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?
   Answer: Both A and B
2. Which of the following statements is true about outliers in linear regression?
   Answer: Linear regression is sensitive to outliers
3. A line falls from left to right if a slope is _____?
   Answer: Negative
4. Which of the following will have symmetric relation between dependent variable and independent variable?
   Answer: Correlation
5. Which of the following is the reason for over fitting condition?
   Answer: Low bias and high variance
6. If output involves label, then that model is called as:
   Answer: Predictive model
7. Lasso and Ridge regression techniques belong to _____?
   Answer: Regularization
8. To overcome with imbalance dataset which technique can be used?
   Answer: SMOTE
9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____to make graph?
   Answer: TPR and FPR
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
    Answer: False
11. Pick the feature extraction from below:
    Answer: Apply PCA to project high dimensional data

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?

    Answer:
    A) We don't have to choose the learning rate.
    B) It becomes slow when number of features is very large.

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 to prevent **overfitting** by adding a penalty term to the loss function. The goal of regularization is to reduce the complexity of a model, making it more generalizable to new, unseen data.

Overfitting occurs when a model is too complex and learns the noise and random fluctuations in the training data, rather than the underlying patterns. As a result, the model performs well on the training data but poorly on new data.

Regularization works by adding a term to the loss function that penalizes large weights or complex models. This encourages the model to find a simpler solution that still fits the training data well, but is less prone to overfitting.

There are several types of regularization techniques, including:

1. **L1 Regularization (Lasso)**: adds a term to the loss function that is proportional to the absolute value of the model's weights. This encourages the model to set some weights to zero, resulting in a sparse model.
2. **L2 Regularization (Ridge)**: adds a term to the loss function that is proportional to the square of the model's weights. This encourages the model to have smaller weights, resulting in a smoother model.
3. **Dropout**: randomly sets some neurons to zero during training, effectively creating an ensemble of models. This encourages the model to be more robust and less prone to overfitting.
4. **Early Stopping**: stops training when the model's performance on the validation set starts to degrade, preventing overfitting.

Regularization is an important technique in machine learning, as it helps to:

- Reduce overfitting
- Improve model generalization
- Increase model interpretability
- Reduce the risk of model complexity

MACHINE LEARNING

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

There are several algorithms that are commonly used for regularization:

1. **Ridge Regression**: This algorithm uses L2 regularization, which adds a penalty term to the loss function that is proportional to the square of the model's weights. This encourages the model to have smaller weights, resulting in a smoother model.
2. **Lasso Regression**: This algorithm uses L1 regularization, which adds a penalty term to the loss function that is proportional to the absolute value of the model's weights. This encourages the model to set some weights to zero, resulting in a sparse model.
3. **Elastic Net Regression**: This algorithm uses a combination of L1 and L2 regularization, which adds both L1 and L2 penalty terms to the loss function. This encourages the model to have smaller weights and to set some weights to zero, resulting in a sparse and smooth model.
4. **Dropout**: This algorithm is commonly used in neural networks and randomly sets some neurons to zero during training, effectively creating an ensemble of models. This encourages the model to be more robust and less prone to overfitting.

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

Error in linear regression refers to the difference between the observed value of the target variable (y) and the predicted value of the target variable based on the linear regression model.

In a linear regression equation, the error is represented by the Greek letter epsilon (ε). The linear regression equation can be written as:

$y = \beta 0 + \beta 1x + \varepsilon$

Where:

- y is the target variable (dependent variable)
- x is the feature variable (independent variable)
- β0 is the intercept or constant term
- β1 is the slope coefficient
- ε is the error term

The error term (ε) represents the random fluctuations or variations in the data that are not explained by the linear relationship between x and y. In other words, it is the difference between the observed value of y and the predicted value of y based on the linear regression model.

The error term can be thought of as the "noise" or "residual" in the data that is not captured by the linear model. The goal of linear regression is to minimize the error term, which means finding the best-fitting line that minimizes the difference between the observed and predicted values of y.

There are several types of errors that can occur in linear regression, including:

1. Residual error: The difference between the observed value of y and the predicted value of y based on the linear regression model.
2. Measurement error: Errors in measuring the values of x and y.
3. Model specification error: Errors in specifying the correct functional form of the linear regression model.
4. Omitted variable error: Errors due to omitting important variables from the linear regression model.