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.

ASSIGNMENT – 39

MACHINE LEARNING

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

13. Explain the term regularization?

When we use regression model to train some data, there is a good chance the model will overfit the given training data set. Overfitting occurs when the model fits the training data too well, capturing noise along with the underlying pattern, which leads to poor performance on unseen data (test set). Regularization in machine learning refers to a set of techniques used to reduce this overfitting by penalizing or constraining the complexity of the model.

14. Which particular algorithms are used for regularization?

Lasso (Least Absolute Shrinkage and selection operator) Regression (L1 form): it penalizes the model based on the sum of magnitude of the coefficient. Applies L1 regularization, leading to sparse models.

 Ridge Regression (L2 Regularization): Penalizes the model based on the sum of squares of magnitude of the coefficients. Applies L2 regularization, which shrinks coefficients to prevent overfitting but keeps all features in the model.

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

In linear regression, the error represents the difference between the observed values (actual data points) and the predicted values produced by the model. The error quantifies how far off the model's predictions are from the actual outcomes.

The total error in linear regression is typically decomposed into three main types:

- **Residual Error (ε)**: The difference between the observed value and the predicted value for each data point. It reflects the variability in the data that the model cannot explain.
- **Mean Squared Error (MSE)**: The average of the squared differences between the actual and predicted values. Squaring the residuals ensures that both positive and negative errors contribute equally and that larger errors are penalized more heavily.
- **Bias and Variance**: These are sources of error that affect model performance:
    - **Bias**: Error due to overly simplistic models that underfit the data.
    - **Variance**: Error due to overly complex models that overfit the data.