Introduction to Machine Learning

Fall 21-22

Assignment 1 – Linear Regression

STUDENT INFORMATION		
Student No	Name and Surname	Dept.
63180072	Onur TOSUN	IE

LINEAR REGRESSION

OBJECTIVE

1. To find finds the best linear line between two variables
2. To estimate the relationship between two variables
3. Learning to construct linear functions

THEORY

Linear regression is a supervised learning algorithm. It is used for forecasting or estimating the relationship between independent and dependent variable in especially statistics and machine learning. There are two types of linear regression which are called “simple linear regression” and “multiple linear regression”. In Simple Linear Regression, there is only one independent variable, and the model must find its linear relationship with the dependent variable. Whereas the multiple linear regression has more independent variable than one as understood from its name. Both regression types are doing the same process. There is an independent variable which usually is called “X”, and dependent variable which usually is called “y”. The linear regression operation predicts the dependent variable y according to independent variable X and calculates the best fit line for given model. Thus, it called as linear regression. The linear regression model defines as the following formula.

$$Y = mx + b$$

For multiple regression,

$$Y = m_0 + m_1x_1 + m_2x_3 + m_3x_3 \ldots + m_nx_n$$

where is the Y is a dependent variable, x is an independent variable, m is a slope(coefficient), b is an intercept, ϵ is an error.

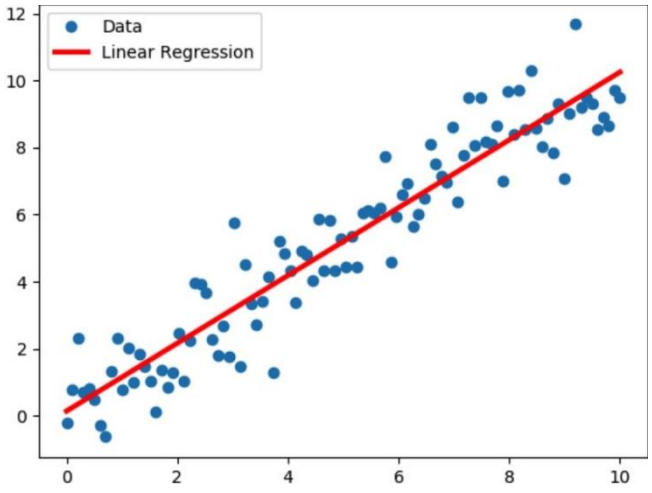


Figure 1. Linear Regression Graph Example

LINEAR REGRESSION FUNDAMENTALS IN MACHINE LEARNING

Hypothesis

In Machine Learning, Hypothesis space is a manageable subset of all possible solutions. [4]

$$h^w(x) = \sum_{j=0} w_j x_j = w^t x$$

Loss Function

Loss function shows that the difference between input and output values. If the predicted values are different too much from actual values, that's mean the model is not good enough.

$$Squared\ error: L\ (\ y\ ,h(x)\ )\ =\ (\ y\ -\ h(x)\ )2$$

Empirical loss

Empirical loss is the average loss over the data points.[1]

$$(1/N) \sum_{i=1}^{N} L((x^i, y^i), h(\cdot))$$

Where N is the length of the data

Training

Used for minimizing the empirical loss to finding the best predictor. In training, A certain part of the data is taken for training and tested. If the loss is high, new data point are added or the model is changed. [4]

Expected Loss

Because the training set is optimized, the empirical loss may not be representative of how well the model will perform on new samples. Therefore, the model is tested on a new set which is not used in during the training.[4]

PROCEDURE

Preparation

First, we import the data (Sklearn.diabetes) we will use, and the libraries such as numpy, matplotlib. After that we split diabetes_X and diabetes_y for 80% X values for training and 20% y values for testing. There must be four data which are named training and testing for X, training, and testing for y.

We used training data for train to data, also we used the data we trained for testing.

Q1. Direct Solution

Direct solution

We calculate the coefficient and intercept by using direct solution formula. Our coefficient is 957.00838947 and Intercept is 152.08225581

```
******** Question 1 *********
 --- Direct Solution ---
Coefficient : 957.0083894795963
Intercept : 152.0822558150509
****************************
```

Q2. Gradient Solution

Our step size (learning rate) is 0.1, iteration is 100000. When iteration size increased, the cost values are reducing as we already expected. The coefficient is 957.00838899, Intercept is 152.08225581, Cost is 4124.82.

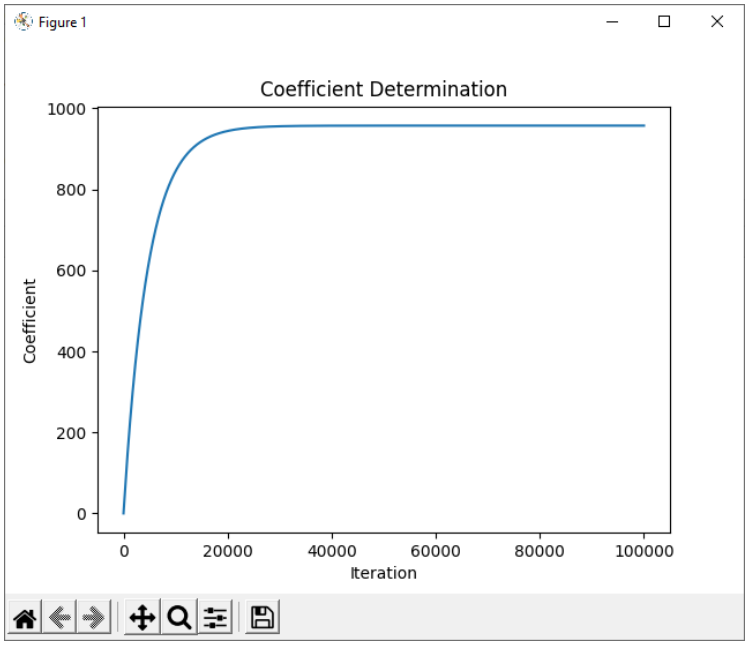
```
******** Question 2 and 3 *********
Iteration 0: 28701.9011299435
Iteration 10000: 3859.483123787407
Iteration 20000: 3832.7383353511755
Iteration 30000: 3832.3683765354513
Iteration 40000: 3832.363258920643
Iteration 50000: 3832.3631881290303
Iteration 60000: 3832.3631871497746
Iteration 70000: 3832.3631871362286
Iteration 80000: 3832.3631871360412
Iteration 90000: 3832.3631871360385

--- Gradient Descent Solution ---
Coefficient:  [957.00838899]
Intercept:  [152.08225581]
Mean squared error: 4124.82
```

Q3. Coefficient and Cost Graphics

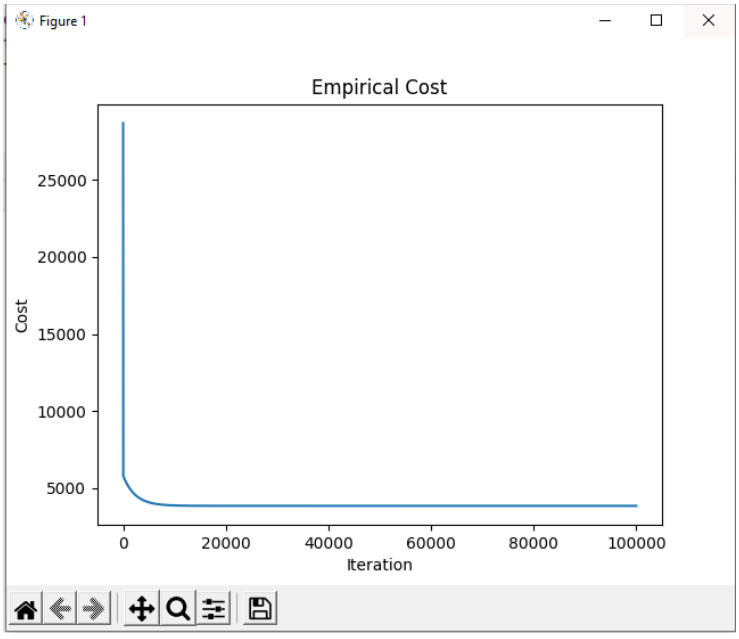
Coefficient Graph

As the number of iterations increases, the coefficient value increases rapidly up to a certain point and after approximately 20000 iterations, it slows down in the same course. So, We can determine the our coefficient value.



Cost Graph

As the number of iterations increases, the cost value are rapidly decreases up to a certain point and after 10000 iterations, the values are starting to change very slowly.



Q4. Sklearn Calculations

The coefficient, intercept, MSE and R2 score are values as the same as we calculated except for the slightly deviation in coefficient.

My Calculation

```
******** Question 5 *********
 --- My Calculation ---
Coefficient of determination: 0.36
*****************************

Coefficient:  [957.00838899]
Intercept:  [152.08225581]
Mean squared error: 4124.82
```

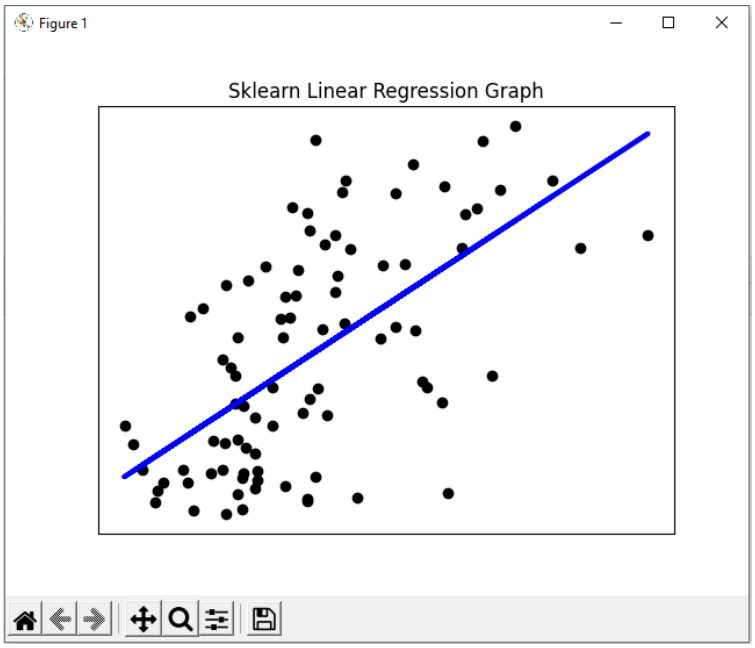
(From gradient question)

Sklearn Calculation

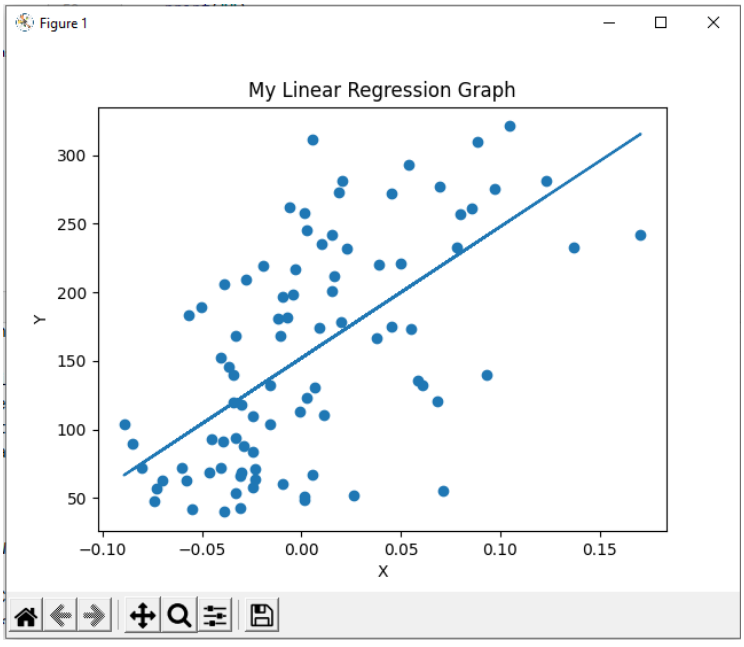
```
******** Question 4 *********
 --- Sklearn Calculation ---
Coefficients:  [957.00838948]
Intercept:  152.08225581505087
Mean squared error: 4124.82
Coefficient of determination: 0.36
*****************************
```

Linear Regression Graphs

Sklearn Graph



My Graph



CONCLUSION

In this assignment, we leant what is linear regression, what benefit does it give us, what the fundamentals of the regression, how we apply the regression etc.

Linear Regression is an algorithm to find the relationship between two variables. These are called independent and dependent variables. Dependent variables denoted by “y”, as understood from the name, it is an output variable. So, there are some variables enter which is name independent variables, denoted by “X”, a linear regression model and sets the value of y. So, we build our model by calculating the relationship between the input and output values as well. There are two methods to construct a model in regression. One of them is direct solution. This method finds the coefficient and intercept directly but not always working. Because to apply this method, our matrix multiplication must be square matrix to get inverse of them. If we do not always have a square matrix, we cannot apply this method.

Other method is gradient descent method. Gradient descent is an iterative optimization algorithm to find the minimum of a function. Gradient descent widely used algorithms in machine learning, mainly because it can be applied to any function to optimize it. [5] In this method, we can calculate all values such as coefficient, intercept, empirical loss, MRE (Cost) etc. The method uses learning rate and iteration. In our experiment, when we increased the learning rate to 0.1 and increased the iteration number, we got a smoother result.

In both methods, our model was formed as follows.

$$y = 957.00838899 * \mathrm{x} + 152.08225581$$

In this model, the coefficient value is 957.00838899, and the intercept value is 152.08225581. This model forms the linear regression line in our graph. So, when we give x a number, this result will give us the closest value to the real value of y.

REFERANCES

1. Sanjay Lall and Stephen Boyd, Supervised Learning via Empirical Risk Minimization, EE104 Stanford University, EE104 Spring 2018
2. Mohit Gupta, geeksforgeeks.org, ML | Linear Regression, 13 Sep 2018
3. Deepanshi, analyticsvidhya.com, All you need to know about your first Machine Learning model – Linear Regression, May 25, 2021
4. Dr. Bahadir K. Gunturk, Introduction, Elements of Machine Learning, Istanbul Medipol University
5. Adarsh Menon, towardsdatascience.com, Linear Regression using Gradient Descent, Sep 16, 2018