Stat 600 derivation

We start with:

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

We know that Damped Newtons method is $\beta^{(t+1)} = \beta^{(t)} - \eta \left(\nabla^2 f(\beta^{(t)})\right)^{-1} \nabla f(\beta^{(t)})$

So it suffices to show $\nabla^2 f(\beta^{(t)}) = (X^T W_k X + \lambda I)$

$$\nabla f(\beta^{(t)}) = \left[X^T \{p_k - \mathbb{1}(y=k)\} + \lambda \beta_k^{(t)} \right]$$

First let's do the gradient:

Regulation term:

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

$$\frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} 2\beta_{kj}$$

$$\lambda \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{kj}$$

which: $\lambda \beta_k^{(t)}$

for an individual

likelihood term:

$$-\sum_{k=0}^{K-1} \mathbb{1}_{\{y_i=k\}} \log p_k(x_i;\beta) \frac{\partial L_i(\beta)}{\partial \beta_k}$$

$$= -\mathbb{1}_{\{y_i=k\}} \frac{1}{p_k(x_i;\beta)} \cdot p_k(x_i;\beta) \cdot x_i + \sum_{l=0}^{K-1} \mathbb{1}(y_i=l) \cdot p_k(x_i;\beta) x_i$$

$$= -\mathbb{1}_{\{y_i=k\}} x_i + p_k(x_i;\beta) x_i$$

Summing over all n

$$= \sum_{i=1}^{n} (p_k(x_i;\beta) - \mathbb{1}(y_i=k)) x_i$$

$$= X^T (p_k - \mathbb{1}_{\{y=k\}})$$

Putting these together we get the

$$\nabla_f(\beta^{(t)}) = \left[X^T \{p_k - \mathbb{1}_{\{y=k\}}\} + \lambda \beta_k^{(t)} \right]$$

Now let's get the Hessian

let's split up into two parts:

$$\frac{\partial^2}{\partial \beta_k}\left(-\log p_k(x_i;\beta)\right) = -\underbrace{p_k(x_i;\beta)\left(1-p_k(x_i;\beta)\right)}_{\text{the form of } W_k} x_i x_i^T$$

Now for the regularizing term:

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

we saw the gradient was

$$\frac{\partial}{\partial \beta_k} \quad \lambda \beta_k^{(t)}$$

$$\lambda I$$

which we can put together as

$$(X^T W_k X + \lambda I)$$

which is what we wanted to show.

Scanned with CamScanner