

# CS3390 Assignment 1

## Problem 4

Gautam Singh (CS21BTECH11018)  
Jaswanth Beere (BM21BTECH11007)

### CONTENTS

1	Newton-Raphson Update Equation	1
2	Relation to Weighted Least Squares	2
3	Convexity of Error Function	2

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

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

The likelihood function for logistic regression is given by

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

Hence, the log likelihood is given by

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

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

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

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

$$\frac{\partial E}{\partial w_r} = - \sum_{i=1}^N \left( y_i - \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)} \right) x_{ir} \quad (6)$$

$$= - \sum_{i=1}^N (y_i - p_i) x_{ir}. \quad (7)$$

The second derivatives are therefore

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \sum_{i=1}^N x_{ir} x_{is} \frac{\exp(-\mathbf{w}^\top \mathbf{x}_i)}{(1 + \exp(-\mathbf{w}^\top \mathbf{x}_i))^2} \quad (8)$$

$$= \sum_{i=1}^N x_{ir} x_{is} p_i (1 - p_i). \quad (9)$$

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

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

and from (9), the Hessian becomes

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

where we define

$$\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_N) \quad (12)$$

$$\mathbf{y} = (y_1 \quad y_2 \quad \dots \quad y_N)^\top \quad (13)$$

$$\mathbf{p} = (p_1 \quad p_2 \quad \dots \quad p_N)^\top \quad (14)$$

$$\mathbf{W} = \text{diag}(p_1(1 - p_1) \quad \dots \quad p_N(1 - p_N)). \quad (15)$$

Thus, the Hessian update equation becomes

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \mathbf{H}_E(\mathbf{w}_n)^{-1} \nabla E(\mathbf{w}_n) \quad (16)$$

$$= \mathbf{w}_n + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p}). \quad (17)$$

The algorithm for performing such a computation is depicted below.

---

**Algorithm 1** Newton-Raphson Update Algorithm for Logistic Regression

---

```

1: function NR-LOGISTIC-REGRESSION( $\mathbf{X}$ ,  $\mathbf{y}$ )
2:    $(r, c) \leftarrow \dim(\mathbf{X})$ 
3:   let  $\mathbf{w}$  be an initial guess for  $\hat{\mathbf{w}}_{ML}$ 
4:   repeat
5:     compute  $\mathbf{p}$  and  $\mathbf{W}$  as in (14) and (15)
6:      $\mathbf{w} \leftarrow \mathbf{w} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p})$ 
7:   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

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

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).