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

Correct Option: 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

Correct Option: A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____?

A) Positive B) Negative

C) Zero D) Undefined

Correct Option: B) Negative

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

Correct Option: B) Correlation

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

Correct Option: C) Low bias and high variance

MACHINE LEARNING

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

Correct Option: B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

A) Cross validation B) Removing outliers

C) SMOTE D) Regularization

Correct Option: D) Regularization

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation B) Regularization

C) Kernel D) SMOTE

Correct Option: 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

Correct Option: A) TPR and FPR

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True B) False

Correct Option: 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

Correct Option: A) Construction bag of words from a email

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

Correct Option: Both options A and B are correct

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 statistics to prevent overfitting and improve the generalization of models. Overfitting occurs when a model fits the training data too closely, capturing noise and random fluctuations rather than the underlying patterns. This can lead to poor performance on new, unseen data. Regularization methods introduce additional constraints or penalties to the model's optimization process, discouraging it from becoming too complex and fitting the noise in the data. One common form of regularization is L2 regularization, which adds a penalty term to the loss function of a model. The penalty term is proportional to the square of the magnitude of the model's coefficients. This encourages the model to have smaller coefficient values, effectively shrinking them towards zero. The result is a simpler model that is less sensitive to small variations in the training data.

Real-time Example:

Imagine you're building a linear regression model to predict housing prices based on various features like square footage, number of bedrooms, and location. If your model has too many features or very high coefficients, it might fit the training data very well, but it could also capture noise or minor fluctuations that aren't representative of the true underlying

relationship between the features and the prices. This could lead to poor predictions for new houses. To apply regularization in this scenario, you could use L2 regularization. By adding the squared magnitudes of the coefficients as a penalty term to the loss function, you encourage the model to prioritize smaller coefficients, effectively reducing the impact of less significant features. This helps prevent the model from overfitting and provides a more balanced fit to the data, improving its ability to make accurate predictions for new houses.

14. Which particular algorithms are used for regularization?

Several machine learning algorithms can be used with regularization techniques to prevent overfitting. Some popular algorithms that incorporate regularization are:

1. Ridge Regression (L2 Regularization): Ridge regression adds a penalty term proportional to the square of the coefficients to the linear regression loss function. It helps prevent large coefficient values and reduces model complexity.

2. Lasso Regression (L1 Regularization): Lasso regression also adds a penalty term to the linear regression loss function, but the penalty is proportional to the absolute values of the coefficients. Lasso tends to drive some coefficients to exactly zero, effectively performing feature selection.

3. Elastic Net: Elastic Net combines both L2 (ridge) and L1 (lasso) regularization, providing a balance between the two. It helps mitigate some of the limitations of individual L1 and L2 regularization.

4. Logistic Regression (with L1 or L2 Regularization): Similar to linear regression, logistic regression can also be regularized using L1 or L2 regularization to prevent overfitting in classification tasks.

5. Support Vector Machines (SVM): SVMs can use L2 regularization to control the trade-off between maximizing the margin and minimizing the classification error.

6. Neural Networks: Regularization techniques like dropout, L1/L2 weight regularization, and batch normalization can be applied to neural networks to prevent overfitting.

7. Decision Trees and Random Forests: Pruning techniques can be seen as a form of regularization in decision trees, where branches are removed to prevent the tree from becoming too deep and overfitting.

8. Gradient Boosting (e.g., XGBoost, LightGBM): Many gradient boosting algorithms allow for L1 and L2 regularization on the weak learners (individual decision trees) to control their complexity and prevent overfitting.

These algorithms and techniques help in striking a balance between model complexity and the fit to the training data, leading to better generalization and performance on new, and unseen data.

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

In the context of linear regression, the term "error" refers to the difference between the actual observed values of the dependent variable (also known as the response variable) and the predicted values generated by the linear regression model. These errors, also known as residuals, represent the variability in the data that is not explained by the linear relationship between the independent variables and the dependent variable.

The linear regression equation is typically represented as:

$y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta pxp+\varepsilon$

Where:

y is the actual observed value of the dependent variable.

$\beta 0,\beta 1,\beta 2,\ldots,\beta p$ are the coefficients (parameters) estimated by the model.

$x1,x2,\ldots,xp$ the independent variables.

ε represents the error term.

The error term captures the variability in the dependent variable that is not explained by the linear relationship with the independent variables. It accounts for factors that the model does not capture, such as measurement errors, unobserved variables, and inherent randomness in the data. The goal of linear regression is to minimize the sum of squared errors (residuals) between the predicted and actual values, leading to a model that best fits the observed data.

In summary, the error term in the linear regression equation represents the discrepancy between the model's predictions and the actual observations, highlighting the inherent uncertainty and variability in real-world data.