Assignment - 1

Challa Akshay Santoshi, CS21BTECH11012
Cheekatla Hema Sri, CS21BTECH11013

October 8, 2023

Question 4

For logistic regression, there is no longer a closed-form solution, due to the nonlinearity of the logistic sigmoid function. the error function can be minimized by an efficient iterative technique based on the Newton-Raphson iterative optimization scheme. Let us first derive the error function of the logistic function. Consider the problem of two class classification. We have defined the logistic sigmoid function as follows

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{1}$$

In logistic regression, we can write the posterior probability of class C_1 as follows

$$p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w^T}\phi) \tag{2}$$

Where ϕ is the feature vector, and the posterior of the class C_2 is given as follows

$$p(C_2|\phi) = 1 - p(C_1|\phi) \tag{3}$$

For the dataset ϕ_n, t_n where $t_n \in 0, 1$ and $\phi_n = \phi(x_n)$ with $n = 1, 2, \ldots, N$, the likelihood function is given as follows

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n} \tag{4}$$

To get the error function, take negative log of the likelihood

$$E(\mathbf{w}) = -ln(p(\mathbf{t}|\mathbf{w})) \tag{5}$$

$$E(\mathbf{w}) = \sum_{n=1}^{N}[t_n log(y_n) + (1-t_n)log(1-y_n)] \tag{6}$$

a

Update equation for the Newton-Raphson optimization technique used to obtain the parameters in the logistic regression model

$$\boxed{\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1}\nabla\mathbf{E}(\mathbf{w})} \tag{7}$$

where $\mathbf{w}^{(new)}$ is new value of parameter vector and $\mathbf{w}^{(old)}$ is the old value of the parameter vector and $\mathbf{H}$ is the hessian matrix, where $\mathbf{H} = \nabla\nabla\mathbf{E}(\mathbf{w})$ and $\nabla\mathbf{E}(\mathbf{w})$ is the derivative error function.

Now let us compute the terms in this update equation to get the corresponding equations, First let us compute the gradient equation ($\nabla \mathbf{E}(\mathbf{w})$)

$$E(\mathbf{w}) = -\sum_{n=1}^{N}[t_n log(y_n) + (1-t_n)log(1-y_n)] \tag{8}$$

$$\nabla \mathbf{E}(\mathbf{w}) = -\sum_{n=1}^{N}[\frac{t_n}{y_n}y'_n + \frac{1-t_n}{1-y_n}(-y'_n)] \tag{9}$$

$$\text{expanding } y_n \text{ as } \frac{1}{1+e^{-\mathbf{w}^T\phi_n}}, \implies \nabla E[\mathbf{w}] = -\sum_{n=1}^{N} y'_n \frac{[t_n(1+e^{-\mathbf{w}^T\phi_n}) - (1-t_n)(1+e^{-\mathbf{w}^T\phi_n})]}{e^{-\mathbf{w}^T\phi_n}} \tag{10}$$

$$\text{we have } y'_n = \frac{(e^{-\mathbf{w}^T\phi_n})\phi_n}{(1+e^{-\mathbf{w}^T\phi_n})^2} \tag{11}$$

$$\implies \nabla E[\mathbf{w}] = -\sum_{n=1}^{N}(1+e^{-\mathbf{w}^T\phi_n})\frac{(e^{-\mathbf{w}^T\phi_n})\phi}{(1+e^{-\mathbf{w}^T\phi_n})^2}[\frac{t_n e^{-\mathbf{w}^T\phi_n} - 1 + t_n}{e^{-\mathbf{w}^T\phi_n}}] \tag{12}$$

$$= -\sum_{n=1}^{N}\frac{\phi_n}{(1+e^{-\mathbf{w}^T\phi_n})}[t_n(1+e^{-\mathbf{w}^T\phi_n}) - 1] \tag{13}$$

$$\implies \nabla \mathbf{E}(\mathbf{w}) = -\sum_{n=1}^{N}\phi_n[t_n - y_n] \tag{14}$$

$$\boxed{\implies \nabla \mathbf{E}(\mathbf{w}) = \Phi^T(\mathbf{y} - \mathbf{t})} \tag{15}$$

Now let us compute the Hessian matrix, which is the derivative of the above function.

$$\nabla\nabla \mathbf{E}(\mathbf{w}) = \nabla(\sum_{n=1}^{N}\phi_n[y_n - t_n]) \tag{16}$$

$$= \sum_{n=1}^{N}\phi_n y'_n \tag{17}$$

$$= \sum_{n=1}^{N}\phi_n \frac{(e^{-\mathbf{w}^T\phi_n})\phi_n}{(1+e^{-\mathbf{w}^T\phi_n})^2} \tag{18}$$

$$\implies \nabla\nabla \mathbf{E}(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T \tag{19}$$

$$\boxed{\nabla\nabla \mathbf{E}(\mathbf{w}) = \Phi^T\mathbf{R}\Phi} \tag{20}$$

Where $\mathbf{R}$ is an $n \times n$ matrix such that

$$R_{ij} = \begin{cases} y_n(1-y_n), & i = j \\ 0, & i \neq j \end{cases} \tag{21}$$

Now let us re-write the update equation of the Newton-Raphson iterative optimization by substituting the computed values

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1}\nabla \mathbf{E}(\mathbf{w}) \tag{22}$$

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\Phi^T\mathbf{R}\Phi)^{-1}\Phi^T(\mathbf{y} - \mathbf{t}) \tag{23}$$

$$= (\Phi^T\mathbf{R}\Phi)^{-1}(\Phi^T\mathbf{R}\Phi\mathbf{w}^{(old)} - \Phi^T(\mathbf{y} - \mathbf{t})) \tag{24}$$

$$\implies \mathbf{w}^{(new)} = (\Phi^T\mathbf{R}\Phi)^{-1}\Phi^T\mathbf{R}\mathbf{z} \tag{25}$$

$$\text{where } z = \Phi\mathbf{w}^{(old)} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t}) \tag{26}$$

Algorithm for this methodology is given as follows

1. **initialization:** Start with an initial guess for the coefficients, denoted as $\mathbf{w}^{(old)}$
2. **iteration:**
   - Calculate the gradient vector $\nabla \mathbf{E}(\mathbf{w})$ and Hessian matrix $\mathbf{H}$
   - Now update the coefficient vector using the update equation $\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1}\nabla \mathbf{E}(\mathbf{w})$
   - Repeat this until the convergence is reached
3. **termination:** Terminate the iteration loop when the $\nabla E(w)$ comes close to 0

b

Now we see that equation for $\mathbf{w}$ for this method (i.e., $(\Phi^T \mathbf{R}\Phi)^{-1}\Phi^T \mathbf{R}\mathbf{z}$) is similar to the equation for $\mathbf{w}$ that we got for the weighted least squares problem described in question 3c (i.e., $(\Phi R \Phi^T)^{-1}(\Phi^T R\mathbf{t})$)

Because the weighing matrix $\mathbf{R}$ is not constant but depends on the parameter vector $\mathbf{w}$, we must apply the normal equations iteratively, each time using the new weight vector $\mathbf{w}$ to compute a revised weighing matrix $\mathbf{R}$. For this reason, the algorithm is known as iterative reweighted least squares

c

The error function for this logistic regression is $\mathbf{E}(\mathbf{w})$. If we observe the Hessian matrix $\mathbf{H}$, it is no longer constant but depends on $\mathbf{w}$ through the weighting matrix $\mathbf{R}$, corresponding to the fact that the error function is no longer quadratic. Using the property $0 < y_n < 1$, which follows from the form of the logistic sigmoid function, we see that $\mathbf{u}^T \mathbf{H}\mathbf{u} > 0$ for an arbitrary vector $\mathbf{u}$, and so the Hessian matrix $\mathbf{H}$ is positive definite. It follows that the error function is a convex function of $\mathbf{w}$ and hence has a unique minimum