Machine Learning

1)

A) Least Square Error is used to find the best fit line for data in Linear Regression.

2)

A) Linear regression is sensitive to outliers as the slope of the regression line will change due to outliners.

3)

B) Negative

4)

A)

Regression will have symmetric relation between dependent variable and independent variable.

5)

C) Low bias and high variance is the reason for over fitting condition

6)

B) Predictive model

7)

Lasso and Ridge regression techniques belong to Regularization

8)

D) SMOTE Synthetic Minority oversampling technique can be used to overcome with imbalance dataset.

9)

A) TPR and FPR is used in The AUC Receiver Operator Characteristic (AUCROC) curve to make graph.

10)

B) False

11)

B) Apply PCA to project high dimensional data

Answer Q.12)

A) We don’t have to choose the learning rate.

B) It becomes slow when number of features is very large.

D) It does not make use of dependent variable.

Answer Q.13)

Regularization

Regularization is a form of regression which avoids the risk of overfitting.It shrinks the co efficient value to zero

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

Here

Y = learned relationship

β = coefficient estimates

X= Predictors

Following are some factors used in Regularization

Data Fitting –

The process of plotting a series of data points and drawing the best fit line to understand the relationship between the variables is called Data Fitting

Noise - Our model is the best fit when it can find all necessary patterns in our data and avoid the random data points and unnecessary patterns called Noise.

Over fitting - A scenario where the machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point on the curve is called Over fitting.

Under fitting - A scenario where a machine learning model can neither learn the relationship between variables in the testing data nor predict or classify a new data point is called under fitting.

Bias –

A Bias occurs when an algorithm has limited flexibility to learn from data. Such models pay very little attention to the training data

Such models always lead to a high error on training and test data. High Bias causes underfitting in our model.

Variance : It defines the algorithm's sensitivity to specific sets of data. High Variance causes overfitting in our model.

Regularization Techniques

There are two main types of regularization techniques:

Ridge Regularization

Lasso Regularization

Ridge Regularization

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

In the cost function, the penalty term is represented by Lambda λ. By changing the values of the penalty function, we are controlling the penalty term. The higher the penalty, it reduces the magnitude of coefficients

Lasso Regression

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.

Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients.

Answer Q.14)

Which particular algorithms are used for regularization?

L2 and L1 Regularization

L2 and L1 are the most common types of regularization. Regularization works on the premise that smaller weights lead to simpler models which in results helps in avoiding overfitting. So to obtain a smaller weight matrix, these techniques add a 'regularization term' along with the loss to obtain the cost function.

Cost function = Loss + Regularization term

Dropout
Another most frequently used regularization technique is dropout. It essentially means that during the training, randomly selected neurons are turned off or 'dropped' out.

So if neurons are randomly dropped out of the network during training, the other neurons step in and make the predictions for the missing neurons.

Early Stopping
In this validation strategy where one part of the training set is used as a validation set, and the performance of the model is gauged against this set. So if the performance on this validation set gets worse, the training on the model is immediately stopped.

Answer Q.15)

Linear regression most often uses mean-square error (MSE) to calculate the error of the model.

MSE is calculated by:

1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. Calculating the mean of each of the squared distances.

Formula to calculate MSE

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where

MSE – mean square error

n- number of data points

Yi= observed value

ŷi = predicted value

Following are some properties of MSE-

1. The values can be less than or more than actual value

2. As the values are obtained as square, all errors are positive and the mean value is also positive

3. Data points are not constant in finding MSE

4. Some time to get more accurate results we obtain the Root mean squared error as (RMSE)