Logistic Regression

Zongheng Yang

May 18, 2012

1 Problem of Linear Regression in Binary Classfication

In linear regression, $y = 0$ or 1, yet $h_\theta(x)$ can be greater than one or less than zero. In logistic regression, $0 \leq h_\theta(x) \leq 1$.

2 Hypothesis

To make $0 \leq h_\theta(x) \leq 1$, define it as:

$$g(x) = \frac{1}{1+e^{-x}} \quad \text{(the Sigmoid function)} \tag{1}$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{2}$$

Interpret $h_\theta(x)$ as the probability that $y = 1$. The **decision boundary**: predict $y = 1$ if $h_\theta(x) \geq 0.5$ which is equivalent to $\theta^T x \geq 0$.

3 Cost Function

In linear regression, cost function is defined as

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}). \end{aligned}$$

It can be shown that the above *Cost* function is not a convex one in the case of logistic regression. Therefore, define the *Cost* function for logistic regression as

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

therefore the cost function for logistic regression is

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) \tag{3}$$

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]. \tag{4}$$

It can be shown via the principle of maximum likelihood estimation that the cost function $J(\theta)$ is convex. In addition, notice that $J(\theta) \geq 0$ at all times.

Taking the partial derivative:

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

so the gradient descent algorithm for logistic regression is

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{5}$$

$$= \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}. \tag{6}$$

4 Optimization Algorithms

gradient descent, and more advanced ones (faster, no need to manually pick α, more complex): conjugate gradient, BFGS, L-BFGS.