MACHINE LEARNING

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.

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.

3. A line falls from left to right if a slope is _____?
A) Positive B) Negative
C) Zero D) Undefined
Answer: B) Negative

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

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

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
Answer: B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation
B) Removing outliers
C) SMOTE
D) Regularization
Answer: D) Regularization

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

A) Cross validation

B) Regularization

C) Kernel

D) SMOTE

Answer: 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

Answer: C) Sensitivity and Specificity

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

11. Pick the feature extraction from below:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

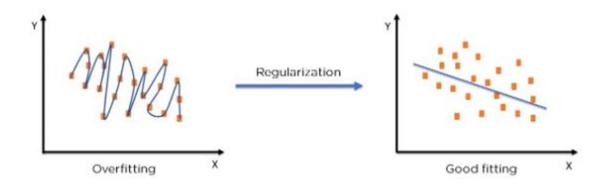C) Removing stop words

D) Forward selection

Answer: Option A,B,C

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.

Answer: Option A,B,C

13. Explain the term regularization?

Regularization: Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

14. Which particular algorithms are used for regularization?

There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization.

Ridge Regularization : 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. This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present.

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. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.

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

In Linier Regression the term error is known as metrics. There are four metrics namely

1. Mean Squared Error (MSE)
2. Mean Absolute Error (MAE)
3. Root Mean Squared Error (RMSE)
4. R2 Score

1. **Mean Squared Error (MSE):** It is the mean of squared difference between actual value and predicted value.

$$MSE = \frac{\sum(Actual\ Value - Predicted\ Value)^2}{n}$$

2. **Mean Absolute Error (MAE):** It is the algebraic mean of the difference between actual value and predicted value.

$$MAE = \frac{\sum(Actual\ Value - Predicted\ Value)}{n}$$

3. **Root Mean Squared Error (RMSE):** It is the square root of mean squared error.
$$RMSE = \sqrt{MSE}$$

4. **R2 Score:** It is also known as coefficient of determination. It is the proportion of the variance in the dependent variables. i.e. change coming in output / label whenever input variable is being changed.

   from sklearn.metrics import r2_score
   r2_score(actual_value, predicted_value)

   Suppose we got an output of r2_score is 0.55592 then it means 55% change occurs in output with the change in input variable.