Multivariate Linear Regression from scratch

Introduction:

Multivariate linear regression is a statistical technique that is used to model the relationship between a dependent variable and multiple independent variables. It helps in making predictions about the dependent variable based on the values of the independent variables. In this report, we will implement multivariate linear regression from scratch using the Sklearn diabetes dataset.

Dataset:

The Sklearn diabetes dataset contains information about 10 baseline variables that were collected for each of the 442 patients, along with a quantitative measure of the progression of the disease one year later. The variables include age, sex, body mass index, average blood pressure, and six blood serum measurements. The dependent variable is the quantitative measure of disease progression one year later.

Data Preprocessing:

The first step in the implementation of multivariate linear regression is to load the data and preprocess it. The data was split into features and labels, then the dataset was shuffled randomly using the Numpy shuffle method. The data was then split into 60% training set, 20% test set and 20% development set using the Sklearn train\_test\_split method.

Cost Function and Hypothesis Function:

The next step is to define the cost function, mean squared error, and hypothesis function. The cost function represents the error between the actual and predicted values. The mean squared error was used as the cost function in this project. The hypothesis function represents the linear relationship between the dependent and independent variables.

Gradient Descent Function:

The gradient descent function was defined to minimize the cost function. It calculates the derivative of the cost function with respect to the weight and bias and updates the weight and bias accordingly. This process is repeated until the cost function reaches its minimum value.

Training and Predicting Function:

The training and predicting functions were defined to train the model and make predictions using the dev set and test set. The model was trained using the training set and the predicted values were compared with the actual values to calculate the mean squared error.

Learning Rate Plot:

The learning rate is a crucial parameter in the gradient descent function. To determine the best and most stable learning rate, a plotting learning rates function was defined. The function plots the mean squared error for different learning rates and the best learning rate was selected based on the plot. The best learning rate for this project was 0.01.

Results:

The model was trained using the training set and predicted using the dev set which had a mean squared error of 2323.32. The model was finally tested using the test dataset which had a mean squared error of 3165.42.

Conclusion:

This project provides a good understanding of the implementation of multivariate linear regression from scratch.