Let X be training data matrix with additional column of 1s on the first column, β be the coefficient matrix. $\vec{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{im})$ is a datum on row i in X. $\overrightarrow{\beta_k} = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{mk})$ is a vector in β of class k in which there are K classes in response y. Let $\hat{y_{ik}}$ be the prediction (or probability) of $\vec{x}_i$ on class k, then $\hat{y}_k = (\hat{y_{1k}}, \hat{y_{2k}}, \ldots, \hat{y_{nk}})^T$. Since

$$\hat{y_{ik}} = P[Y = k \,|\, X = \vec{x}_i] = \frac{e^{\vec{\beta}_k \cdot \vec{x}_i}}{\sum_{l=1}^{K} e^{\vec{\beta}_l \cdot \vec{x}_i}}$$

$$\rightarrow \ln(\hat{y_{ik}}) = \vec{\beta}_k \cdot \vec{x}_i - \ln\Big(\sum_{l=1}^{K} e^{\vec{\beta}_l \cdot \vec{x}_i}\Big)$$

$$\rightarrow \begin{cases} \nabla_{\vec{\beta}_k} \ln(\hat{y_{ik}}) = \vec{x}_i - \dfrac{\nabla_{\vec{\beta}_k}[\sum_l^K e^{\vec{\beta}_l \cdot \vec{x}_i}]}{\sum_l^K e^{\vec{\beta}_l \cdot \vec{x}_i}} = \vec{x}_i - \dfrac{e^{\vec{\beta}_k \cdot \vec{x}_i}}{\sum_l^K e^{\vec{\beta}_l \cdot \vec{x}_i}}\vec{x}_i = \vec{x}_i - \hat{y}_{ik}\vec{x}_i = (1 - \hat{y}_{ik})\vec{x}_i \\ \nabla_{\vec{\beta}_j} \ln(\hat{y_{ik}}) = -\dfrac{e^{\vec{\beta}_j \cdot \vec{x}_i}}{\sum_l^K e^{\vec{\beta}_l \cdot \vec{x}_i}}\vec{x}_i = -\hat{y}_{ij}\vec{x}_i \end{cases}$$

Now we use likelihood function l to build loss function and try to maximize it to find matrix β in which

$$l = \prod_{i_1:Y=1} \hat{y_{i_1 1}} \prod_{i_2:Y=2} \hat{y_{i_2 2}} \cdots \prod_{i_K:Y=K} \hat{y_{i_K K}}$$

$$\rightarrow \ln(l) = \sum_{i_1} \ln(\hat{y_{i_1 1}}) + \sum_{i_2} \ln(\hat{y_{i_2 2}}) + \ldots + \sum_{i_K} \ln(\hat{y_{i_K K}})$$

To maximize l is to maximize $\ln(l)$, then the goal is to find β for which

$$\nabla_{\overrightarrow{\beta_k}} \ln(l) = 0 \quad \text{for } k = 1 \ldots K$$

Let $\mathit{1}_k$ be a n x 1 array where

$$\mathit{1}_k = \begin{cases} 1 & \text{Y=k} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \nabla_{\overrightarrow{\beta_k}} \ln(l) &= -\sum_{i_1} \hat{y_{i_1 k}}\vec{x}_{i_1} - \sum_{i_2} \hat{y_{i_2 k}}\vec{x}_{i_2} - \ldots + \sum_{i_k} (1 - \hat{y_{i_k k}})\vec{x}_{i_k} - \ldots - \sum_{i_K} \hat{y_{i_K k}}\vec{x}_{i_K} \\ &= \sum_{i_k} \vec{x}_{i_k} - \sum_{i_1} \hat{y_{i_1 k}}\vec{x}_{i_1} - \sum_{i_2} \hat{y_{i_2 k}}\vec{x}_{i_2} - \ldots - \sum_{i_K} \hat{y_{i_K k}}\vec{x}_{i_K} \\ &= \sum_{i_k} \vec{x}_{i_k} - \sum_{i}^{n} \hat{y}_{ik}\vec{x}_i \\ &= \sum_{i}^{n} (\mathit{1}_k - \hat{y}_{ik})\vec{x}_i \\ &= X^T \cdot (\mathit{1}_k - \hat{y}_k) \end{aligned}$$

So

$$\begin{cases} \nabla_{\vec{\beta}_1} \ln(l) &= \sum_i^n (\mathit{1}_1 - \hat{y_{i1}})\vec{x}_i = X^T \cdot (\mathit{1}_1 - \hat{y}_1) \\ \nabla_{\vec{\beta}_2} \ln(l) &= \sum_i^n (\mathit{1}_2 - \hat{y_{i2}})\vec{x}_i = X^T \cdot (\mathit{1}_2 - \hat{y}_2) \\ &\vdots \\ \nabla_{\vec{\beta}_K} \ln(l) &= \sum_i^n (\mathit{1}_K - \hat{y_{iK}})\vec{x}_i = X^T \cdot (\mathit{1}_K - \hat{y}_K) \end{cases}$$

$$\rightarrow \nabla_{\vec{\beta}} \ln(l) = X^T \cdot (\mathit{1} - \hat{y})$$

Now we know which direction to renew matrix β via $\nabla_{\vec{\beta}} \ln(l)$, and with the help of learning rate α, we then have

$$\color{red}{\beta_{new} = \beta_{old} + \alpha X^T \cdot (\mathit{1} - \hat{y})}$$

Below are matrices definition used in this document

$$X = \begin{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [x_1]\,[x_2] \dots [x_m] \end{bmatrix} \qquad \beta = \begin{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \vdots \\ \beta_{m1} \end{bmatrix} [\beta_2] \dots [\beta_K] \end{bmatrix}$$

$$\mathit{1} = \begin{bmatrix} [\mathit{1}_1]\,[\mathit{1}_2] \dots [\mathit{1}_K] \end{bmatrix} \qquad \hat{y} = \begin{bmatrix} [\hat{y}_1]\,[\hat{y}_2] \dots [\hat{y}_K] \end{bmatrix}$$