Derivations: STAT 600

Brigham Halverson

September 2024

First we find that $\log p_k(x_i;\beta) = \log e^{x_i^T\beta_k} - \log\sum_{l=0}^{K-1} e^{x_i^T\beta_k} = x_i^T\beta_k - \log\sum_{l=0}^{K-1} e^{x_i^T\beta_l}$
Next we find,

$$\begin{aligned}\frac{\partial \log p_k(x_i;\beta)}{\partial\beta_{kb}} &= x_{ib} - \frac{1}{\sum_{l=0}^{K-1} e^{x_i^T\beta_l}}(x_{ib}e^{x_i^T\beta_k}) \\ &= x_{ib}(1-p_k(x_i;\beta))\end{aligned}$$

It is also important to note that $\frac{\partial \log p_k(x_i;\beta)}{\partial\beta_{ab}}$ for $k \neq a$ we get the derivative is $-p_a(x_i;\beta)x_{ib}$
To begin this I will find the gradient of $f(\beta)$. We can find for $z = 0, ..., K-1$,

$$\begin{aligned}\frac{\partial f}{\partial\beta_{ab}} &= -\sum_{i=1}^{n}\{1(y_i = a)x_{ib}(1-p_a(x_i;\beta)) - \sum_{k\neq a} 1(y_i = k)p_a(x_i;\beta)x_{ib}\} + \lambda\beta_{ab} \\ &= \sum_{i=1}^{n}[x_{ib}\{\sum_{k\neq a} 1(y_i = k)p_a(x_i;\beta) + p_a(x_i;\beta)1(y_i = a) - 1(y_i = a)\}] + \lambda\beta_{ab} \\ &= \sum_{i=1}^{n}[x_{ib}(p_a(x_i;\beta)(\sum_{k=1}^{K-1} 1(y_i = k)) - 1(y_i = a))] + \lambda\beta_{ab} \\ &= \sum_{i=1}^{n}[x_{ib}(p_a(x_i;\beta) - 1(y_i = a))] + \lambda\beta_{ab} \\ &= x_b(P_a - 1(Y = a)) + \lambda\beta_{ab}\end{aligned}$$

where x_b is the b^{th} column of X.

This means that for β_k its gradient is in $\mathbb{R}^k$ and can be written as follows using the terms described in README as follows:

$$\frac{\partial f}{\partial\beta_k} = X^T\{P_k - 1(Y = k)\} + \lambda\beta_k$$

Now we will find the hessian, by focusing on $\frac{\partial}{\partial\beta_z}(\frac{\partial f}{\partial\beta_k})$. We find for:

$$\begin{aligned}\frac{\partial p_k(x_i;\beta)}{\partial\beta_{ac}} &= \frac{x_{ic}e^{x_i^T\beta_a}\sum_{l=0}^{K-1} e^{x_i^T\beta_l} - x_{ic}e^{x_i^T\beta_a}e^{x_i^T\beta_a}}{(\sum_{l=0}^{K-1} e^{x_i^T\beta_l})^2} \\ &= x_{ic}p_a(x_i;\beta)(\frac{\sum_{l=0}^{K-1} e^{x_i^T\beta_l} - e^{x_i^T\beta_a}}{(\sum_{l=0}^{K-1} e^{x_i^T\beta_l})}) \\ &= x_{ic}p_a(x_i;\beta)(1-p_a(x_i;\beta))\end{aligned}$$

So, we find for $z \neq k$,

$$\frac{\partial}{\partial \beta_z}(\frac{\partial f}{\partial \beta_k}) = \frac{\partial}{\partial \beta_z}(X^T\{P_k - 1(Y=k)\} + \lambda\beta_k)$$
$$= X^T[(P_k(1-P_k))X] + \mathbf{0}$$

Where $[(P_k(1-P_k))X]$ has rows of $p_k(x_i;\beta)(1-p_k(x_i;\beta)) * x_i$. Further for $z = k$ we find,

$$\frac{\partial}{\partial \beta_z}(\frac{\partial f}{\partial \beta_k}) = \frac{\partial}{\partial \beta_z}(X^T\{P_k - 1(Y=k)\} + \lambda\beta_k)$$
$$= X^T(P_k(1-P_k))X + \lambda\mathbf{1}$$

Thus we can write the full Hessian for β_k as follows:

$$X^T W_k X + \lambda I.$$

So then using the Damped Newton's update formula we find that the equation given in the slides will hold. when we plug in our answers we find:

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