1.

$$h(x_1, x_2) = \sigma(b + w_1x_1 + w_2x_2) = \sigma(z), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

$$(x_1, x_2, y) = (1, 2, 3)$$

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

$$\theta' = \theta^0 - \alpha \nabla_\theta L(\theta^0), \quad L(\theta) = \frac{1}{2}|y - h(x_1, x_2; \theta)|^2$$

$\nabla_\theta L(\theta)$:

$$\frac{\partial L}{\partial \theta} = (y - h(x_1, x_2)) \cdot - \frac{\partial h}{\partial \theta}$$

$$\frac{\partial h}{\partial \theta} = \frac{\partial \sigma(z)}{\partial z} \cdot \frac{\partial z}{\partial \theta}, \quad z = b + w_1x_1 + w_2x_2$$

$$= (1 - \sigma(z))\sigma(z)$$

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

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

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

$$\Rightarrow \theta' = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix} + \alpha \begin{bmatrix} \frac{\partial L}{\partial b} \\ \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} + \alpha \begin{bmatrix} (\sigma(21) - 3)(1 - \sigma(21))\sigma(21) \\ (\sigma(21) - 3)(1 - \sigma(21))\sigma(21) \cdot 1 \\ (\sigma(21) - 3)(1 - \sigma(21))\sigma(21) \cdot 2 \end{bmatrix}$$

2.

(a)

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

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

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

$$\sigma'''(x) = 6\sigma^2(x)\sigma'(x) - 6\sigma(x)\sigma'(x) + \sigma'(x)$$

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

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

(b)

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

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

$$\Rightarrow \sigma(x) = \frac{\tan(\frac{x}{2}) + 1}{2}$$

3.

1. In class, we discussed the Mean Squared Error (MSE) as a loss function. After doing some research, I found that MSE is more suitable for regression problems, while classification tasks often use different loss functions, such as Cross Entropy. I would like to ask: will we have the opportunity to learn about classification-specific loss functions in the future? Also, how does MSE perform in terms of numerical stability for classification problems?

2. Why is the sigmoid activation function prone to the vanishing gradient problem? Are there any known solutions or alternative activation functions to address this problem?

- The third question involves using AI tools and searching for information online.