Assignment 2 (Machine learning) [313652008 黃睿帆]

1. ($n_L = 1$) Calculate $\nabla a^{[L]}(x) = \begin{bmatrix} \frac{\partial a^{[L]}}{\partial x_1} \\ \vdots \\ \frac{\partial a^{[L]}}{\partial x_{n_1}} \end{bmatrix}_{n_1 \times 1}$ $\left(\text{Notation: } \begin{cases} a^{[1]} = x \in \mathbb{R}^{n_1} \\ a^{[\ell]} = \sigma(\underbrace{W^{[\ell]} a^{[\ell-1]} + b^{[\ell]}}_{= \boxed{z^{[\ell]}}}) \in \mathbb{R}^{n_\ell}, \ \ell = 2, \dots, L \end{cases} \right)$

$\boxed{\text{sol:}}$ Define $\delta_j^{[\ell]} = \frac{\partial a^{[L]}}{\partial z_j^{[\ell]}}$, $\ell = 2, \dots L$. $(n_L = 1 \Rightarrow z^{[L]} \in \mathbb{R})$

$\because a^{[L]} = \sigma(z^{[L]}) \Rightarrow \delta_j^{[L]} = \frac{\partial a^{[L]}}{\partial z_j^{[L]}} = \sigma'(z_j^{[L]})$, define $D^{[\ell]} = \text{diag}(\sigma'(z^{[\ell]})) \in \mathbb{R}^{n_\ell \times n_\ell}$

$$= \begin{bmatrix} \sigma'(z_1^{[\ell]}) & & 0 \\ & \sigma'(z_2^{[\ell]}) & \\ 0 & & \ddots \end{bmatrix} \quad \left(\frac{\partial a^{[\ell]}}{\partial z^{[\ell]}} = D^{[\ell]} \right)$$

Then, for other ℓ,

$$\delta_j^{[\ell]} = \frac{\partial a^{[L]}}{\partial z_j^{[\ell]}} = \sum_{k=1}^{n_{\ell+1}} \frac{\partial a^{[L]}}{\partial z_k^{[\ell+1]}} \frac{\partial z_k^{[\ell+1]}}{\partial z_j^{[\ell]}}, \text{ where } z_k^{[\ell+1]} = \sum_m W_{km}^{[\ell+1]} a_m^{[\ell]} + b_k^{[\ell+1]}$$

so $\frac{\partial z_k^{[\ell+1]}}{\partial a_m^{[\ell]}} = W_{km}^{[\ell+1]}$; $\frac{\partial a_m^{[\ell]}}{\partial z_j^{[\ell]}} = \begin{cases} \sigma'(z_j^{[\ell]}), & m = j \\ 0, & m \neq j \end{cases}$ $\Rightarrow \frac{\partial z_k^{[\ell+1]}}{\partial z_j^{[\ell]}} = \sum_m \frac{\partial z_k^{[\ell+1]}}{\partial a_m^{[\ell]}} \frac{\partial a_m^{[\ell]}}{\partial z_j^{[\ell]}} = \underline{W_{kj}^{[\ell+1]} \cdot \sigma'(z_j^{[\ell]})}$

$$\therefore \delta_j^{[\ell]} = \sum_{k=1}^{n_{\ell+1}} \delta_k^{[\ell+1]} W_{kj}^{[\ell+1]} \cdot \sigma'(z_j^{[\ell]}) = \sigma'(z_j^{[\ell]}) \sum_k W_{kj}^{[\ell+1]} \delta_k^{[\ell+1]} \Leftrightarrow \underline{\delta^{[\ell]} = D^{[\ell]} (W^{[\ell+1]})^T \delta^{[\ell+1]}, \ \ell = L-1, L-2, \dots, 2}$$

(計算順序: $\delta^{[L]} \to \delta^{[L-1]} \to \dots \to \delta^{[2]}$) Recall: $z^{[2]} = W^{[2]} a^{[1]} + b^{[2]} = W^{[2]} x + b^{[2]}$

$$\Rightarrow \frac{\partial a^{[L]}}{\partial x} = \frac{\partial a^{[L]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial x} = \underbrace{\delta^{[2]}}_{n_2 \times 1} \cdot \underbrace{W^{[2]}}_{n_2 \times n_1} \Rightarrow \therefore \underline{\underbrace{\nabla a^{[L]}(x)}_{n_1 \times 1} = (W^{[2]})^T \cdot \delta^{[2]}} \ \#$$

2. Problems:

- In supervised learning, we often using classification when target has discrete value, what is the common loss function $L(\theta)$ in classification? (We mentioned that MSE loss is uncommon in classification.)

- Like above, can we use classification when target has continuous value? why we often using regression instead of classification?