Stat 600: Homework 3

For Multi-class logistic regression, we know that,

$$P(y_i = k | x_i) = p_k(x_i ; \beta), \sum_{k=0}^{K-1} p_k(x_i ; \beta) = 1$$

$$p_k(x_i ; \beta) = \frac{e^{x_i^T \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T \beta_l}}$$

$$f(\beta) = \left[-\sum_{i=1}^{n} \left\{ \sum_{k=0}^{K-1} 1(y_i = k) \log p_k(x_i ; \beta) \right\} + \frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{k,j}^2 \right] \leftarrow ①$$

Taking derivatives of different ~~par~~ parts and then combining them to form a gradient,

$$\frac{\partial}{\partial \beta_k} \log p_k(x_i ; \beta) = \frac{1}{p_k(x_i;\beta)} \cdot \frac{\partial p_k(x_i;\beta)}{\partial \beta_k} \leftarrow ②$$

$$\frac{\partial}{\partial \beta_k} p_k(x_i ; \beta) = \frac{\partial}{\partial \beta_k} \left(\frac{e^{x_i^T \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T \beta_l}} \right)$$

$$\frac{\partial}{\partial \beta_k} p_k(x_i ; \beta) = p_k(x_i;\beta)(1 - p_k(x_i;\beta)) x_i \leftarrow ③$$

When $l \neq k$,

$$\frac{\partial}{\partial \beta_l} p_k(x_i;\beta) = -p_k(x_i;\beta)\, p_l(x_i;\beta)\, x_i \leftarrow ④$$

Therefore, from ②, ③, and ④,

$$\frac{\partial}{\partial \beta_k}\left(-1(y_i=k)\log p_k(x_i;\beta)\right)$$

$$= -\left(1(y_i=k) - p_k(x_i;\beta)\right) x_i \leftarrow ⑤$$

$$\frac{\partial}{\partial \beta_k} \frac{\lambda}{2}\sum_{k=0}^{K-1}\sum_{j=1}^{p} \beta_{k_j}^2 = \lambda \beta_k \leftarrow ⑥$$

Therefore from ①, ⑤ and ⑥, we get the gradient term,

$$\nabla_{\beta_k} f(\beta) = \underbrace{-\sum_{i=1}^{n}\left(1(y_i=k) - p_k(x_i;\beta)\right)x_i}_{\text{likelihood term}} + \underbrace{\lambda \beta_k}_{\text{Regularization term}} \leftarrow ⑦$$

For $k = l$,

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_k} = p_k(x_i;\beta)(1 - p_k(x_i;\beta))\, x_i$$

$$\frac{\partial^2 p_k(x_i;\beta)}{\partial \beta_k^2} = p_k(x_i;\beta)(1 - p_k(x_i;\beta))\, x_i x_i^T \leftarrow (8)$$

For $k \neq l$,

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_l} = -p_k(x_i;\beta)\, p_l(x_i;\beta)\, x_i$$

$$\frac{\partial^2 p_k(x_i;\beta)}{\partial \beta_k \partial \beta_l} = -p_k(x_i;\beta)\, p_l(x_i;\beta)\, x_i x_i^T \leftarrow (9)$$

The full Hessian matrix is obtained by summing over all samples i:

$$H_{kl} = \sum_{i=1}^{n} H_{kl}^{(i)}$$

The second derivative of (6) gives us λI_p where I_p is a $p \times p$ identity matrix.

Thus,

$$H_{kk} = \sum_{i=1}^{n} p_k(x_i;\beta)\left(1 - p_k(x_i;\beta)\right) x_i x_i^T + \lambda I_p$$

$$H_{kl} = -\sum_{i=1}^{n} p_k(x_i;\beta)\, p_l(x_i;\beta)\, x_i x_i^T \quad , \quad (k \neq l)$$

Here, we approximate by assuming that the off-diagonal interactions between different classes are small and can be ignored.

Thus, we approximate the Hessian for class k by:

$$H_k \approx \sum_{i=1}^{n} p_k(x_i;\beta)\left(1 - p_k(x_i;\beta)\right) x_i x_i^T + \lambda I_p$$

The matrix form of this approximation is written as:

$$H_k = X^T W_k X + \lambda I_p \leftarrow (10)$$

The Damped Newton's update rule is given by,

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta H^{-1} \nabla f(\beta_k) \leftarrow (11)$$

where

η = learning rate (eta)

λ = ridge parameter

From (7), (10), and (11), we get the desired equation,

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left(X^T W_k X + \lambda I\right)^{-1} \left[X^T \{p_k - 1(Y=k)\} + \lambda \beta_k^{(t)}\right]$$

where $k = 0, 1, \ldots, K-1$