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 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

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 performance of models. Overfitting occurs when a model learns the noise and random fluctuations in the training data too well, resulting in poor performance on unseen data. Regularization helps to address this issue by adding a penalty term to the model's cost function, which penalizes large coefficients or overly complex models.

The main idea behind regularization is to find a balance between fitting the training data well and keeping the model simple enough to generalize effectively to new, unseen data. By penalizing overly complex models, regularization encourages the model to prioritize simpler explanations of the data, which are less likely to be influenced by noise and therefore more likely to generalize well.

Two common types of regularization techniques used in linear regression and other models are Ridge regression and Lasso regression. Ridge regression adds a penalty term proportional to the square of the magnitude of the coefficients, while Lasso regression adds a penalty term proportional to the absolute value of the coefficients. These techniques help to control the complexity of the model and reduce the risk of overfitting, ultimately leading to more robust and reliable models.

14. Which particular algorithms are used for regularization?

Ans: Regularization techniques are often applied to various machine learning algorithms to prevent overfitting and improve model generalization. Some of the algorithms commonly associated with regularization include:

1. Linear Regression: Regularization techniques like Ridge regression and Lasso regression are commonly used to regularize linear regression models.

2. Logistic Regression: Similar to linear regression, logistic regression models can be regularized using Ridge or Lasso regression to prevent overfitting.

3. Support Vector Machines (SVM): SVM models can be regularized using techniques such as the L2 norm penalty in the objective function.

4. Neural Networks: Regularization techniques like L1 and L2 regularization, dropout, and early stopping are commonly used in neural networks to prevent overfitting and improve generalization.

5. Decision Trees and Random Forests: Pruning techniques can be considered a form of regularization in decision trees and random forests, where overly complex trees are pruned to improve generalization.

6. Gradient Boosting Machines (GBM): Regularization parameters can be tuned in gradient boosting models to control model complexity and prevent overfitting.

MACHINE LEARNING

7. Bayesian Methods: Bayesian regression models inherently incorporate regularization through the choice of prior distributions, which helps to prevent overfitting and improve generalization.

These are just a few examples, and regularization techniques can be applied to a wide range of machine learning algorithms to improve their performance and robustness.

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

Ans: In linear regression, the error term represents the difference between the observed values of the dependent variable and the values predicted by the regression model. It quantifies the discrepancy or residual between the actual data points and the estimated values produced by the linear regression equation.

Mathematically, the error term (\(\epsilon \)) for each data point \(i \) can be expressed as:

\[\epsilon_i = y_i - \hat{y}_i \]

Where:

- \(y_i \) represents the observed value of the dependent variable for data point \(i \).
- \(\hat{y}_i \) represents the predicted value of the dependent variable for data point \(i \) based on the linear regression model.

The goal of linear regression is to minimize the error term across all data points. Typically, this is achieved by minimizing the sum of the squared errors, known as the residual sum of squares (RSS) or the sum of squared residuals (SSR):

\[RSS = \sum_{i=1}^{n} \epsilon_i^2 \]

Minimizing the error term implies finding the best-fitting line or hyperplane that describes the relationship between the independent and dependent variables in the dataset. This best-fitting line minimizes the overall discrepancy between the observed values and the values predicted by the regression model, allowing for accurate predictions of the dependent variable based on the independent variables.