Question : 4

linear combination of features and weights.

$$y(x) = w^T x + w_0$$

for logistic Regression
[To get values in Range $[0,1]$]

Sigmoid function

$$h(x) = \frac{1}{1+e^{-y(x)}} = \frac{1}{1+e^{-(w^T x + w_0)}}$$

→ Using $h(x)$ to formulate the likelihood

$$P(y=1 \mid x; w) = h_w(x)$$
$$P(y=0 \mid x; w) = 1 - h_w(x)$$
$$P(y \mid x; w) = h_w(x)^y \left(1-h_w(x)\right)^{1-y}$$

→ likelihood function : likelihood that our model will correctly predict any given y values given its corresponding feature vector

$$L(w) = \prod_{i=1}^{n} P(y_i \mid x_i; w)$$
$$L(w) = \prod_{i=1}^{n} h_w(x_i)^{y_i} \left(1-h_w(x_i)\right)^{1-y_i}$$

log likelihood of our hypothesis function.

$$l(w) = \log L(w)$$
$$= \sum_{i=1}^{n} y_i \log\left(h_w(x_i)\right) + (1-y_i)\log\left(1-h_w(x_i)\right)$$

Newton Raphson Method :

→ Basically a method to find the Root of function which converges faster. The equation is denoted as below.

$$x_{new} = x_{old} - \frac{f(x)}{f'(x)}$$ } finding roots of equation of $f(x)$.

Stop until the difference of two consecutive values of x's $\cong 0$.

→ As we want to find value of 'w' such that it minimize the log likelihood function.

(A.) Expression for Gradient

→ Hence using Newton Raphson Method to find the value of 'w'.

$\nabla l(w)$ = Gradient of log-likelihood — 1st order partial derivative of $\log L(w)$

H^{-1} = Hessian — 2nd order partial derivative of $\log L(w)$

→ $\nabla l(w)$ Partial derivative of log likelihood function with respect to all the Parameters.

$$\nabla l(w) = \begin{bmatrix} \frac{\partial l(w)}{\partial w_1} \\ \frac{\partial l(w)}{\partial w_2} \\ \vdots \\ \frac{\partial l(w)}{\partial w_n} \\ \frac{\partial l(w)}{\partial w_0} \end{bmatrix}_{(n+1)\times 1} = \begin{bmatrix} \sum_{i=1}^{n} (y_i - h_w(x_i))\, x_1 \\ \sum_{i=1}^{n} (y_i - h_w(x_i))\, x_2 \\ \vdots \\ \sum_{i=1}^{n} (y_i - h_w(x_i)).1 \end{bmatrix}$$

$\frac{\partial l(w)}{\partial w_j} = (y - h_w(x_i))\, x_j$

∵ Gradient in matrix form can be written as.

$$\nabla l(w) = X^T (y - h(X)) \quad \text{—— (1)}$$

Expression for Hessian

→ Hessian is Second order partial derivative of Gradient w.r.t all the parameters.

$$H(w) = \nabla\nabla l(w) = \begin{bmatrix} \frac{\partial^2 l(w)}{\partial^2 w_1} & \frac{\partial^2 l(w)}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 l(w)}{\partial w_1 \partial w_n} & \frac{\partial^2 l(w)}{\partial w_1 \partial w_0} \\ \vdots & & & & \vdots \\ \frac{\partial^2 l(w)}{\partial w_0 \partial w_1} & \frac{\partial^2 l(w)}{\partial w_0 \partial w_2} & \cdots & \frac{\partial^2 l(w)}{\partial w_0 \partial w_n} & \frac{\partial^2 l(w)}{\partial w_0^2} \end{bmatrix}_{(n+1)\times(n+1)}$$

$$\frac{\partial^2 l(w)}{\partial w_1 \partial w_2} = h_w(x_i)(1 - h_w(x_i))\, x_1 x_2$$

$$H^{-1}(w) = \nabla\nabla l(w) = \sum_{i=1}^{n} h_w(x_i)(1-h_w(x_i))\, x_i x_i^T$$

$$H^{-1}(w) = \nabla\nabla l(w) = X^T R X \quad \text{---(2)} \qquad R = h_w(x_i)(1-h_w(x_i))$$
Diagonal Matrix

Expression of Update equations for Newton Raphson optimization Technique used to obtain the parameters in the logistic regression model.

$$w_{new} = w_{old} - H^{-1}\nabla l(w_{old}) \quad \text{---(3)}$$

ALGORITHM PSEUDOCODE :

1. Define log likelihood function.
2. initialize the parameters
3. Convergence conditions
4. Loop which iterates till the difference of 2 consecutive values of w becomes less then the convergence condition.
   - → EVAL Gradient
   - → EVAL Hessian
   - → (Hessian)$^{-1}$
   - → Update your parameters
   - → Update the log-likelihood at each iteration
   - → Update the difference which helps in checking the condition at the beginning of loop.
5. Return the parameters.

B. From equation (1), (2) and (3)

$$w_{new} = w_{old} - H^{-1}\nabla l(w_{old})$$
$$= w_{old} - (X^T R X)^{-1} X^T (y - h(X))$$
$$= (X^T R X)^{-1} \{ X^T R X\, w_{old} - X^T (y - h(X)) \}$$

$$w_{new} = (X^T R X)^{-1} X^T R z \qquad \text{[Taking } X^T R \text{ common]}$$

Here, $z = X w_{old} - R^{-1}(y - h(X))$

Thus the w_{new} takes the form same as the solution of Weighted least squares.

→ But here in the solution the R (diagonal matrix) is not constant it depends on the parameter vector w. And in every iteration we are updating it.

Hence, it is called iterative reweighted least squares method.

(C.) Error function of logistic Regression.

$$l(w) = \sum_{i=1}^{n} y_i \log(h_w(x_i)) + (1-y_i)\log(1-h_w(x_i))$$

→ A function is convex if and only if its H (Hessian) matrix is positive for all point. i.e. $X^T R X > 0$.

→ Since $h(x)$ and $(1-h(x))$ falls in the range of $[0,1]$ because they are the probabilities. Therefore H itself is +ve.

→ Since H is +ve it implies that the error function of logistic Regression is convex with respect to parameter w. Hence, it ensures that using Newton Raphson method will converges to unique minimum.