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 model
   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.

Q.13 Explain the term regularization?

Answer: -

Regularization is a technique used in machine learning to prevent overfitting and improve the generalization ability of a model. It involves adding a regularization term to the loss function during model training, which penalizes complex or large parameter values.

The goal of regularization is to find a balance between fitting the training data well and avoiding overemphasis on the specific patterns or noise present in the training data. It helps to control the complexity of the model and reduce the impact of irrelevant or redundant features.

There are different types of regularization techniques commonly used:

1. L1 Regularization (Lasso): It adds the sum of the absolute values of the coefficients as the regularization term. L1 regularization encourages sparsity by pushing less important features to have zero or near-zero coefficients. It can be used for feature selection.

2. L2 Regularization (Ridge): It adds the sum of the squared values of the coefficients as the regularization term. L2 regularization shrinks the coefficients towards zero without forcing them to be exactly zero. It helps to reduce the impact of multicollinearity and can stabilize the model.

3. Elastic Net Regularization: It combines both L1 and L2 regularization by adding a linear combination of the absolute and squared values of the coefficients. It provides a balance between feature selection and coefficient shrinkage.

Regularization helps to prevent overfitting by reducing the model's reliance on individual training samples and reducing the sensitivity to noise in the data. It encourages the model to generalize well to unseen data by promoting smoother and simpler models.

The strength of regularization is controlled by a hyperparameter, often denoted as λ or alpha, which determines the extent of the penalty imposed on the coefficients. The optimal value of the regularization parameter is typically determined through techniques like cross-validation.

MACHINE LEARNING

Overall, regularization is an essential technique in machine learning to improve model performance, prevent overfitting, and enhance the model's ability to generalize to new data.

Q.14 Which particular algorithms are used for regularization?

Answer: -

Several algorithms use regularization techniques to improve model performance and prevent overfitting. Some popular algorithms that incorporate regularization are:

1. Ridge Regression: Ridge regression applies L2 regularization to linear regression. It adds the sum of squared coefficients multiplied by a regularization parameter (alpha) to the ordinary least squares loss function. Ridge regression shrinks the coefficients towards zero, reducing their impact on the model without eliminating them completely. This helps to handle multicollinearity and improve the model's stability.

2. Lasso Regression: Lasso regression applies L1 regularization to linear regression. It adds the sum of absolute coefficients multiplied by a regularization parameter (alpha) to the ordinary least squares loss function. Lasso regression encourages sparsity by driving some coefficients to exactly zero, effectively performing feature selection. It is particularly useful when dealing with datasets with a large number of features and can identify the most important features for prediction.

3. Elastic Net: Elastic Net combines both L1 and L2 regularization. It adds a linear combination of the absolute and squared coefficients to the ordinary least squares loss function, weighted by regularization parameters (alpha and l1_ratio). Elastic Net combines the benefits of both L1 and L2 regularization, providing both feature selection and coefficient shrinkage.

4. Logistic Regression: Logistic regression can also utilize regularization techniques such as L1 and L2 regularization. Regularized logistic regression helps prevent overfitting in binary classification problems and improves the model's ability to generalize to new data.

5. Support Vector Machines (SVM): SVM algorithms can incorporate regularization through the use of regularization parameters, such as C or gamma. These parameters control the trade-off between achieving a low training error and a low complexity model. Regularization helps SVM models to find an optimal balance between fitting the training data and avoiding overfitting.

It's important to note that regularization techniques can be applied to various machine learning algorithms, not just the ones listed above. The choice of the regularization technique depends on the problem at hand and the specific requirements of the dataset.

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

Answer: -

In linear regression, the term "error" refers to the difference between the actual dependent variable values and the predicted values generated by the linear regression equation. It represents the unexplained or residual variation in the data that cannot be captured by the linear relationship between the independent variables and the dependent variable.

Mathematically, the error term is denoted as "ε" (epsilon) and is given by the equation:

$\varepsilon = y - \hat{y}$

where:

- ε is the error term
- y is the actual value of the dependent variable
- $\hat{y}$ is the predicted value of the dependent variable based on the linear regression equation

The error term captures the discrepancy between the observed data points and the predictions made by the linear regression model. It represents the extent to which the model is unable to perfectly fit the training data.

The goal of linear regression is to minimize the sum of squared errors (SSE) or mean squared error (MSE), which is obtained by squaring the individual errors and summing them up. Minimizing the errors helps to find the best-fit line or hyperplane that optimally represents the relationship between the independent variables and the dependent variable.

By analyzing the errors, we can assess the accuracy and performance of the linear regression model. Common techniques to evaluate errors include computing metrics such as mean squared error (MSE), root mean squared error (RMSE), or assessing the residuals (the differences between the actual and predicted values) through residual analysis.

In summary, the error term in linear regression quantifies the discrepancy between the actual and predicted values, representing the unexplained variation that the linear model is unable to capture. Minimizing these errors helps to find the best-fit line or hyperplane that represents the relationship between the variables in the dataset.