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 $\delta_L = \nabla L$

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:

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

+) TH 2: $a \neq y$ (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:

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

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

áp dụng chuỗi đạo hàm:

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

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

vì $(AB)^T = B^T A^T$, ~~rút gọn~~:

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

* Yêu cầu 3: Gradient theo trọng số W^{ℓ} và bias b^{ℓ}

Bước 1: Gradient theo phần tử.

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

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

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

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

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

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

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

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

Gradient theo bias b^{ℓ}_{a} :

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

Bước 2: Gradient dạng ma trận và vector.

Đặt: $\delta^{\ell} = [\delta^{\ell}_{1}, \ldots, \delta^{\ell}_{d_\ell}]^{T} \in \mathbb{R}^{d_\ell}$.

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

Gradient theo ma trận W^{ℓ} :

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

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

Gradient theo bias b^{ℓ}:

$$\nabla_{b^{\ell}} L = \delta^{\ell} \in \mathbb{R}^{d_\ell}$$

⊛ 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

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