1. $h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2)$, where $\sigma(x)$ is the sigmoid function.

Use SGD,

$$\theta' = \theta^0 - \alpha \nabla_\theta \left(\frac{1}{2} \| y - h(x_1, x_2) \|^2 \right), \quad \text{where } \alpha \text{ is learning rate.}$$

For $\nabla_\theta \left(\frac{1}{2} \| y - h(x_1, x_2) \|^2 \right) = \begin{bmatrix} \frac{\partial}{\partial b} \frac{1}{2} |y - h(x_1, x_2)|^2 \\ \frac{\partial}{\partial w_1} \frac{1}{2} |y - h(x_1, x_2)|^2 \\ \frac{\partial}{\partial w_2} \frac{1}{2} |y - h(x_1, x_2)|^2 \end{bmatrix}$,

Note $\frac{\partial}{\partial x} \sigma(x) = \sigma(x)(1 - \sigma(x))$

where $\frac{\partial}{\partial b} \frac{1}{2} |y - h(x_1, x_2)|^2 = |y - \sigma(b + w_1 x_1 + w_2 x_2)| \frac{\partial}{\partial b} \left(- \underline{\sigma(b + w_1 x_1 + w_2 x_2)} \right)$

$$= |y - \sigma(b + w_1 x_1 + w_2 x_2)| \left(-\sigma(b + w_1 x_1 + w_2 x_2)(1 - \sigma(b + w_1 x_1 + w_2 x_2)) \right)$$

$$\frac{\partial}{\partial w_1} \frac{1}{2} |y - h(x_1, x_2)|^2 = |y - \sigma(b + w_1 x_1 + w_2 x_2)| \left(-\sigma(b + w_1 x_1 + w_2 x_2)(1 - \sigma(b + w_1 x_1 + w_2 x_2)) \right) \cdot x_1$$

$$\frac{\partial}{\partial w_2} \frac{1}{2} |y - h(x_1, x_2)|^2 = |y - \sigma(b + w_1 x_1 + w_2 x_2)| \left(-\sigma(b + w_1 x_1 + w_2 x_2)(1 - \sigma(b + w_1 x_1 + w_2 x_2)) \right) \cdot x_2$$

Take $(x_1, x_2, y) = (1, 2, 3)$ and $\theta^0 = (b, w_1, w_2) = (4, 5, 6)$, then

$$\theta' = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \alpha \begin{bmatrix} |3 - \sigma(21)| (-\sigma(21)(1 - \sigma(21))) \\ |3 - \sigma(21)| (-\sigma(21)(1 - \sigma(21))) \\ 2|3 - \sigma(21)| (-\sigma(21)(1 - \sigma(21))) \end{bmatrix} \#$$

2. (a) For $k = 1$, $\frac{d}{dx} \sigma(x) = \frac{d}{dx} \frac{1}{1 + e^{-x}} = \frac{-e^{-x} \cdot (-1)}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} = \sigma(x)(1 - \sigma(x))_\#$

For $k = 2$, $\frac{d^2}{dx^2} \sigma(x) = \frac{d}{dx} (\sigma(x)(1 - \sigma(x))) = (1 - \sigma(x)) \frac{d}{dx} \sigma(x) + \sigma(x) \frac{d}{dx} (1 - \sigma(x))$

$$= \sigma(x)(1 - \sigma(x))^2 - \sigma^2(x)(1 - \sigma(x))_\#$$

For $k = 3$, $\frac{d^3}{dx^3} \sigma(x) = \frac{d}{dx} \left(\sigma(x)(1 - \sigma(x))^2 - \sigma^2(x)(1 - \sigma(x)) \right)$

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

$$= \sigma(x)(1 - \sigma(x))^3 - 2\sigma^2(x)(1 - \sigma(x))^2 - 2\sigma^2(x)(1 - \sigma(x))^2 + \sigma^3(x)(1 - \sigma(x))$$

$$= \sigma(x)(1 - \sigma(x))^3 - 4\sigma^2(x)(1 - \sigma(x))^2 + \sigma^3(x)(1 - \sigma(x))_\#$$

(b) Note $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

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

3. How to choose a suitable loose function ?

MSE is suitable for regression problems, but if it is classification problems or unbalanced data, should cross-entropy or weighted loss be used ?