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
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
3. A line falls from left to right if a slope is_____?
   A) Positive B) Negative
   C) Zero D) Undefined
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
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
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
7. Lasso and Ridge regression techniques belong to_____?
   A) Cross validation B) Removing outliers
   C) SMOTE D) Regularization
8. To overcome with imbalance dataset which technique can be used?
   A) Cross validation B) Regularization
   C) Kernel 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 B) Sensitivity and precision
   C) Sensitivity and Specificity D) Recall and precision
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
    A) True B) False
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

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.

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?

ANSWERS: 1.A

2.A

3.B

4.B

5.C

6.A

7.D

8.D

9.C

10.B

11.B

12.A&B

Q.13 Explain the term regularization?

Ans. Regularization technique help to reduce the chance of overfitting and underfitting and help us to get an optimal model. Now let us discuss over fitting and underfitting

Now to train our machine learning model, we give it some data points and drawing the best fit line to understand the relationship between the variables is called Data fitting. Our model is best fit when it can find all the necessary patterns in our data and avoid the random data points and unnecessary patterns called noise. If we look at our data multiple time we will find patterns which are unnecessary also.

Now let us understand Overfitting and Underfitting

OVERFITTING: 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 overfitting. Now, conversely if we look at our model but wont be able to find pattern in our dataset, then

UNDERFITTING: 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 underfitting.

In order to make these point more clear let us learn BIAS AND VARIANCE

A bias occurs when an algorithm has limited flexibility to learn from the data. Such model pay very less attention to the training data and thus cause high error on training and test data. High bias cause underfitting in our model. On the other hand Variance defined the algorithm sensitivity to specific set of data. A model with high variance pays a lot of attention to the training model and does not generalize. Such model generally performs very well on training data but have high error rates and hence cause overfitting in our model.

REGULARIZATION TECHNIQUES

There are two main types of regularization techniques: RIDGE AND LASSO REGULARIZATION

RIDGE REGULARIZATION: it modifies the over fitted or under fitted models by adding the squares of the magnitude of the coefficient. This means that mathematical function representing our machine learning model is minimized and coefficient are calculated. Then these are squared and added.

LASSO REGULARISATION: it modifies the over fitted and under fitted models by adding the penalty

MACHINE LEARNING

equivalent to the sum of the absolute values of coefficient.
Lasso regularization also performs coefficient minimization but instead of squaring the magnitude of the coefficient , it takes the true values of the coefficient .this means that the coefficient sum can also be zero due to the presence of the negative coefficient.
CONCLUSION
We have noticed that there are numerous ways through which our model become unstable by being under fitted and over fitted . so through regularization we can easily deal with under fitting and over fitting of a working model . we have also saw the role of bias and variance in model optimization. So In order to uplift the performance of the machine learning model we regularize it .
Q.14 Which particular algorithms are used for regularization?
ANS: The most common problem in machine learning is the overfitting of the data. So regularization helps in overcoming the problem of overfitting and also increases the model interpretability.
Sometimes what happens, a training model performs well on the training data but not on the test data. It means the model is not able to predict the output or target column for the unseen data by introducing noise in the output and hence the model is called overfitted model.
So in simple word, we can say that regularization is a technique through which we reduce the magnitude of the independent variables by keeping the same number of variables. It maintains the accuracy and generalization of the model.
ALGORITHM USED FOR REGULARIZATION
RIDGE REGRESSION
LASSO REGRESSION
NOW, RIDGE REGRESSION: it is a type of linear regression in which we introduce a small amount of bias known as penalty so that we can get better long term prediction. Its uses are as follows:

1. When we have independent variables having high collinearity between them and polynomial regression fails to solve such problems then ridge regression can be used.
2. If we have more parameters than samples.

LIMITATIONS:

1. It decreases the complexity of a model but does not reduces the number of independent variables since it never leads to a coefficient being zero rather only minimizes.
2. Its disadvantage is model interpretability since it will shrink the coefficient for least important predictors, very close to zero but it will never make them exactly zero.

NOW LASSO REGRESSION: it is an another variant of regularization technique used to reduce the complexity of the model. It stand for least absolute and selection operator. It is similar to ridge regression except the penalty term includes the absolute weights instead of the square of weights.
In this technique penalty has forcefully estimates the value of coefficient to be equal to zero and hence results in the removal of some important features of the model.
LIMITATIONS:

1. if the number of predictor is greater than the number of data points ,lasso will pick at most n predictors even if all the predictor are relevant.
2. If there are two or more highly collinear variable then lasso regression selects one of them randomly even if they are not good for the interpretation of the model .

CONCLUSION:

1. In simple regression, the standard least square model tends to have some variance in it. This wont generalize well for future dataset.
2. Regularization tries to reduce the variance of the model without a substantial increase in bias.

MACHINE LEARNING

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

ANS: Regression is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. While dealing with linear regression we have multiple lines for different values of slopes and intercepts. But the main question arises is which of those lines actually represents the right relationship between the X&Y and in order to find that we can use the MEAN SQUARE ERROR or MSE . for linear regression this MSE is nothing but the cost function .

MEAN SQUARED ERROR is the sum of the squared differences between the predicted and the true value. And the output is a single number representing the cost.

It actually measures the amount of error in statistical models. Is assesses the average squared differences between the observed and predicted values. When the model has no error , the MSE equals zero. As model error increases its value increases. The mean squared error is also known as the mean squared deviation MSD.

For example , in regression , the mean squared error represents the average squared residual.

The data points fall closer to the regression line, the model has less error, decreasing the MSE. A model with less error produces more precise prediction.

To find the MSE, take the observed Value, and square that difference. Repeat that for all observations. Then sum all the squared values and divide by the number of observations.

INTERPRETING THE MEAN SQUARE ERROR

The mean square error is the average squared distance between the observed and the predicted values. Because, it uses squared units rather than the natural data units. Squaring the difference serves several purposes.

Squaring the difference eliminates negative values for the difference and ensures that the mean squared error is always greater than or equal to zero. It is almost a positive value. Only a perfect model with no error produces an MSE of zero.

The least sum of squared error is also for the line having minimum MSE. So many best-fit algorithm use the least sum of squared error methods to find a regression line. MSE unit order is higher than the error unit as the error is squared. To get the same unit order, many times the square root of MSE is taken. It is called the ROOT MEAN SQUARED ERROR (RMSE).

RMSE=sqrt(MSE)

This is also used for the model evaluation. There are other measures like MSE, R2 used for regression model evaluation.

Mean Absolute Error is the sum of the absolute difference between actual and predicted values. R2 or R squared is the coefficient of determination. It is the total variance explained by the model.

MACHINE LEARNING