1.

Model:

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

Loss function (squared error):

$$L = \frac{1}{2}\big(y - h(x_1, x_2)\big)^2$$

Let $z = b + w_1 x_1 + w_2 x_2$, $h = \sigma(z)$,

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

Gradients with respect to parameters:

$$\frac{\partial L}{\partial b} = (h - y)\sigma'(z)$$

$$\frac{\partial L}{\partial w_1} = (h - y)\sigma'(z)x_1$$

$$\frac{\partial L}{\partial w_2} = (h - y)\sigma'(z)x_2$$

SGD Formula:

$$\theta^{n+1} \coloneqq \theta^n - \alpha \nabla_\theta J(\theta^n)$$

where $J: \mathbb{R}^m \to \mathbb{R}$ is the loss function, $\alpha > 0$ is the learning rate, $n \in \mathbb{N} \cup \{0\}$.

Thus,

$$\theta^1 = \theta^0 - \alpha \nabla_\theta J(\theta^0)$$

$$\Rightarrow \theta^1 = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial b} \\ \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \end{bmatrix}$$

Since $z = 4 + 5 \cdot 1 + 6 \cdot 2 = 21, h = \sigma(21), \sigma'(21) = \sigma(21)(1 - \sigma(21))$

Thus,

$$\theta^1 = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \alpha \begin{bmatrix} (\sigma(21) - 3)\sigma(21)\big(1 - \sigma(21)\big) \\ (\sigma(21) - 3)\sigma(21)\big(1 - \sigma(21)\big) \cdot 1 \\ (\sigma(21) - 3)\sigma(21)\big(1 - \sigma(21)\big) \cdot 2 \end{bmatrix}$$

where α is the learning rate.

2.

(a)

Since σ is the sigmoid function, we have

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

Then, when $k = 1$,

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

When $k = 2$,

$$\begin{aligned}\frac{d^2}{dx^2}\sigma(x) &= \frac{d}{dx}\sigma(x)\big(1-\sigma(x)\big)\\ &= \sigma'(x)\big(1-\sigma(x)\big) - \sigma(x)\sigma'(x)\\ &= \sigma(x)\big(1-\sigma(x)\big)\big(2-\sigma(x)\big)\end{aligned}$$

When $k = 3$,

$$\begin{aligned}\frac{d^3}{dx^3}\sigma(x) &= \frac{d}{dx}\sigma(x)\big(1-\sigma(x)\big)\big(2-\sigma(x)\big)\\ &= \sigma'(x)\big(1-\sigma(x)\big)\big(2-\sigma(x)\big)\\ &\quad -\sigma(x)\sigma'(x)\big(2-\sigma(x)\big)\\ &\quad -\sigma(x)\big(1-\sigma(x)\big)\cdot 2\sigma'(x)\\ &= \sigma(x)\big(1-\sigma(x)\big)\big(1-6\sigma(x)+6\sigma^2(x)\big)\end{aligned}$$

(b)

We know

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

And we also observe that the hyperbolic functions

$$\begin{aligned}\tanh(x) &= \frac{e^x-e^{-x}}{e^x+e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1}\\ &= \frac{2e^{2x}}{e^{2x}+1} - \frac{e^{2x}+1}{e^{2x}+1} = \frac{2e^{2x}}{e^{2x}+1} - 1\\ &= 2\sigma(2x) - 1\end{aligned}$$

Thus, we can also write as

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

3.

Question: In what situations should I use each type of loss function?