For some multiclass Logistic Regression model,

we know that $J = -\frac{1}{N}\sum_{i=1}^{N}(L_i + L2)$ where $L_i = \sum_{k=1}^{K} y_{ik}\log_e(\hat{y}_{ik})$

$$L2 = \sum_{k=1}^{k}\lambda\sum_{j=1}^{d}\omega_{kj}^2$$

For $L2$, $\frac{\partial L2}{\partial \omega_{a,b}} = 2\lambda\omega_{ab}$ — (A)

For L_i, $\frac{\partial L_i}{\partial \omega_{a,b}} = \frac{\partial L_i}{\partial \hat{y}_{ia}} \cdot \frac{\partial \hat{y}_{ia}}{\partial \omega_{ab}}$

(1) $$\frac{\partial L_i}{\partial \hat{y}_{ia}} = \frac{\partial}{\partial \hat{y}_{ia}}\left[y_{i1}\ln(\hat{y}_{i1}) + \dots + y_{ia}\ln(\hat{y}_{ia}) + \dots\right] = y_{ia}\frac{1}{\hat{y}_{ia}}$$

(2) $$\frac{\partial \hat{y}_{ia}}{\partial \omega_{ab}} = \frac{\partial}{\partial \omega_{ab}}\left(\frac{e^{\hat{y}_{ia}}}{\sum_j e^{\hat{y}_{ij}}}\right)$$

a) when $k=a$,

$$= \frac{\hat{y}_{ia}\sum_j e^{\hat{y}_{ij}} - e^{\hat{y}_{ia}}\hat{y}_{ia}}{\left(\sum_j e^{\hat{y}_{ij}}\right)^2} \cdot x_b$$

$$= \hat{y}_{ia}(1-\hat{y}_{ia})x_b$$

b) when $k \neq a$, similarly,

$$= -\hat{y}_{ia}\hat{y}_{ik}x_b$$

$$\Rightarrow \Delta\omega_{ab} \approx \frac{\partial L_i}{\partial \omega_{ab}} = \sum_{k=1}^{K}(1 \cdot 2) = 1 \cdot 2a + \sum_{k}^{K} 1 \cdot 2b$$

$$= \cancel{\hat{y}_{ia}}(1-\hat{y}_{ia})x_b y_{ia}\frac{1}{\cancel{\hat{y}_{ia}}}$$

$$+ \sum_{k}^{K} -\cancel{\hat{y}_{ia}}\hat{y}_{ik}x_b y_{ia}\frac{1}{\cancel{\hat{y}_{ia}}}$$

$$= -x_b\left(y_{ia}\hat{y}_{ia} - y_{ia} + \sum_{k}^{K}\hat{y}_{ik}y_{ia}\right)$$

$$= -x_b\left(\sum_{k}^{K}\hat{y}_{ik}y_{ia} - y_{ia}\right)$$

$$= -x_b(\hat{y}_{ia} - y_{ia})$$

$$= (y_{ia} - \hat{y}_{ia})x_b \quad \text{— (B)}$$

Combining (A) & (B),

$$\boxed{\Delta\omega_{ab} \approx \frac{\partial L}{\partial \omega_{ab}} = -\frac{1}{N}\sum_{i=1}^{N}\left[(y_{ia} - \hat{y}_{ia})x_b + 2\lambda\omega_{ab}\right]}$$

In matrix form,

$$\boxed{\Delta\omega = \frac{1}{N}\left[X^T(Y - \hat{Y})\right] + 2\lambda\omega}$$

Agamdeep S. Chopra