

 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression

IDE: Google Colab

Theory:

Linear regression is a method for determining the best linear relationship between two variables X and Y . If variables X and Y are uncorrelated, it is pointless embarking upon linear regression. However, if a reasonable degree of correlation exists between X and Y then linear regression may be a useful means to describe the relationship between the two variables. The usual approach is to use the *least-squares* method, which minimizes the squared difference between the actual data points and a straight line. Let $[x_i, y_i]$, $i = 1, 2, 3, \dots, N$ be the N pairs of data values of the variables X and Y . The straight-line relating X and Y is $y = mx + c$, where m and c are the gradient and constant values (to be determined) defining the straight line. Thus, $y(x_i) - y_i$ is the difference between the line and data point i (see Fig. 1). Taking all the data points, we seek values of m and c that minimize the squared difference SD .

$$\sum_1^N [y(x_i) - y_i]^2$$

This is achieved by calculating the partial derivatives of SD with respect to m and c and finding the pair $[m, c]$ such that SD is at a minimum.

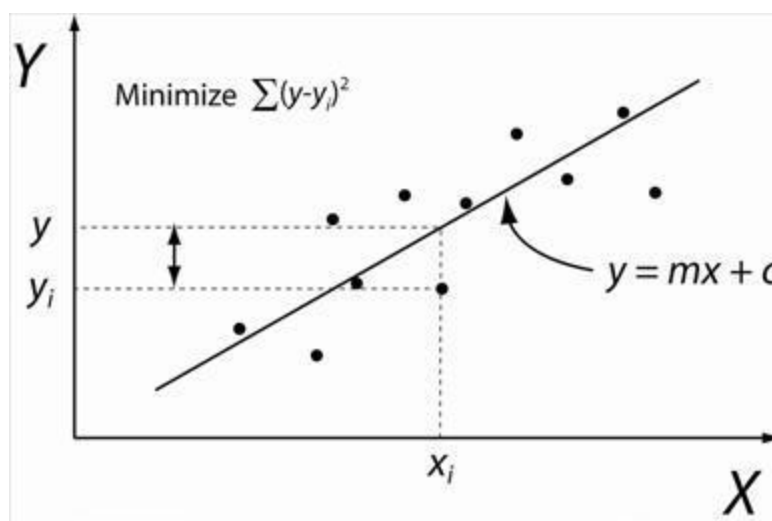



Figure 1: Illustration of Linear Regression. Linear least squares regression, the idea is to find the line $y = mx + c$ that minimizes the mean squared difference between the line and the data points

 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

Batch Gradient Descent:

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model. Batch gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive.

Let m be the number of training examples. Let n be the number of features.

Algorithm for batch gradient descent :

Let $h_{\theta}(x)$ be the hypothesis for linear regression. Then, the cost function is given by:

Let Σ represents the sum of all training examples from $i=1$ to m .

$$J_{\text{train}}(\theta) = (1/2m) \Sigma (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Repeat {

$$\theta_j = \theta_j - (\text{learning rate}/m) * \Sigma (h_{\theta}(x^{(i)}) - y^{(i)})x_j^{(i)}$$

For every $j = 0 \dots n$

}

Where $x_j^{(i)}$ Represents the j^{th} feature of the i^{th} training example. So if m is very large (e.g. 5 million training samples), then it takes hours or even days to converge to the global minimum. That's why for large datasets, it is not recommended to use batch gradient descent as it slows down the learning.

Program (Code):

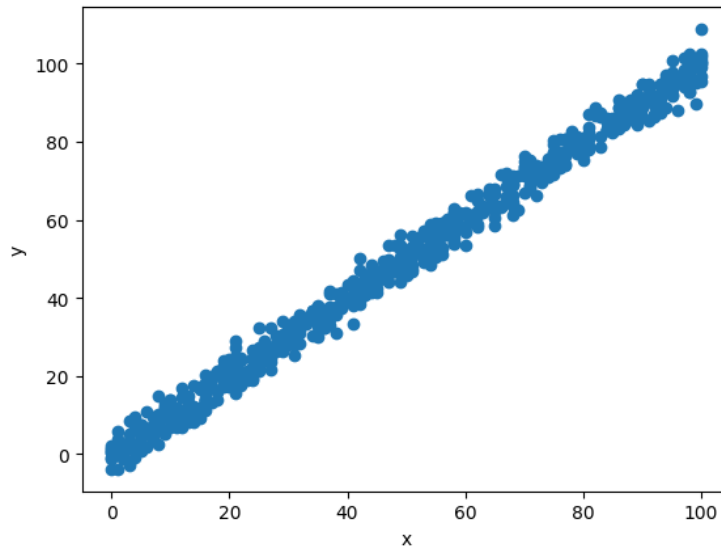
1. Load the basic libraries and packages
2. Load the dataset
3. Analyse the dataset
4. Pre-process the data
5. Visualize the Data
6. Separate the feature and prediction value columns
7. Write the Hypothesis Function
8. Write the Cost Function
9. Write the Gradient Descent optimization algorithm
10. Apply the training over the dataset to minimize the loss
11. Find the best fit line to the given dataset
12. Observe the cost function vs iterations learning curve

 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

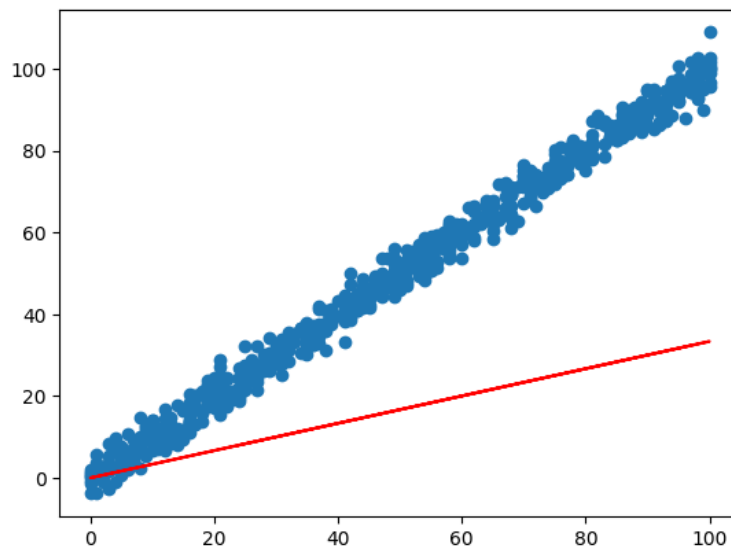
Results:


To be attached with

- Datapoints scattering (without best fit line)

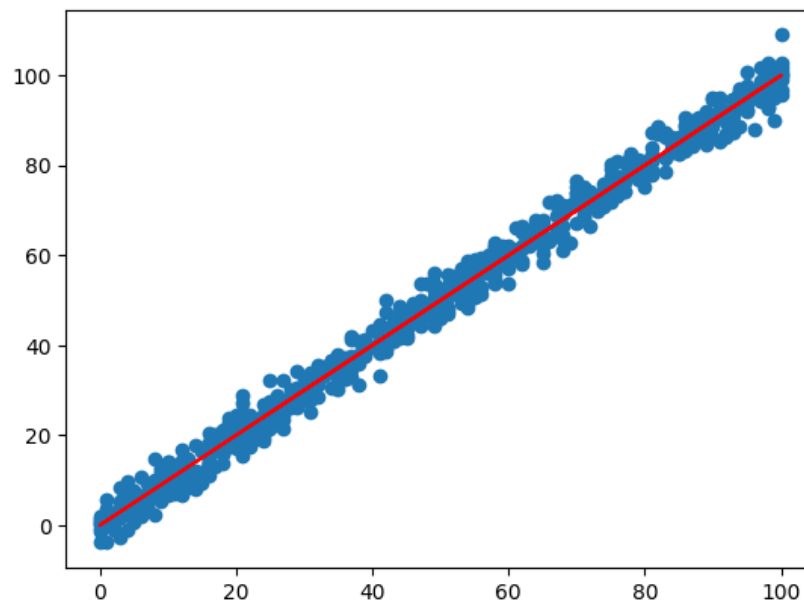


- Scatter plot showing the Best fit line in the 1st iteration of training

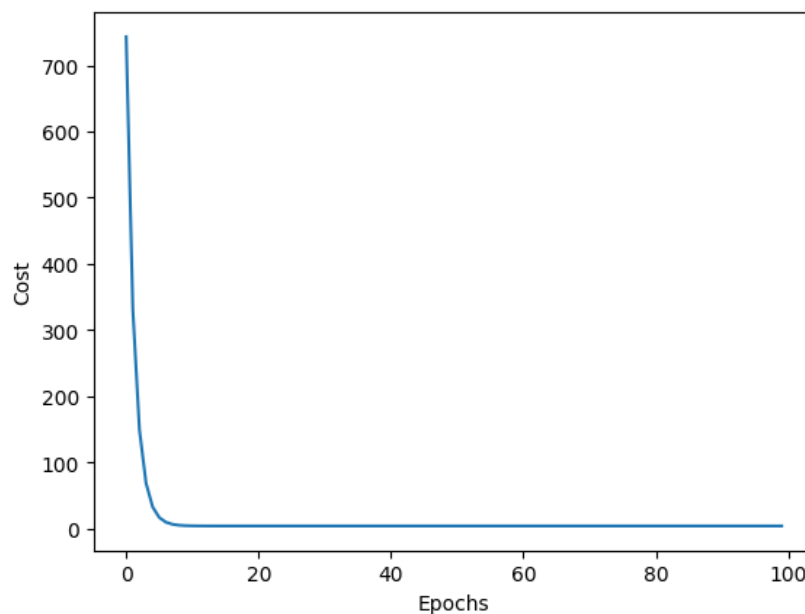



 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

c. Scatter plot showing the Best fit line in the last iteration of training



d. Learning Curve (Cost function vs iterations)



 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

Observation and Result Analysis:

a. Nature of the dataset

b. During Training Process

c. After the training Process

d. Observation over the Learning Curve

 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

Post Lab Exercise:

- a. Write three applications of linear regression

- b. Write three advantages of linear regression


- c. Write three limitations of linear regression

- d. What are the major assumptions considered in linear regression

- e. Why MSE is used instead of MAE for calculating the loss function

- f. How can the behavior of outliers be understood while dealing with the unseen dataset

- g. Derive the Normal Equation for the Linear Regression.

 Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Machine Learning (01CT0519)	Aim: To obtain the best fit line over single feature scattered datapoints using Linear Regression	
Experiment No: 02	Date: 13-08-2024	Enrollment No: 92200133030

Post Lab Activity:

Consider any dataset from <https://archive.ics.uci.edu/ml/datasets> and perform the linear regression analysis over the dataset and obtain the best fit line. Make sure that the dataset is not matching with your classmates. You can also select the dataset from other ML repositories with prior permission from your concerned subject faculty.