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
   ans 1: D

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 2: A

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

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 4: A

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 5: C

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 6: B

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

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

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 9: A

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

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 11: A

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.
    ans 12: A,B

MACHINE LEARNING

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

13. Explain the term regularization?
14. Which particular algorithms are used for regularization?
15. Explain the term error present in linear regression equation?

Ans 13:

Regularization is a technique used in machine learning to prevent overfitting & underfitting,

Overfitting occurs when a model is too complex and captures noise or irrelevant patterns in the training data. Regularization involves adding a penalty term to the cost function of a model to discourage large values of the model parameters, which can lead to overfitting.

Underfitting means that it is too simple and cannot capture the underlying patterns in the data. In this case, regularization can be used to encourage the model to learn a more complex representation of the data by allowing larger weights. This is typically done by reducing the strength of the regularization penalty.

Ans 14:

For regularization of Linear or Logistic Regression: Lasso, Ridge & ElasticNet algorithms are used

Support vector regression, neural networks regression, and decision trees regression are also used for regularization.

Ans 15:

In linear regression, the error refers to the difference between the predicted values and the actual values of the target variable. It is also known as the residual, and can be computed as the difference between the actual value of the target variable and the predicted value obtained using the linear regression equation.The linear regression equation is typically of the form: $y = \beta0 + \beta1 * x1 + \beta2 * x2 + ... + \beta n * xn + \varepsilon$where y is the target variable, x1, x2, ..., xn are the predictor variables, β0, β1, β2, ..., βn are the coefficients of the regression equation, and ε is the error term.The error term represents the variability in the target variable that cannot be explained by the predictor variables included in the model. It can be caused by various factors such as measurement errors, omitted variables, non-linear relationships, or random noise in the data.The goal of linear regression is to minimize the sum of squared errors (SSE) between the predicted values and the actual values of the target variable. This is done by finding the optimal values of the coefficients that minimize the SSE. The SSE is a measure of how well the regression line fits the data points, and a lower SSE indicates a better fit.In summary, the error term in linear regression represents the part of the target variable that is not explained by the predictor variables, and the goal of linear regression is to minimize this error term by finding the best coefficients for the regression equation.