Bài 1:

* Yêu cầu 1: Viết công thức tính đạo hàm của hàm mất mát theo đầu ra O_a, tức là $\frac{\partial L}{\partial O_a}$. Sau đó, viết ở dạng vectơ $\delta_L = \Delta L$

Bước 1: Viết lại ~~log~~ loss theo softmax.

Ta có:

$$L(o,y) = -\sum_{c=1}^{C} I(y=c)\log P(c|x)$$

$$= -\sum_{c=1}^{C} I(y=c)\log \frac{e^{o_c}}{\sum_{c'} e^{o_{c'}}}$$

$$= -\sum_{c=1}^{C} I(y=c)\left(o_c - \log\sum_{c'} e^{o_{c'}}\right)$$

$$= -\sum_{c=1}^{C} I(y=c)o_c + \sum_{c=1}^{C} I(y=c)\log\sum_{c'} e^{o_{c'}}$$

mà $\sum_c I(y=c)o_c = o_y$; $\sum_c I(y=c)\log\sum_{c'} e^{o_{c'}} = \log\sum_{c'} e^{o_{c'}}$

$$\Rightarrow L(o,y) = -o_y + \log\sum_{c'=1}^{C} e^{o_{c'}}$$

Bước 2: Đạo hàm theo o_a.

+) TH1: $a = y$ (lớp đúng).

$$\frac{\partial L}{\partial o_a} = \frac{\partial}{\partial o_y}\left(-o_y + \log\sum_{c'} e^{o_{c'}}\right)$$

$$= -1 + \frac{1}{\sum_{c'} e^{o_{c'}}}\cdot\frac{\partial}{\partial o_y}\sum_{c'} e^{o_{c'}}$$

chỉ có $c' = y$ phụ thuộc o_y nên:

$$\frac{\partial}{\partial o_y} \sum_{c'} e^{o_{c'}} = e^{o_y}$$

Vậy:

$$\frac{\partial L}{\partial o_y} = -1 + \frac{e^{o_y}}{\sum_{c'} e^{o_{c'}}} = p(y|x) - 1$$

+) TH2 : $a \neq y$ (lớp sai)

$$\frac{\partial L}{\partial o_a} = \frac{\partial}{\partial o_a}\left(-o_y + \log \sum_{c'} e^{o_{c'}}\right)$$

$$= 0 + \frac{1}{\sum_{c'} e^{o_{c'}}} \cdot e^{o_a} = p(a|x)$$

Bước 3: Gộp:

$$\frac{\partial L}{\partial o_a} = p(a|x) - I(y = a), a = 1, 2, \dots C$$

Bước 4: Viết dạng vector.

$$\delta^L = \nabla_{o^L} L = p - y_{vec}$$

* Yêu cầu 2: Gradient theo đầu ra tầng l:

Ta biết:

$$\delta^{l+1} = \frac{\partial L}{\partial o^{l+1}}$$

$z^{l+1} = W^{l+1} o^{l} + b^{l+1}$ và $o^{l+1} = \sigma^{l+1}(z^{l+1})$, ~~đặt~~

áp dụng chuỗi đạo hàm:

$$\frac{\partial L}{\partial o^{l}} = \frac{\partial L}{\partial z^{l+1}} \frac{\partial z^{l+1}}{\partial o^{l}} = \cancel{(\delta^{l+1} \odot \sigma'^{l+1}}$$

$$= (\delta^{l+1} \odot \sigma'^{l+1}(z^{l+1}))(W^{l+1})^{T}$$

vì $(AB)^T = B^T A^T$, ~~viết gọn~~:

$$\delta^{l} = \nabla_{o^l} L = (W^{l+1})^T (\delta^{l+1} \odot \sigma'^{l+1}(z^{l+1}))$$

⊛ Yêu cầu 3: Gradient theo trọng số W^l và bias b^l

Bước 1: Gradient theo phần tử:

với tầng l, đầu ra $o^l = o^l(z^l)$, đầu vào tầng l là o^{l-1}, ta có:

$$z^{l} = W^{l} o^{l-1} + b^{l}, \quad o^{l} = o^{l}(z^{l}).$$

· Gradient theo trọng số phần tử $w^{l}_{a,b}$:

$$\frac{\partial L}{\partial w^{l}_{a,b}} = \frac{\partial L}{\partial z^{l}_{a}} \frac{\partial z^{l}_{a}}{\partial w^{l}_{a,b}}$$

~~Nhưng $z^{l}_{a} = \Sigma_{k} w^{l}_{a,k} o^{l-1}_{k}$~~

Nhưng $z^{l}_{a} = \sum_{k} w^{l}_{a,k} o^{l-1}_{k} + b^{l}_{a}$

$$\Rightarrow \frac{\partial z^{l}_{a}}{\partial w^{l}_{a,b}} = o^{l-1}_{b}$$

$$\Rightarrow \frac{\partial L}{\partial w^{l}_{a,b}} = \delta^{l}_{a} \cdot o^{l-1}_{b}$$

Gradient theo bias b^{l}_{a} :

$$\frac{\partial L}{\partial b^{l}_{a}} = \frac{\partial L}{\partial z^{l}_{a}} \frac{\partial z^{l}_{a}}{\partial b^{l}_{a}} = \delta^{l}_{a} \cdot 1 = \delta^{l}_{a}$$

Bước 2: Gradient dạng ma trận và vector

Đặt: $\delta^{l} = [\delta^{l}_{1}, \dots, \delta^{l}_{d_l}]^{T} \in \mathbb{R}^{d_l}$.

Đầu ra tầng trước $o^{l-1} \in \mathbb{R}^{d_{l-1}}$

Gradient theo ma trận W^{l} :

$$\nabla_{W^{l}} L = \delta^{l} (o^{l-1})^{T} \in \mathbb{R}^{d_l \times d_{l-1}}$$

. Mỗi phần tử: $(\nabla_{W^{l}} L)_{a,b} = \delta^{l}_{a} \cdot o^{l-1}_{b}$

Gradient theo bias b^{l}:

$$\nabla_{b^{l}} L = \delta^{l} \in \mathbb{R}^{d_l}$$

⊛ Yêu cầu 4: Cập nhật tham số bằng SGD.

Với kích thước lô bằng 1 thì là cập nhật sau mỗi mẫu

tốc độ học η, công thức SGD là:

$$w^l \leftarrow w^l - \eta \Delta_{w^l} L$$
$$b^l \leftarrow b^l - \eta \Delta_{b^l} L$$

Thay $\Delta_{w^l} L = \delta^l (o^{l-1})^T$ và $\Delta_{b^l} L = \delta^l$ vào, ta được

$$\boxed{\begin{aligned} w^l &\leftarrow w^l - \eta \delta^l (o^{l-1})^T \\ b^l &\leftarrow b^l - \eta \delta^l \end{aligned}}$$