Linear Regression Report Using Gradient Descent

Introduction

Linear regression is a machine learning technique that models the relationship between a quantitative dependent variable and one or more quantitative independent variables. This technique fits a straight line (y = mx + b) to the data in a way that reduces the error between actual and predicted values. This method works best on data with a clear linear relationship. Using the created linear regression model, future data can be predicted reliably.

In this Python notebook, two methods were used in order to find a linear regression model for a sample dataset. This dataset included 25 students and compared the hours they dedicated to studying to their test score. This dataset showed a clear linear relationship.

Sklearn

To start, the model was created using the built-in linear regression method imported from Sklearn. This method found a slope value of 9.91 and an intercept value of 1.12. The MSE and MAE for the training data were 29.17 and 5.03, respectively.

Gradient Descent

Next, using gradient descent, a second linear regression model was created using the same dataset. Gradient descent works by gradually adjusting the slope and intercept of the line over each iteration in a way that reduces error. It starts with slope and intercept being initialized as 0.0. Each iteration, the MSE is calculated and the partial derivatives of how much the MSE changes with respect to slope and intercept are found. The partial derivative of the cost function with respect to slope is multiplied by the learning rate and subtracted from slope. This is repeated with

intercept, and an iteration is complete. This model had a learning rate of 0.00001 and ran for 100,000 iterations. However, after only 3,000 iterations, the model had converged on a slope and intercept that reduced cost as much as the model was able to. This model had a MSE and MAE of 29.33 and 5.07, respectively.

Conclusion

Both models were able to create similar results. It suggests that within the sample dataset, every hour of studying results in an additional 9.91 grade points on the test. This model can be used to predict future test scores given a student's study time. This demonstration shows the effectiveness and utility of supervised machine learning in a variety of use cases.