Derivation of Gradient and Hessian for Multi-class Logistic Regression

Objective Function

We start with the objective function for multi-class logistic regression with ridge regularization:

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

where

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

Partial derivation formula for class probability

Definition of $p_k(x_i; \beta)$

In multi-class logistic regression, the probability of the i-th sample x_i belonging to class k is given by:

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

where

- x_i is the feature of the i -th sample.
- β_k is the parameter vector for class $k \in 0, 1, \ldots, K-1$

Derivative with respect to β_k

First, we take the derivative with respect to β_k, which is the parameter vector for the class itself. The derivative is given by:

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_k} \tag{2}$$

1. For the numerator $e^{x_i^\top \beta_k}$, the derivative with respect to β_k is:

$$\frac{\partial e^{x_i^\top \beta_k}}{\partial \beta_k} = e^{x_i^\top \beta_k} \cdot x_i \tag{3}$$

2. For the denominator $\sum_{l=0}^{K-1} e^{x_i^\top \beta_l}$, the derivative with respect to β_k is:

$$\frac{\partial \sum_{l=0}^{K-1} e^{x_i^\top \beta_l}}{\partial \beta_k} = e^{x_i^\top \beta_k} \cdot x_i \tag{4}$$

3. Using the quotient rule:

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_k} = \frac{\partial \left(\frac{e^{x_i^\top \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^\top \beta_l}} \right)}{\partial \beta_k} = \frac{e^{x_i^\top \beta_k} \cdot x_i \cdot \sum_{l=0}^{K-1} e^{x_i^\top \beta_l} - e^{x_i^\top \beta_k} \cdot \sum_{l=0}^{K-1} e^{x_i^\top \beta_l} \cdot x_i}{\left(\sum_{l=0}^{K-1} e^{x_i^\top \beta_l} \right)^2} \tag{5}$$

4. Simplifying, we get:

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_k} = p_k(x_i;\beta)\,(1 - p_k(x_i;\beta))x_i \tag{6}$$

Derivative with respect to β_l (where $l \neq k$)

Next, we calculate the derivative with respect to another class parameter β_l (where $l \neq k$)

1. The numerator $e^{x_i^\top \beta_k}$, so its derivative is 0.
2. The denominator $\sum_{l=0}^{K-1} e^{x_i^\top \beta_l}$ contains the term $e^{x_i^\top \beta_l}$, so its derivative with respect to β_l is:

$$\frac{\partial \sum_{l=0}^{K-1} e^{x_i^\top \beta_l}}{\partial \beta_l} = e^{x_i^\top \beta_l} \cdot x_i \tag{7}$$

3. Using the quotient rule:

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_l} = \frac{0 \cdot \sum_{l=0}^{K-1} e^{x_i^\top \beta_l} - e^{x_i^\top \beta_k} \cdot e^{x_i^\top \beta_l} \cdot x_i}{\left(\sum_{l=0}^{K-1} e^{x_i^\top \beta_l} \right)^2} \tag{8}$$

4. Simplifying, we get:

$$\frac{\partial p_k(x_i;\beta)}{\partial \beta_l} = -p_k(x_i;\beta)p_l(x_i;\beta)x_i \tag{9}$$

Gradient Derivation

To calculate the gradient, we find the derivative with respect to β_k. We'll use the gradient notation ∇_{β_k} to denote this:

$\nabla_{\beta_k} f = -\sum_{i=1}^{n} \left\{ \sum_{m=0}^{K-1} 1(y_i = m) \cdot \nabla_{\beta_k} \log p_m(x_i; \beta) \right\} + \lambda \beta_k$

We have calculated $\nabla_{\beta_k} p_m(x_i; \beta)$:

$\nabla_{\beta_k} p_m(x_i; \beta) = \begin{cases} -x_i \cdot p_m(x_i; \beta) \cdot p_k(x_i; \beta) & \text{if } m \neq k \\ x_i \cdot p_k(x_i; \beta) \cdot (1 - p_k(x_i; \beta)) & \text{if } m = k \end{cases}$

Substituting this result back into the gradient expression:

$$
\begin{aligned}
\nabla_{\beta_k} f &= -\sum_{i=1}^{n} \left\{ 1(y_i = k) \cdot x_i \cdot (1 - p_k(x_i; \beta)) + \sum_{m \neq k} 1(y_i = m) \cdot x_i \cdot (-p_k(x_i; \beta)) \right\} + \lambda \beta_k \\
&= -\sum_{i=1}^{n} \left\{ 1(y_i = k) \cdot x_i - 1(y_i = k) \cdot x_i \cdot p_k(x_i; \beta) - \sum_{m \neq k} 1(y_i = m) \cdot x_i \cdot p_k(x_i; \beta) \right\} + \lambda \beta_k \\
&= -\sum_{i=1}^{n} \left\{ 1(y_i = k) \cdot x_i - x_i \cdot p_k(x_i; \beta) \cdot \left(1(y_i = k) + \sum_{m \neq k} 1(y_i = m) \right) \right\} + \lambda \beta_k \\
&= -\sum_{i=1}^{n} \left\{ 1(y_i = k) \cdot x_i - x_i \cdot p_k(x_i; \beta) \cdot \left(\sum_{m=0}^{K-1} 1(y_i = m) \right) \right\} + \lambda \beta_k \\
&= -\sum_{i=1}^{n} \left\{ 1(y_i = k) \cdot x_i - x_i \cdot p_k(x_i; \beta) \cdot 1 \right\} + \lambda \beta_k \\
&= -\sum_{i=1}^{n} \left\{ x_i \cdot (1(y_i = k) - p_k(x_i; \beta)) \right\} + \lambda \beta_k \\
&= X^\top (P_k - 1(Y = k)) + \lambda \beta_k
\end{aligned}
$$

where X is the design matrix, $1(Y = k)$ is an indicator vector for class k, and P_k is a vector containing all $p_k(x_i; \beta)$.

Hessian Derivation

To derive W_k directly, we calculate the second derivative of f with respect to β_k. Here, we'll use partial derivative notation as it's more common for Hessian calculations:

$\frac{\partial^2 f}{\partial \beta_k \partial \beta_k} = -X^\top \frac{\partial}{\partial \beta_k} (1(Y = k) - P_k) + \lambda I$

Since the indicator function 1(Y = k) doesn't depend on β_k, we only need to differentiate P_k:

$\frac{\partial^2 f}{\partial \beta_k \partial \beta_k} = X^\top \frac{\partial P_k}{\partial \beta_k} + \lambda I$

Now, we calculate $\frac{\partial p_k(x_i;\beta)}{\partial \beta_k}$. For a single sample:

$$\begin{aligned}\frac{\partial p_k(x_i;\beta)}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k}\left(\frac{e^{x_i^\top \beta_k}}{\sum_{m=0}^{K-1} e^{x_i^\top \beta_m}}\right)\\ &= p_k(x_i;\beta)\cdot x_i - p_k(x_i;\beta)\cdot x_i \cdot p_k(x_i;\beta)\\ &= p_k(x_i;\beta)\cdot(1-p_k(x_i;\beta))\cdot x_i\end{aligned}$$

Substituting this back into our Hessian calculation:

$$\begin{aligned}\frac{\partial^2 f}{\partial \beta_k \partial \beta_k} &= X^\top \mathrm{diag}(p_k(x_i;\beta)\cdot(1-p_k(x_i;\beta)))X + \lambda I\\ &= X^\top W_k X + \lambda I\end{aligned}$$

where Wk is a diagonal matrix with entries:

$$W_{k,ii} = p_k(x_i;\beta)\cdot(1-p_k(x_i;\beta))$$

Conclusion

We have derived the gradient:

$$\nabla_{\beta_k} f = X^\top (P_k - 1(Y=k)) + \lambda \beta_k$$

and the Hessian (in terms of W_k):

$$\frac{\partial^2 f}{\partial \beta_k \partial \beta_k} = X^\top W_k X + \lambda I$$

where $W_{k,ii} = p_k(x_i;\beta)\cdot(1-p_k(x_i;\beta))$

These expressions can be used in the damped Newton's method update:

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

where η is the learning rate.