Stat 600 HW3: Jivoung Park 531006793

Task 1. Prove $\beta_k^{(t+1)} = \beta_k^{(t)} - \eta(X^t W_k X + \lambda I)^{-1}[X^t(p_k - \mathbb{1}_{\{Y=k\}}) + \lambda \beta_k^{(t)}]$.
is a damped Newton method for multi-class Ridge regression.

Sol). Damped Newton's method is written as
$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta(\nabla^2_{k\times k} f(\beta))^{-1} \nabla_k f(\beta)$

In this case, $f(\beta) = (-\sum_i^n (\sum_k^{K-1} \mathbb{1}_{\{y_i=k\}} \log p_k(x_i)) + \frac{\lambda}{2}\sum_k^{K-1}\sum_j^p \beta_{k,j}^2$

$= -\sum_i^n (\sum_k^{K-1} \mathbb{1}_{\{y_i=k\}} (\log(e^{x_i^t\beta_k}) - \log(\sum_{l=0}^{K-1} e^{x_i^t\beta_l})) + \frac{\lambda}{2}\sum_k^{K-1}\sum_j^p \beta_{k,j}^2$

$= -\sum_i^n (\sum_k^{K-1} \mathbb{1}_{\{y_i=k\}} x_i^t\beta_k - \log(\sum_{l=0}^{K-1} e^{x_i^t\beta_l}) \underbrace{\sum_k^{K-1}\mathbb{1}_{\{y_i=k\}}}_{=1}) + \cdots$

$= -\sum_i^n (\sum_k^{K-1} \mathbb{1}_{\{y_i=k\}} x_i^t\beta_k - \log(\sum_{l=0}^{K-1} e^{x_i^t\beta_l})) + \frac{\lambda}{2}\sum_k^{K-1}\sum_j^p \beta_{k,j}^2$

① derivative

$\frac{\partial f(\beta)}{\partial \beta_m} = -\sum_i^n (\underbrace{\frac{\partial}{\partial \beta_m}\sum_k^{K-1}\mathbb{1}_{\{y_i=k\}} x_i^t\beta_k}_{\| \leftarrow \frac{\partial a^t x}{\partial x} = a} - \underbrace{\frac{\partial}{\partial \beta_m}\log(\sum_{l=0}^{K-1} e^{x_i^t\beta_l})}_{\|}) + \frac{\lambda}{2}\underbrace{\frac{\partial}{\partial \beta_m}\sum_k^{K-1}\sum_j^p \beta_{k,j}^2}_{\beta_k^t\beta_k}$

$x_i \mathbb{1}_{\{y_i=m\}} \qquad x_i \cdot \underbrace{\frac{e^{x_i^t\beta_m}}{\sum_{l=0}^{K-1} e^{x_i^t\beta_l}}}_{\| p_m(x_i)} \qquad \underbrace{\frac{\partial}{\partial \beta_m}\beta_m^t\beta_m}_{\| 2\beta_m} \rightarrow \underbrace{\lambda\beta_m}$

$\therefore \frac{\partial f(\beta)}{\partial \beta_k} = \underbrace{\sum_i^n x_i (p_k(x_i) - \mathbb{1}_{\{y_i=k\}})}_{= X^t(p_k - \mathbb{1}_{\{Y=k\}})} + \lambda\beta_k$ ■ ①

② hessian

$\frac{\partial f(\beta)}{\partial \beta_k \beta_m} = \frac{\partial}{\partial \beta_m^t} X^t(p_k - \mathbb{1}_{\{Y=k\}}) + \lambda\frac{\partial}{\partial \beta_m^t}\beta_k$

$= \sum_i^n \frac{\partial x_i(p_k(x_i) - \mathbb{1}_{\{y_i=k\}})}{\partial \beta_m^t} + \lambda\frac{\partial}{\partial \beta_m^t}\beta_k$

$= \sum_i^n x_i \frac{\partial}{\partial \beta_m^t} p_k(x_i) + \lambda\frac{\partial}{\partial \beta_m^t}\beta_k$

$= \sum_i^n x_i \underbrace{\frac{\partial \beta_m}{\partial \beta_m^t}}_{= \text{map}: \mathbb{R}^p \to (\mathbb{R}^p)^* \; x_i \mapsto x_i^t} \cdot \underbrace{\frac{\partial p_k(x_i)}{\partial \beta_m}}_{\otimes} + \lambda\frac{\partial \beta_m}{\partial \beta_m^t}\cdot\frac{\partial \beta}{\partial \beta_m}$

$(\because$ dual map's derivative is transposed identity map$)$

$\otimes = \frac{\mathbb{1}_{\{m=k\}} x_i \exp(x_i^t\beta_k)\cdot\sum_l^{K-1}\exp(x_i^t\beta_l) - \exp(x_i^t\beta_k)\cdot x_i\exp(x_i^t\beta_m)}{(\sum_{l=0}^{K-1}\exp(x_i^t\beta_l))^2}$

$= \mathbb{1}_{\{m=k\}} x_i p_k(x_i) - x_i p_k(x_i) p_m(x_i) = x_i p_k(x_i)(\mathbb{1}_{\{m=k\}} - p_m(x_i))$

plugging ⊛ $= x_i p_k(x_i)(\mathbb{1}_{\{m=k\}} - p_m(x_i))$ into original eq.

$$\frac{\partial f}{\partial \beta_k \beta_k} = \sum_i^n x_i \Big(\underbrace{x_i p_k(x_i)(\mathbb{1}_{\{m=k\}} - p_m(x_i))}_{\text{scalar. } \swarrow \text{ take out}} \Big)^t + \lambda \underbrace{\Big(\frac{\partial \beta_k}{\partial \beta_m}\Big)^t}_{= (\mathbb{1}_{\{m=k\}} I_\cdot)^t = \mathbb{1}_{\{m=k\}} I}$$

$$= \sum_i^n p_k(x_i)(\mathbb{1}_{\{m=k\}} - p_m(x_i)) \cdot x_i x_i^t \cdot + \lambda \mathbb{1}_{\{m=k\}} I.$$

Now, note $\nabla^2_{k\times k} f = \dfrac{\partial f}{\partial \beta_k \beta_k} = \sum_i^n p_k(x_i)(1 - p_k(x_i)) \cdot x_i x_i^t + \lambda I.$

$$= \sum_i^n W_{ki} x_i x_i^t + \lambda I = X^t W_k X + \lambda I$$

since $X^t W_k X = \begin{pmatrix} x_{11} \cdots x_{n1} \\ \vdots \\ x_{1p} \cdots x_{np} \end{pmatrix} \begin{pmatrix} p(x_1)(1-p(x_1)) & & \\ & \ddots & \\ & & \end{pmatrix} X$

$$= \Big(x_{ij} W_i^k \Big)_{\substack{i=1\cdots n \\ j=1\cdots p}} \begin{pmatrix} x_{11} \cdots x_{1p} \\ \vdots \\ x_{n1} \cdots x_{np} \end{pmatrix}$$

$$= \sum_k^n x_{ik} W_i^k x_{kj} = \sum_i^n W_i^k x_i x_i^t \quad \blacksquare\ ②$$

∴ $\begin{cases} \nabla_k f = X^t(P_k - \mathbb{1}_{\{Y=k\}}) + \lambda \beta_k \\ \nabla^2_{k\times k} f = X^t W_k X + \lambda I. \end{cases}$

∴ $\beta_k^{(t+1)} = \beta_k^{(t)} - \eta (\nabla^2_{k\times k} f)^{-1} \nabla_k f$ (dumped Newton)

$$= \beta_k^{(t)} - \eta (X^t W_k X + \lambda I)^{-1} \cdot (X^t(P_k - \mathbb{1}_{\{Y=k\}}) + \lambda(\beta_k)$$

(target form).

■ (Task 1)