Softmax loss

M = # training example

m = dimention of each training example

C = # class

$$X^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ x_2^{(i)} \\ \vdots \end{bmatrix} \in R^{(m+1)\times 1}$$ { i^{th} training example with bias dimension }

$y^{(i)} \in \{1, 2, \cdots, C\}$ { class lable associated with i^{th} training example }

$X \in R^{M\times(m+1)}$ { Design matrix }

$$X = \begin{bmatrix} X^{(1)T} \\ X^{(2)T} \\ \vdots \\ X^{(M)T} \end{bmatrix}$$

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{bmatrix}$$ { Label vector }

$S^{(i)} = WX^{(i)} \in R^{C\times 1}$ { Score vector for i^{th} train example }

{ $S_j^{(i)}$ = { Score given to j^{th} class lable }

where,

$W \in R^{C\times(m+1)}$ { Weight matrix } W_j = { j^{th} row of W }

$S = XW^T \in R^{M\times C}$ { Score matrix }

⇒ In Softmax we assume score to be **unnormalized log probability.**

→ Let Y be the random variable of label associated with input example random variable X.

$$\boxed{P(Y=k \mid X=x) = \frac{e^{S_k}}{\sum_j e^{S_j}}} \quad \text{where, } \boxed{S = Wx}$$

⇒ For every training example, we want to maximise the assigned probability to the correct class.

i.e. → Maximize log likelihood

i.e. → Minimize negative log likelihood

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

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

★ Calculating Gradient

$$\nabla_w L(w) = \frac{1}{n} \sum_{i=1}^{n} \boxed{\nabla_w L_i(w)}$$

$$\nabla_w L_i(w) = -\nabla_w \log \left(\frac{e^{s_{y^{(i)}}^{(i)}}}{\sum_j e^{s_j^{(i)}}} \right)$$

$$= -\nabla_w \log \left(\frac{e^{w_{y^{(i)}} x^{(i)}}}{\sum_j e^{w_j x^{(i)}}} \right)$$

$$= -\frac{\sum_j e^{w_j x^{(i)}}}{e^{w_{y^{(i)}} x^{(i)}}} \nabla_w \left(\frac{e^{w_{y^{(i)}} x^{(i)}}}{\sum_j e^{w_j x^{(i)}}} \right)$$

$$\underbrace{\frac{\left(\sum_j e^{w_j x^{(i)}}\right) \underbrace{\left(\nabla_w e^{w_{y^{(i)}} x^{(i)}}\right)}_{①} - \left(e^{w_{y^{(i)}} x^{(i)}}\right) \underbrace{\left(\sum_j \nabla_w e^{w_j x^{(i)}}\right)}_{②}}{\left(\sum_j e^{w_j x^{(i)}}\right)^2}}_{③}$$

2

$$\nabla_W e^{W_j x^{(i)}}$$

$$= e^{W_j x^{(i)}} \cdot \boxed{\nabla_W \left(W_j x^{(i)}\right)}$$

$$\left(\nabla_W \left(W_j x^{(i)}\right)\right)_{a,b} = \frac{\delta}{\delta W_{ab}} \left(W_j x^{(i)}\right)$$

$$= \begin{cases} 0 & \text{if } a \neq j \\ x_b^{(i)} & \text{if } a = j \end{cases}$$

$$\left(\nabla_W \left(W_j x^{(i)}\right)\right)_a = \begin{cases} x^{(i)} & \text{if } a = j \\ 0 & \text{if } a \neq j \end{cases}$$

$$\begin{array}{cl} \begin{bmatrix} \leftarrow \ 0 \ \rightarrow \\ \leftarrow \ x^{(i)T} \ \rightarrow \\ \leftarrow \ 0 \ \rightarrow \end{bmatrix} & \begin{array}{l} \\ \leftarrow j^{th} \text{ row} \\ \\ \end{array} \end{array}$$

$$\sum_j \nabla_W e^{W_j x^{(i)}} = \begin{bmatrix} \leftarrow e^{W_1 x^{(i)}} x^{(i)T} \rightarrow \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \leftarrow e^{W_2 x^{(i)}} x^{(i)T} \rightarrow \\ 0 \\ 0 \end{bmatrix}$$

$$+ \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \leftarrow e^{W_c x^{(i)}} x^{(i)T} \rightarrow \end{bmatrix}$$

$$\sum_j \nabla_w e^{w_j x^{(i)}} = \begin{bmatrix} \leftarrow e^{w_1 x^{(i)}} x^{(i)T} \rightarrow \\ \leftarrow e^{w_2 x^{(i)}} x^{(i)T} \rightarrow \\ \vdots \\ \leftarrow e^{w_c x^{(i)}} x^{(i)T} \rightarrow \end{bmatrix}$$

(1) $\nabla_w e^{w_{y^{(i)}} x^{(i)}}$

$$= e^{w_{y^{(i)}} x^{(i)}} \cdot \nabla w_{y^{(i)}} x^{(i)}$$

$$= \begin{array}{cl} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \leftarrow e^{w_{y^{(i)}} x^{(i)}} x^{(i)T} \rightarrow \\ 0 \\ \vdots \\ 0 \end{bmatrix} & \\ & \end{array} \longrightarrow y^{(i)th} \text{ row}$$

(3rd)

$$\begin{array}{rl} & \\ y^{(i)th} \text{ row} \leftarrow & \begin{bmatrix} 0 \\ 0 \\ \left(e^{w_{y^{(i)}} x^{(i)}} \sum_j e^{w_j x^{(i)}}\right) x^{(i)T} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{array} - \begin{bmatrix} e^{(w_1 + w_{y^{(i)}}) x^{(i)}} x^{(i)T} \\ \vdots \\ e^{2 w_{y^{(i)}} x^{(i)}} x^{(i)T} \\ \vdots \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} -e^{(w_1 + w_{y^{(i)}}) x^{(i)}} x^{(i)T} \\ \vdots \\ e^{w_{y^{(i)}} x^{(i)}} \left(\sum_{j \neq y^{(i)}} e^{w_j x^{(i)}} \right) x^{(i)T} \\ \vdots \\ \vdots \end{bmatrix}$$

$$\nabla_\omega L_i(\omega) = \frac{1}{e^{\omega_{y^{(i)}} x^{(i)}} \left(\sum_j e^{\omega_j x^{(i)}}\right)} \begin{bmatrix} e^{(\omega_1 + \omega_{y^{(i)}}) x^{(i)}} \; x^{(i)T} \\ \vdots \\ -e^{\omega_{y^{(i)}} x^{(i)}} \left(\sum_{j \neq y^{(i)}} e^{\omega_j x^{(i)}}\right) x^{(i)T} \\ \vdots \\ e^{(\omega_c + \omega_{y^{(i)}}) x^{(i)}} \; x^{(i)T} \end{bmatrix}$$

$$\boxed{\nabla_\omega L_i(\omega) = \frac{1}{\sum_j e^{\omega_j x^{(i)}}} \begin{bmatrix} e^{\omega_1 x^{(i)}} \; x^{(i)T} \\ e^{\omega_2 x^{(i)}} \; x^{(i)T} \\ \vdots \\ -\left(\sum_{j \neq y^{(i)}} e^{\omega_j x^{(i)}}\right) x^{(i)T} \\ \vdots \\ e^{\omega_c x^{(i)}} \; x^{(i)T} \end{bmatrix}}$$

$$\boxed{\nabla_\omega L_i(\omega) = \begin{bmatrix} P(Y=1 \mid X = x^{(i)}) \; x^{(i)T} \\ P(Y=2 \mid X = x^{(i)}) \; x^{(i)T} \\ \vdots \\ -1 + P(Y = y^{(i)} \mid X = x^{(i)}) \; x^{(i)T} \\ \vdots \\ P(Y = c \mid X = x^{(i)}) \; x^{(i)T} \end{bmatrix}}$$