good job of explaining the variation in the dependent variable.

The process of fitting a linear regression model involves finding the values of the coefficients that minimize the sum of the squared errors. The sum of squared errors is often referred to as the "residual sum of squares" (RSS) or the "sum of squared residuals." The idea is to make the predicted values as close as possible to the actual observed values, minimizing the discrepancies between them.