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 model

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.

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

13. Explain the term regularization?

Answer: As we train a model in ML, overfitting or underfitting is generally a problem where a model performs well on a training data but when it comes to new data it doesn't perform well. To Overcome this issue of overfitting in a ML model to an extend we generally use the set of methods in Regularization. The need of regularization in machine learning is that the model does not become excessively complex and overfit the training data as well as the model gets to capture the pattern from the training data which can be inexplicit to unseen data. To minor changes in the training set the regularization reduces the sensitivity of modal outputs. While dealing with limited data, regularization prevents the model to get into complex form. This also helps to get a balance between the error made by the models(bias) and how much the model changes with the change in the training data(variance).

14. Which particular algorithms are used for regularization?

Answer: Lasso (Least Absolute Shrinkage and selection operator) is a formula whose main purpose is the feature(variable) selection and regularization of data models. This method regularizes the model parameters by shrinking the coefficient of regression to an extent to zero, which will reduce the changes in the training data(variance) and minimizes the error made by the models(bias). The feature selection occurs after the shrinkage, i.e. every non-zero value is selected to be used in the model where choosing the right variable helps to determine the accuracy of the model. This can perform feature selection, simplifying the model and potentially improving interpretability.

Ridge Regression is a formula whose main purpose is to regularize the data models that has the high multicollinearity (meaning that some of the features are highly correlated to each other) by shrinking the coefficient without the feature selection i.e. it does not force the coefficient to be zero. This deals particularly well with multicollinearity and performs well when there are more features than observations.

Elastic Net Regression is the technique that combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of models i.e. it first finds the ridge regression coefficients and then conducts the second step by using a lasso sort of shrinkage of the coefficients. This method performs feature selection and regularization simultaneously. The grouping effect helps the features to be easily identified using correlation, enhancing the sampling procedure. It also increases the number of features selected. When one feature is sampled in a highly correlated group, all the other features in that group are automatically added to the sample.

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

Answer: An error term appears in a regression model, to indicate the uncertainty in the model. An error term represents the way observed data differs from the actual population i.e. a variable which represents how a given ML model differs from reality. An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. An error term is generally unobservable.