Two layer neural net

n = # training example

m = dimention of each training example

C = # class

$x^{(i)} \in R^{m \times 1}$ {ith training example}

$y^{(i)} \in \{1, 2, \dots C\}$ {class associated with ith training example}

$S^{(i)} = f(W_1, W_2, b_1, b_2, x^{(i)})$ {Score vector for ith training example}

$$= W_2 \max(0, W_1 x^{(i)} + b_1) + b_2$$

where, $W_1 \in R^{h \times m}$ $\quad b_1 \in R^{h}$

$W_2 \in R^{C \times h}$ $\quad b_2 \in R^{C}$

★ Softmax loss

$$L_i = -\log\left(P(Y = y^{(i)} \mid X = x^{(i)})\right)$$

$$P(Y = y^{(i)} \mid X = x^{(i)}) = \frac{e^{f_{y(i)}}}{\sum_j e^{f_j}}$$

$$\boxed{L = \frac{1}{n} \sum_{i=1}^{n} L_i}$$

★ Softmax loss gradient

$$\nabla L = \frac{1}{n}\sum_{i=1}^{n} \boxed{\nabla L_i}$$

$$\nabla L_i = -\nabla \log\left(\frac{e^{f_{y(i)}}}{\sum e^{f_j}}\right)$$

$$= \frac{-\sum_j e^{f_j}}{e^{f_{y(i)}}} \nabla\left(\frac{e^{f_{y(i)}}}{\sum_j e^{f_j}}\right)$$

$$= \frac{-\sum_j e^{f_j}}{e^{f_{y(i)}}} * \frac{\left(\sum_j e^{f_j}\right)\left(e^{f_{y(i)}}\nabla f_{y(i)}\right) - \left(e^{f_{y(i)}}\right)\left(\sum_j e^{f_j}\nabla f_j\right)}{\left(\sum_j e^{f_j}\right)^2}$$

$$\boxed{\nabla L_i = \frac{\sum_j e^{f_j}\nabla f_j}{\sum_j e^{f_j}} - \nabla f_{y(i)}}$$

Calculating gradient of f_j wrt $\omega_1, \omega_2, b_1, b_2$

ⓐ $\nabla_{\omega_1} f_j$

$$\left(\nabla_{\omega_1} f_j\right)^{a,b} = \frac{\delta f_j}{\delta \omega_1^{a,b}} = \begin{cases} 0 & \text{if } \omega_1^a x^{(i)} + b_1^a < 0 \\ \omega_2^{j,a} x_b^{(i)} & \text{else} \end{cases}$$

$$
\begin{array}{c c}
 & \begin{array}{cc} b=1 & b=2 \end{array} \\
\begin{array}{c} a=1 \\ a=2 \\ \\ \end{array} &
\begin{bmatrix}
I(\omega_1^1 x^{(i)} + b_1^1)\, \omega_2^{j,1} x_1^{(i)} & I(\omega_1^1 x^{(i)} + b_1^1)\, \omega_2^{j,1} x_2^{(i)} & \cdots \\
I(\omega_1^2 x^{(i)} + b_1^2)\, \omega_2^{j,2} x_1^{(i)} & I(\omega_1^2 x^{(i)} + b_1^2)\, \omega_2^{j,2} x_2^{(i)} & \cdots \\
\vdots & \vdots &
\end{bmatrix}
\end{array}
$$

$$
= \begin{bmatrix}
I(\omega_1^1 x^{(i)} + b_1^1)\, \omega_2^{j1} \\
I(\omega_1^2 x^{(i)} + b_1^2)\, \omega_2^{j2} \\
\vdots
\end{bmatrix}
\begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \cdots \end{bmatrix}
$$

$$
= \left(I(\omega_1 x^{(i)} + b_1) * \omega_2^{j^T} \right) x^{(i)^T}
$$

@ $\underline{\nabla_{\omega_2} f_j}$

$$
(\nabla_{\omega_2} f_j)^{a,b} = \frac{\delta f_j}{\delta \omega_2^{a,b}} = \begin{cases} 0 & \text{if } j \neq a \\ \max(0, \omega_1^b x^{(i)} + b_1^b) & \text{if } j = a \end{cases}
$$

$$
\nabla_{\omega_2} f_j = \begin{bmatrix}
\leftarrow 0 \rightarrow \\
\max(0, \omega_1 x^{(i)} + b_1)^T \\
\vdots \\
\leftarrow 0 \rightarrow
\end{bmatrix} \rightarrow j^{th} \text{ row}
$$

@ $\nabla_{b_1} f_j$

$$(\nabla_{b_1} f_j)^a = \frac{\delta f_j}{\delta b_1^a} = \begin{cases} 0 & \text{if } \omega_1^a x^{(i)} + b_1^a < 0 \\ \omega_2^{ja} & \text{else} \end{cases}$$

$$= \mathbb{I}(\omega_1 x^{(i)} + b_1) * \omega_2^{j^T}$$

@ $\nabla_{b_2} f_j$

$$(\nabla_{b_2} f_j)^a = \frac{\delta f_j}{\delta b_2^a} = \begin{cases} 0 & \text{if } j \neq a \\ 1 & \text{if } j = a \end{cases}$$

$$= \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow j^{th} \text{ element}$$

Calculating gradient of L_i wrt $\omega_1, \omega_2, b_1, b_2$

@ $\nabla_{\omega_1} L_i$

$$= \frac{\sum_j e^{f_j} \left(\mathbb{I}(\omega_1 x^{(i)} + b_1) * \omega_2^{j^T}\right) x^{(i)^T}}{\sum_j e^{f_j}} - \left(\mathbb{I}(\omega_1 x^{(i)} + b_1) * \omega_2^{y^{(i)T}}\right) x^{(i)T}$$

$$= \left(\mathbb{I}(\omega_1 x^{(i)} + b_1) * \frac{\sum_j e^{f_j} \omega_2^{j^T}}{\sum_j e^{f_j}}\right) x^{(i)^T} - \quad ''$$

$$= \left(I(\omega_1 x^{(i)} + b) * \frac{\omega_2^T e^f}{\sum_j e^{f_j}} \right) x^{(i)T} - \quad ''$$

$$= \boxed{\left(I(\omega_1 x^{(i)} + b) * \left(\frac{\omega_2^T e^f}{\sum_j e^{f_j}} - \omega_2^{y^{(i)}T} \right) \right) x^{(i)T}}$$

@ $\underline{\nabla_{\omega_2} L_i}$

$$\sum_j e^{f_j} \nabla f_j = \begin{bmatrix} e^{f_1} \max(0, \omega_1 x^{(i)} + b_1)^T \\ \vdots \\ e^{f_{y^{(i)}}} \max(0, \omega_1 x^{(i)} + b_1)^T \\ \vdots \\ e^{f_c} \max(0, \omega_1 x^{(i)} + b_1)^T \end{bmatrix}$$

$$\nabla_{\omega_2} L_i = \begin{bmatrix} \frac{e^{f_1}}{\sum_j e^{f_j}} \max(0, \omega_1 x^{(i)} + b_1)^T \\ \vdots \\ \left(\frac{e^{f_{y^{(i)}}}}{\sum_j e^{f_j}} - 1 \right) \max(0, \omega_1 x^{(i)} + b_1)^T \\ \vdots \\ \frac{e^{f_c}}{\sum_j e^{f_j}} \max(0, \omega_1 x^{(i)} + b_1)^T \\ \vdots \end{bmatrix} \rightarrow j^{th} \text{ row}$$

$$= \begin{bmatrix} \frac{e^{f_1}}{\sum_j e^{f_j}} \\ \vdots \\ \left(\frac{e^{f_{y^{(i)}}}}{\sum_j e^{f_j}} - 1\right) \\ \vdots \\ \frac{e^{f_c}}{\sum_j e^{f_j}} \end{bmatrix} \max(0, \omega_1 x^{(i)} + b_1)^T$$

@ $\nabla_{b_1} L_i$

$$= I(\omega_1 x^{(i)} + b) * \left(\frac{\omega_2^T e^f}{\sum_j e^{f_j}} - \omega_2^{y^{(i)}T}\right)$$

@ $\nabla_{b_2} L_i$

$$\sum_j e^{f_j} \nabla_{b_2} f_j = e^f$$

$$\nabla_{b_2} L_i = \begin{bmatrix} \frac{e^{f_1}}{\sum_j e^{f_j}} \\ \frac{e^{f_2}}{\sum_j e^{f_j}} \\ \vdots \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow y^{(i)\text{th}} \text{ element}$$

$$\nabla_{b_2} L_i = \begin{bmatrix} \dfrac{e^{f_1}}{\sum_j e^{f_j}} \\ \vdots \\ \left(\dfrac{e^{f_{y(i)}}}{\sum_j e^{f_j}} - 1 \right) \\ \vdots \\ \dfrac{e^{f_c}}{\sum_j e^{f_j}} \end{bmatrix}$$