ASSIGNMENT-01

Machine Learning

Question No.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

Answer:- (A) Least Square Error

Question No.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

Answer:- (A) Linear Regression is sensitive to outliers.

Question No.3:- A line falls from left to right if a slope is____?

(A) Positive (B) Negative

(C) Zero (D) Undefined

Answer:- (A) Positive

Question No.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

Answer:- (B) Correlation

Question No.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

Answer:- (C) Low bias and High variance.

Question No.6:- If Output involves label that model is called as:

(A) Descriptive model (B) Predictive modal

(C) Reinforcement learning (D) All of the above

Answer:- (B) Predictive Model

Question No.7:- Lasso and Ridge regression techniques belong to_____?

(A) Cross validation (B) Removing outliers

(C) SMOTE (D) Regularization

Answer:- (D) Regularization.

Question No.8:- To overcome with imbalance dataset which technique can be used?

(A)Cross validation (B) Regularization

(C) Kernel (D) SMOTE

Answer:- (C) SMOTE

Question No.9:- The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problem. It uses____ to make graph?

(A) TPR and FPR (B) Sensitivity and precision

(C) Sensitivity and Specificity (D) Recall and precision

Answer:- (A) TPR and FPR

Question No.10:- In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

(A) True **(B)** False

Answer:- (B) False

Question No. 11:- Pick the feature extraction.

(A) Construction bag of words from a email (B) Apply PCA to project high dimensional data
(C) Removing stop words (D) Forward selection

Answer:- (A) Construction bag of words from a email.

Question No. 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

Answer:- (A) We don’t have to choose the learning rate. & (B) It becomes slow when number of features is very large.

Question No. 13:- Explain the term regularization?

Answer:- As a data scientist, whenever we will build model we need check that whether if the model is overfitted model or not? How do we check? There comes regularization techniques comes into picture. Because data is changing by every second in real time, so we have to restrict our model to handle this type of problem also.

When we use regression models to train some data, there is a good chance that the model will overfit the given training data set. Regularization helps us to sort this overfitting problem by restricting the degrees of freedom of a given equation i.e. simply reducing the number of degrees of a polynomial function by reducing their corresponding weights.

In Linear equation, we do not want huge weights as a small change in weight can make a large difference for the dependent variable. So regularization constraints the weights of such features to avoid overfitting.

There are different type of regularization techniques:

1- **Lasso:-** lasso regression penalizes the model based on the sum of magnitude of the coefficients.

Example :- Height, Email, Phone No, GRE, TOFEL, SOP,….Chance of Admission

(**Features)** **(Label)**

In this example we can see that some features like(Height, Email, Phone no.) are not related to the label. There is no relation between candidates heights to his chance of Admission. So In LASSO, it will nullify all those features or variables , and it will give them 0% importance whose are not related to the label.

2- **Ridge :-** Ridge regression penalizes the model based on the sum of squares of magnitude of the coefficients.

In above example, Where lasso find no relation between feature and label it nullify them but in Ridge it will not nullify them, rather what it will do , it will give very very less importance like 0.000213%.But these features will be there, they will carry very very less importance .

Question No. 14:- Which particular algorithms are used for regularization?

Answer:- 1- LASSO (Least Absolute Shrinkage and Selection Operator) Regression:- It is also called as L1 Form.

2 - Ridge Regression :- It is also called as L2 form.
3 **-** Elasticent : It is very less popular that's why people does not use this.

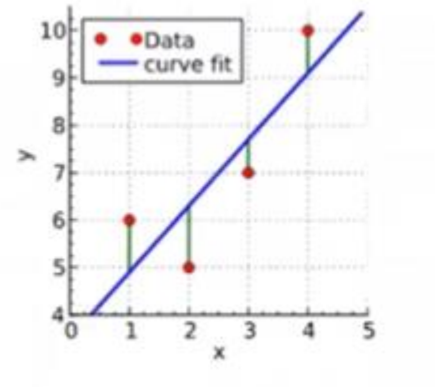
Question No.15:- Explain the term error present in linear regression equation?

Answer:- The error term is also known as the residual, disturbance or remainder term. When the model does not fully represent the actual relationship between the independent variables and the dependent variables, it called the error term in regression.

The distance between actual data points and the best fit line is called as residual. Because we can not predict always 100% correct, we should agree with that nobody is going to be 100% precise but it is very important how close we are from actual number.

For Example in T20 cricket match India will play against Australia. We may predict the score of first inning based on previous matches . Suppose we predict 180 runs, but at the end the team scored 176 runs only, but there is very less we make exactly 180 but we will be happy because the margin of error is very very less and we are ok with that. But how much error we made, that error is called as Residual.

Now look into the plot create , now consider each point , and know that each of them has coordinate in the form of (X,Y). Now draw a imaginary line between each point and the current "best fit line". We'll called distance between each point and the current best fit line as D.



A- The red point are observed values of X and Y.

B- The blue line is Least Squares Line.

C- The green lines are the residuals, which is the distance between the observed values and the least square line.

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a multiple linear regression function that takes the following form:

$Y=aX+\beta\rho+\epsilon$

where:

a,β=Constant parameters X,

ρ=Independent variables

ϵ=Error term