Machine Learning – WORKSHEET 1

Q1 to Q11 have only one correct answer.

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

   Ans: A) Least Square Error

2. Which of the following statement is true about outliers in linear regression?

   Ans: A) Linear regression is sensitive to outliers

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

   Ans: B) Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?

   Ans: C) Both of them (Regression & Correlation)

5. Which of the following is the reason for over fitting condition?

   Ans: C) Low bias and high variance

6. If output involves label then that model is called as:

   Ans: B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

   Ans: D) Regularization

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

   Ans: D) SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?
   Ans: A) TPR and FPR

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less

    Ans: B) False

11. Pick the feature extraction from below:

    Ans: B) Apply PCA to project high dimensional data

In Q12, more than one options are correct

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Ans: 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.

Q13and Q15 are subjective answer type questions,

13.Explain the term regularization?

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. The problem with overfitting model is that the model may predict with great accuracy for training data, but on test data the model may not perform well. That is the model is biased towards training data and may have high variance with respect to test data. On the other hand, if the prediction is low on training data as well as on the test data then it is a case of under fitting the model. In ML the objective must be to arrive at a generalized model. Usually, we have to build models that make accurate prediction using training set. A generalized model is a good model when the model has low bias and low variance. This may be achieved by introducing a small amount of bias in the train model by readjusting the slope of the line. Lasso and Ridge regression techniques may be used in regularization.

Overfitting: It is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data.

Underfitting: Choosing too simple a model is called underfitting. Underfit model performs badly even on training data. The prediction score is very less.

14.Which particular algorithms are used for regularization?

The algorithms used in regularization are

Lasso Regression: A regression model which uses L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression. It is used to reduce the complexity of the model.

Ridge Regression: Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function(L). Ridge regression is one of the types of linear regression in which we introduce a small amount of bias, known as Ridge regression penalty so that we can get better long-term predictions. In Statistics, it is known as the L-2 norm.

Elasticnet: It's a combination of both Lasso & Ridge Regression

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

- The error is defined as the difference between the actual value and the predicted value. i.e. **error = actual value – predicted value.**
- The objective is to minimize the error so that model performs well on the test data.
- The error in case of linear regression is the difference between the actual data points from that of best fit linear curve from the data points.