1) Forward.

$$a_1(x) = h(\alpha^{(1)}(x; W^{(1)}, b^{(1)})), \quad \text{where} \quad \alpha^{(1)}(x; W^{(1)}, b^{(1)}) = (W^{(1)})^T x + b^{(1)}$$

↳ (hidden) second layer

$$a_2(x) = o(W^{(2)T} a_1(x) + b^{(2)})$$

↳ output softmax activation function.

Like this, 일반화 한다면 $L+1$개 층을 가진. {Input = 1, hidden = $L-1$, output = 1}

(from $k = 1$ to $L-1$)

① $a_k(x) = h_k(x) = h^{(k)}(\alpha^{(k)}(x; W^{(k)}, b^{(k)}))$

where, $\alpha^{(k)}(x; W^{(k)}, b^{(k)}) = (W^{(k)})^T h^{(k-1)}(x) + b^{(k)}$

② $h_L(x) = o(\alpha^{(L)}(x)) = a_L(x)$

Loss function.

$$l(a_L(x), y) = -\sum_c I_{(y=c)} \ln a_L(x)_c = -\ln a_L(x)_y$$

↑ predicted ↑ real

∴ with m sample ⇒ negative log-likelihood loss

$$L = -\frac{1}{m}\sum_{i=1}^{m} -\ln a_L(x_i)_{y_i} = -\frac{1}{m}\sum_{i=1}^{m}\sum_c I_{(y_i=c)} \ln a_L(x_i)_c$$

partial derivative)

$$\frac{\partial L}{\partial a_L(x)_c} = \frac{\partial(-\ln a_L(x)_y)}{\partial a_L(x)_c} = \frac{-1_{(y=c)}}{a_L(x)_c}$$

$$\therefore \frac{\partial L}{\partial a_L} = -\begin{bmatrix} \frac{1_{(y=0)}}{a_L(x)_0} \\ \vdots \\ \frac{1_{(y=C-1)}}{a_L(x)_{C-1}} \end{bmatrix} = -e(y) \oslash a_L(x)$$

$$e(y) = \begin{bmatrix} 1_{(y=0)} \\ \vdots \\ 1_{(y=C-1)} \end{bmatrix}$$

$\oslash$: element wise division.

$\therefore$ Loss: $\ell = -\ln a_L(x)_y$

$a_L(x) = \text{Softmax}(\alpha^{(L)}(x))$

$$\therefore \frac{\partial \mathcal{L}}{\partial \alpha^{(L)}(x)_c} = \frac{\partial \mathcal{L}}{\partial a_L(x)_y} \cdot \frac{\partial a_L(x)_y}{\partial a_L(x)_c} = \frac{-1}{a_L(x)_y} \cdot \frac{\partial a_L(x)_y}{\partial a_L(x)_c}$$

$\sigma(x)(1-\sigma(x))$ for sigmoid

$$= \frac{-1}{a_L(x)_y} a_L(x)_y \left(\mathbb{1}_{(y=c)} - a_L(x)_c \right) = a_L(x)_c - \mathbb{1}_{(y=c)}$$

$\therefore$ to be general

$$\frac{\partial \mathcal{L}}{\partial \alpha^{(L)}} = a_L(x) - \begin{bmatrix} \mathbb{1}_{(y=0)} \\ \vdots \\ \mathbb{1}_{(y=c)} \end{bmatrix} = a_L(x) - e(y)$$

---

at hidden layer (need W, b)

from $\alpha^{(k)}(x) = (W^{(k)})^T a_{k-1}(x) + b^{(k)}$

i) $$\frac{\partial \mathcal{L}}{\partial W_{ij}^{(k)}} = \frac{\partial \mathcal{L}}{\partial \alpha^{(k)}(x)_j} \frac{\partial \alpha^{(k)}(x)_j}{\partial W_{ij}^{(k)}} = \frac{\partial \mathcal{L}}{\partial \alpha^{(k)}(x)_j} \cdot a_{k-1}(x)_i$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial W^{(k)}} = \boxed{\frac{\partial \alpha^{(k)}}{\partial W^{(k)}} \left(\frac{\partial \mathcal{L}}{\partial \alpha^{(k)}} \right)^T} = a_{k-1}(x) \left(\frac{\partial \mathcal{L}}{\partial \alpha^{(k)}} \right)^T$$

dimension matching.

ii) $$\frac{\partial \mathcal{L}}{\partial b_i^{(k)}} = \frac{\partial \ell}{\partial \alpha^{(k)}(x)_i} \cdot \frac{\partial \alpha^{(k)}(x)_i}{\partial b_i^{(k)}} = \frac{\partial \ell}{\partial \alpha^{(k)}(x)_i} \cdot 1$$

$\Rightarrow$ to general

$$\frac{\partial \mathcal{L}}{\partial b^{(k)}} = \boxed{\left(\frac{\partial \alpha^{(k)}}{\partial b^{(k)}} \right)^T \frac{\partial \ell}{\partial \alpha^{(k)}}} = 1 \cdot \frac{\partial \ell}{\partial \alpha^{(k)}} = \frac{\partial \mathcal{L}}{\partial \alpha^{(k)}}$$

Then, what is $\frac{d\mathcal{L}}{d\alpha_k}$?

From $\alpha^{(k+1)}(x) = (w^{(k+1)})^T \cdot a_k(x) + b^{(k+1)}$

i) $$\frac{d\mathcal{L}}{d a_k(x)_j} = \sum_i \frac{d\mathcal{L}}{d\alpha^{(k+1)}(x)_i} \cdot \boxed{\frac{d\alpha^{(k+1)}(x)}{d a_k(x_j)}}$$

$$= \sum_i \frac{d\mathcal{L}}{d\alpha^{(k+1)}(x)_i} w_{ij}^{(k+1)}$$

$\therefore$ to general

$$\frac{d\mathcal{L}}{d a_k} = \boxed{\left(\frac{d\alpha^{(k+1)}}{d a_k}\right)^T} \cdot \frac{d\mathcal{L}}{d\alpha^{(k+1)}} = w^{(k+1)} \cdot \frac{d\mathcal{L}}{d\alpha^{(k+1)}}$$

Dimension matching.

$$\frac{d\mathcal{L}}{d\alpha^{(k)}(x)_j} = \frac{d\mathcal{L}}{d a_k(x)_j} \cdot \frac{d a_k(x)_j}{d\alpha^{(k)}(x)_j} = \frac{d\mathcal{L}}{d a_k(x)_j} \cdot h'(\alpha^{(k)}(x)_j)$$

to general

$$\frac{d\mathcal{L}}{d\alpha^{(k)}} = \left(\frac{d a_k}{d\alpha^{(k)}}\right)^T \left(\frac{d\mathcal{L}}{d a_k}\right) = \begin{bmatrix} h'(\alpha^{(k)}(x)_1 & & \\ & \ddots & \\ & & h'(\alpha^{(k)}(x)_j) \end{bmatrix} \left(\frac{d\mathcal{L}}{d a_k}\right)$$

$$= h'(\alpha^{(k)}(x)) \odot \left(\frac{d\mathcal{L}}{d a_k}\right)$$

$$= h'(\alpha^{(k)}(x)) \odot w^{(k)} d\alpha^{(k)}$$

Summarise:
$$\begin{cases} \text{Loss} = \mathcal{L} = -\ln a_L(x)_y \\ a_L = \text{Softmax}(\alpha^{(L)}(x)) \\ a_{k-1}(x) = h(\alpha^{(k-1)}(x)) \\ \alpha^{(k)}(x) = b^{(k)} + (w^{(k)})^T a_{k-1}(x) \end{cases}$$

$$\frac{\partial L}{\partial a_L} = - \begin{bmatrix} 1_{(y=0)} / a_L(x)_0 \\ \vdots \\ 1_{(y=c-1)} / a_L(x)_{c-1} \end{bmatrix}$$

$\therefore$ at output level $\Rightarrow$ $d\alpha^{(L)} = \frac{\partial L}{\partial \alpha^{(L)}} = a_L(x) - e(y)$

From layer L to 1

$$\Rightarrow \quad dw^{(k)} = \frac{\partial L}{\partial w^{(k)}} = a_{k-1}(x) \left(\frac{\partial L}{\partial \alpha^{(k)}} \right)^T = a_{k-1}(x) \left(d\alpha^{(k)} \right)^T$$

$$db^{(k)} = \frac{\partial L}{\partial b^{(k)}} = \frac{\partial L}{\partial \alpha^{(k)}} = d\alpha^{(k)}$$

where $d\alpha^{(k-1)} = \frac{\partial L}{\partial \alpha^{(k-1)}} = h'\left(\alpha^{(k-1)}(x)\right) \odot \left(\frac{\partial L}{\partial a_{k-1}} \right)$

$$= h'\left(\alpha^{(k-1)}(x)\right) \odot W^{(k)} d\alpha^{(k)}$$