Tony Stone

Week One Report

Linear Regression

Linear Regression is a process in finding a relationship between existing data (x, y) where x is a predictor, and y is the response. In linear regression, this relationship is represented by a linear equation of the form $y = \beta_0 + \beta_1 x$, where β_0 is the bias, and β_1 is the weight. The goal here is to find optimal values for the bias and the weight in order to minimize the error in predicted values against known values. The overall goal with linear regression is to make predictions for future values.

Experimental Results

In this experiment, we aim to use linear regression in order to determine a relationship between the number of hours a student spends studying and resulting scores. The number of hours is the predictor x, and the score is the response y. With this, we are able to make predictions on a score a student will achieve based on the number of hours they spend studying. **Figure 1**. Shows a scatter plot of the initial dataset.

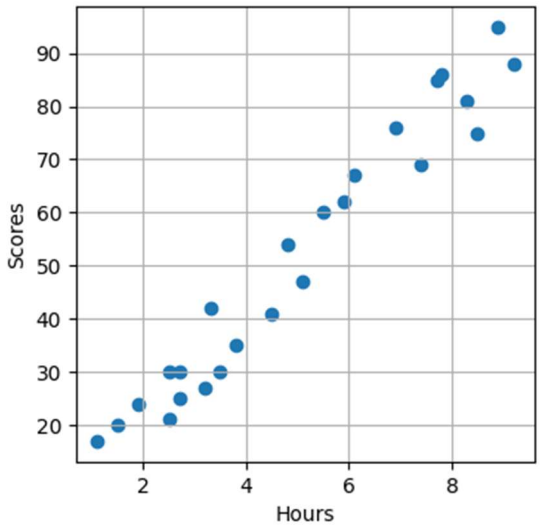

Figure 1. Plot of Hours and Scores

The model used here is the LinearRegression' model by scikit-learn. The metrics used to measure the model's performance are the 'mean_squarred_error' and 'mean_absolute_error' functions, also by scikit-learn.

The data is first split into training and testing data, 80% allocated for training and 20% allocated for testing. After the model is fit with the training data, we can determine the model's bias to be: 2.8269, and the weight to be: 9.6821. Therefore, this linear regression model is of the form $y = 2.8269 + 9.6821\,X$.

The model is then used to make predictions for the training case, and testing case as seen in **Figure 2**, and its performance is evaluated using mean squared error (MSE) and mean absolute error (MAE) as shown in **Figure 3**.

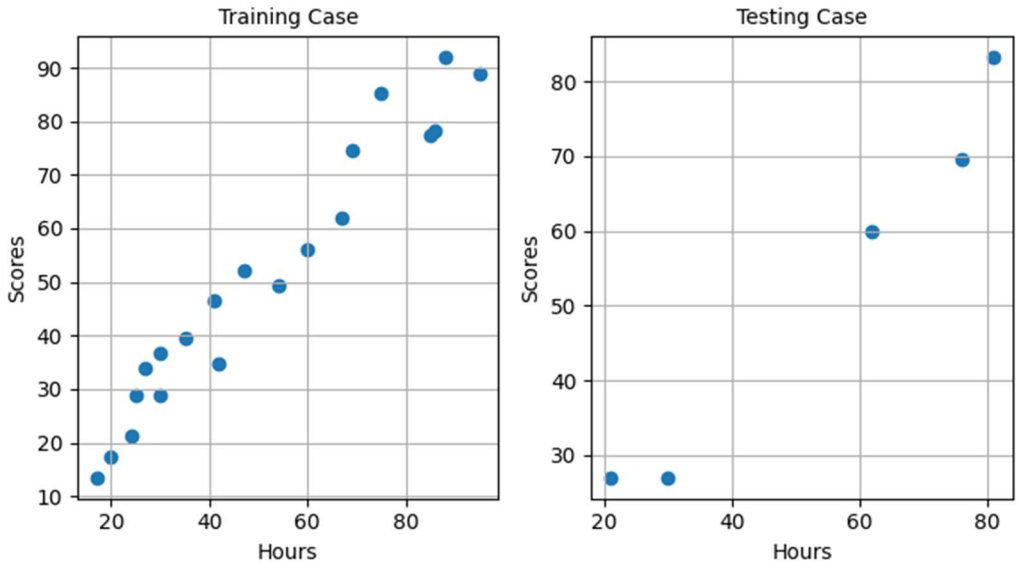

Figure 2. Predictions in Training and Testing Cases

	0	1	2	3
Metric	MSE	MAE	MSE	MAE
DataType	train	test	train	test
Values	31.454836	5.221357	18.943212	3.920751

Figure 3. Model Evaluation

It is important to note that these values are determined with a set random state. Repeating this process with a different random state will result in slightly different results. Especially given that this dataset contains only twenty-five measurements (x, y). There is more variation in how training and testing data gets split among different random states. **Figure 4** shows the same performance metrics with the same model yet run with a different random state. The model performs similarly, though the specific metrics have slightly different values.

	0	1	2	3
Metric	MSE	MAE	MSE	MAE
DataType	train	test	train	test
Values	27.700906	4.767634	34.300751	5.632882

Figure 4. Model Performance with an Alternative Random State

The MAE the model achieved in both random states, as well as in the training and testing case indicates that a prediction for a student's test score for the number of hours they have put into studying is off by about five points on average.