Machine learning

1. **Which of the following methods do we use to find the best fit line for data in Linear Regression?**

Answer: - D) Both A and B (A) Least Square Error B) Maximum Likelihood)

2. **Which of the following statement is true about outliers in linear regression?**

Answer: - A) Linear regression is sensitive to outliers

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

Answer: - B) Negative

4. **Which of the following will have symmetric relation between dependent variable and independent variable?**

Answer: - B) Correlation

5. Which of the following is the reason for over-fitting conditions?

Answer: - C) Low bias and high variance.

5. **If output involves label, then that model is called as:**

Answer: - B) Predictive modal.

7. Lasso and Ridge regression techniques belong to _____?

Answer: - D) Regularization.

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

Answer: - D) SMOTE.

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

Answer: - A) 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: - A) Construction bag of words from a email.

B) Apply PCA to project high dimensional data.

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 the number of features is very large.

13. Explain the term regularization?

Answer: - **Regularization** is a technique used in machine learning and statistical modeling to prevent overfitting by adding a penalty term to the model's loss function. Overfitting occurs when a model becomes too complex and fits the noise in the training data, leading to poor generalization on new, unseen data.

Regularization discourages the model from fitting too closely to the training data by penalizing large model coefficients. It helps in controlling the model's complexity, making it more robust.

There are two common types of regularization:

1. **L1 Regularization (Lasso)**: Adds the absolute value of the coefficients to the loss function, which can lead to some coefficients becoming zero, effectively performing feature selection.
2. **L2 Regularization (Ridge)**: Adds the square of the coefficients to the loss function, shrinking them toward zero without eliminating any features.

14. Which algorithms are used for regularization?

Answer: - Ridge Regression (L2 Regularization): Adds an L2-norm penalty (the square of the magnitude of the coefficients) to the loss function. This helps in shrinking the coefficients, but it doesn't eliminate any features.

Lasso Regression (L1 Regularization): Adds an L1-norm penalty (the absolute value of the coefficients) to the loss function. This can result in some coefficients becoming exactly zero, effectively performing feature selection.

Elastic Net: A combination of both L1 (Lasso) and L2 (Ridge) regularization. It combines the strengths of both methods and can handle multicollinearity better while also performing feature selection.

Regularized Logistic Regression: Applies L1, L2, or Elastic Net regularization to logistic regression to handle classification problems while preventing overfitting.

Support Vector Machines (SVM): Uses regularization by applying a penalty term in the optimization process to maximize the margin between classes and prevent overfitting.

Neural Networks: Techniques like L2 regularization (also called weight decay) and Dropout (a form of regularization that randomly drops units during training) are commonly used in deep learning models to avoid overfitting.

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

Answer: - In Linear Regression, the term "error" refers to the difference between the actual values (observed data) and the predicted values** (model's output). This difference is also known as the **residual**.

Types of Errors in Linear Regression:

1. Residual (Prediction Error):

- The residual for a data point is the difference between the actual value y and the predicted value $\hat{y}$ given by the model.

- Mathematically:

$$
\text{Residual (Error)} = y - \hat{y}
$$

- This represents how much the model's prediction deviates from the actual data for that particular point.

2. Mean Squared Error (MSE):

To evaluate the overall performance of the model, we often use the **Mean Squared Error (MSE), which is the average of the squared residuals across all data points.

Squaring the residuals ensures that both positive and negative errors are treated equally and prevents cancellation of errors.

Mathematically:

\[

MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2

\]

- Where \(n\) is the number of data points, \(y_i\) is the actual value, and \(\hat{y}_i\) is the predicted value for the \(i\)-th data point.

3. Bias and Variance (Error components):

- In linear regression, error can also be decomposed into **bias** and **variance**:

Bias: Error introduced by the simplifying assumptions made by the model (e.g., assuming a linear relationship when the true relationship is more complex).

Variance: Error due to the model being too sensitive to small fluctuations in the training data (overfitting).

Role of Error:

The goal in linear regression is to minimize the overall error, typically measured through metrics like MSE or RMSE (Root Mean Squared Error), by adjusting the coefficients of the regression equation. Reducing this error helps improve the model's accuracy and its ability to generalize new data.