

The Derivative of Cost Function for Logistic Regression

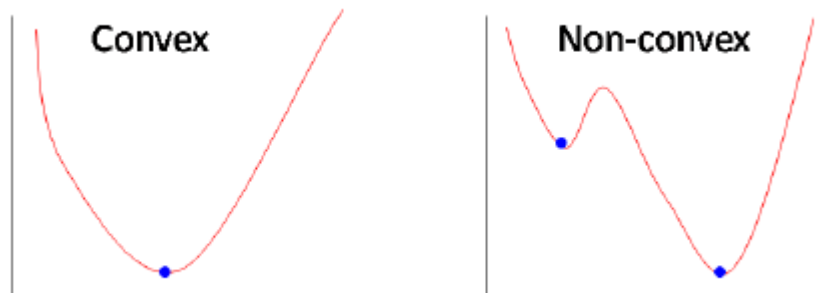


Saket Thavanani [Follow](#)

Dec 13, 2019 · 3 min read

Introduction:

Linear regression uses **Least Squared Error** as a loss function that gives a convex loss function and then we can complete the optimization by finding its vertex as a global minimum. However for logistic regression, the hypothesis is changed, the Least Squared Error will result in a non-convex loss function with local minimums by calculating with the sigmoid function applied on raw model output.



Left(Linear Regression mean square loss), Right(Logistic regression mean square loss function)

However, we are very familiar with the gradient of the cost function of linear regression it has a very simplified form given below, But I wanted to mention a point here that gradient for the loss function of logistic regression also comes out to have the same form of terms in spite of having a complex log loss error function.

$$\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \end{bmatrix}$$

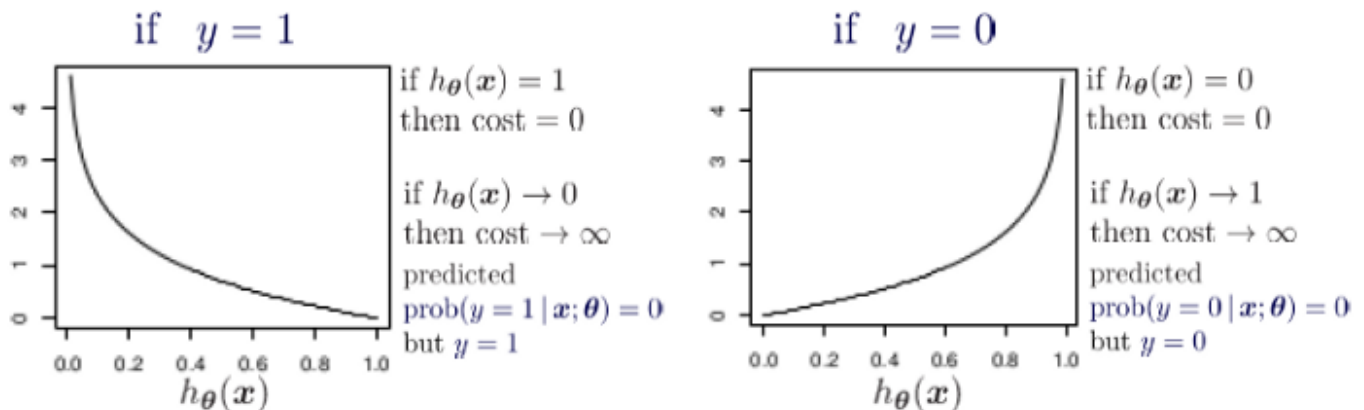
$$\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \frac{1}{m} x^T (h(x) - y)$$

Gradient for Linear Regression Loss Function

In order to preserve the convex nature for the loss function, a log loss error function has been designed for logistic regression. The cost function is split for two cases $y=1$ and $y=0$.

For the case when we have $y=1$ we can observe that when hypothesis function tends to 1 the error is minimized to zero and when it tends to 0 the error is maximum. This criterion exactly follows the criterion as we wanted

$$\text{cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$



Cost Function

Combining both the equation we get a convex log loss function as shown below-

$$\text{cost}(h_{\theta}(x), y) = -y \log(h_{\theta}(x)) - (1 - y) \log(1 - h_{\theta}(x))$$

$$\text{if } y = 1: \quad \text{cost}(h_{\theta}(x), y) = -\log(h_{\theta}(x))$$

$$\text{if } y = 0: \quad \text{cost}(h_{\theta}(x), y) = -\log(1 - h_{\theta}(x))$$

Combined Cost Function

In order to optimize this convex function, we can either go with gradient-descent or newtons method. For both cases, we need to derive the gradient of this complex loss function. The mathematics for deriving gradient is shown in the steps given below

The Derivative of Cost Function:

Since the hypothesis function for logistic regression is sigmoid in nature hence, The First important step is finding the gradient of the sigmoid function. We can see from the derivation below that gradient of the sigmoid function follows a certain pattern.

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Hypothesis Function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\frac{d(\sigma(x))}{dx} = \frac{0 * (1 + e^{-x}) - (1) * (e^{-x} * (-1))}{(1 + e^{-x})^2}$$

$$\frac{d(\sigma(x))}{dx} = \frac{(e^{-x})}{(1 + e^{-x})^2} = \frac{1 - 1 + (e^{-x})}{(1 + e^{-x})^2} = \frac{1 + e^{-x}}{(1 + e^{-x})^2} - \frac{1}{(1 + e^{-x})^2}$$

$$\frac{d(\sigma(x))}{dx} = \frac{1}{1 + e^{-x}} * \left(1 - \frac{1}{1 + e^{-x}}\right) = \sigma(x)(1 - \sigma(x))$$

Derivative of Sigmoid Function

Step 1:

Applying Chain rule and writing in terms of partial derivatives.

$$\frac{\partial(J(\theta))}{\partial(\theta_j)} = -\frac{1}{m} * \sum_{i=1}^m \left[y^{(i)} * \frac{1}{h_{\theta}(x^{(i)})} * \frac{\partial(h_{\theta}(x^{(i)}))}{\partial(\theta_j)} \right]$$

$$+ \sum_{i=1}^m \left[(1 - y^{(i)}) * \frac{1}{(1 - h_{\theta}(x^{(i)}))} * \frac{\partial (1 - h_{\theta}(x^{(i)}))}{\partial (\theta_j)} \right]$$

$$\begin{aligned} \frac{\partial(J(\theta))}{\partial(\theta_j)} = & -\frac{1}{m} * \left(\sum_{i=1}^m \left[y^{(i)} * \frac{1}{h_{\theta}(x^{(i)})} * \sigma(z)(1 - \sigma(z)) * \frac{\partial(\theta^T x)}{\partial(\theta_j)} \right] \right. \\ & \left. + \sum_{i=1}^m \left[(1 - y^{(i)}) * \frac{1}{(1 - h_{\theta}(x^{(i)}))} * (-\sigma(z)(1 - \sigma(z))) * \frac{\partial(\theta^T x)}{\partial(\theta_j)} \right] \right) \end{aligned}$$

Step 2:

Evaluating the partial derivative using the pattern of the derivative of the sigmoid function.

$$\begin{aligned} \frac{\partial(J(\theta))}{\partial(\theta_j)} = & -\frac{1}{m} * \left(\sum_{i=1}^m \left[y^{(i)} * \frac{1}{h_{\theta}(x^{(i)})} * \sigma(z)(1 - \sigma(z)) * \frac{\partial(\theta^T x)}{\partial(\theta_j)} \right] \right. \\ & \left. + \sum_{i=1}^m \left[(1 - y^{(i)}) * \frac{1}{(1 - h_{\theta}(x^{(i)}))} * (-\sigma(z)(1 - \sigma(z))) * \frac{\partial(\theta^T x)}{\partial(\theta_j)} \right] \right) \end{aligned}$$

$$\begin{aligned} \frac{\partial(J(\theta))}{\partial(\theta_j)} = & -\frac{1}{m} * \left(\sum_{i=1}^m \left[y^{(i)} * \frac{1}{h_{\theta}(x^{(i)})} * h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)})) * x_j^i \right] + \right. \\ & \left. \sum_{i=1}^m \left[(1 - y^{(i)}) * \frac{1}{(1 - h_{\theta}(x^{(i)}))} * (-h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)}))) * x_j^i \right] \right) \end{aligned}$$

Step 3:

Simplifying the terms by multiplication

$$\frac{\partial(J(\theta))}{\partial(\theta_j)} = -\frac{1}{m} * \left(\sum_{i=1}^m \left[y^{(i)} * (1 - h_{\theta}(x^{(i)})) * x_j^i - (1 - y^{(i)}) * h_{\theta}(x^{(i)}) * x_j^i \right] \right)$$

$$\frac{\partial J(\theta_j)}{\partial \theta_j} = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} - h_{\theta}(x^{(i)})] x_j^{(i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} - h_{\theta}(x^{(i)})] x_j^{(i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} - h_{\theta}(x^{(i)})] x_j^{(i)}$$

Step 4:

Removing the summation term by converting it into a matrix form for the gradient with respect to all the weights including the bias term.

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m} X^T [h_{\theta}(x) - y]$$

Conclusion:

This little calculus exercise shows that both linear regression and logistic regression (actually a kind of classification) arrive at the same update rule. What we should appreciate is that the design of the cost function is part of the reasons why such “coincidence” happens.

Thank you for reading!!!!

If you like my work and want to support me:

1-The BEST way to support me is by following me on **Medium**.

2-Follow me on **LinkedIn**.

Machine Learning

Calculus

Derivatives

Data Science

Mathematics

[About](#) [Help](#) [Legal](#)

Get the Medium app

