MACHINE LEARNING - WORKSHEET1

1) (a) Least square error
2) (a) Linear regression is sensitive to outliers
3) (b) Negative
4) (b) Correlation
5) (c) Low bias & high variance
6) (b) Predictive model
7) (d) Regularization
8) (d) SMOTE
9) (a) TPR AND FPR
10) (b) False

11) (b) Apply PCA to project high dimensional data
12) (a) we don't have to choose the learning rate.
(b) it becomes slow when number of features is very large.

13) Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all 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.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features.*"

Techniques of Regularization:

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

- **Ridge Regression**
- **Lasso Regression**

14) Ridge Regularization :

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present. The function depicted below shows the cost function of ridge regression :

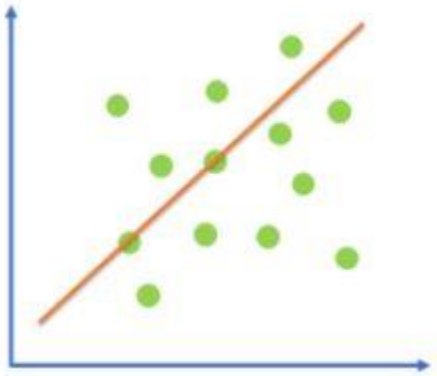
Cost function = Loss + λ x $\sum \|w\|^2$

Here,

Loss = Sum of the squared residuals

λ = Penalty for the errors

W = slope of the curve/ line



Lasso Regression :

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.

Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients. Consider the cost function for Lasso regression :

Cost function = Loss + λ x $\sum \|w\|$

Here,

Loss = Sum of the squared residuals

λ = Penalty for the errors

w = slope of the curve/ line

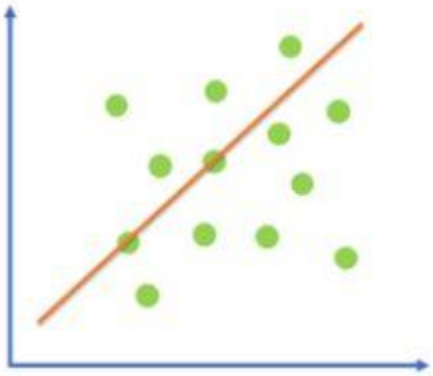


Figure 11: Cost function for Lasso Regression

15) 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 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.

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.

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 function that takes the following form:

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

where:

α,β=Constant parameters

X,ρ=Independent variables

ϵ=Error term