1. Read Deep Learning: An Introduction for Applied Mathematicians. Consider a network as defined in (3.1) and (3.2). Assume that $n_L = 1$, find an algorithm to calculate $\nabla a^{[L]}(x)$.

$$\nabla a^{[L]}(x) = \begin{bmatrix} \frac{\partial a^{[L]}}{\partial w_{ij}^{[L]}} \\ \frac{\partial a^{[L]}}{\partial w_{ij}^{[L-1]}} \\ \vdots \\ \frac{\partial a^{[L]}}{\partial w_{ij}^{[2]}} \\ \frac{\partial a^{[L]}}{\partial b_j^{[L]}} \\ \vdots \\ \frac{\partial a^{[L]}}{\partial b_j^{[2]}} \end{bmatrix} \begin{array}{l} \rightarrow 1 \le i \le \overset{1}{\overset{\parallel}{n_L}},\ 1 \le j \le n_{L-1} \leftarrow \text{將 } n_L \times n_{L-1} \text{ 個數排成一行} \\ \rightarrow 1 \le i \le n_{L-1},\ 1 \le j \le n_{L-2} \leftarrow \text{將 } n_{L-1} \times n_{L-2} \text{ 個數排成一行} \\ \\ \rightarrow 1 \le i \le n_2,\ 1 \le j \le n_1 \leftarrow \text{將 } n_2 \times n_1 \text{ 個數排成一行} \\ \rightarrow 1 \le j \le n_L \\ \\ \rightarrow 1 \le j \le n_2 \end{array}$$

Note: $a^{[k]} = \sigma(z^{[k]})$ where $z^{[k]} \in \mathbb{R}^{n_k}$

$z^{[k]} = W^{[k]} a^{[k-1]} + b^{[k]}$ where $W^{[k]} \in M_{n_k \times n_{k-1}}$, $a \in \mathbb{R}^{n_{k-1}}$, $b^{[k]} \in \mathbb{R}^{n_k}$

$$\underset{\mathbb{R}}{\underset{\cap}{a^{[L]}}} = \sigma(\underset{\mathbb{R}}{\underset{\cap}{z^{[L]}}}) = \sigma(\underset{M_{1\times n_{L-1}}}{\underset{\cap}{W^{[L]}}}\, \underset{\mathbb{R}^{n_{L-1}}}{\underset{\cap}{a^{[L-1]}}} + \underset{\mathbb{R}}{\underset{\cap}{b^{[L]}}}) = \sigma(W^{[L]}\sigma(z^{[L-1]}) + b^{[L]})$$

$$= \sigma(W^{[L]}\sigma(W^{[L-1]}a^{[L-2]} + b^{[L-1]}) + b^{[L]})$$

$$= \sigma(W^{[L]}\sigma(W^{[L-1]}\sigma(z^{[L-2]}) + b^{[L-1]}) + b^{[L]})$$

$$= \sigma\left(W^{[L]}\sigma\left(W^{[L-1]}\sigma\left(W^{[L-2]}a^{[L-3]}+b^{[L-2]}\right)+b^{[L-1]}\right)+b^{[L]}\right)$$

$\because n_L = 1$

$\therefore W^{[L]} \in M_{1\times n_{L-1}}$ $\quad \dfrac{\partial a^{[L]}}{\partial w_{1j}^{[L]}} = \dfrac{\partial a^{[L]}}{\partial z^{[L]}} \cdot \dfrac{\partial z^{[L]}}{\partial w_{1j}^{[L]}}$

$$\frac{\partial a^{[L]}}{\partial w_{ij}^{[L-1]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \frac{\partial a_j^{[L-1]}}{\partial z_i^{[L-1]}} \frac{\partial z_i^{[L-1]}}{\partial w_{ij}^{[L-1]}}$$

$$\frac{\partial a^{[L]}}{\partial w_{ij}^{[L-2]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \frac{\partial a_j^{[L-1]}}{\partial z_i^{[L-1]}} \frac{\partial z_i^{[L-1]}}{\partial a_j^{[L-2]}} \frac{\partial a_j^{[L-2]}}{\partial z_i^{[L-2]}} \frac{\partial z_i^{[L-2]}}{\partial w_{ij}^{[L-2]}}$$

$$\Rightarrow \frac{\partial a^{[L]}}{\partial w_{ij}^{[L-k]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \cdot \cdots \cdot \frac{\partial a_j^{[L-k+1]}}{\partial z_i^{[L-k+1]}} \frac{\partial z_i^{[L-k]}}{\partial w_{ij}^{[L-k]}}, \quad \text{for } k = 0, \cdots, L-2$$

$$W^{[L-k]} \in M_{n_{L-k}\times n_{L-k+1}}$$

$$\frac{\partial a^{[L]}}{\partial b^{[L]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \cdot \underbrace{\frac{\partial z^{[L]}}{\partial b^{[L]}}}_{=1} = \frac{\partial a^{[L]}}{\partial z^{[L]}}$$

$$\frac{\partial a^{[L]}}{\partial b_j^{[L-1]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \frac{\partial a_j^{[L-1]}}{\partial z_j^{[L-1]}} \frac{\partial z_j^{[L-1]}}{\partial b_j^{[L-1]}}$$

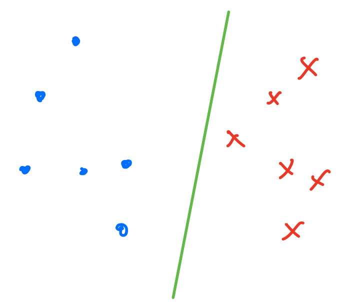
$$\frac{\partial a^{[L]}}{\partial b_j^{[L-2]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \frac{\partial a_j^{[L-1]}}{\partial z_j^{[L-1]}} \cdot \frac{\partial z_j^{[L-1]}}{\partial a_j^{[L-2]}} \frac{\partial a_j^{[L-2]}}{\partial z_j^{[L-2]}}$$

$$\Rightarrow \frac{\partial a^{[L]}}{\partial b_j^{[L-k]}} = \frac{\partial a^{[L]}}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial a_j^{[L-1]}} \cdot \cdots \cdot \frac{\partial a_j^{[L-k]}}{\partial z^{[L-k]}}, \quad \text{for } k=0, \cdots, L-2. \quad b^{[L-k]} \in \mathbb{R}^{n_{L-k}}.$$

2. There are unanswered questions during the lecture, and there are likely more questions we haven't covered. Take a moment to think about them and write them down here.

在 classification 問題中，

data : $\{(\vec{x}_i, c_i)\}$, $c_i \in \{0,1\}$



方法1: Find a function $H : \mathbb{R}^2 \to \mathbb{R}$ s.t $H(\vec{x}_i) = c_i$

方法2: (One-hot coding)

Let $c_i \in \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$

Find a function $\hat{H} : \mathbb{R}^2 \to \mathbb{R}$ s.t $\hat{H}(\vec{x}_i) = c_i$

Q: 在方法1中 找的 H 會是 discontinuous function，在分界會學不好

為什麼方法2比較好? (可以克服方法1的問題?)