Derivation of Gradient Descent for Logistic Regression (multi-class)

Derive $W = W - \frac{\alpha}{n}X_b^T(Y-\hat{Y})$

เรามี $Cost = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{k} y_{i,c}\log\hat{y}_{i,c}$ และ $W = W - \alpha\nabla Cost$ โดยที่ $W = \begin{bmatrix} w_{0,1} & w_{0,2} & \cdots & w_{0,k} \\ w_{1,1} & w_{1,2} & \cdots & w_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p,1} & w_{p,2} & \cdots & w_{p,k} \end{bmatrix}$

$$\nabla Cost = \begin{bmatrix} \frac{\partial Cost}{\partial w_{0,1}} & \frac{\partial Cost}{\partial w_{0,2}} & \cdots & \frac{\partial Cost}{\partial w_{0,k}} \\ \frac{\partial Cost}{\partial w_{1,1}} & \frac{\partial Cost}{\partial w_{1,2}} & \cdots & \frac{\partial Cost}{\partial w_{1,k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Cost}{\partial w_{p,1}} & \frac{\partial Cost}{\partial w_{p,2}} & \cdots & \frac{\partial Cost}{\partial w_{p,k}} \end{bmatrix}$$

พิจารณา $\frac{\partial Cost}{\partial w_{d,s}}$ เมื่อ $d \in \{0,1,...,p\}, s \in \{1,2,...,k\}$ จะได้ว่า

$$\frac{\partial Cost}{\partial w_{d,s}} = -\frac{1}{n}\sum_{i=1}^{n} x_{i,d}(y_{i,s}-\hat{y}_{i,s}) \tag{1}$$

ซึ่งสามารถเขียนให้อยู่ในรูป matrix ได้ดังนี้

$$\frac{\partial Cost}{\partial w_{d,s}} = -\frac{1}{n}\begin{bmatrix} x_{1,d} & x_{2,d} & \cdots & x_{n,d} \end{bmatrix}\left(\begin{bmatrix} y_{1,s} \\ y_{2,s} \\ \vdots \\ y_{n,s} \end{bmatrix} - \begin{bmatrix} \hat{y}_{1,s} \\ \hat{y}_{2,s} \\ \vdots \\ \hat{y}_{n,s} \end{bmatrix}\right) \tag{2}$$

จาก $\nabla Cost$ เราสามารถเขียนให้อยู่ในรูปต่อไปนี้

$$\begin{bmatrix} \frac{\partial Cost}{\partial w_{0,1}} & \frac{\partial Cost}{\partial w_{0,2}} & \cdots & \frac{\partial Cost}{\partial w_{0,k}} \\ \frac{\partial Cost}{\partial w_{1,1}} & \frac{\partial Cost}{\partial w_{1,2}} & \cdots & \frac{\partial Cost}{\partial w_{1,k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Cost}{\partial w_{p,1}} & \frac{\partial Cost}{\partial w_{p,2}} & \cdots & \frac{\partial Cost}{\partial w_{p,k}} \end{bmatrix} = -\frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,p} & x_{2,p} & \cdots & x_{n,p} \end{bmatrix}\left(\begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,k} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,k} \end{bmatrix} - \begin{bmatrix} \hat{y}_{1,1} & \hat{y}_{1,2} & \cdots & \hat{y}_{1,k} \\ \hat{y}_{2,1} & \hat{y}_{2,2} & \cdots & \hat{y}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{n,1} & \hat{y}_{n,2} & \cdots & \hat{y}_{n,k} \end{bmatrix}\right)$$

โดยที่ $x_{i,0} = 1,$ สำหรับ $i \in \{1,2,...,n\}$

กำหนดให้

$$X_b = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}, Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,k} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,k} \end{bmatrix}, \hat{Y} = \begin{bmatrix} \hat{y}_{1,1} & \hat{y}_{1,2} & \cdots & \hat{y}_{1,k} \\ \hat{y}_{2,1} & \hat{y}_{2,2} & \cdots & \hat{y}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{n,1} & \hat{y}_{n,2} & \cdots & \hat{y}_{n,k} \end{bmatrix}$$

นั่นคือ $\nabla Cost = -\frac{1}{n}X_b^T(Y - \hat{Y})$

$$\because W = W - \alpha\nabla Cost$$

ดังนั้น $W = W + \frac{\alpha}{n}X_b^T(Y - \hat{Y})$

Derive Equation 1

$$
\begin{aligned}
\frac{\partial Cost}{\partial w_{d,s}} &= \frac{\partial \frac{-1}{n}\sum_{i=1}^{n}\sum_{c=1}^{k} y_{i,c}\log \hat{y}_{i,c}}{\partial w_{d,s}} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \sum_{c=1}^{k} y_{i,c}\log \hat{y}_{i,c}}{\partial w_{d,s}} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial[y_{i,1}\log \hat{y}_{i,1} + y_{i,2}\log \hat{y}_{i,2} + \cdots + y_{i,s}\log \hat{y}_{i,s} + \cdots + y_{i,k}\log \hat{y}_{i,k}]}{\partial w_{d,s}} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[\frac{\partial y_{i,1}\log \hat{y}_{i,1}}{\partial w_{d,s}} + \frac{\partial y_{i,2}\log \hat{y}_{i,2}}{\partial w_{d,s}} + \cdots + \frac{\partial y_{i,s}\log \hat{y}_{i,s}}{\partial w_{d,s}} + \cdots + \frac{\partial y_{i,k}\log \hat{y}_{i,k}}{\partial w_{d,s}}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[y_{i,1}\frac{\partial \log \hat{y}_{i,1}}{\partial w_{d,s}} + y_{i,2}\frac{\partial \log \hat{y}_{i,2}}{\partial w_{d,s}} + \cdots + y_{i,s}\frac{\partial \log \hat{y}_{i,s}}{\partial w_{d,s}} + \cdots + y_{i,k}\frac{\partial \log \hat{y}_{i,k}}{\partial w_{d,s}}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[y_{i,1}\frac{\partial \log \hat{y}_{i,1}}{\partial \hat{y}_{i,1}} \cdot \frac{\partial \hat{y}_{i,1}}{\partial z_{i,s}} \cdot \frac{\partial z_{i,s}}{\partial w_{d,s}} + y_{i,2}\frac{\partial \log \hat{y}_{i,2}}{\partial \hat{y}_{i,2}} \cdot \frac{\partial \hat{y}_{i,2}}{\partial z_{i,s}} \cdot \frac{\partial z_{i,s}}{\partial w_{d,s}} + \cdots \right. \\
&\qquad \left. + y_{i,s}\frac{\partial \log \hat{y}_{i,s}}{\partial \hat{y}_{i,s}} \cdot \frac{\partial \hat{y}_{i,s}}{\partial z_{i,s}} \cdot \frac{\partial z_{i,s}}{\partial w_{d,s}} + \cdots + y_{i,k}\frac{\partial \log \hat{y}_{i,k}}{\partial \hat{y}_{i,k}} \cdot \frac{\partial \hat{y}_{i,k}}{\partial z_{i,s}} \cdot \frac{\partial z_{i,s}}{\partial w_{d,s}}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[y_{i,1}\left(\frac{1}{\hat{y}_{i,1}}\right)\frac{\partial \hat{y}_{i,1}}{\partial z_{i,s}}(x_{i,d}) + y_{i,2}\left(\frac{1}{\hat{y}_{i,2}}\right)\frac{\partial \hat{y}_{i,2}}{\partial z_{i,s}}(x_{i,d}) + \cdots \right. \\
&\qquad \left. + y_{i,s}\left(\frac{1}{\hat{y}_{i,s}}\right)\frac{\partial \hat{y}_{i,s}}{\partial z_{i,s}}(x_{i,d}) + \cdots + y_{i,k}\left(\frac{1}{\hat{y}_{i,k}}\right)\frac{\partial \hat{y}_{i,k}}{\partial z_{i,s}}(x_{i,d})\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[\frac{y_{i,1}}{\hat{y}_{i,1}}\frac{\partial \hat{y}_{i,1}}{\partial z_{i,s}}x_{i,d} + \frac{y_{i,2}}{\hat{y}_{i,2}}\frac{\partial \hat{y}_{i,2}}{\partial z_{i,s}}x_{i,d} + \cdots + \frac{y_{i,s}}{\hat{y}_{i,s}}\frac{\partial \hat{y}_{i,s}}{\partial z_{i,s}}x_{i,d} + \cdots + \frac{y_{i,k}}{\hat{y}_{i,k}}\frac{\partial \hat{y}_{i,k}}{\partial z_{i,s}}x_{i,d}\right]
\end{aligned}
$$

เนื่องจาก

$$
\frac{\partial \hat{y}_{i,m}}{\partial z_{i,s}} = \hat{y}_{i,s}(1-\hat{y}_{i,s}) \qquad \text{เมื่อ } m = s \tag{3}
$$

$$
\frac{\partial \hat{y}_{i,m}}{\partial z_{i,s}} = -\hat{y}_{i,m}\hat{y}_{i,s} \qquad \text{เมื่อ } m \neq s \tag{4}
$$

ดังนั้น

$$
\begin{aligned}
\frac{\partial Cost}{\partial w_{d,s}} &= -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{y_{i,1}}{\hat{y}_{i,1}}(-\hat{y}_{i,1}\hat{y}_{i,s})x_{i,d} + \frac{y_{i,2}}{\hat{y}_{i,2}}(-\hat{y}_{i,2}\hat{y}_{i,s})x_{i,d} + \cdots \\
&\qquad + \frac{y_{i,s}}{\hat{y}_{i,s}}(\hat{y}_{i,s}(1-\hat{y}_{i,s}))x_{i,d} + \cdots + \frac{y_{i,k}}{\hat{y}_{i,k}}(-\hat{y}_{i,k}\hat{y}_{i,s})x_{i,d}\Bigg] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[-y_{i,1}\hat{y}_{i,s}x_{i,d} - y_{i,2}\hat{y}_{i,s}x_{i,d} - \cdots + y_{i,s}(1-\hat{y}_{i,s})x_{i,d} - \cdots - y_{i,k}\hat{y}_{i,s}x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(-y_{i,1}\hat{y}_{i,s} - y_{i,2}\hat{y}_{i,s} - \cdots + y_{i,s}(1-\hat{y}_{i,s}) - \cdots - y_{i,k}\hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(-y_{i,1}\hat{y}_{i,s} - y_{i,2}\hat{y}_{i,s} - \cdots + y_{i,s} - y_{i,s}\hat{y}_{i,s} - \cdots - y_{i,k}\hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(y_{i,s} - y_{i,1}\hat{y}_{i,s} - y_{i,2}\hat{y}_{i,s} - \cdots - y_{i,s}\hat{y}_{i,s} - \cdots - y_{i,k}\hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(y_{i,s} - (y_{i,1} + y_{i,2} + \cdots + y_{i,s} + \cdots + y_{i,k})\hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(y_{i,s} - (1)\hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}[(y_{i,s} - \hat{y}_{i,s})x_{i,d}] \\
&= -\frac{1}{n}\sum_{i=1}^{n}x_{i,d}(y_{i,s} - \hat{y}_{i,s})
\end{aligned}
$$

Derive Equation 2

จาก $\frac{\partial Cost}{\partial w_{d,s}} = -\frac{1}{n}\sum_{i=1}^{n} x_{i,d}(y_{i,s} - \hat{y}_{i,s})$

พิจารณา $-\frac{1}{n}\sum_{i=1}^{n} x_{i,d}(y_{i,s} - \hat{y}_{i,s})$

$$
\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n} x_{i,d}(y_{i,s} - \hat{y}_{i,s}) &= -\frac{1}{n}\left[x_{1,d}(y_{1,s} - \hat{y}_{1,s}) + x_{2,d}(y_{2,s} - \hat{y}_{2,s}) + \cdots + x_{n,d}(y_{n,s} - \hat{y}_{n,s})\right] \\
&= -\frac{1}{n}\begin{bmatrix} x_{1,d} & x_{2,d} & \cdots & x_{n,d} \end{bmatrix}\begin{bmatrix} y_{1,s} - \hat{y}_{1,s} \\ y_{2,s} - \hat{y}_{2,s} \\ \vdots \\ y_{n,s} - \hat{y}_{n,s} \end{bmatrix} \\
&= -\frac{1}{n}\begin{bmatrix} x_{1,d} & x_{2,d} & \cdots & x_{n,d} \end{bmatrix}\left(\begin{bmatrix} y_{1,s} \\ y_{2,s} \\ \vdots \\ y_{n,s} \end{bmatrix} - \begin{bmatrix} \hat{y}_{1,s} \\ \hat{y}_{2,s} \\ \vdots \\ \hat{y}_{n,s} \end{bmatrix}\right)
\end{aligned}
$$

ดังนั้น $\frac{\partial Cost}{\partial w_d} = -\frac{1}{n}\begin{bmatrix} x_{1,d} & x_{2,d} & \cdots & x_{n,d} \end{bmatrix}\left(\begin{bmatrix} y_{1,s} \\ y_{2,s} \\ \vdots \\ y_{n,s} \end{bmatrix} - \begin{bmatrix} \hat{y}_{1,s} \\ \hat{y}_{2,s} \\ \vdots \\ \hat{y}_{n,s} \end{bmatrix}\right)$

Derive Equation 3

$$
\begin{aligned}
\frac{\partial \hat{y}_{i,s}}{\partial z_{i,s}} &= \frac{\partial \left(\frac{e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}}\right)}{\partial e^{z_{i,s}}} \cdot \frac{\partial e^{z_{i,s}}}{\partial z_{i,s}} \\
&= \frac{\sum_{c=1}^{k} e^{z_{i,c}} \frac{\partial e^{z_{i,s}}}{\partial e^{z_{i,s}}} - e^{z_{i,s}} \frac{\partial \sum_{c=1}^{k} e^{z_{i,c}}}{\partial e^{z_{i,s}}}}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} \frac{\partial e^{z_{i,s}}}{\partial z_{i,s}} \\
&= \frac{\sum_{c=1}^{k} e^{z_{i,c}}(1) - e^{z_{i,s}}(1)}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} e^{z_{i,s}} \\
&= \frac{e^{z_{i,s}} \left(\sum_{c=1}^{k} e^{z_{i,c}} - e^{z_{i,s}}\right)}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} \\
&= \frac{e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}} \frac{\sum_{c=1}^{k} e^{z_{i,c}} - e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}} \\
&= \frac{e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}} \left(\frac{\sum_{c=1}^{k} e^{z_{i,c}}}{\sum_{c=1}^{k} e^{z_{i,c}}} - \frac{e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}}\right) \\
&= \hat{y}_{i,s}(1 - \hat{y}_{i,s})
\end{aligned}
$$

Derive Equation 4

$$
\begin{aligned}
\frac{\partial \hat{y}_{i,m}}{\partial z_{i,s}} &= \frac{\partial \left(\frac{e^{z_{i,m}}}{\sum_{c=1}^{k} e^{z_{i,c}}}\right)}{\partial e^{z_{i,s}}} \cdot \frac{\partial e^{z_{i,s}}}{\partial z_{i,s}} \\
&= \frac{\sum_{c=1}^{k} e^{z_{i,c}} \dfrac{\partial e^{z_{i,m}}}{\partial e^{z_{i,s}}} - e^{z_{i,m}} \dfrac{\partial \sum_{c=1}^{k} e^{z_{i,c}}}{\partial e^{z_{i,s}}}}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} \frac{\partial e^{z_{i,s}}}{\partial z_{i,s}} \\
&= \frac{\sum_{c=1}^{k} e^{z_{i,c}}(0) - e^{z_{i,m}}(1)}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} e^{z_{i,s}} \\
&= \frac{-e^{z_{i,m}} e^{z_{i,s}}}{\left(\sum_{c=1}^{k} e^{z_{i,c}}\right)^2} \\
&= \left(\frac{-e^{z_{i,m}}}{\sum_{c=1}^{k} e^{z_{i,c}}}\right) \left(\frac{e^{z_{i,s}}}{\sum_{c=1}^{k} e^{z_{i,c}}}\right) \\
&= -\hat{y}_{i,m} \hat{y}_{i,s}
\end{aligned}
$$