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

2. Which of the following statement is true about outliers in linear regression?
   A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is_______?
   B) **Negative**

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

5. Which of the following is the reason for over fitting condition?
   C) **Low bias and high variance**

6. If output involves label then that model is called as:
   B) **Predictive modal**

7. Lasso and Ridge regression techniques belong to_________?
   D) **Regularization**

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

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

11. Pick the feature extraction from below:
    B) **Apply PCA to project high dimensional data**

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**.

MACHINE LEARNING

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

Q-13) Explain the term regularization?

A-13) Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

Q-14 Which particular algorithms are used for regularization?

Ans-14) There are mainly two types of regularization techniques, which are given below:

- **Ridge Regression**
- **Lasso Regression**

Ridge Regression

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
- The equation for the cost function in ridge regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
- As we can see from the above equation, if the values of **λ tend to zero, the equation becomes the cost function of the linear regression model.** Hence, for the minimum value of λ, the model will resemble the linear regression model.
- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

Lasso Regression:

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator.**
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization.** The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|^{\square}$$

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

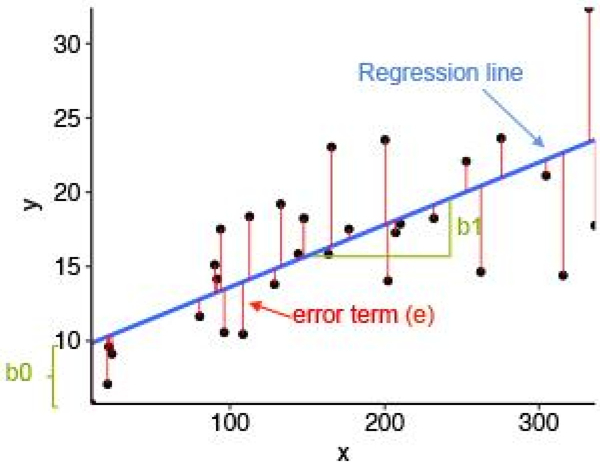
MACHINE LEARNING

Q-15 Explain the term error present in linear regression equation?

A-15) In the most simple words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.



○

In the above diagram,

- x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_0 is the intercept which is 10 and b1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.

The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.