Simple Linear Regression

Thenu Kaluarachchi

1 Introduction

Linear Regression is used to calculate the best fit line of a given set of data. This document will only cover the derivation of the simple case, i.e where there is only **one** independent variable. However the equation can be iterated over multiple times in a loop where there are multiple variables.

2 Derivation

Let $f(x)$ be the equation of the line that estimates the actual dependant variable y.

$f(x) = \theta \cdot x + \epsilon$

Where θ and ϵ are the **weight** and the **bias** respectively. They are, unknowns.

The summed residual error is,

$$J = \sum_{r=i}^{n} (y_i - f(x_i))^2 = \sum_{r=i}^{n} (y_i - \theta \cdot x_i - \epsilon)^2 \tag{1}$$

The purpose of linear regression is to find suitable values of θ and ϵ, such that the residual error is minimized. Thus, we will take the partial derivatives of J and equate them to zero.

$$\frac{\partial J}{\partial \theta} = \sum 2(y_i - \theta \cdot x_i - \epsilon) \cdot (-x_i) = 0$$

$$\frac{\partial J}{\partial \epsilon} = \sum 2(y_i - \theta \cdot x_i - \epsilon) \cdot (-1) = 0$$

We can simplify this to get,

$$(1)\ -\sum x_i \cdot y_i + \theta \cdot \sum (x_i)^2 + \epsilon \cdot \sum x_i = 0$$

$$(2)\ -\sum y_i + \theta \cdot \sum x_i + \epsilon \cdot \sum \epsilon = 0$$

Note that we can replace $\sum \epsilon$ with $n \cdot \epsilon$.

Solving further simultaneously, we can come to the following conclusions,

$$\theta \ = \ \frac{n \cdot \sum x_i \cdot y_i \ - \ \sum y_i \cdot \sum x_i}{n \cdot \sum (x_i)^2 \ - \ (\sum {x_i}^2)}$$

$$\epsilon \ = \ \frac{1}{n} \cdot (\sum y_i - \theta \cdot \sum x_i)$$