1.

$h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$

given one data point $(x_1, x_2, y) = (1, 2, 3)$ and assuming that

$\theta^0 = (b, w_1, w_2) = (4, 5, 6)$

Use the stochastic gradient descent method to evaluate θ'

$$\theta' = \theta^0 - \alpha \nabla_\theta \text{Loss} \quad , \text{ where } \text{Loss}(\theta) = (y - h(x_1, x_2))^2 .$$

$$= \theta^0 + 2\alpha (y - h(x_1, x_2)) \cdot \nabla_\theta h \qquad \alpha > 0 = \text{learning rate.}$$

Note that $\sigma(x) = \frac{1}{1+e^{-x}} \Rightarrow \sigma'(x) = \frac{d(1+e^{-x})^{-1}}{dx} = -(1+e^{-x})^{-2} \cdot (-e^{-x})$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$= \sigma(x) \cdot (1 - \sigma(x))$$

$$\Rightarrow \nabla_\theta h = \begin{bmatrix} \frac{\partial h}{\partial b} \\ \frac{\partial h}{\partial w_1} \\ \frac{\partial h}{\partial w_2} \end{bmatrix} , \begin{cases} \frac{\partial h}{\partial b} = \frac{\partial h}{\partial \theta} \cdot \frac{\partial \theta}{\partial b_1} = h(1-h) \cdot 1 \\ \frac{\partial h}{\partial w_1} = \frac{\partial h}{\partial \theta} \cdot \frac{\partial \theta}{\partial w_1} = h(1-h) \cdot x_1 = h \cdot (1-h) \cdot 1 \\ \frac{\partial h}{\partial w_2} = \frac{\partial h}{\partial \theta} \cdot \frac{\partial \theta}{\partial w_2} = h(1-h) \cdot x_2 = h(1-h) \cdot 2 \end{cases}$$

$$\Rightarrow \theta' = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} + 2 \cdot \alpha \cdot (3-h) \cdot \begin{bmatrix} h(1-h) \\ h(1-h) \\ h(1-h) \cdot 2 \end{bmatrix}$$

$$\Rightarrow \theta' = \begin{bmatrix} 4 + 2\alpha \cdot (3-h) \cdot h(1-h) \\ 5 + 2\alpha \cdot (3-h) \cdot h(1-h) \\ 6 + 2\alpha \cdot (3-h) \cdot h(1-h) \cdot 2 \end{bmatrix} , \text{ where } h = \sigma(4 + 5 \cdot 1 + 6 \cdot 2) = \sigma(21)$$

*

2.

(a) $\sigma'(x) = \frac{d\,(1+e^{-x})^{-1}}{dx} = -(1+e^{-x})^{-2}\cdot(-e^{-x})$

$= (1+e^{-x})^{-2}\cdot e^{-x}$

$= (1+e^{-x})^{-1}\cdot\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\cdot(1-\sigma(x))$ ✡

$\sigma''(x) = \frac{d\,(\sigma(x)\cdot(1-\sigma(x))}{dx}$

$= \sigma'(x)\cdot(1-\sigma(x)) - \sigma(x)\cdot\sigma'(x)$

$= \sigma'(x)\cdot(1-2\sigma(x))$

$= \sigma(x)\cdot(1-\sigma(x))\cdot(1-2\sigma(x))$ ✡

$\sigma'''(x) = \frac{d\,\sigma''(x)}{dx} = \sigma(x)\cdot(1-\sigma(x))\cdot(1-2\sigma(x))\cdot(1-2\sigma(x)) - 2\sigma(x)\cdot(1-\sigma(x))\cdot\sigma(x)\cdot(1-\sigma(x))$

$= \sigma(x)\cdot(1-\sigma(x))\cdot(1-4\sigma(x)+4\sigma(x)^2-2\sigma(x)+2\sigma(x)^2)$

$= \sigma(x)\cdot(1-\sigma(x))\cdot(1-6\sigma(x)+6\sigma(x)^2)$ ✡

(b) $\sigma(x) = \frac{1}{1+e^{-x}}$, $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

let $\tanh(x) = C_1\sigma(C_2x+C_3)+C_4$

take $x=0 \Rightarrow \tanh(0)=0 \Rightarrow C_1\sigma(C_3)+C_4=0$

$\because \tanh(x)$ is an odd function, to make RHS be an odd function $\Rightarrow C_3=0$

$\because \sigma(0)=\frac{1}{2} \Rightarrow C_1\cdot\frac{1}{2}+C_4=0$

$\tanh(x)\longrightarrow 1$ as $x\to\infty \Rightarrow C_1\cdot 1+C_4=1$

$\Rightarrow C_1=2,\ C_4=-1$

$\because \tanh'(0)=1 \Rightarrow \frac{d}{dx}[C_1\cdot\sigma(C_2x)] = C_1\cdot C_2\cdot\sigma'(x) \Rightarrow C_1\cdot C_2\cdot\sigma'(0)=1 \Rightarrow C_2=2$

$\Rightarrow \tanh(x) = 2\cdot\sigma(2x)-1$ ✡

3、

1、Mini-Batch GD 的 size m 通常怎麼選

2、為什麼要用非線性函數 (如 sigmoid)

3、用線性函數的話，結果會有什麼變化

4、上課時老師有時候 Loss function 前面有乘 $\frac{1}{2}$

有時候沒有，想知道什麼情況下要乘 $\frac{1}{2}$，why?