Logistic Regression Notes

Louie Dinh

December 27, 2015

1 Classification And The Logistic Function

We define a classification as follows: Given X $\in \mathbb{R}^{m \times n}$ and y $\in \{0, 1\}$, find a function $p : \mathbb{R}^m \to \{0, 1\}$, parameterized by θ, that maximizes the likelihood function

$$\mathcal{L}(X, y, \theta) = \prod_{y^{(i)}=1} p_\theta(x^{(i)}) \prod_{y^{(i)}=0} 1 - p_\theta(x^{(i)}) \tag{1}$$

where p is a learned function that returns $P(y = 1 \mid X = x)$

In general it is difficult to search the entire function space, so we limit our attention to p's in some particular form. Specifically, in logistic regression we want to look at the logistic function. See Figure 1

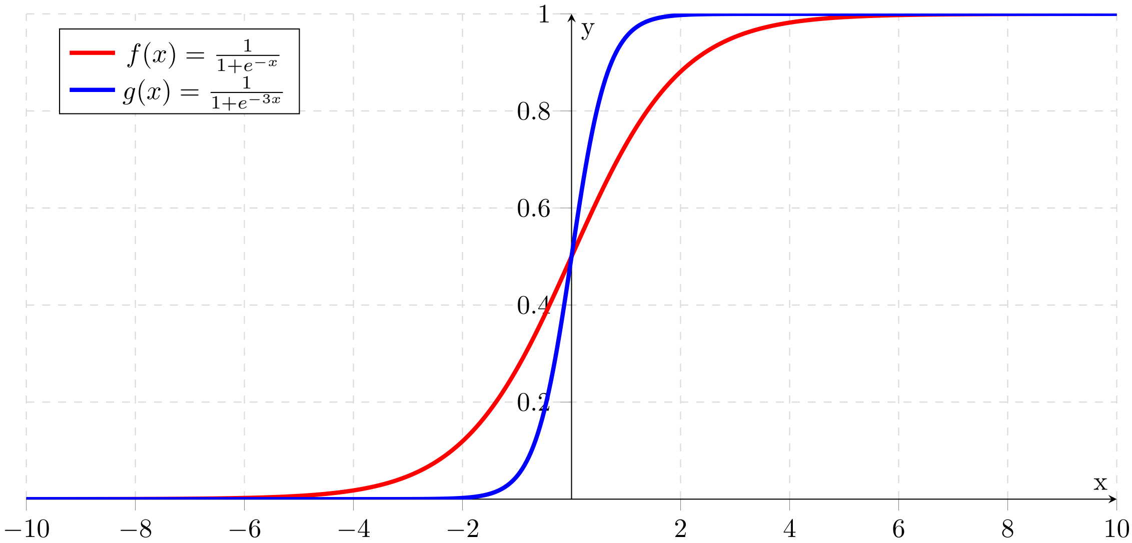

Figure 1: Graph of the logistic/sigmoid function

2 Linear Regression Review

Recall that in linear regression, we estimate y with $\hat{y} = \theta^\intercal X$. We fit θ by minimizing the following quantity:

$$\arg\min_\theta \sum_{i=1}^{m} (y^{(i)} - \theta^\intercal X^{(i)})^2 \tag{2}$$

We have a closed form solution for (2) in the normal equations.

$$\theta = (X^\intercal X)^{-1} X^\intercal y \tag{3}$$

3 Logistic Function and Decision Boundaries

The logistic function $f : \mathbb{R} \to [0, 1]$ is defined by:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Note that the logistic function has several nice properties for our application. It maps the real line onto [0,1] and takes on the value of 0.5 when x=0. We will use the logistic function to map our covariates into a probability and call the positive case then $p(x) > 0.5$.

Before, we estimated the response with $\theta^\intercal X$ but this returns a real number. Given $X^{(i)}$, we'd like to estimate the probability that y=1. From our previous discussion, we can just pass our linear estimator through the logistic function.

$$p_\theta(x) = \frac{1}{1 + e^{-\theta^\intercal X}} = g(\theta^\intercal X) \tag{5}$$

4 Fitting Parameters

Maximizing the likelihood function in (1) is the same thing as maximizing the log likelihood. The utility function is then

$$log\mathcal{L}(\theta) = \sum_{y^{(i)}=1} log(p_\theta(X^{(i)})) + \sum_{y^{(i)}=0} log(1 - p_\theta(X^{(i)}) \tag{6}$$

which can be rewritten as

$$\sum_{i=1}^{m} y^{(i)} log(g(\theta^\intercal X^{(i)})) + (1 - y^{(i)}) log(1 - g(\theta^\intercal X^{(i)})) \tag{7}$$

Before we continue, note that $g'(z) = g(z)(1 - g(z))$. Then taking the partial derivative of (7), we get

$$\frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} y^{(i)} log(g(\theta^\intercal X^{(i)})) + (1 - y^{(i)}) log(1 - g(\theta^\intercal X^{(i)})) \tag{8}$$

$$= \sum_{i=1}^{m} \left(\frac{y^{(i)}}{g(\theta^\intercal X^{(i)})} - \frac{1 - y^{(i)}}{1 - g(\theta^\intercal X^{(i)})} \right) \frac{\partial}{\partial \theta_j} g(\theta^\intercal X^{(i)}) \tag{9}$$

$$= \sum_{i=1}^{m} \left(\frac{y^{(i)}}{g(\theta^\intercal X^{(i)})} - \frac{1 - y^{(i)}}{1 - g(\theta^\intercal X^{(i)})} \right) g(\theta^\intercal X^{(i)})(1 - g(\theta^\intercal X^{(i)}) \frac{\partial}{\partial \theta_j} \theta^\intercal X^{(i)} \tag{10}$$

$$= \sum_{i=1}^{m} y^{(i)}(1 - g(\theta^\intercal X^{(i)})) - (1 - y^{(i)}) g(\theta^\intercal X^{(i)}) X_j^{(i)} \tag{11}$$

$$= \sum_{i=1}^{m} (y^{(i)} - p_\theta(X^{(i)})) X_j^{(i)} \tag{12}$$

This is surprisingly the exact same form as the gradient of our linear least squares problem!