ML Assignment No. 1

Name: Mayuresh R. Dindorkar	Name: Sanyam Kaul.
Roll No.: CS23 MTECH 14007	Roll No: CS23 MTECH 14011

Q.4) For logistic regression, there is no longer a closed form solution due to nonlinearity of logistic sigmoid function. The error function can be minimized by an efficient iterative technique based on Newton-Raphson iterative optimization scheme.

(a) Provide the expressions of gradient, Hessian and update equations for Newton-Raphson optimization technique used to obtain the parameters in logistic regression model. Provide the algorithm, describing the methodology.

→

Inorder to obtain the ~~expressions of~~ Hessian, gradient & update equations, we need the error function $E(w)$ of logistic regression. Hence, first we will find the error function for logistic regression.

(*) **Logistic Regression:-**

Consider the dataset $\{\phi_n, t_n\}$,
where $t_n \in \{0,1\}$ and $\phi_n = \phi(x_n)$ with $n = 1, 2 \ldots N$.
w = parameter vector.
In logistic regression,
we model the probability of positive class C_1 as,

~~of positive class (C1)~~

$$p(C_1|\phi) = y(\phi) = \sigma(w^T\phi) = \frac{1}{1+e^{-w^T\phi}}$$

$$\therefore p(C_2|\phi) = 1 - p(C_1|\phi) = \frac{1}{1+e^{w^T\phi}}$$

The likelihood function for logistic regression, can be written as,

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

where, $\vec{t} = (t_1, t_2, \cdots, t_N)^T$ and $y_n = p(C_1|\phi_n)$.

To find the error function $E(w)$, we take negative logarithm of the likelihood as below:-

$$E(w) = -\log p(\vec{t}|w)$$

$$\boxed{E(w) = -\sum_{n=1}^{N} \{t_n \cdot \log y_n + (1-t_n) \log(1-y_n)\}} \leftarrow (1)$$

which is also called "cross entropy error" function,

where $y_n = \sigma(a_n)$ and $a_n = w^T \phi_n$.

(I) Equation for gradient: (Denoted by $\nabla E(w)$)

The equation for gradient is obtained by taking derivate of $E(w)$ wrt w,

$$\therefore \nabla E(w) = \sum_{n=1}^{N} (y_n - t_n) \phi_n$$

$$\boxed{\nabla E(w) = \Phi^T (\vec{y} - \vec{t})} \leftarrow (2)$$

where Φ is "design matrix" of size $N \times M$,

$\Phi \in R^{N \times M}$ and

$\vec{y} = [y_1, y_2, \cdots, y_N]$

Equation (2) represents the equation of gradient.

(II) Equation for Hessian:- (Denoted by H).

Hessian matrix H is obtained by taking double derivative of $E(w)$.

$$\therefore H = \nabla\nabla E(w)$$

$$= \frac{\partial (E(w))}{\partial w_j \cdot \partial w_i}$$

$$= \frac{\partial}{\partial w}\left[\sum_{n=1}^{N} (y_n - t_n)\,\phi(x_n)\right] \quad \text{... from (2)}$$

$$= \sum_{n=1}^{N} \phi(x_n)\left(\frac{\partial y_n}{\partial w}\right)$$

$$= \sum_{n=1}^{N} y_n(1-y_n)\cdot \phi(x_n)\cdot \phi(x_n)$$

$$\therefore H = \sum_{n=1}^{N} y_n(1-y_n)\,\phi_n \cdot \phi_n^T.$$

$$\boxed{\therefore H = \Phi^T \cdot R \cdot \Phi} \qquad \text{..... (3)}$$

i.e, have introduced the diagonal matrix R of dimensions N x N with elements,

$$R_{nn} = y_n(1-y_n)$$

Equation (3) represents the equation of Hessian.

(III) Equation for update:-

The update equation according to Newton Rapson for minimizing error function $E(w)$ is given by,

$$w^{new} = w^{old} + \Delta w.$$

where Δw is called stepsize.

We want to minimize the objective function $\nabla E(w)^{new}$

We can expand $E(w^{new})$ using Taylor series around current estimate w^{old} as:

$$E(w^{new}) \approx E(w^{old}) + (w - w^{old})^T \nabla E(w^{old}) + \frac{1}{2}(w - w^{old})^T \cdot H \cdot (w - w^{old}).$$

where,

$\nabla E(w^{old})$ is gradient vector and H is Hessian matrix.

Now, we differentiate it and equate to zero, to obtain w^{new} as:

$$\frac{\partial}{\partial w}\left[E(w^{old}) + (w - w^{old})^T \cdot \nabla E(w^{old}) + \frac{1}{2}(w - w^{old})^T \cdot H \cdot (w - w^{old})\right] = 0$$

Solving this we get, $\boxed{w^{new} = w^{old} - H^{-1} \cdot \nabla E(w^{old})}$ ... (4)

which represents the 'update equation' for newton rapson technique.

After substituting values of H & ∇E, the update equation becomes, $w^{new} = w^{old} - (\Phi^T R \Phi)^{-1} \Phi^T (\vec{y} - \vec{t})$.

$$= (\Phi^T R \Phi)^{-1} \{\Phi^T \cdot R \cdot \Phi \cdot w^{old} - \Phi^T (\vec{y} - \vec{t})\}$$

$$w^{new} = (\Phi^T \cdot R \cdot \Phi)^{-1} \Phi^T \cdot R \cdot z.$$

where $z = \Phi w^{old} - R^{-1}(\vec{y} - \vec{t})$.

z is vector with N dimensions.

(*) Algorithm:-

(1) Initialize ω^0. with some random weights.

(2) For each $\tau = 1, 2, 3, \ldots N$ (N == no. of iterations)

(i) Compute $\nabla E(\omega^\tau)$ // gradient

(ii) Compute H (Hessian matrix) using ω^τ

(iii) Determine $\omega^{\tau+1}$ using Newton-Rapson's update expression.

$$\omega^{\tau+1} = \omega^\tau - H^{-1} \cdot \nabla E(\omega^\tau)$$

(iv) Check whether convergence criteria met or not. If met, stop & come out of loop.

(3) Return the estimated parameter vector ω^τ.

Page No.
Date

(Q.4) (b) Show that the Newton Rapson update scheme is related to the weighted least squares problem described in the question 3(c). and explain why it is called the iterative reweighted least squares. method.

→

In logistic regression, the Newton-Rapson update technique updates the parameter vector w using inverse of H ie Hessian matrix. This update operation resembles a weighted least squares problem.

The update equation for logistic regression using Newton Rapson is given by,

$$w^{new} = w^{old} - H^{-1} \nabla E(w^{old}) \quad \text{—— (1)}$$

$$\therefore w^{new} = (\Phi^T R \Phi)^{-1} \Phi^T R \cdot z \quad \text{—— (2)}$$

where,

z is a vector with N dimensions.

R = Diagonal matrix with $N \times N$ dimensions with elements $R_{nn} = y_n(1-y_n)$.

Φ = Design matrix of size $N \times M$.

In (1), the H^{-1} acts as a weight matrix where the weight are determined by curvature of error function.

~~specifically~~, In particular, data points with steeper sigmoid function gradients contributes less to the update.

The weight matrix H^{-1} resembles R having elements r_n as

$r_n = \delta(y_n)[1-\delta(y_n)]$,

where,

$y_n = w^T x_n$; in context of WLS. method.

Equation (2) closely resembles the form of normal equation of weighted least squares (WLS) problem.

As the matrix R is dependant on parameter vector w, the normal equation needs to be applied iteratively, using new weight vector w with each iteration to obtain the revised weight matrix R.

Therefore, this algorithm is often referred as 'Iterative Reweighted least squares' (IRLS) method.

Q.4) (c) Show that the error function of logistic regression is a convex function of w and has unique minimum with help of Hessian matrix.

→ First we will prove positive definiteness of Hessian matrix H and then will prove that the solution is unique minimum.

(I) Positive definiteness of Hessian matrix(H):-

We will show that R is a positive definite matrix.

Hence, as per definition of positive definiteness,

$$u^T \cdot R \cdot u = \sum_n u_n \cdot y_n (1-y_n) \cdot u_n$$

$$\therefore \boxed{u^T \cdot R \cdot u = \sum_n u_n^2 \cdot y_n (1-y_n)}$$

where,

u = Any arbitrary vector

R = diagonal matrix.

y_n = output of logistic regression model for input x_n

This equation will be positive for any arbitrary nonzero vector u, because sigmoid ensures $0 < y_n < 1$.

By using same analogy,

Consider a Hessian matrix H & arbitrary vector u,

$$u^T \cdot H \cdot u = u^T \cdot \Phi^T \cdot R \cdot \Phi \cdot u$$

$$\therefore H = \Phi^T \cdot R \cdot \Phi$$

where Φ = design matrix.

$$\therefore u^T \cdot H \cdot u = v^T \cdot H \cdot v \quad \text{--- } v = u^T \cdot \Phi^T$$

$$= \sum_n v_n^2 \cdot y_n (1-y_n) > 0.$$

Page No.
Date

Hence this show that hessian matrix H is also positive definite. Hence, error function of logistic regression is a convex function.

(II) Unique minimum:-

After expanding the taylor series around minimum point w^*, we get our objective function as:

$$E(w^* + \Delta w) = E(w^*) + \Delta_w E(w^*) \Delta w + (\Delta w)^T \cdot H(w) \cdot (\Delta w) + Z(\|\Delta w\|^3)$$

where, H is function of w because all y_n's are function of w.

At minimum, the first order term $\Delta E(w^*) \cdot \Delta w = 0$ because derivative is zero at minimum point.

Similarly, the higher order terms denoted by $Z(\|\Delta w\|^3)$ are negligible.

The remaining quadratic term $(\Delta w)^T \cdot H \cdot (\Delta w)$ gets minimized when $\Delta w = 0$, hence our objective function also gets minimized at $\Delta w = 0$.

Hence, by using properties of positive definiteness & Taylor series expansion we have shown that E(w) has unique minimum at w^*.