1.

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

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

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

$\theta' = \theta^0 - \alpha \nabla_\theta Loss(\theta^0)$

$Loss : L(\theta) = \frac{1}{2} \| y - h(x_1, x_2) \|^2$

$$\theta' = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial b} \\ \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \end{bmatrix}$$

$$\frac{\partial h(x_1, x_2)}{\partial b} = \sigma'(u) \cdot 1 = (1-\sigma(u))\sigma(u) = (1-\sigma(21))\sigma(21)$$

$$\frac{\partial h(x_1, x_2)}{\partial w_1} = \sigma'(u) \cdot x_1 = (1-\sigma(21))\sigma(21) \cdot 1$$

$$\frac{\partial h(x_1, x_2)}{\partial w_2} = \sigma'(u) \cdot x_2 = (1-\sigma(21))\sigma(21) \cdot 2$$

$$\frac{\partial L}{\partial b} = \| y - h(x_1, x_2) \| \cdot (-1) \cdot \frac{\partial h(x_1, x_2)}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \| y - h(x_1, x_2) \| \cdot (-1) \cdot \frac{\partial h(x_1, x_2)}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \| y - h(x_1, x_2) \| \cdot (-1) \cdot \frac{\partial h(x_1, x_2)}{\partial w_2}$$

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

2.

(a) $\frac{d^k}{dx^k} \sigma$, $k = 1, 2, 3$

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

$k=1$: $\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = -(1+e^{-x})^{-2} \cdot e^{-x} \cdot (-1) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x) \cdot (1-\sigma(x))$ ※

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

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

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

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

$k=3$: $\frac{d^3}{dx^3}\sigma(x) = \frac{d}{dx}[\sigma(x)(1-\sigma(x))(1-2\sigma(x))]$

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

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

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

(b) Sigmoid function : $\frac{1}{1+e^{-x}}$

Hyperbolic function : choose $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

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

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

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

3. Besides the sigmoid function, is there any other commonly used activation functions? What are their respective limitations, pros and cons?

Δ Answer from Chat GPT :

- ReLU and its variants are widely used in hidden layers due to efficiency and reduced vanishing gradient issues.
- tanh is sometimes preferred for small networks or zero-centered outputs.
- Softmax is specialized for output layers in classification tasks.
- Each activation function has trade-offs between gradient behavior, output range, and computational cost.