Derivations

Arpan Ghosh

UIN: 236002102

Damped Newton's Method for multi-class Logistic Regression:

In the question, the multiclass logistic regression problem with ridge regularization,

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

where $P_k(x_i) = \dfrac{e^{x_i^T \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T \beta_l}}$ and $\lambda > 0$

We want to find the damped Newton's update step with learning rate η,

Note that,

$$f(\beta) = \left[- \sum_{i=1}^{n} \left\{ \sum_{l=0}^{K-1} \mathbb{1}(y_i = l)\, x_i^T \beta_l - \log\left(\sum_{l=0}^{K-1} e^{x_i' \beta_l} \right) \right\} + \frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{k,j}^2 \right]$$

Observe that, $-\mathbb{1}(y_i = l)\, x_i^T \beta_l$ is linear in β_l for each β_l and hence convex.

Also, taking $g(z) = \log\left(\sum_{k \neq l} e^{x_k^T \beta} + e^z\right)$

$$g'(z) = \frac{e^z}{\sum_{k \neq l} e^{x_k^T \beta} + e^z}$$

$$\Rightarrow g''(z) = \frac{e^z \sum_{k \neq l} e^{x_k^T \beta}}{\left(\sum_{k \neq l} e^{x_k^T \beta} + e^z\right)^2} \geq 0$$

$\Rightarrow g(z)$ is convex.

$\Rightarrow \log\left(\sum_{k=0}^{K-1} e^{x_k^T \beta}\right) = g(x_i^T \beta_l)$ is convex.

is convex in β_k.

further, $\sum_{j=1}^{p} \beta_{k,j}^2 = \|\beta_k\|^2$ is convex

So, addition of two also gives convex in each β_k.

So, $f(\beta)$ is convex in each β_k, it is enough

To find $\hat{\beta} = \arg\min_{\beta} f(\beta)$

to solve for $\frac{\partial f(\beta)}{\partial \beta_k} = 0 \quad \forall k$.

So, update step for damped Newton's method with learning rate η, is,

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left[\frac{\partial^2 f(\beta)}{\partial \beta_k^2} \Big|_{\beta_k = \beta_k^{(t)}}\right]^{-1} \left(\frac{\partial f(\beta)}{\partial \beta_k} \Big|_{\beta = \beta_k^{(t)}}\right)$$

Gradient:

From the above expression $f(B)$,

$$\frac{\partial f(B)}{\partial B_k} = \left[- \sum_{i=1}^{n} \left\{ \mathbb{1}(y_i = k) x_i - \frac{1}{\sum_{l=0}^{k-1} e^{x_i^T B_l}} \cdot e^{x_i^T B_k} x_i \right\} + \lambda B_k \right]$$

$$= \left[- \sum_{i=1}^{n} x_i \left\{ \mathbb{1}(y_i = k) - P_k(x_i) \right\} + \lambda B_k \right]$$

$$= \left[\sum_{i=1}^{n} X^T e_i e_i^T \left\{ P_k - \mathbb{1}(Y = k) \right\} + \lambda B_k \right]$$

$$= \left[X^T I_n \left\{ P_k - \mathbb{1}(Y = k) \right\} + \lambda B_k \right]$$

$$= \left[X^T \left\{ P_k - \mathbb{1}\{Y = k) \right\} + \lambda B_k \right]$$

[Here, e_i is the ith cannonical basis vector for $\mathbb{R}^n$ and I_n is the identity matrix]

$$\Rightarrow \left. \frac{\partial f(B)}{\partial B_k} \right|_{B = B_k^{(k)}} = \left[X^T \left\{ P_k^T - \mathbb{1}(Y = k) \right\} + \lambda B_k^{(k)} \right]$$

Hessian:

Now, $$\frac{\partial^2 f(B)}{\partial B_k^2} = \frac{\partial}{\partial B_k} \frac{\partial f(B)}{\partial B_k}$$

$$= \frac{\partial}{\partial B_k} \left[- \sum_{i=1}^{n} \left\{ \mathbb{1}(y_i = k) x_i - \frac{1}{\sum_{l=0}^{k-1} e^{x_i^T B_l}} \cdot e^{x_i^T B_k} x_i \right\} + \lambda B_k \right]$$

$$= \left[\sum_{i=1}^{n} x_i \left\{ \frac{\left(\sum_{l=0}^{k-1} e^{x_i^T \beta_l}\right) e^{x_i^T \beta_k} - e^{x_i^T \beta_k} \cdot e^{x_i^T \beta_k}}{\left(\sum_{l=0}^{k-1} e^{x_i^T \beta_l}\right)^2} \right\} x_i^T + \lambda I_k \right]$$

$$= \left[\sum_{i=1}^{n} X^T e_i \left\{ P_k(x_i) - P_k(x_i)^2 \right\} e_i^T X + \lambda I_k \right]$$

$$= \left[X^T \left\{ \sum_{i=1}^{n} e_i P_k(x_i) (1 - P_k(x_i)^2) e_i^T \right\} X + \lambda I_k \right]$$

$$= \left[X^T W_k X + \lambda I_k \right]$$

[Since, it is given, $\sum_{i=1}^{n} e_i P_k(x_i)(1-P_k(x_i)^2) e_i^T = W_k$,

$$\Rightarrow \left. \frac{\partial^2 f(\beta)}{\partial \beta_k^2} \right|_{\beta = \beta_k^{(t)}} = \left[X^T W_k^{(t)} X + \lambda I_k \right]$$

So, the final update eqn becomes,

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left[X^T W_k^{(t)} X + \lambda I_k \right]^{-1} \left[X^T \left\{ P_k^{(t)} - \mathbb{1}(y=k) \right\} + \lambda \beta_k^{(t)} \right]$$

(proved).