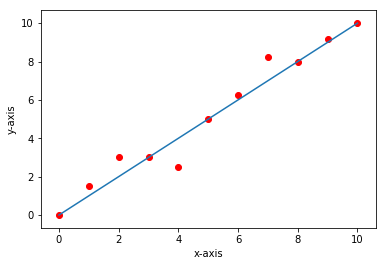
Linear Regression from Scratch Statistical Approch



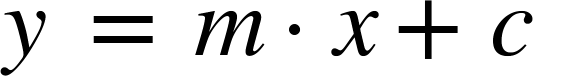
**The Theory**

Linear Regression is the process of fitting a line to the dataset.

**Single Variable Linear Regression**

**The Mathematics**

The equation of Line is



Where,

y = dependent variable

X = independent variable

C = intercept

The algorithm is trying to fit a line to the data by adjusting the values of m and c. Its Objective is to attain to a value of m such that for any given value of x it would be properly predicting the value of y.

There are various ways in which we can attain the values of m and c

1. Statistical approach
2. Iterative approach

In this post we are discussing Statistical approach

We were given data, set of x and y values and we were asked to find a line which best fits that which means mathematically we should be able to fine the slope and c values of the line to describe the line.

The derivation to the slope and intercept values of the line

The equation of the Line is

Which means for a particular value of x, let us say xi the value of y would be yi = m\*xi + c

For x1 value of x y1 value of y is obtained,

For x2 value of x y2 value of y is obtained,

And goes on

On summating all these values into one equation, it can be written as,

The objective is to solve equation 1 and 2 to attain the values of c and m

On solving the c term in Equation 1, it can also be written as,

Rewriting it as,

Substituting the value of m in Equation 2

Now only c is there in Equation 3, solving for c

Step1,

Step2,

Step 3,

On multiplying both numerator and denominator by -1

To calculate the value of m taking the value of c and substituting in Equation -1,

Equation 4 have only one parameter m, so solving for m,

On expanding RHS and cancelling terms,

The value of m,

The values of c and m are,

We use the m and c formulas obtained in derivation in the code.

**The Dataset**

Dataset consists of two columns namely X and y

Where

X = pH of well water

Y = Bicarbonate (parts per million) of well water

The data is by water well from a random sample of wells in Northwest Texas

[Link to the dataset](http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/frame.html)

**The Code**

The Code was written in three phases

1. Data preprocessing phase
2. Training
3. Prediction and plotting

The data preprocessing phase

|  |
| --- |
| **#Imports**  **import** numpy **as** np **import** pandas **as** pd **import** matplotlib.pyplot **as** plt |

Numpy import for array processing ,python doesn’t have built in array support. The feature of working with native arrays can be used in python with the help of numpy library.

Pandas is a library of python used for working with tables , on importing the data , mostly data will be of table format , for ease manipulation of tables pandas library is imported

Matplotlib is a library of python used to plot graphs, for the purpose of visualizing the results we would be plotting the results with the help of matplotlib library.

|  |
| --- |
| #Reading the dataset from data dataset = pd.read\_excel(r'data\\slr09.xls') |

In this line of code using the read\_excel method of pandas library, the dataset have been imported from data folder and stored in dataset variable.

On visualising the dataset , it contains of two columns X and Y where X is dependent variable and Y is Independent Variable



X stands for pH Level of the water

Y stands for Bicarbonate ppm in the water

Inference

For ph of 7.6 ,157 ppm of bicarbonate

for ph of 7.1 ,174 ppm of bicarbonate

And goes on

|  |
| --- |
| #Creating Dependent and Independent variables dataset = dataset.sort\_values('X')  X = dataset['X'].values  y = dataset['Y'].values |

Sorting is done for better visualization of data. The X Column from the dataset is extracted into an X variable of type numpy, similarly the y variable.



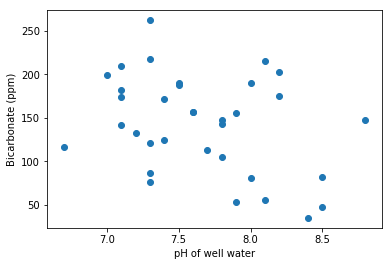
On input 10 it would result in a pandas Series object

So, values attribute is used to attain an numpy array

|  |
| --- |
| #Visualizing the data  plt.scatter(X,y)  plt.xlabel('pH of well water')  plt.ylabel('Bicarbonate (ppm)')  plt.show() |

The step is to just see how the dataset is

On visualization the data would appear something like this



Each point on the plot is a data point showing the respective ph value on x-axis and bicarbonate value on y-axis

The Training phase

|  |
| --- |
| #Importing the linear model from sklearn framework **from** sklearn.linear\_model **import** LinearRegression lr = LinearRegression() lr.fit(X = X\_train, y = y\_train) |

From scikit learn Library LinearRegression is imported. Lr is an object of LinearRegression.

The process of training is done in the fit method, our dependent and independent variable are fed into to the fit method in which it would try to fit a line to the data provided.

The Prediction phase

|  |
| --- |
| #Predicting the results  y\_pred = lr.predict(X\_test) |

By the trained linear regression model we are trying to predict the values of test data. Y\_pred variable contains all the predicted bicarbonate values of the test ph values.

|  |
| --- |
| #Visualizing the results  plt.scatter(X\_test,y\_test,c='red')  plt.plot(X\_test,y\_pred)  plt.xlabel('pH of well water')  plt.ylabel('Bicarbonate (ppm)')  plt.show() |

As we have predicted the bicarbonate values for a set of ph values we are visualizing the results to check how good did our line fit for our predictions.