Machine Learning Assignment

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 ?
   B) Correlation
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 model
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 to make a graph?
   B) Sensitivity and Specificity
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the Should be less
    B) False
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.
    C) We need to iterate.
    D) It does not make use of dependent variable.

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

13. Explain the term regularization?

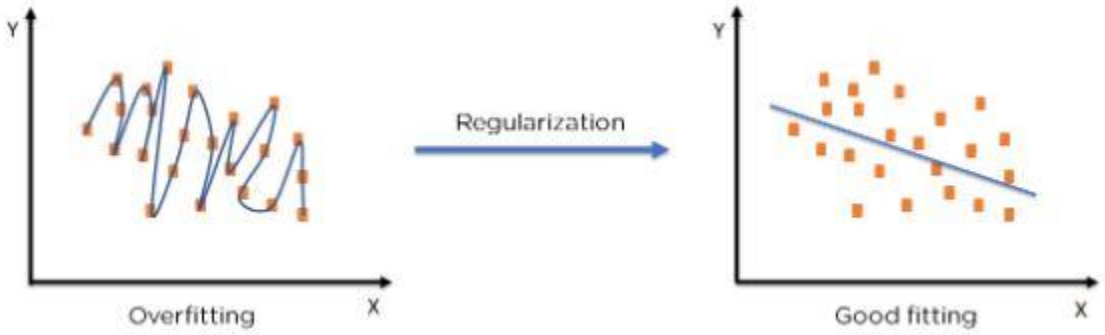
Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting or under fitting.
The commonly used regularization techniques are :

1. L1 regularization
2. L2 regularization

14. Which particular algorithms are used for regularization?

This is a form of regression, that regularizes or shrinks the coefficient estimates towards zero. Regularization helps in avoiding the over fitting or under fitting of the curve. The commonly used regularization techniques are :

1. L1 regularization
2. L2 regularization



A regression model which uses L1 Regularization technique is called LASSO(Least Absolute Shrinkage and Selection Operator) regression. A regression model that uses L2 regularization technique is called Ridge regression.

Lasso regression is very similar to the concept of Ridge regression. We can understand Lasso regression by considering an example. Suppose we have a bunch of mice. We can start by making a graph of the weight and size of individual mice. On the vertical line of the graph, we take the size, and on the horizontal line, we will take the weight.

Now split this data on the graph into two different sets for better classification. We will highlight the training data as red dots on the graph, and we will highlight the testing data with the green dots. Now, we will use the Least Squares and place a line on the training data. Now, we can use ridge regression and fit the line on the data. By doing this, we are minimizing the sum of the squared ridge regression and lambda times the slope squared. Ridge regression is the Least-squares plus the Ridge Regression Penalty. The sum of squared ridge regression + λ x the slope^2

The ridge regression line and least-squares do not fit each other as well as the training data. We can say that the Least Squares has lower Bias than Ridge Regression. However, due to the small Bias, you will see a huge drop in the variance of the ridge regression. If we remove the square on the slope, we take the absolute value, we will find Lasso Regression.

The sum of squared ridge regression + λ x |the slope|

Lasso regression also has little Bias, just like Ridge Regression but has less Variance than Least Squared. Both these types of regressions look similar and perform the same function of making the size of the training data less sensitive.

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

An error is a value which represents how observed data differs from the actual population data. The error term is also known as the residual, disturbance, or remainder term, and is variously represented in models by the letter ε. It appears in a statistic model, like a regression model to indicate the uncertainty in the model. It will reflect the non-linearities, unpredictable effects, measurement errors and omitted variables.

Error of the data set is the differences between the observed values and the true / unobserved values. Residual is calculated after running the regression model and is the differences between the observed values and the estimated values.

When you perform simple linear regression (or any other regression analysis), you get a line of best fit. The data points usually don't fall exactly on this regression equation line; they are scattered around. A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are:

- Positive if they are above the regression line,
- Negative if they are below the regression line,
- Zero if the regression line actually passes through the point.