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: 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

   Answer: A

3. A line falls from left to right if a slope is _____?

   A) Positive

   B) Negative

   C) Zero

   D) Undefined

   Answer: B

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

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

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: A

7. Lasso and Ridge regression techniques belong to _____?

   A) Cross validation

   B) Removing outliers

   C) SMOTE

   D) Regularization

   Answer:D

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

   A) Cross validation

   B) Regularization

   C) Kernel

   D) SMOTE

   Answer:A

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: A

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

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:B

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: A,B

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

13. . Explain the term regularization?

Answer: While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model. Or 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.

14. Which particular algorithms are used for regularization?

Answer:

1.Ridge Regression

2.LASSO (Least Absolute Shrinkage and Selection Operator) Regression

Ridge Regression

Ridge regression is a method for analyzing data that suffer from multi-collinearity.

$$Loss = \sum_{i=1}^{n} (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^{n} w_i^2$$

Loss Function for Ridge Regression

Ridge regression adds a penalty ***(L2 penalty)*** to the loss function that is equivalent to the square of the magnitude of the coefficients.

The regularization parameter *(λ)* regularizes the coefficients such that if the coefficients take large values, the loss function is penalized.

i. ***λ → 0***, the penalty term has no effect, and the estimates produced by ridge regression will be equal to least-squares i.e. the loss function resembles the loss function of the Linear Regression algorithm. Hence, a lower value of λ will resemble a model close to the Linear regression model.

ii. ***λ → ∞,*** the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will **approach zero** (coefficients are close to zero, but not zero).

*Note: Ridge regression is also known as the **L2 Regularization.***

To sum up, Ridge regression shrinks the coefficients as it helps to reduce the model complexity and multi-collinearity.

LASSO Regression

LASSO is a regression analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model.

$$Loss = \sum_{i=1}^{n} (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^{n} |w_i|$$

Loss Function for LASSO Regression

LASSO regression adds a penalty *(L1 penalty)* to the loss function that is equivalent to the magnitude of the coefficients.

In LASSO regression, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the regularization parameter λ is sufficiently large.

Note: LASSO regression is also known as the L1 Regularization (L1 penalty).

To sum up, LASSO regression converts coefficients of less important features to zero, which indeed helps in feature selection, and it shrinks the coefficients of remaining features to reduce the model complexity, hence avoiding overfitting.

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

Answer:

An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.

The error term is also known as the residual, disturbance, or remainder term, and is variously represented in models by the letters *e,* ε, or u.

An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

For example, assume there is a multiple linear regression function that takes the following form:

$$Y=\alpha X+\beta \rho+\epsilon$$

where:

α,β=Constant parametersX,ρ=Independent variablesϵ=Error term

When the actual Y differs from the expected or predicted Y in the model during an empirical test, then the error term does not equal 0, which means there are other factors that influence Y.