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

FLIP ROBO

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?**

Regularization is a technique used in machine learning and statistical modeling to prevent overfitting and improve the generalization performance of a model. Overfitting occurs when a model fits the training data too closely, capturing noise or irrelevant patterns, which leads to poor performance on new, unseen data.

Regularization introduces a penalty term into the model's objective function, which discourages complex or extreme parameter values. This penalty term helps to control the model's complexity and reduce the impact of individual features or coefficients. By doing so, regularization helps the model generalize better to new data by finding a balance between fitting the training data and avoiding overfitting.

There are different types of regularization techniques, such as L1 regularization (Lasso), L2 regularization (Ridge), and Elastic Net regularization, each with its own specific penalty terms and effects on the model. These techniques shrink or constrain the parameter values, making them less sensitive to individual data points and reducing the likelihood of overfitting.

In summary, regularization is a technique used to control the complexity of a model and prevent overfitting by adding a penalty term to the objective function, ultimately improving the model's generalization performance on unseen data.

14. **Which particular algorithms are used for regularization?**

Several algorithms utilize regularization techniques to improve model performance and prevent overfitting. Here are some commonly used algorithms that incorporate regularization:

1. Ridge Regression: This algorithm applies L2 regularization by adding a penalty term to the regression objective function. It helps in shrinking the coefficients, reducing their impact and preventing overfitting.

2. Lasso Regression: Lasso Regression employs L1 regularization by adding a penalty term to the regression objective function. It encourages sparsity in the coefficients, effectively performing feature selection and reducing the impact of irrelevant features.

3. Elastic Net: Elastic Net combines both L1 and L2 regularization, providing a balance between feature selection (L1) and coefficient shrinkage (L2). It addresses some of the limitations of Ridge and Lasso regression by incorporating both penalties.

4. Logistic Regression: Logistic Regression is a classification algorithm that can be regularized using L1 or L2 regularization. Regularized Logistic Regression helps in controlling model complexity and preventing overfitting in binary or multi-class classification problems.

5. Support Vector Machines (SVM): SVM algorithms can be regularized using techniques like L1 or L2 regularization. Regularization in SVM helps in controlling the margin and reducing the impact of individual support vectors, improving generalization.

6. Neural Networks: Regularization techniques, such as Dropout and L1/L2 regularization, can be applied to neural networks to prevent overfitting and improve their generalization capabilities.

These are some examples of algorithms that incorporate regularization techniques. The choice of which algorithm to use depends on the specific problem and the nature of the data.

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

The error in the context of linear regression refers to the discrepancy or the difference between the observed values of the dependent variable and the predicted values by the linear regression equation.

In linear regression, the goal is to find a line that best fits the given data points by minimizing the overall error. The error, also known as the residual, represents the deviation between the predicted values and the actual observed values of the dependent variable.

Mathematically, the error at each data point is calculated as the difference between the observed value (y) and the predicted value ($\hat{y}$) for a given set of independent variables (x):

Error = $y - \hat{y}$

The linear regression equation estimates the relationship between the independent variables and the dependent variable by determining the slope and intercept of the line of best fit. However, due to the inherent variability and noise present in real-world data, the predicted values will not be an exact match for the observed values, resulting in residual errors.

The objective of linear regression is to minimize the sum of squared errors (SSE) or the mean squared error (MSE) across all data points. By minimizing the errors, the linear regression model aims to capture the underlying trend and pattern in the data and make accurate predictions.

In summary, the error in linear regression refers to the discrepancy between the observed values and the predicted values of the dependent variable, which is minimized during the model fitting process to find the best-fitting line.