1 Introduction

In this tutorial we will get an insight into how linear regression works an see how we can apply it. We will be using libraries in python to deploy linear regression algorithm.

2 Theory

Many data sets have an approximately linear relationship between variables. In these cases, we can predict one variable using a known value for another using a best-fit line, a line of the form y=mx+c, where m denotes the slope of the best fit line and c is the intercept.

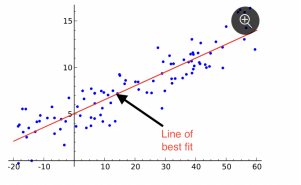

Figure 1: Scatter plot of of data points arranged approximately in a linear manner.

Given data points $(x1, y1), (x2, y2).....(x_n, y_n)$ we need to find the best fit line that roughly covers all the n data points. In order to plot the best fit line it is sufficient to compute its slope and y intercept.

To compute the slope of the best fit line we need to minimise the sum of squared errors. As the name suggests sum of squared errors is the difference between the value of the label predicted by the line of best fit and the value in the data set for the same x value.

Let us define matrix X,Y and L as follows:

$$\text{X}=\begin{bmatrix} x1 & 1 \\ x2 & 1 \\ . & . \\ . & . \\ x_n & 1 \end{bmatrix}$$

where $x1, x2....x_n$ are the x coordinates of data points in the given data set.

$$\text{Y}=\begin{bmatrix} y1 \\ y2 \\ . \\ . \\ y_n \end{bmatrix} \quad \text{L}=\begin{bmatrix} m \\ c \end{bmatrix}$$

where $y1, y2....y_n$ are the y coordinates of data points in the given data set

m and c are the slope and y-intercept respectively of the best fit line.

To minimise the sum of squared distance it necessary for $x.L \approx Y$(Eq. 1)

On pre-multiplying X_T on the L.H.S and R.H.S of Eq.1 we get

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} . \begin{bmatrix} m \\ c \end{bmatrix}$$

$$\approx \begin{bmatrix} \sum_{i=1}^{n} x_i.y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

On simplifying the above expression we get two equations. On solving the two equations we can compute the value of m and c.

$$m.\sum_{i=1}^{n} x_i^2 + c.\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i.y_i$$

$$m.\sum_{i=1}^{n} x_i + c.n = \sum_{i=1}^{n} y_i$$

Note that we could even solve the system of linear of equation using gaussian elimination technique.