IT 166 Lab 9

In this lab, we will implement a simple linear regression modeul from scratch for a dataset from a CSV file.

Objectives

- Be able to develop a simple linear regression model
- Be able to evaluate the performance of the predictive model

Preparation

- Launch the Jupyter notebook.
- Rename the notebook page as “Lab9”

Simple Linear Regression

Linear regression assumes a linear or straight line relationship between the input variables (X) and the single output variable (y).

More specifically, that output (y) can be calculated from a linear combination of the input variables (X). When there is a *single input variable*, the method is referred to as a **Simple Linear Regression**.

In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.

The line for a simple linear regression model can be written as: $y = b_0 + b_1 \times x$

where b_0 and b_1 are the coefficients we must estimate from the training data.

Once the coefficients are known, we can use this equation to estimate output values for y given new input examples of x.

It requires that you calculate statistical properties from the data such as mean, variance and covariance.

All the algebra has been taken care of and we are left with some arithmetic to implement to estimate the simple linear regression coefficients.

Briefly, we can estimate the coefficients as follows:

$$\begin{aligned} b_1 &= \sum((x_i - \bar{x}) \times (y_i - \bar{y}))/\sum((x_i - \bar{x})^2) \\ &= Covariance(x, y)/Variance(x) \\ b_0 &= \bar{y} - b_1 \times \bar{x} \end{aligned}$$

where i refers to the value of the i^{th} value of the input x or output y. $\bar{x}$ and $\bar{y}$ are the mean value of the input x or output y.

Background:

Covariance & Variance

The covariance of two groups of numbers describes how those numbers change together.

Covariance is a generalization of correlation. Correlation describes the relationship between two groups of numbers, whereas covariance can describe the relationship between two or more groups of numbers.

We can calculate the covariance between two variables as follows:

$covariance = \sum((x_i - \bar{x}) \times (y_i - \bar{y}))/(n-1)$

$variance = \sum((x_i - \bar{x})^2)/(n-1)$

The calculation of covariance between x and y, and the variance of x is the estimation based on the samples of x and y.

Performance Evaluation

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit (i.e., the regression model). Root mean square error is commonly used in forecasting, and regression analysis to verify experimental results.

$$RMSE = \sqrt{\frac{\sum(y_i^{predicted} - y_i^{actual})^2}{n}}$$

Another commonly used performance metric is the Mean Absolute Error(MAE), which is the average of all absolute errors. The formula is:

$$MAE = \frac{\sum|y_i^{predicted} - y_i^{actual}|}{n}$$

The Task

In this exercise, we are provided with a CSV (Comma Separated Values) file called salary.csv. The file is essentially a plaintext file, which contains 30 lines of information. Each line consists of two numbers (i.e., years of experience and salary) separated by a comma, such as "1.1,39343". Assuming an employee's salary (y) is approximately linearly increased with the number of years of working experience (x). Our task is to build a Simple Linear Regression (i.e., $y = b_0 + b_1 \times x$).

Here are the steps we intend to do:

1. Open/Read the CSV file to a List, in which each item is also a List containing one pair of "years" and "salary". So after successfully loading the dataset, it will be saved to a List as [["1.1", "39343"], ["1.3", "46205"], ...]. This list is similar to a $[2 \times 30]$ array, representing 30 samples.
2. Split the dataset into a training set and a test set based on a ratio (e.g., 75% for training and 25% for testing).
3. Calculate Covariance and Variance of the training dataset to find out coefficients (i.e., b_0, b_1)
4. Evaluate the model performance (e.g., using Root Mean Square Error or RMSE)