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

Answer: A) Least Square Error
Explanation: The least square error method is used to minimize the error in linear regression by minimizing the sum of the squared differences between the observed and predicted values.

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

Answer: A) Linear regression is sensitive to outliers
Explanation: Linear regression is highly sensitive to outliers as they can skew the line of best fit.

3. A line falls from left to right if a slope is _____?
   A) Positive   B) Negative C) Zero   D) Undefined

   Answer: 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

   Answer: B) Correlation
   Explanation: Correlation measures the symmetric relationship between variables, while regression focuses on predicting one variable from the other.

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

   Answer: C) Low bias and high variance
   Explanation: Overfitting occurs when a model has low bias but high variance, meaning it fits the training data well but doesn't generalize to new data.

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

Answer: B) Predictive model

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

Answer: D) Regularization

Explanation: Lasso and Ridge regression are both regularization techniques used to prevent overfitting by penalizing large coefficients.

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

Answer: D) SMOTE

Explanation: SMOTE (Synthetic Minority Over-sampling Technique) is used to handle imbalanced datasets by generating synthetic samples for the minority class.

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

Answer: A) TPR and FPR

Explanation: The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

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

Answer: B) False

Explanation: A better model should have a higher area under the curve (closer to 1), not less.

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

Answer: A) Construction bag of words from an email

Explanation: Creating a bag of words from an email is a feature extraction technique used in natural language processing.

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.

Answer: A) We don't have to choose the learning rate.
B) It becomes slow when the number of features is very large.
Explanation: In the Normal Equation, there is no need to set a learning rate since it directly computes the parameters. However, it can become slow for large feature sets due to matrix inversion complexity.

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 to the model for having too complex or large weights. This penalty term is typically added to the loss function, discouraging the model from learning overly complex patterns in the training data, which might not generalize to new data. The two most common types of regularization are L1 regularization (Lasso) and L2 regularization (Ridge), where Lasso tends to shrink some coefficients to zero, effectively selecting a subset of features, and Ridge penalizes large coefficients, making the model less sensitive to noise in the data.

14. Which particular algorithms are used for regularization?

    The specific algorithms used for regularization include:

    - Lasso Regression (L1 Regularization): Adds an L1 penalty (sum of absolute values of coefficients), encouraging sparsity in the model.

    - Ridge Regression (L2 Regularization): Adds an L2 penalty (sum of squared coefficients), shrinking the coefficients but typically not to zero.

    - Elastic Net: Combines L1 and L2 regularization, giving a balance between feature selection and coefficient shrinkage.

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

The error in a linear regression equation is the difference between the observed value and the predicted value, known as the residual.

In linear regression, the goal is to minimize the sum of squared errors (SSE), which is the sum of the squares of all the residuals. This ensures that the model fits the data as closely as possible.