一. 对 softmax 求导

设 z_j 是输出

$$y_j = \text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad \text{Loss}(\hat{y}, y) = -\sum y_j \log(\hat{y}_j)$$

设正确的类为 t，$y_t = 1$，$y_{j \neq t} = 0$ $\quad \text{Loss}(\hat{y}, y) = -\log(\hat{y}_t)$ $\quad \frac{\partial L}{\partial \hat{y}_t} = -\frac{1}{\hat{y}_t}$ $\quad$ **$j = t$ 时**

$$\frac{\partial \hat{y}_t}{\partial z_t} = \frac{\partial \left(\frac{e^{z_t}}{\sum_{k} e^{z_k}}\right)}{\partial z_t} = \frac{e^{z_t} \cdot \sum - e^{z_t} \cdot e^{z_t}}{\sum^2} = \frac{e^{z_t}}{\sum} - \left(\frac{e^{z_t}}{\sum}\right)^2 = \hat{y}_t (1 - \hat{y}_t)$$

$j \neq t$ 时

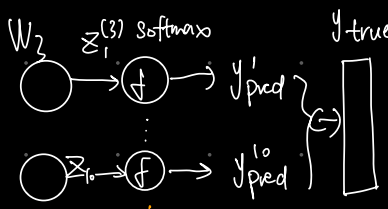
$$\frac{\partial \hat{y}_t}{\partial z_j} = -\hat{y}_t \hat{y}_j$$

则链式：$\frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial z_t} = \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t} = -\frac{1}{\hat{y}_t} \cdot \hat{y}_t (1 - \hat{y}_t) = \hat{y}_t - 1 = \hat{y}_t - y_t$

$j = t$

$j \neq t$ $\quad \frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_j} = -\frac{1}{\hat{y}_t} (-\hat{y}_t \cdot \hat{y}_j) = \hat{y}_j = \hat{y}_j - y_j = \hat{y}_j - 0$

$\therefore \frac{\partial J}{\partial z_j} = \hat{y}_j - y_j$ ← $\frac{\partial L}{\partial z^{(3)}}$ $\quad$ grad_y_pred = y_pred − y_true $\in R^{m \times 10}$



$$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_3}$$

$z^{(3)} = W_3 a^{(2)} + b_3$ $\quad$ 64×10 $\quad$ $m \times 64$

$\therefore \frac{\partial z^{(3)}}{\partial W_3} = a^{(2)} \in R^{m \times 64}$

$\therefore \frac{\partial L}{\partial W_3} \in (64 \times m) \cdot (m \times 10) = R^{64 \times 10}$

$\therefore \frac{\partial L}{\partial W_3} = \frac{1}{m} \cdot np.dot(a_2^T \cdot grad\_y\_true)$

$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial b_3} = \frac{1}{m} \cdot np.sum(grad\_y\_pred)$

$R^{m \times 10}$, axis=0 $\Rightarrow R^{10}$

$R^{m \times 64} \in \frac{\partial L}{\partial a^{(2)}} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} = \frac{1}{m} \cdot np.dot(grad\_y\_true, W_3^T) \in R^{m \times 64}$

$grad\_y\_true, \in R^{m \times 10}$ $\quad$ $W_3 \in R^{64 \times 10}$

矩阵乘法本来就是 sum 的过程

$$\frac{\partial L}{\partial z^{(2)}} = \frac{\partial L}{\partial a^{(2)}} \cdot \underbrace{\frac{\partial a^{(2)}}{\partial z^{(2)}}}_{\text{Grelu}}$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W_2} \rightarrow a^{(1)} \in R^{m\times 128},\ W_2 \in R^{128\times 64}$$

$\downarrow$ grad_z2 $\in R^{m\times 64}$ — $\frac{1}{m}$ np.dot(a_1^T, grad_z2)

总结: $\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_3} =$ grad_y_pred $\cdot\ a^{(2)} = \frac{1}{m}$ np.dot(a_2^T, grad_y_pred)

$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W_2} =$ grad_z2 $\cdot\ a^{(1)} = \frac{1}{m}(a_1^T$, grad_z2)

算 $\frac{\partial L}{\partial W_l} = \frac{\partial L}{\partial z^{(l)}} \cdot \frac{\partial z^{(l)}}{\partial W_l} =$ grad_zl $\cdot\ a^{(l-1)}$

算 grad_zl 用 $\frac{\partial L}{\partial a^{(l)}}$, 而 $\frac{\partial L}{\partial a^{(l)}} = \frac{\partial L}{\partial z^{(l+1)}} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(l)}}$ ($\rightarrow R^{m,d_l}$)

$=$ grad_z(l+1) $\cdot$ $W^{(l+1)}$ $=$ np.dot(grad_z(l+1), $W_{(l+1)}^T$)

($\downarrow$ $R^{m,d_{l+1}}$, $\downarrow$ $R^{d_l,d_{l+1}}$)

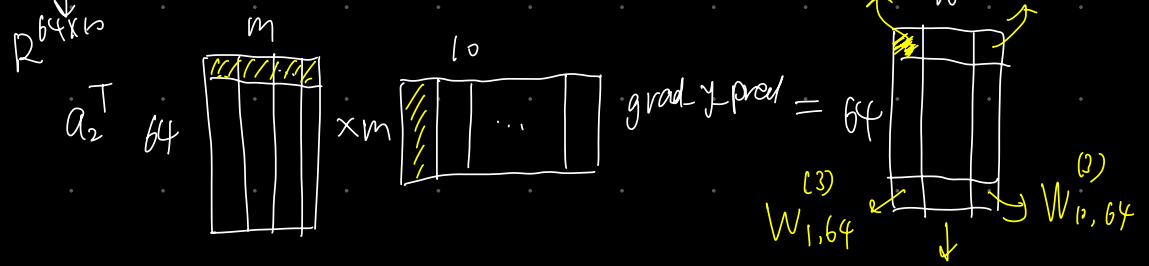
二. 为什么矩阵版 $\frac{\partial L}{\partial W^{(l)}}$ 要乘 $1/m$?

在SGD中, 一个batch共 m 个样本, 乘 $1/m$ 是这一批样本的梯度均值.

以 $W_3 \in R^{64,10}$ 为例.

$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_3} =$ grad_y_pred $\cdot\ a^{(2)}$ ($\downarrow R^{m\times 10}$, $\downarrow R^{m\times 64}$)

($\downarrow R^{64\times 10}$)



为什么 $\frac{\partial L}{\partial a^{(1)}}$ 不用 $\times 1/m$？

这个矩阵中每个元素是 m 个样本梯度的累加，∴ $\times 1/m$

$$\frac{\partial L}{\partial a^{(1)}} = \frac{\partial L}{\partial z^{(l+1)}} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(1)}} = grad\_z^{(l+1)} \cdot W^{(l+1)}$$

以 $a^{(1)}$ 为例，$\in R^{m \times 64}$

$z^{(3)} \in R^{m \times 10}$　$W^3 \in R^{64 \times 10}$

$z^{(1)} \cdot W^{(1)T} \Rightarrow np.dot(grad\_z_3, W_3^T)$

$m \times 10,\ 10 \times 64 \Rightarrow m \times 64$

m的维度抹掉了要 不用 $\times 1/m$