Damped Newton's Method Derivation for Multi-class Logistic Regression

YANG Chen

2024-09-23

Objective Function

The objective function for multi-class logistic regression with ridge regularization is given by:

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

where

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

Gradient Derivation

To derive the gradient, we need to calculate $\frac{\partial f}{\partial \beta_k}$ for each k.

First, let's calculate $\frac{\partial \log p_k}{\partial \beta_k}$:

$$\begin{aligned}
\frac{\partial \log p_k}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \left(x_i^\top \beta_k - \log \sum_{l=0}^{K-1} e^{x_i^\top \beta_l} \right) \\
&= x_i - \frac{e^{x_i^\top \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^\top \beta_l}} x_i \\
&= x_i(1 - p_k)
\end{aligned}$$

Similarly, for $m \neq k$:

$$\frac{\partial \log p_m}{\partial \beta_k} = - \frac{e^{x_i^\top \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^\top \beta_l}} x_i = -p_k x_i$$

Now, we can calculate the gradient:

$$
\begin{aligned}
\frac{\partial f}{\partial \beta_k} &= -\sum_{i=1}^{n}\left[1(y_i = k)x_i(1-p_k) - \sum_{m \neq k} 1(y_i = m)p_k x_i\right] + \lambda\beta_k \\
&= -\sum_{i=1}^{n} x_i\left[1(y_i = k) - p_k\right] + \lambda\beta_k \\
&= -X^\top\left[1(Y = k) - P_k\right] + \lambda\beta_k
\end{aligned}
$$

where $1(Y = k)$ is a vector of indicators and P_k is a vector of probabilities $p_k(x_i; \beta)$.

Hessian Derivation

Now, let's calculate the Hessian. We need to compute $\frac{\partial^2 f}{\partial \beta_k \partial \beta_m}$ for all k and m.

For $k = m$:

$$
\begin{aligned}
\frac{\partial^2 f}{\partial \beta_k^2} &= \frac{\partial}{\partial \beta_k}\left(-X^\top\left[1(Y = k) - P_k\right] + \lambda\beta_k\right) \\
&= X^\top \text{diag}(p_k(1-p_k))X + \lambda I \\
&= X^\top W_k X + \lambda I
\end{aligned}
$$

For $k \neq m$:

$$
\begin{aligned}
\frac{\partial^2 f}{\partial \beta_k \partial \beta_m} &= \frac{\partial}{\partial \beta_m}\left(-X^\top\left[1(Y = k) - P_k\right] + \lambda\beta_k\right) \\
&= X^\top \text{diag}(p_k p_m)X
\end{aligned}
$$

Newton's Method

The Newton's method update is given by:

$$
\beta^{(t+1)} = \beta^{(t)} - H^{-1}g
$$

where H is the Hessian and g is the gradient.

For the multi-class case, we can write this as a system of equations:

$$
\begin{bmatrix}
H_{00} & H_{01} & \cdots & H_{0,K-1} \\
H_{10} & H_{11} & \cdots & H_{1,K-1} \\
\vdots & \vdots & \ddots & \vdots \\
H_{K-1,0} & H_{K-1,1} & \cdots & H_{K-1,K-1}
\end{bmatrix}
\begin{bmatrix}
\Delta\beta_0 \\
\Delta\beta_1 \\
\vdots \\
\Delta\beta_{K-1}
\end{bmatrix}
= -
\begin{bmatrix}
g_0 \\
g_1 \\
\vdots \\
g_{K-1}
\end{bmatrix}
$$

where $H_{km} = X^\top \text{diag}(p_k p_m)X$ for $k \neq m$, and $H_{kk} = X^\top W_k X + \lambda I$.

Simplification

The system above is complicated to solve directly. However, we can simplify it by noting that:

$$
\sum_{k=0}^{K-1} p_k = 1 \implies \sum_{k=0}^{K-1} \frac{\partial p_k}{\partial \beta_m} = 0
$$

This allows us to treat each β_k update independently:

$$(X^\top W_k X + \lambda I)\Delta\beta_k = -g_k$$

Damped Newton's Method

The damped Newton's method introduces a learning rate η:

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta(X^\top W_k X + \lambda I)^{-1} g_k$$

Substituting the expression for g_k, we get:

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

This completes the derivation of the damped Newton's method update for multi-class logistic regression.