MACHINE LEARNING Worksheet_1

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: - **B) Correlation** (Symmetric = Correlation and Asymmetric =Regression)

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: - **A) Cross validation**

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**

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.

13) Explain the term regularization?

Ans: - Regularization is a technique used to reduce the appropriately errors by fitting a function to a given training set and avoid overfitting.

Using Regularization, we can fit our machine learning model appropriately on a given test set and thus reduce the errors in it.

Overfitting   Good Fit   Underfitting
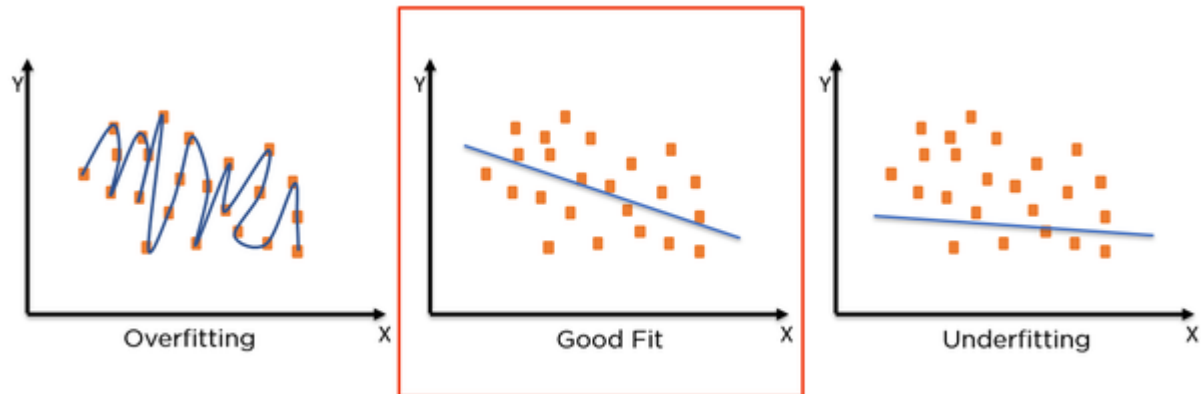
This technique allows to maintain all the variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model. In this we reduce the magnitude of the features by keeping the same number of features.

There are mainly two types of regularization techniques, which are given below:

- o **Ridge Regression**
- o **Lasso Regression**
- o **Dropout**

14) Which particular algorithms are used for regularization?

Ans: -

- **Ridge Regression: -** It is also called L2 Norm or Regularization. Ridge regression is a method for analyzing data that suffer from multi-collinearity.

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

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

- **Lasso Regression: -** This analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model. It is also known as the L1 Regularization

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

It 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?

Ans: - **Error term** in statistics is a value which represents how observed data differs from actual population data. The error term is often written 'ε'.

The error term includes everything that separates your model from actual reality. This means that it will reflect nonlinearities, unpredictable effects, measurement errors, Functional Misspecification, Model Misspecification and omitted variables. This can range from being relatively small to huge, even within one model, across the observed data points.

This error term helps in the calculation of the R-squared value, that is, it tells us how good the model is overall. If the R-squared value of the model is 0.8, then your model explains 80% of the variation in your target variable.