FOML Assignment 1

Vishal Vijay Devadiga (CS21BTECH11061)
Abhay Kumar (BM21BTECH11001)

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**.

Part A

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

Provide an algorithm describing the methodology.

Solution

The cost function for logistic regression is given by:

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

where:

- $y_n = y(x_n, \theta)$ is the model prediction for input x_n and θ is the set of parameters of the model.
- t_n is the target value for input x_n.

The gradient of the cost function is given by:

$$\nabla E(\theta) = [\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}, \dots, \frac{\partial E(\theta)}{\partial \theta_N}]^T$$

$$\implies \frac{\partial E(\theta)}{\partial \theta_j} = -\sum_{n=1}^{N}\left[t_n \frac{1}{y_n}\frac{\partial y_n}{\partial \theta_j} + (1-t_n)\frac{1}{1-y_n}\frac{\partial (1-y_n)}{\partial \theta_j}\right]$$

$$\implies \frac{\partial E(\theta)}{\partial \theta_j} = -\sum_{n=1}^{N}\left[t_n \frac{1}{y_n}\frac{\partial y_n}{\partial \theta_j} - (1-t_n)\frac{1}{1-y_n}\frac{\partial y_n}{\partial \theta_j}\right]$$

$$\implies \frac{\partial E(\theta)}{\partial \theta_j} = -\sum_{n=1}^{N}\left[\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right]\frac{\partial y_n}{\partial \theta_j}$$

$$\implies d\frac{\partial E(\theta)}{\partial \theta_j} = -\sum_{n=1}^{N}\left[\frac{t_n - y_n}{y_n(1-y_n)}\right]\frac{\partial y_n}{\partial \theta_j}$$

$$y_n = \frac{1}{1+\exp(-\theta^T x_n)}$$

$$\implies \frac{\partial y_n}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left(\frac{1}{1+\exp(-\theta^T x_n)}\right)$$

$$\implies \frac{\partial y_n}{\partial \theta_j} = \frac{\exp(-\theta^T x_n)}{(1+\exp(-\theta^T x_n))^2}\frac{\partial}{\partial \theta_j}(-\theta^T x_n)$$

$$\implies \frac{\partial y_n}{\partial \theta_j} = \frac{\exp(-\theta^T x_n)}{(1+\exp(-\theta^T x_n))^2}(-x_{nj})$$

$$\implies \frac{\partial y_n}{\partial \theta_j} = y_n(1-y_n)(-x_{nj})$$

$$\frac{\partial E(\theta)}{\partial \theta_j} = -\sum_{n=1}^{N}\left[\frac{t_n - y_n}{y_n(1-y_n)}\right] y_n(1-y_n)(-x_{nj})$$

$$\implies \boxed{\frac{\partial E(\theta)}{\partial \theta_j} = \sum_{n=1}^{N}(y_n - t_n)x_{nj}}$$

Let:

$$X = [x_1, x_2, \ldots, x_N]^T$$

$$Y - T = [y_1 - t_1, y_2 - t_2, \ldots, y_N - t_N]^T$$

$$\implies \boxed{\nabla E(\theta) = X^T(Y - T)}$$

The Hessian of the cost function is given by:

$$H = \nabla^2 E(\theta) = \begin{bmatrix} \frac{\partial^2 E(\theta)}{\partial \theta_1^2} & \frac{\partial^2 E(\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 E(\theta)}{\partial \theta_1 \partial \theta_N} \\ \frac{\partial^2 E(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 E(\theta)}{\partial \theta_2^2} & \dots & \frac{\partial^2 E(\theta)}{\partial \theta_2 \partial \theta_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E(\theta)}{\partial \theta_N \partial \theta_1} & \frac{\partial^2 E(\theta)}{\partial \theta_N \partial \theta_2} & \dots & \frac{\partial^2 E(\theta)}{\partial \theta_N^2} \end{bmatrix}$$

$$H_{ij} = \frac{\partial^2 E(\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\left(\frac{\partial E(\theta)}{\partial \theta_j}\right)$$

$$\implies H_{ij} = \frac{\partial}{\partial \theta_i}\left(\sum_{n=1}^{N}(y_n - t_n)x_{nj}\right)$$

$$\implies H_{ij} = \sum_{n=1}^{N} \frac{\partial}{\partial \theta_i}\left((y_n - t_n)x_{nj}\right)$$

$$\implies H_{ij} = \sum_{n=1}^{N} \frac{\partial y_n}{\partial \theta_i} x_{nj}$$

$$\implies H_{ij} = \sum_{n=1}^{N} y_n(1 - y_n)x_{ni}x_{nj}$$

$$\implies \boxed{H = \sum_{n=1}^{N} y_n(1 - y_n)x_n x_n^T}$$

Let:

$$S = \begin{bmatrix} y_1(1 - y_1) & 0 & \dots & 0 \\ 0 & y_2(1 - y_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_N(1 - y_N) \end{bmatrix}$$

$$\implies \boxed{H = X^T S X}$$

The update equation for the Newton-Raphson optimization technique is given by:

$$\boxed{\theta^{(new)} = \theta^{(old)} - H^{-1}\nabla E(\theta^{(old)})}$$

$$\implies \boxed{\theta^{(new)} = \theta^{(old)} - (X^T S X)^{-1} X^T (Y - T)}$$

where:

- $\theta^{(new)}$ is the new value of θ.
- $\theta^{(old)}$ is the old value of θ.
- H is the Hessian of the cost function.
- $\nabla E(\theta)$ is the gradient of the cost function.

Algorithm:

```
Input: X, t, theta
Output: theta

while (Not Converged):
    g = gradient(X, t, theta)
    H = hessian(X, t, theta)
    theta = theta - inverse(H) * g

return theta
```

Part B

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

Solution

The Newton-Raphson update scheme is given by:

$$\theta^{(new)} = \theta^{(old)} - H^{-1}\nabla E(\theta)$$

where:

- $\theta^{(new)}$ is the new value of θ.

- $\theta^{(old)}$ is the old value of θ.
- H is the Hessian of the cost function.
- $\nabla E(\theta)$ is the gradient of the cost function.

$$\implies \theta^{(new)} = \theta^{(old)} - (X^T S X)^{-1} X^T (Y - T)$$

$$\implies \theta^{(new)} = (X^T S X)^{-1} (X^T S X \theta^{(old)} - X^T (Y - T))$$

$$\implies \theta^{(new)} = (X^T S X)^{-1} X^T S (X \theta^{(old)} - S^{-1} (Y - T))$$

Let $\tilde{Y} = X\theta^{(old)} - S^{-1}(Y - T)$.

$$\theta^{(new)} = (X^T S X)^{-1} X^T S \tilde{Y}$$

This update scheme is related to the weighted least squares problem described in question $3(c)$ because the weighted least squares problem is given by:

$$\theta^{(new)} = (X^T R X)^{-1} X^T R Y$$

The matrix S and $\tilde{Y}$ are recalculated in each iteration of the Newton-Raphson update scheme. Hence, it is called the **iterative reweighted least squares method**.

Part C

Show that the error function of the logistic regression is a convex function of w and hence has a unique minimum with the help of the Hessian matrix.

Solution

If a function f has a Hessian matrix H such that H is positive semidefinite, then f is a convex function

A matrix H is positive semidefinite if:

$$P^T H P \geq 0 \quad \forall P \in \mathbb{R}^n - \{0\}$$

The Hessian matrix of the error function of the logistic regression is given by:

$$H = X^T S X$$

$$P^T H P = P^T X^T S X P$$

$$\implies P^T H P = (XP)^T S (XP)$$

$$\implies P^T H P = \sum_{n=1}^{N} (XP)_n^2 S_{nn}$$

S_{nn} is positive because y_n is a probability and hence $y_n \in [0, 1]$.

$$\implies P^T H P \geq 0 \quad \forall P \in \mathbb{R}^n - \{0\}$$

Hence, the error function of the logistic regression is a convex function of w and hence has a unique global minimum and provides the optimal solution.