Labwork 2: Linear Regression

1. Introduction

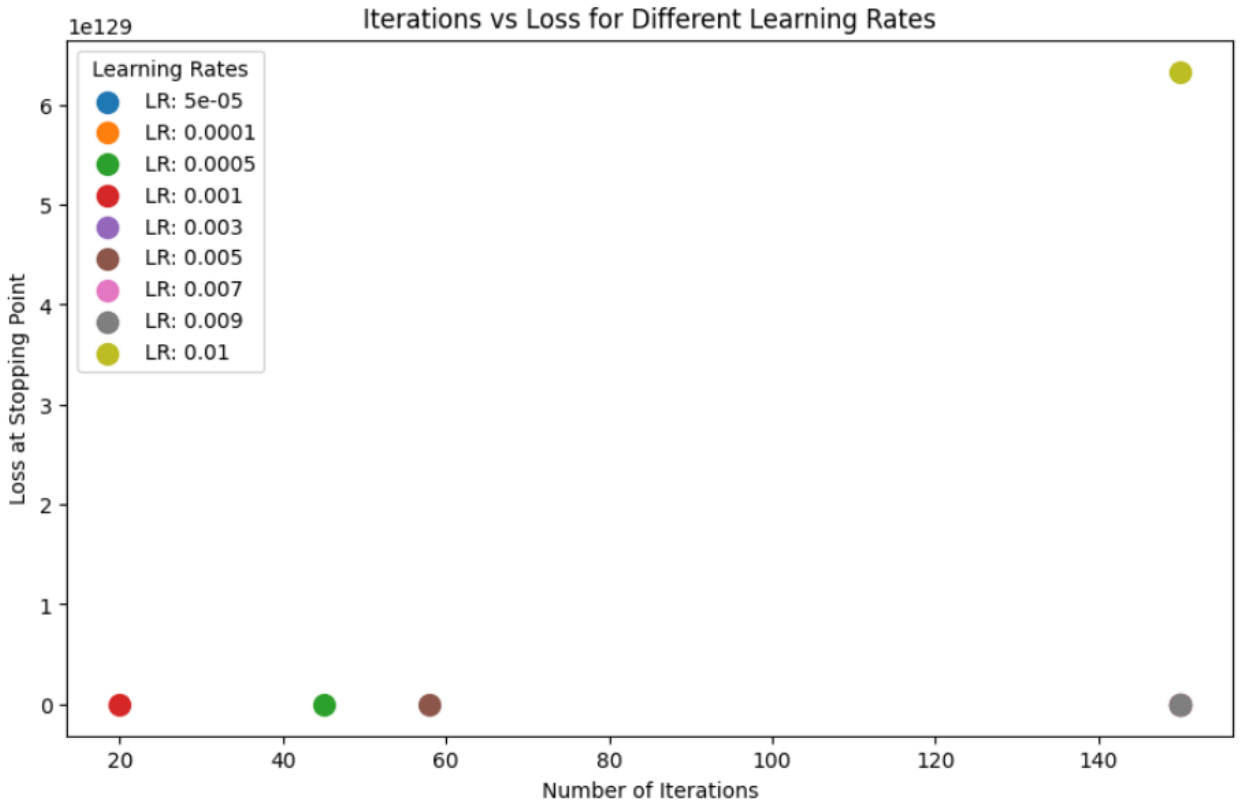
Linear regression: Linear regression analysis is used to predict the value of a variable based on the value of another variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

2. Implementation

- Load 2 dimensional data. Then assign values of x and y to 2 arrays.
- Define f(x) – that calculate y from x, with y = w1*x + w0.
- Define loss function – mean squared error
- Define the function for calculating the partial derivative of loss following w0 and w1.
- Define gradient descent function to update the values of w0 and w1 to find the opimized value of w1 and w0. In this function, we need to set input parameters are x, y, w0, w1, maximum of iterations, learning rate, and threshold. Loop until we have error smaller than threshold.
- Define print and plot functions to see the result, visualize the result, and make comparisons.
- Define the list of learning rates then evaluate the corresponding error.
- Change the value of w1 and w0, then evaluate the effect of learning rates on new parameters.

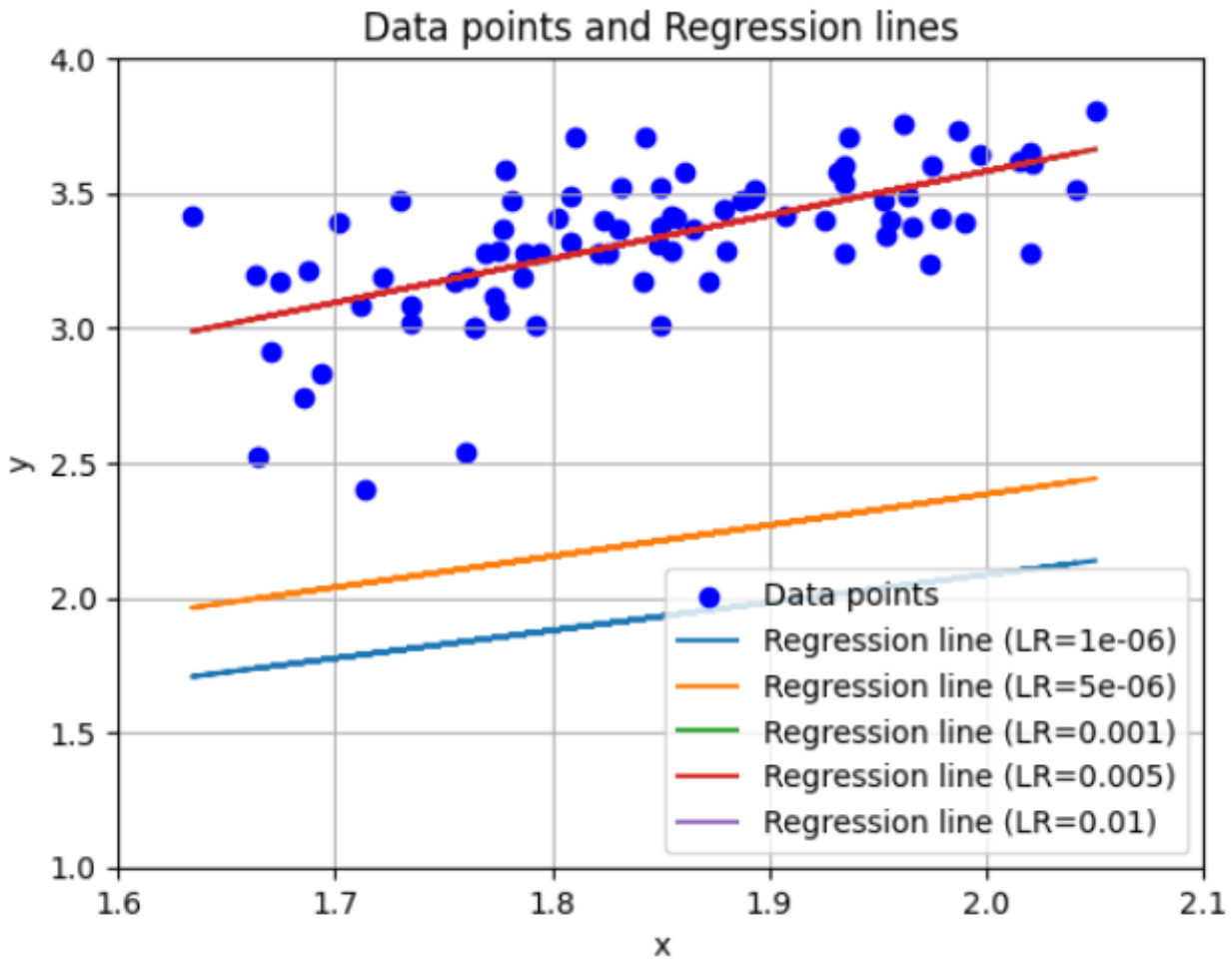
3. Evaluation

We need to choose an appropriate learning rate to minimize the error of the model. It means that we need to choose the learning rate having the lowest error and lowest iteration.
With w1 = 1 and w0 = 0, and a set of learning rates, we have the errors illustrated below:



From the above evaluation of Loss and Number of Iterations, we can easily see that the best learning rate is 0.001, which is represented by the red point in the graph.
With too small or too big value of learning rates, we can see that the loss can not decrease to the expected value, or they waste a lot of iterations which leads to low performance.

To have more comparison, we should show the regression lines on the graph by different values of learning rate.



Data points are represented by blue points, and we can see the regression lines with a set of learning rates. The best one has learning rate = 0.001, and the line really fit to the data. The other lines belong to other learning rates and have worse results, which lead to bigger errors.

4. Conclusion

Linear regression can be optimized by gradient descent, but we need to choose the appropriate value of learning rate by selecting from comparison. With more dimensions of the data, we need to calculate more partial derivatives following the number of data's dimensions.