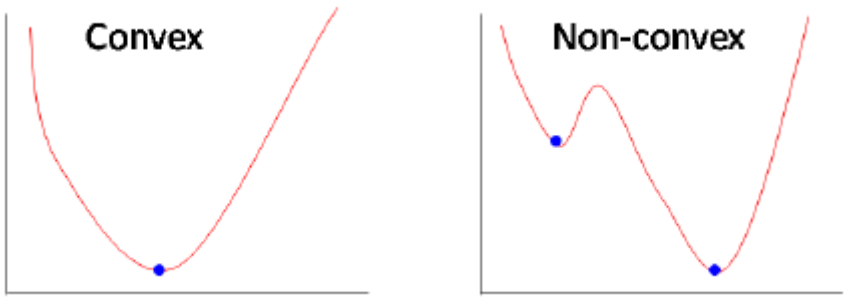
Derivative of Cost Function for Logistic Regression

Saket Thavanani
Dec 14, 2019 · 3 min read

Introduction:

Linear regression uses **Least Squared Error** as loss function that gives a convex loss function and then we can complete the optimization by finding its vertex as global minimum. However for logistic regression the hypothesis is changed, Least Squared Error will result in a non-convex loss function with local minimums by calculating with 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 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 come come out to have the same form of terms inspite of have a complex log loss error function.

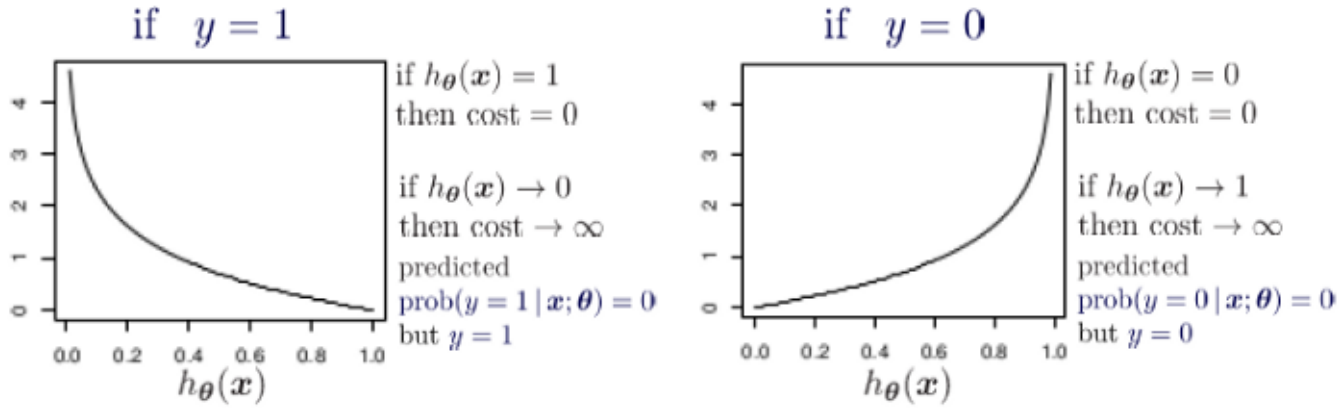
$$\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \frac{1}{m} x^T (h(x) - y)$$

Gradient for Linear Regression Loss Function

In order to preserve the convex nature for 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 the 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

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



Cost Function

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

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

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

Combined Cost Function

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

Derivative of Cost Function:

Since the hypothesis function for logistic regression is sigmoid in nature hence, First important step is finding the gradient of sigmoid fucntion. We can see from the

derivation below that gradient of the sigmoid function follows a certain pattern.

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

Hypothesis Function

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

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

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

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

Derivative of Sigmoid Function

Step 1:

Applying Chain rule and writing in terms of partial derivatives.

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

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

$$\frac{\partial(J(\theta))}{\partial(\theta j)} = -\frac{1}{m} * \left(\sum_{\mathrm{i}=1}^{\mathrm{m}} \left[\mathrm{y}^{(\mathrm{i})} * \frac{1}{\mathrm{h}_\theta\big(\mathrm{x}^{(\mathrm{i})}\big)} * \sigma(\mathrm{z})\big(1 - \sigma(\mathrm{z})\big) * \frac{\partial(\theta^{\mathrm{T}}\mathrm{x})}{\partial(\theta \mathrm{j})} \right] \right.$$

$$\left. + \sum^{m} \left[\big(1 - y^{(i)}\big) * \frac{1}{(\quad)} * \big(-\sigma(\mathrm{z})\big(1 - \sigma(\mathrm{z})\big) * \frac{\partial(\theta^{\mathrm{T}}x)}{\partial(\theta j)} \right] \right)$$

$$\sum_{i=1}\Big[\qquad \big(1-h_\theta(x^{(i)})\big) \qquad\qquad \Big]$$

Step 2:

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

$$\frac{\partial(J(\theta))}{\partial(\theta j)} = -\frac{1}{m}*\Big(\sum_{i=1}^{m}\Big[y^{(i)} * \frac{1}{h_\theta(x^{(i)})} * \sigma(z)\big(1-\sigma(z)\big) * \frac{\partial(\theta^T x)}{\partial(\theta j)}\Big] + \sum_{i=1}^{m}\Big[\big(1-y^{(i)}\big) * \frac{1}{\big(1-h_\theta(x^{(i)})\big)} * (-\sigma(z)\big(1-\sigma(z)\big) * \frac{\partial(\theta^T x)}{\partial(\theta j)}\Big]\Big)$$

$$\frac{\partial(J(\theta))}{\partial(\theta j)} = -\frac{1}{m}*\Big(\sum_{i=1}^{m}\Big[y^{(i)} \frac{1}{h_\theta(x^{(i)})} h_\theta(x^{(i)})\big(1-h_\theta(x^{(i)})\big) * x_j^i\Big] + \sum_{i=1}^{m}\Big[\big(1-y^{(i)}\big) * \frac{1}{\big(1-h_\theta(x^{(i)})\big)} * \big(-h_\theta(x^{(i)})\big(1-h_\theta(x^{(i)})\big) * x_j^i\Big]\Big)$$

Step 3:

Simplifying the terms by multiplication

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

$$\frac{\partial(J(\theta))}{\partial(\theta j)} = -\frac{1}{m}*\Big(\sum_{i=1}^{m}\big[\, y^{(i)} - y^{(i)} * h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)} * h_\theta(x^{(i)})\big] * x_j^i\Big)$$

$$\frac{\partial(J(\theta))}{\partial(\theta j)} = -\frac{1}{m}*\Big(\sum_{i=1}^{m}\big[\, y^{(i)} - h_\theta(x^{(i)})\big] * x_j^i\Big)$$

Step 4:

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

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

$\partial(\theta)$ m

Conclusion:

This little calculus exercise shows that both linear regression and logistic regression (actually a kind of classification) arrive 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.

Thanks for Reading!!!!

If you like my work and want to support me. The BEST way to support me is by following me on **Medium.**

Machine Learning Calculus Derivatives Data Science Mathematics