Closed-form derivation of Least Square regression

Maxwell Aladago
January 23, 2019

We have learned a lot about linear regression over the past two weeks in class. We spoke about choosing a model, and an error criterion. You also saw your first learning algorithm, gradient descent. Of all these exciting things, perhaps, the most exciting of all was learning that least square regression can be solved in closed-form. That is you can perform a complete linear least square regression without executing gradient descent or any of its variants, and be guaranteed an optimal solution. We learned this closed-form solution is the so-called normal equation in Eq 1.

$$(X^T X)\theta = X^T y \tag{1}$$

where θ is our vector of weights.

But where did this normal equation come from? How is it connected to the model and error choices we made? Is it related to the error term at all? I seek to answer precisely these questions in this document. We'll derive the normal equation starting from our model choice.

Let's stick to our linear model such that

$$f_\theta(x) = \theta_0 + \theta_1 x_1 +, ..., +\theta_n x_n$$

and our objective is to minimize the error function

$$E(\theta) = \frac{1}{2}\sum_{i=m}^{m}\left(f_\theta(x^{(i)}) - y^{(i)})\right)^2 \tag{2}$$

where $\left(x^{(i)}, y^{(i)}\right)$ is the i-th training example, n is the number of true features and m the number of examples.

To derive the normal equation, lets define a few more variables.

Let

$$\theta = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \qquad x = \begin{bmatrix}1\\ x_1\\ \vdots\\ x_n\end{bmatrix} \qquad y = \begin{bmatrix}y^{(0)}\\ y^{(1)}\\ \vdots\\ y^{(m)}\end{bmatrix} \qquad \text{and} \qquad X = \begin{bmatrix}1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)}\\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)}\end{bmatrix}$$

Noticed the 1 we've added at the beginning of each example to match with the bias?

With these new definitions, we can write our hypothesis as

$$f_\theta(x) = \theta^T x$$

Also, we can write our cost function

$$\begin{aligned} E(\theta) &= \frac{1}{2}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}\left(\left(x^{(i)}\right)^T \theta - y^{(i)}\right)^2 \end{aligned}$$

Notice that we swapped the positions of θ and $x^{(i)}$. This is just for convenience and to accommodate our definition of $X = \left[\left(x^{(1)}\right)^T, ..., \left(x^{(m)}\right)^T\right]^T$ above.

Now, we can write our cost function over all examples as

$$
\begin{aligned}
E(\theta) &= \frac{1}{2}\left(X\theta - y\right)^T\left(X\theta - y\right) \\
&= \frac{1}{2}\left((X\theta)^T - y^T\right)\left(X\theta - y\right) \\
&= \frac{1}{2}\left((X\theta)^T(X\theta) - (X\theta)^T y - y^T(X\theta) + y^T y\right) \\
&= \frac{1}{2}\left((X\theta)^T(X\theta) - 2(X\theta)^T y + y^T y\right) \\
&= \frac{1}{2}\left(\theta^T X^T X\theta - 2\theta^T X^T y + y^T y\right)
\end{aligned} \tag{3}
$$

We now get to the exciting part. Since we are interested in the values of θ which gives the minimum value of the cost function in Eq.3, we can simply take the partial derivative of Eq. 3 with respect to θ. We can then set this value to 0 and solve for θ. That partial derivative is given by

$$
\begin{aligned}
\frac{\partial E(\theta)}{\partial \theta} &= \frac{1}{2}\left(2X^T X\theta - 2X^T y + 0\right) \\
&= (X^T X)\theta - X^T y
\end{aligned} \tag{4}
$$

Can see how the $\frac{1}{2}$ in our cost function in Eq. 2 came in handy to help cancel everything out neatly? Setting the derivative to zero, we have

$$
\begin{aligned}
(X^T X)\theta - X^T y &= 0 \\
(X^T X)\theta &= X^T y
\end{aligned} \tag{5}
$$

Eq.5 is exactly our normal equation as defined before. Hopefully, you see that the normal equation is really easy to derive and depends on our error function.