1. Consider stochastic gradient descent method to learn the house price model

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

where σ is the sigmoid function.

Given one single data point $(x_1, x_2, y) = (1, 2, 3)$, and assuming that the current parameter is $\theta^0 = (b, w_1, w_2) = (4, 5, 6)$, evaluate θ^1.

> Just write the expression and substitute the numbers; no need to simplify or evaluate.

$$\theta' = \theta^\circ - \alpha \nabla_\theta L(h(x_1, x_2), y)$$

$$= (4, 5, 6) - \alpha \nabla_\theta L(\sigma(b + w_1 x_1 + w_2 x_2), y)$$

$$= \Big(4 - \alpha \frac{\partial}{\partial b} L(\sigma(4 + 5 \cdot 1 + 6 \cdot 2), 3),$$

$$5 - \alpha \frac{\partial}{\partial w_1} L(\sigma(4 + 5 \cdot 1 + 6 \cdot 2), 3),$$

$$6 - \alpha \frac{\partial}{\partial w_2} L(\sigma(4 + 5 \cdot 1 + 6 \cdot 2), 3) \Big)$$

2. (a) Find the expression of $\frac{d^k}{dx^k}\sigma$ in terms of $\sigma(x)$ for $k = 1, \cdots, 3$ where σ is the sigmoid function.

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

(b) Find the relation between sigmoid function and hyperbolic function.

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

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

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

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

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

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

$$= \sigma(x)(1 - \sigma(x)) \cdot [(1 - 2\sigma(x))^2 - 2\sigma(x)(1 - \sigma(x))]$$

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

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

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

3. There are unanswered questions during the lecture, and there are likely more questions we haven't covered. Take a moment to think about them and write them down here.

How to choose an appropriate learning rate α?

A scheduler is a mechanism that can dynamically adjust the learning rate during model training, including methods such as StepLR, cosine annealing and exponential decay.

Exponential Decay:

$\alpha_t = \alpha_0 \times e^{-kt}$, where

α_t is the learning rate at the t-th iteration.

α_0 is the initial learning rate.

k is the decay rate, $k \in (0,1)$ (usually).

t is the current training step or iteration number.