SimpleRegression Derivation

Ben Liu

April 2023

1 Introduction

During our conversation, there was a discussion about finding the minimum of the error function. While in deep learning this becomes very hard to do analytically because of the squishification as well as the hundreds of possible input variables, it is very much doable in the simple regression case. In other words, we can find explicit representations for my weights and biases.

In Simple Linear regression, we want to compute the weights and biases relating 1 single independent variable to 1 dependent variable.

In other words, let Y be my dependent variable, and X be my dependent variable. We have a model that I claim will best represent my data, represented as the function

$$Y = w * X + b$$

We are also given a set of observations denoted as (x_i, y_i) where i represents the index of each of my observations. The goal is to produce an error function (E) that measures the difference between my model outputs and my observations. This error function takes as input the model parameters (w, b) and tries to find values of w and b such that w and b are minimums.

2 Error Function

We want an error function that will correlate my model outputs to my actual observed values. A popular error function is the Mean squared error function. This error function is derived from the notion of Euclidean distances, we want to measure the average distance between observations and my model output, and minimize the function that is computing this average.

In other words, my error function can be described as

$$E(w, b) = 1/_n \Sigma_{i=1}^{n} (y_i - Y(x_i))^2$$

An important thing to note. The error when there is no observation of this function is 0. So while there are jumps in my error value as I move along the line – by aggregating and averaging, I don't deal with those discontinuities.

3 Derivation of parameters w and b

4 Derivation of b

We can take the expression for the error function and substitute the expression $Y = w * X + b$ in the error, plugging in values of x_i for X to get the error values. So the expression becomes

$$E(w, b) = 1/_n \Sigma_{i=1}^{n} (y_i - (wx_i + b))^2 \tag{1}$$

First, to compute, b such that I get a critical point, we take

$$dE(w, b)/db$$

So the above expression then becomes

$$dE(w, b) = 1/_n \frac{\mathrm{d}E}{\mathrm{d}b} \Sigma_{i=1}^{n} (y_i - (wx_i + b))^2$$

Using the fact that

$$d/dx(\Sigma f(x)) = \Sigma \frac{\mathrm{d}f}{\mathrm{d}x}(f(x))$$

,

We can write the error expression's partial derivative relative to b as

$$\frac{\mathrm{d}E(w, b)}{\mathrm{d}b} = 1/_n \Sigma_{i=1}^{n} \frac{\mathrm{d}E}{\mathrm{d}b} (y_i - (wx_i + b))^2 \tag{2}$$

Focusing on the inner part of the sum from (fig 1), we can simplify the expression in the parenthesis by distributing the negative sign.

$$(y_i - (wx_i + b))^2 = (y_i - wx_i - b)^2$$

What we want is to compute the derivative of the inner expression:

$$\frac{\mathrm{d}E}{\mathrm{d}b}(y_i - (wx_i + b))^2) = \frac{\mathrm{d}E}{\mathrm{d}b}(y_i - wx_i - b)^2$$

Using the chain rule, derivative of the outer exponentiation function is 2. And the derivative of the inner function with respect to b is -1, so the expression for the partial derivative relative to b is

$$\frac{\mathrm{d}E}{\mathrm{d}b}(y_i - (wx_i + b))^2) = -2(y_i - wx_i - b)$$

Plugging this expression in to the expression in (2) and factoring out the -2, we get.

$$\frac{\mathrm{d}E(w, b)}{\mathrm{d}b} = -2/_n \Sigma_{i=1}^{n} y_i - wx_i - b \tag{3}$$

To find the critical points, we have to find the values of b such that the above expression gives me 0.

In other words, solve for b in:

$$0 = -2/_n \Sigma_{i=1}^{n} y_i - wx_i - b \tag{4}$$

We can multiply both sides of the equation by -n/2 and then distribute the sum so this equation then becomes

$$0 = \Sigma_{i=1}^{n} y_i - \Sigma_{i=1}^{n} wx_i - \Sigma_{i=1}^{n} b \tag{5}$$

Then, we subtract the sum containing the y_i term from both sides and add the sum containing x_i on both sides to get

$$-\Sigma_{i=1}^{n} y_i + \Sigma_{i=1}^{n} wx_i = -\Sigma_{i=1}^{n} b \tag{6}$$

Which then factoring out b from the b term sum and using

$$n = \Sigma_{i=1}^{n} 1$$

,

we get

$$-\Sigma_{i=1}^{n} y_i + \Sigma_{i=1}^{n} wx_i = -bn \tag{7}$$

and then dividing by -n, we get that

$$(\Sigma_{i=1}^{n} y_i)/n - (\Sigma_{i=1}^{n} wx_i)n = b \tag{8}$$

Note: The average of my observations is also going to be a constant, so alot of derivations online will actually substitute the average observation terms with constants for simplicity. Knowing this, we can use the fact that the average is defined as this below:

$$\bar{x} = 1/_n \Sigma_{i=1}^{n} x_i$$

To then simplify the equation in (8) as:

$$\bar{y} - w\bar{x} = b \tag{9}$$

5 Derivation of w

Using this solution for b, we can then use this to solve for w in the error function as defined in (1). My error function now becomes:

$$E(w,b) = 1/_n \Sigma_{i=1}^{n} (y_i - wx_i - \bar{y} + w\bar{x})^2 \tag{10}$$

I repeat the process of taking the derivative of the inner function now to get an expression for the derivative relative to w

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 1/_n \Sigma_{i=1}^{n} \frac{\mathrm{d}E(w,b)}{\mathrm{d}w}(y_i - wx_i - \bar{y} + w\bar{x})^2 \tag{11}$$

Which the derivative of the outer function is still (2), but the derivative of the inner function is now $(-x_i + \bar{x})$

So the equation for the derivative of the above equation becomes:

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 1/_n \Sigma_{i=1}^{n} 2(-x_i + \bar{x})(y_i - wx_i - \bar{y} + w\bar{x}) \tag{12}$$

Which by distributing the terms for the derivative as well as factoring out the 2, we get a very big expression:

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 2/_n \Sigma_{i=1}^{n}(-x_i y_i + w{x_i}^2 + \bar{y}x_i - w\bar{x}x_i + \bar{x}y_i - w\bar{x}x_i - \bar{y}\bar{x} + w\bar{x}^2 \tag{13}$$

Reordering the terms and then combining like terms, we get that this entire large expression then becomes.

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 2/_n \Sigma_{i=1}^{n}(-x_i y_i + \bar{y}x_i + \bar{x}y_i - \bar{y}\bar{x} + w{x_i}^2 - 2w\bar{x}x_i + w\bar{x}^2) \tag{14}$$

We can use the fact that $(a-b)^2 = a^2 - 2ab + b^2$ to then simplify the terms with w in the sum as

$$w{x_i}^2 - 2w\bar{x}x_i + w\bar{x}^2 = w(x_i - \bar{x})^2$$

So the above sum in (14) then becomes

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 2/_n \Sigma_{i=1}^{n}(-x_i y_i + \bar{y}x_i + \bar{x}y_i - \bar{y}\bar{x} + w(x_i - \bar{x})^2) \tag{15}$$

Which can be simplified further using the fact that $(-a+b)(g-h) = -ag + bg + ha - hb$ to get. An important note about this part, $(a-b)(-g+h) = -ag + bg + ha - hb$ also works, but the fact of the matter is that one of the terms will have its signs flipped compared to the others so no matter which term is flipped, we can factor out a -1 from this expression for the next part.

$$\frac{\mathrm{d}E(w,b)}{\mathrm{d}w} = 2/_n \Sigma_{i=1}^{n}(-x_i + \bar{x})(y_i - \bar{y}) + \Sigma_{i=1}^{n} w(x_i - \bar{x})^2 \tag{16}$$

Which when we then set this expression to be 0, then multiplying by $n/_2$, we get

$$-\Sigma_{i=1}^{n}(-x_i + \bar{x})(y_i - \bar{y}) = w\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \tag{17}$$

Which after factoring out a -1 from the $(-x_i + \bar{x})$ term, we can cancel out the -1 and then divide both sides by $\Sigma_{i=1}^{n}(x_i - \bar{x})^2$

$$\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} = w \tag{18}$$

6 Showing w and b are minimums

The second order derivative with respect to b of the error function is:

$$\left(\frac{\mathrm{d}^2 E(w,b)}{\mathrm{d}b^2}\right)_= - 2/_n \Sigma_{i=1}^n - 1 \tag{19}$$

Which since this is always positive because $\Sigma_{i=1}^n - 1 = -n$ and the -n and +n terms cancel and the signs cancel so this error function is always going to be concave up so this b value gives me a minimum with respect to b.

The second order derivative with respect to w is

$$\left(\frac{\mathrm{d}^2 E(w,b)}{\mathrm{d}w^2}\right)_= 2/_n \Sigma_{i=1}^n (x_i - \bar{x})^2 \tag{20}$$

Which will always be positive because this is the sum of squared terms inside multiplied by a quantity that is always postiive so this will also always give me a minimum with respect to w.

7 Conclusion

While the deep learning solution utilizes adjustments, in the simple linear regression case, there are explicit terms you can use to find w and b using just the observations and the average of my observations, so there is no training cycle needed like in the deep learning approach.

Specifically, values for b and w are as follows

$$\bar{y} - w\bar{x} = b \tag{21}$$

$$\frac{\Sigma_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^n (x_i - \bar{x})^2} = w \tag{22}$$