Logistic Regression Derivations

Mickey Chao

February 2016

Intuition Behind Logistic Regression

Note: For the rest of this document, we assume that $\bar{\theta}$ has a built-in bias coefficient θ_0 and all training data, $\bar{x}^{(i)}$, have a built-in bias term x_0.

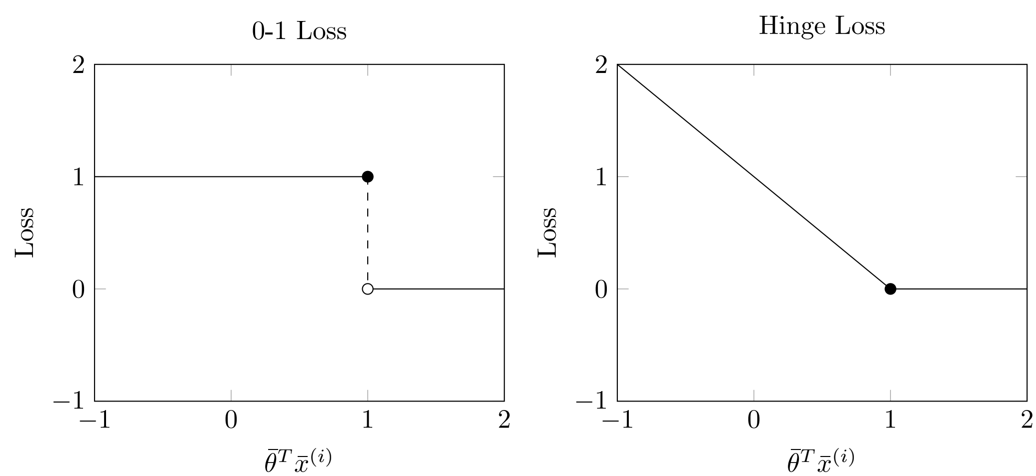
There are many kinds of loss functions that we can use. For example, we have used 0-1 loss and hinge loss. 0-1 loss was defined as

$$\text{Loss}_{0-1}(x^{(i)}) = \begin{cases} 0 & \text{if } \bar{\theta}^T \bar{x}^{(i)} = y^{(i)} \\ 1 & \text{otherwise} \end{cases}$$

and

$$\text{Loss}_h(x^{(i)}) = \max\{1 - \bar{\theta}^T \bar{x}^{(i)}, 0\}$$

Their graphs are shown below



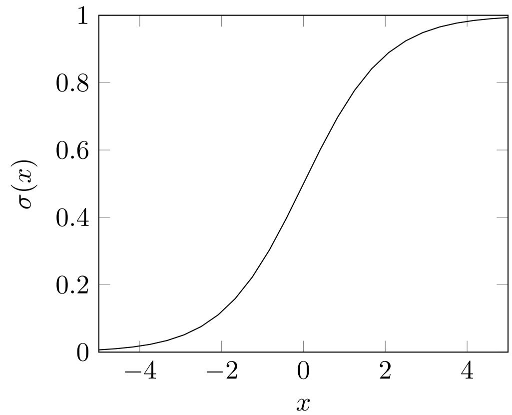
0-1 Loss is difficult to deal with because it is discontinuous at $x = 1$. Therefore, we can't perform any type of hill-climbing optimization on a problem formulated with 0-1 Loss. Hinge loss may also present difficulties because it is not differentiable at $x = 1$.

We may notice that these graphs both have a similar shape, where the cost is 0 beyond $x = 1$ and positive before $x = 1$. We would also like to expand on the hinge loss idea that even if $\bar{\theta}^T \bar{x}^{(i)} > 0$, there is still a significant penalty associated if $\bar{\theta}^T \bar{x}^{(i)}$ is not positive enough.

Now, we go through the idea behind logistic loss, which will give us another type of loss with these properties.

Consider the sigmoid function $\sigma(x) = \frac{1}{1+e^{-z}}$, which is shown below:

Sigmoid Function



We notice that this function has values between 0 and 1. We can think of $\sigma(x)$ as how confident we are that something is positive (1) or negative (0). That is, we define

$$P(Y = y^{(i)} | \bar{\theta}, \bar{x}^{(i)}) = \sigma(y^{(i)}(\bar{\theta}^T \bar{x}^{(i)}))$$

Essentially, we think of P as the probability that our given $\bar{\theta}$ predicts a label of $y^{(i)}$ for a given training datum $\bar{x}^{(i)}$. The question we wish to solve now is the following: "Given $\bar{\theta}$, how likely is it that the training data (X, y) came from the distribution predicted by $\bar{\theta}$?"

The probability that the distribution given by $\bar{\theta}$ predicts the $y^{(i)}$ correctly from the $\bar{x}^{(i)}$ is

$$L(\bar{\theta}) = \prod_{i=1}^{n} P(Y = y^{(i)} | X = x^{(i)}, \bar{\theta})$$

We would like to maximize L and typically, we solve a maximization/minimization problem by taking a partial derivative. However, it is difficult to apply the product rule n times. Instead, we take the base 2 log of all the terms and maximize the sum:

$$l(\bar{\theta}) = \sum_{i=1}^{n} \log_2(P(Y = y^{(i)} | X = x^{(i)}, \bar{\theta})) = \sum_{i=1}^{n} \log_2\left(\frac{1}{1 + e^{-\bar{\theta}^T \bar{x}^{(i)}}}\right)$$

This is also equivalent to minimizing the sum

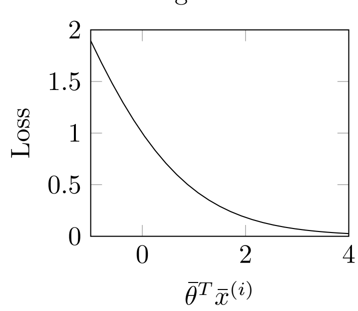
$$l'(\bar{\theta}) = \sum_{i=1}^{n} \log_2(\frac{1}{P(Y = y^{(i)} | X = x^{(i)}}, \bar{\theta})) = \sum_{i=1}^{n} \log_2\left(1 + e^{-\bar{\theta}^T \bar{x}^{(i)}}\right)$$

Notice that l' is basically a combination of losses due to every training datum. From this, we get another type of loss function known as logistic loss:

$$\text{Loss}_{\text{log}} = \log_2(1 + e^{-\bar{\theta}^T \bar{x}})$$

which has the following plot:

Logistic Loss



We see that logistic loss has a similar shape to hinge loss, but it is also differentiable everywhere.