

Notes on IRLS

Deva Ramanan

January 31, 2008

This note is meant to help code up the IRLS algorithm for homework 2. The final equation is given by (7). Assume we can given a data set $\{(x^{(i)}, y^{(i)}) : 1 = 1 \dots m\}$, and we want to build a logistic regression classifier parameterized by θ . Let us assume that each data point $x^{(i)}$ and the model parameter θ is represented as a $n \times 1$ vector.

$$P(y^{(i)}|x^{(i)}, \theta) = \frac{1}{1 + \exp(-\theta^T x^{(i)})} \quad (1)$$

$$l(\theta) = \sum_i \log P(y^{(i)}|x^{(i)}, \theta) \quad (2)$$

We want to fit a model θ that maximizes the log likelihood of the training data.

1 Gradient descent

To optimize the log likelihood, we can use gradient descent.

$$\theta := \theta + \alpha \left[\frac{\partial l(\theta)}{\partial \theta_i} \right] \quad (3)$$

$$\theta := \theta + \alpha \sum_i (y^{(i)} - \mu^{(i)}) x^{(i)} \quad (4)$$

where

$$\mu^{(i)} = \frac{1}{1 + \exp(-\theta^T x^{(i)})} \quad (5)$$

We write $\left[\frac{\partial l(\theta)}{\partial \theta_i} \right]$ for the $n \times 1$ vector of partial derivatives - this is also called the gradient. Note that for stochastic gradient descent, we update θ after *each* training point i .

2 IRLS

We can also optimize the log likelihood using a second-order method - rather than approximating $l(\theta)$ locally with a tangent plane, we approximate it locally with a parabolic curve. We can then step to the optimum of the curve, eliminating the need for the stepsize α .

$$\theta := \theta - \left[\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]^{-1} \left[\frac{\partial l(\theta)}{\partial \theta_i} \right] \quad (6)$$

$$\theta := \theta + \left(\sum_i w^{(i)} x^{(i)} x^{(i)T} \right)^{-1} \sum_i (y^{(i)} - \mu^{(i)}) x^{(i)} \quad (7)$$

where

$$w^{(i)} = \mu^{(i)}(1 - \mu^{(i)}) \quad (8)$$

We write $\left[\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]$ for the $m \times m$ matrix of second-order derivatives - this is also called the hessian.

The Newton-Raphson optimization algorithm, when applied to training a logistic regression model, is also called the iteratively re-weighted least squares (IRLS) algorithm. This is because one can interpret (7) as solving a weighted least squares problem, with weights given by $w^{(i)}$ - see the tutorial by Jordan.