SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)
Re - accredited by NAAC with ‘A’ Grade
Founder: Prof.Dr. S. B. Mujumdar, M.Sc.,Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Name: Naivedya Rai
PRN: 20070122083

Machine Learning Lab Assignment 6

Aim: Apply Linear Regression algorithm on a sample case study and data set. Evaluate results

Theory:

What are linear models?

The Linear Model is one of the most straightforward models in machine learning. It is the building block for many complex machine learning algorithms, including deep neural networks. Linear models predict the target variable using a linear function of the input features. In this article, we will cover two crucial linear models in machine learning: linear regression and logistic regression. Linear regression is used for regression tasks, whereas logistic regression is a classification algorithm.

The linear model is one of the most simple models in machine learning. It assumes that the data is linearly separable and tries to learn the weight of each feature. Mathematically, it can be written as Y=W^T.X, where X is the feature matrix, Y is the target variable, and W is the learned weight vector. We apply a transformation function or a threshold for the classification problem to convert the continuous-valued variable Y into a discrete category. Here we will briefly learn linear and logistic regression, which are the regression and classification task models, respectively.

The two types of linear models are:

1. Linear Regression
2. Logistic Regression

About Linear Regression:
Linear regression is a statistical method used for modeling the relationship between a dependent variable (also known as the target or outcome) and one or more independent variables (also known as predictors or features). It's a fundamental technique in the field of machine learning and statistics, primarily used for predictive analysis and understanding the relationships between variables.

Here's a breakdown of how linear regression works:

1. Model Representation:

In simple linear regression, there is one dependent variable (Y) and one independent variable (X). The relationship between them is represented as:

$Y = \beta 0 + \beta 1X + \varepsilon$

Y represents the dependent variable (the variable you want to predict).
X represents the independent variable (the input or predictor variable).
$\beta 0$ is the intercept, the point where the regression line intersects the Y-axis.
$\beta 1$ is the slope, which represents the change in Y for a unit change in X.
ε represents the error term, which accounts for the variability in Y that cannot be explained by the linear relationship with X.

2. Objective:
The objective of linear regression is to find the best-fitting linear equation that minimizes the sum of the squared errors (the ε term) between the predicted values ($\hat{Y}$) and the actual values (Y) in the dataset.

3. Fitting the Model:
To find the values of $\beta 0$ and $\beta 1$ that minimize the error, various mathematical techniques can be used, but the most common one is the method of least squares. This method minimizes the sum of the squared differences between the observed Y values and the predicted Y values for each data point.

4. Making Predictions:
Once the model parameters ($\beta 0$ and $\beta 1$) are determined, you can use the model to make predictions for new or unseen data. Given a value of X, you can calculate the corresponding predicted value of Y using the linear equation.

5. Model Evaluation:

Linear regression models are often evaluated using metrics such as the coefficient of determination (R-squared), mean squared error (MSE), or mean absolute error (MAE). These metrics help assess the model's performance and how well it fits the data.

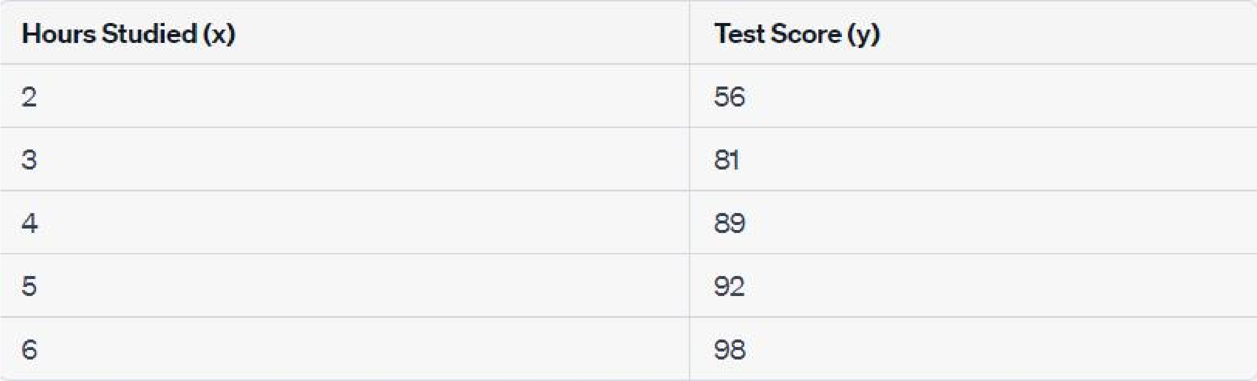
Assumptions of Linear Regression:

Linearity: The relationship between the variables is assumed to be linear.
Independence: The errors (ε) are assumed to be independent of each other.
Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
Normality: The errors are normally distributed.

Example:

To predict the score a student might get on a test based on the number of hours they study using linear regression.

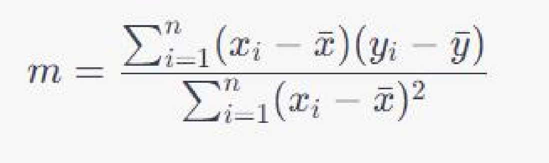
Sample Dataset:



Hours Studied (x)	Test Score (y)
2	56
3	81
4	89
5	92
6	98

The objective is to find the linear relationship between the independent variable, Hours Studied(x), and the dependent variable, Test Score(y). This can be expressed in the form of a linear equation y = mx +b, where m is the slope, b is the y-intercept.

Step 1: Calculate the average of x and y
We get, avg(x)=4 and avg(y)=83.2
Step 2: Calculate the slope(m):
Using the formula,



$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We get m=6.9

Step 3: Calculate the y-intercept

Using the formula,

b = y - mx

We get b=55.6

Therefore, the linear regression equation for the given problem statement will be:

y = 6.9x + 55.6

Metrics for evaluation of Linear Regression Model

1. Mean Absolute Error (MAE)
   MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
   To better understand, let's take an example where you have input data and output data and use Linear Regression, which draws a best-fit line.
   Now you have to find the MAE of your model which is basically a mistake made by the model known as an error. Now find the difference between the actual value and predicted value that is an absolute error but we have to find the mean absolute of the complete dataset. So, sum all the errors and divide them by a total number of observations. This is MAE. And we aim to get a minimum MAE because this is a loss.
2. Mean Squared Error (MSE)
   MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value. So, above we are finding the absolute difference and here we are finding the squared difference.
   What actually the MSE represents? It represents the squared distance between actual and predicted values. We perform squaring to avoid the cancellation of negative terms and it is the benefit of MSE.
3. Root Mean Squared Error (RMSE)
   As RMSE is clear by the name itself, that it is a simple square root of mean squared error.
4. Root Mean Squared Log Error (RMSLE)
   Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

Linear Regression Code:

Dataset used: The dataset used is an “Advertising” dataset. The aim was to find the relationship between the independent variables ,i.e, T.V, Radio, And Newspaper and the dependent variable ,i.e, Sales

Code: https://github.com/Naivedya-Rai/ML-Algo-Implement/blob/main/Lab6_Regression.ipynb