 ASSIGNMENT – 39

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?
   1. Least Square Error B) Maximum Likelihood

C) Logarithmic Loss D) Both A and B

ANS . A

1. Which of the following statement is true about outliers in linear regression?
   1. Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers

C) Can’t say D) none of these

ANS . A

1. A line falls from left to right if a slope is \_\_\_\_\_\_?
   1. Positive B) Negative C) Zero D) Undefined

ANS . B

1. Which of the following will have symmetric relation between dependent variable and independent variable?
   1. Regression B) Correlation C) Both of them D) None of these

ANS . B

1. Which of the following is the reason for over fitting condition?
   1. High bias and high variance B) Low bias and low variance

C) Low bias and high variance D) none of these

ANS . C



1. If output involves label then that model is called as:
   1. Descriptive model B) Predictive modal

C) Reinforcement learning D) All of the above

ANS . B

1. Lasso and Ridge regression techniques belong to \_\_\_\_\_\_\_\_\_?
   1. Cross validation B) Removing outliers

C) SMOTE D) Regularization

ANS . D

1. To overcome with imbalance dataset which technique can be used?
   1. Cross validation B) Regularization

C) Kernel D) SMOTE

ANS . D

1. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?
   1. TPR and FPR B) Sensitivity and precision

C) Sensitivity and Specificity D) Recall and precision

ANS . C

1. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
   1. True B) False

Ans B

1. Pick the feature extraction from below:

A) Construction bag of words from a email

* 1. Apply PCA to project high dimensional data
  2. Removing stop words
  3. Forward selection

Ans

In Q12, more than one options are correct, choose all the correct options:

1. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
   1. We don’t have to choose the learning rate.
   2. It becomes slow when number of features is very large.
   3. We need to iterate.
   4. It does not make use of dependent variable.

Ans A, B, C

 ASSIGNMENT – 39

MACHINE LEARNING

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

1. Explain the term regularization?

ANS) 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.

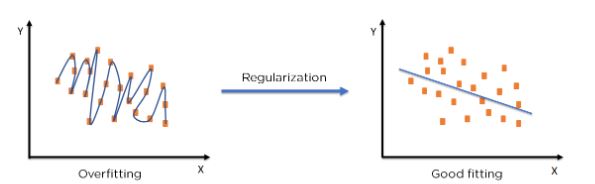


                                                  Figure 5: Regularization on an over-fitted model

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

1. Which particular algorithms are used for regularization?

Ans **There are three main regularization techniques, namely:**

* Ridge Regression (L2 Norm)
* Lasso (L1 Norm)
* Dropout.

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

Ans 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.

### **KEY TAKEAWAYS**

* An error term appears in a statistical model, like a regression model, to indicate the uncertainty in the model.
* The error term is a residual variable that accounts for a lack of perfect goodness of fit.
* Heteroskedastic refers to a condition in which the variance of the residual term, or error term, in a regression model varies widely.

## Understanding an Error Term

An error term represents the margin of error within a statistical model; it refers to the [sum of the deviations](https://www.investopedia.com/terms/s/sum-of-squares.asp) within the [regression line](https://www.investopedia.com/terms/r/regression.asp), 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.

## Error Term Use in a Formula

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a [multiple linear regression](https://www.investopedia.com/terms/m/mlr.asp) function that takes the following form:

​*Y*=*αX*+*βρ*+*ϵ*

**where:**

*α*,*β*=Constant parameters

*X*,*ρ*=Independent variables

*ϵ*=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