CS3390 Assignment 1 Problem 4

Gautam Singh (CS21BTECH11018)
Jaswanth Beere (BM21BTECH11007)

Contents

This document discusses an algorithm to find a solution to a binary classification problem that uses the sigmoid function in logistic regression.

1 Newton-Raphson Update Equation

Define

$$p_i \triangleq \Pr\left(y_i = 1 | \mathbf{x}_i, \mathbf{w}\right) = \frac{1}{1 + \exp\left(-\mathbf{w}^\top \mathbf{x}_i\right)} \tag{1}$$

The likelihood function for logistic regression is given by

$$L\left(\mathbf{w}\right) = \prod_{i:y_i=1} p_i \prod_{j:y_j=0} \left(1 - p_j\right) \tag{2}$$

Hence, the log likelihood is given by

$$l\left(\mathbf{w}\right) = \sum_{i=1}^{N} y_i \log p_i + \left(1 - y_i\right) \log\left(1 - p_i\right) \tag{3}$$

and using (1), the error function is simply the negative of (3), given by

$$\begin{aligned} E\left(\mathbf{w}\right) &= -\sum_{i=1}^{N} y_i \log p_i + \left(1 - y_i\right) \log\left(1 - p_i\right) && (4) \\ &= -\sum_{i=1}^{N} y_i \left(\mathbf{w}^\top \mathbf{x}_i\right) - \log\left(1 + \exp\left(\mathbf{w}^\top \mathbf{x}_i\right)\right) && (5) \end{aligned}$$

Using (1), the first derivatives of (5) are

$$\begin{aligned} \frac{\partial E}{\partial w_r} &= -\sum_{i=1}^{N} \left(y_i - \frac{1}{1 + \exp\left(-\mathbf{w}^\top \mathbf{x}_i\right)}\right) x_{ir} && (6) \\ &= -\sum_{i=1}^{N} \left(y_i - p_i\right) x_{ir}. && (7) \end{aligned}$$

The second derivatives are therefore

$$\begin{aligned} \frac{\partial^2 E}{\partial w_r \partial w_s} &= \sum_{i=1}^{N} x_{ir} x_{is} \frac{\exp\left(-\mathbf{w}^\top \mathbf{x}\right)}{\left(1 + \exp\left(-\mathbf{w}^\top \mathbf{x}_i\right)\right)^2} && (8) \\ &= \sum_{i=1}^{N} x_{ir} x_{is} p_i \left(1 - p_i\right). && (9) \end{aligned}$$

Hence, from (7) the gradient of (5) is

$$\nabla E\left(\mathbf{w}\right) = -\sum_{i=1}^{N} \left(y_i - p_i\right) \mathbf{x}_i = -\mathbf{X}^\top \left(\mathbf{y} - \mathbf{p}\right) \tag{10}$$

and from (9), the Hessian becomes

$$\mathbf{H}_E\left(\mathbf{w}\right) = \mathbf{X}^\top \mathbf{W} \mathbf{X}, \tag{11}$$

where we define

$$\begin{aligned} \mathbf{X} &= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N \end{pmatrix} && (12) \\ \mathbf{y} &= \begin{pmatrix} y_1 & y_2 & \dots & y_N \end{pmatrix}^\top && (13) \\ \mathbf{p} &= \begin{pmatrix} p_1 & p_2 & \dots & p_N \end{pmatrix}^\top && (14) \\ \mathbf{W} &= \operatorname{diag}\left(p_1\left(1 - p_1\right) \quad \dots \quad p_N\left(1 - p_N\right)\right). && (15) \end{aligned}$$

Thus, the Hessian update equation becomes

$$\begin{aligned} \mathbf{w}_{n+1} &= \mathbf{w}_n - \mathbf{H}_E\left(\mathbf{w}_n\right)^{-1} \nabla E\left(\mathbf{w}_n\right) && (16) \\ &= \mathbf{w}_n + \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\mathbf{y} - \mathbf{p}\right). && (17) \end{aligned}$$

The algorithm for performing such a computation is depicted below.

Algorithm 1 Newton-Raphson Update Algorithm for Logistic Regression

function NR-LOGISTIC-REGRESSION($\mathbf{X}$, $\mathbf{y}$)
  $(r, c) \leftarrow \dim(\mathbf{X})$
  let $\mathbf{w}$ be an initial guess for $\hat{\mathbf{w}}_{ML}$
  repeat
    compute $\mathbf{p}$ and $\mathbf{W}$ as in (14) and (15)
    $\mathbf{w} \leftarrow \mathbf{w} + \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p})$
  until (the change in $\mathbf{w}$ is small enough)
return $\mathbf{w}$

2 Relation to Weighted Least Squares

The optimal solution $\hat{\mathbf{w}}_{ML}$, given by

$$\hat{\mathbf{w}}_{ML} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) . \tag{18}$$

Note that the optimal solution is quite similar to the one for reweighted least squares regression in Problem 3, where $\mathbf{y} - \mathbf{p}$ is the "corrected" outputs with the probabilities of success being the correction. However, the diagonal matrix $\mathbf{W}$ is a function of $\mathbf{w}$, hence this algorithm is known as the *iterative reweighted least squares* (IRLS) method.

3 Convexity of Error Function

For all i, $p_i (1 - p_i) \geq 0$, thus the diagonal entries of $\mathbf{W}$ are all positive. Hence, from (11), it follows that the eigenvalues of $\mathbf{H}_E$, which are the diagonal entries in $\mathbf{W}$ are nonnegative for any parameters $\mathbf{w}$. Thus, $\mathbf{H}_E$ is positive semidefinite and therefore, $E(\mathbf{w})$ is convex. Hence, it has a unique minimum at the point defined in (18).