1. For MSE loss, we have $\theta' = \theta^0 + 2\alpha[(y - h(x_1, x_2))\nabla_\theta h]$ for SGD method and $h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2)$. Also, $\nabla_\theta h = (\sigma(1-\sigma), x_1\sigma(1-\sigma), x_2\sigma(1-\sigma))$

Thus, $\theta' = (4,5,6) + 2\alpha[(3 - h(1,2))(1,1,2) \cdot h(1,2)(1-h(1,2))]$

$= (4,5,6) + 2\alpha[(3 - \sigma(21))\sigma(21)(1-\sigma(21))(1,1,2)]$ □

2. (a) For $\sigma(x) = \frac{1}{1+e^{-x}}$, $\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$

$$= \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right)$$

$$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = \underline{\sigma(1-\sigma)}.$$

Then $\sigma''(x) = [\sigma(1-\sigma)]' = \sigma'(1-\sigma) + \sigma(-1)\sigma' = \sigma'(1-2\sigma) = \underline{\sigma(1-\sigma)(1-2\sigma)}$.

Then $\sigma'''(x) = [\sigma(1-\sigma)(1-2\sigma)]' = [\sigma(1-\sigma)]'(1-2\sigma) + \sigma(1-\sigma)(1-2\sigma)'$

$$= \sigma(1-\sigma)(1-2\sigma)(1-2\sigma) + \sigma(1-\sigma)(-2)\sigma(1-\sigma)$$

$$= \sigma(1-\sigma)[1 - 4\sigma + 4\sigma^2 - 2\sigma + 2\sigma^2]$$

$$= \underline{\sigma(1-\sigma)(6\sigma^2 - 6\sigma + 1)}$$

(b) For $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$

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

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

$$= \underline{2\sigma(2x) - 1}$$

$\cosh(x) = \frac{e^x + e^{-x}}{2} = \frac{1+e^{-2x}}{2}e^x = \frac{2}{\sigma(2x)}e^x = \underline{\frac{2(1-\sigma(-x))}{\sigma(2x)\sigma(-x)}}$ since we have

$\sigma(x) = \frac{1}{1+e^{-x}} \Rightarrow 1 = \sigma + \sigma e^{-x} \Rightarrow e^{-x} = \frac{1-\sigma}{\sigma}$.

Then $\sinh(x) = \tanh(x)\cosh(x) = \underline{(2\sigma(2x) - 1)\frac{2(1-\sigma(-x))}{\sigma(2x)\sigma(-x)}}$.

Hence, $\coth(x) = \underline{\frac{1}{2\sigma(2x)-1}}$, $\operatorname{sech}(x) = \underline{\frac{\sigma(2x)\sigma(-x)}{2(1-\sigma(-x))}}$,

$$csch(x) = \frac{\sigma(2x)\sigma(-x)}{2(2\sigma(2x)-1)(1-\sigma(x))}$$ □

3. 如果 data set 裡的 data 出現少數幾點與其它 data 相比偏離很大，使用 MSE loss 可能會導致因這幾點的影響過大，而降低對其它多數點的預測。這種情況怎麼調整 loss function 或 data set ?