MACHINE LEARNING

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
   Ans:- 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
   Ans:- 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
   Ans:- 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
   Ans:- 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
   Ans:- C) Low bias and high variance

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
Ans:- B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?
   A) Cross validation
   B) Removing outliers
   C) SMOTE
   D) Regularization
   Ans:- D) Regularization

8. To overcome with imbalance dataset which technique can be used?
   A) Cross validation
   B) Regularization
   C) Kernel
   D) SMOTE
   Ans:- 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
   Ans:- 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
    Ans:- 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
    Ans:- 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?

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

13. Explain the term regularization?

Ans:- **Regularization** is a technique used in statistical modeling and machine learning to prevent overfitting, which occurs when a model learns the noise in the training data instead of the underlying patterns. Regularization adds a penalty to the loss function used to train the model, discouraging complexity in the model and promoting simpler, more generalizable solutions.

14. Which particular algorithms are used for regularization?

Ans:- Regularization techniques are applied in various machine learning algorithms to prevent overfitting and improve model generalization. Here are some specific algorithms that incorporate regularization:

1. Linear Regression

- Ridge Regression (L2 Regularization): Adds a penalty equal to the square of the magnitude of coefficients to the loss function.
- Lasso Regression (L1 Regularization): Adds a penalty equal to the absolute value of the coefficients, leading to sparse models by shrinking some coefficients to zero.
- Elastic Net: Combines both L1 and L2 regularization techniques, allowing for both feature selection and coefficient shrinkage.

2. Logistic Regression

- Similar to linear regression, logistic regression can also use L1, L2, or Elastic Net regularization to handle binary classification problems.

3. Support Vector Machines (SVM)

- Regularization is inherently part of the SVM formulation, where a regularization parameter (C) controls the trade-off between maximizing the margin and minimizing classification errors.

4. Neural Networks

- L1 and L2 Regularization: These can be applied to the weights of neural network layers to prevent overfitting.

- Dropout: A technique where a fraction of the neurons are randomly turned off during training to reduce dependency on specific nodes and improve generalization.

5. Decision Trees and Ensemble Methods

While traditional decision trees do not include regularization, ensemble methods like Random Forests and Gradient Boosting incorporate mechanisms that can help control overfitting, such as:

- Tree pruning: Reducing the size of the trees after training.
- Limiting tree depth: Setting a maximum depth for trees in ensemble methods.
- Shrinkage: In gradient boosting, a learning rate (shrinkage parameter) can be applied to reduce the contribution of each tree, acting as a form of regularization.

6. Bayesian Methods

- In Bayesian statistics, regularization can be seen as incorporating prior distributions over parameters (e.g., placing a Gaussian prior on weights in a linear regression model).

7. K-Nearest Neighbors (KNN)

- While KNN is not inherently regularized, techniques like feature selection or dimensionality reduction (e.g., PCA) can be employed as a form of regularization to mitigate overfitting.

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

Ans:- In linear regression, the term error (often referred to as residual) quantifies the difference between the observed values (actual outcomes) and the values predicted by the regression model. Understanding this concept is essential for evaluating model performance and making improvements. The error term in a linear regression equation plays a crucial role in understanding how well the model fits the data. By analyzing the errors, practitioners can identify weaknesses in the model, assess its predictive accuracy, and make necessary adjustments to improve its performance. Understanding and minimizing errors is fundamental to building effective predictive models.

1. Components of Error

The error can stem from various sources, including:

- Model Inaccuracy: The linear model may not capture the true relationship between the variables, particularly if the relationship is non-linear.
- Omitted Variables: Important explanatory variables might be missing from the model, leading to biased estimates.
- Measurement Error: Errors in measuring the independent or dependent variables can contribute to the overall error.

- Random Variability: Natural randomness in the data (e.g., fluctuations due to unobserved factors) can also result in errors.

2. Types of Error

Errors can be classified in different ways:

- Residuals: The difference between the observed values and the predicted values from the model. This is what is typically analyzed during model evaluation.
- Bias: The systematic error that occurs when the model consistently predicts a certain way, potentially due to model misspecification.
- Variance: The variability of the model's predictions when applied to different datasets. A high variance can lead to overfitting, where the model captures noise in the training data rather than the underlying pattern.

3. Evaluating Error

To assess the performance of a linear regression model, several metrics can be used, including:

- Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values.
- Mean Squared Error (MSE): The average of the squared differences between predicted and actual values, emphasizing larger errors.
- Root Mean Squared Error (RMSE): The square root of the MSE, providing error in the same units as the dependent variable.
- R-squared: A statistical measure representing the proportion of variance for the dependent variable that's explained by the independent variables in the model.