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.

Let the learning rate be α.

First, we compute ∇h:

Since $\sigma'(x) = \sigma(x)(1-\sigma(x))$ (prove in 2.), we have

$$\frac{\partial h}{\partial b} = \sigma'(4+5x_1+6x_2) \cdot 1$$

$$= \sigma(4+5x_1+6x_2) \cdot [1 - \sigma(4+5x_1+6x_2)]$$

$$\frac{\partial h}{\partial w_1} = \sigma'(4+5x_1+6x_2) \cdot x_1$$

$$= \sigma(4+5x_1+6x_2) \cdot [1 - \sigma(4+5x_1+6x_2)] \cdot x_1$$

$$\frac{\partial h}{\partial w_2} = \sigma'(4+5x_1+6x_2) \cdot x_2$$

$$= \sigma(4+5x_1+6x_2) \cdot [1 - \sigma(4+5x_1+6x_2)] \cdot x_2$$

Hence $\nabla_\theta h(1,2) = \Big(\sigma(21)\cdot[1-\sigma(21)],$

$\sigma(21)\cdot[1-\sigma(21)],$

$\sigma(21)\cdot[1-\sigma(21)]\cdot 2 \Big)$

$$\theta^1 = \theta^0 + 2\alpha\,(y - h(1,2))\,\nabla_\theta h(1,2)$$

$$= (4,5,6) + 2\alpha\Big[(3-\sigma(21))\,\nabla_\theta h(1,2)\Big]$$

$$= \Big(4 + 2\alpha\cdot(3-\sigma(21))\cdot\sigma(21)\cdot[1-\sigma(21)],$$

$$5 + 2\alpha\cdot(3-\sigma(21))\cdot\sigma(21)\cdot[1-\sigma(21)],$$

$$6 + 2\alpha\cdot(3-\sigma(21))\cdot\sigma(21)\cdot[1-\sigma(21)]\cdot 2\Big)$$

*

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.

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

(a) Since $\sigma(x) = \dfrac{1}{1+e^{-x}}$, we have

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

$$= \frac{1}{1+e^{-x}} \times \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \times \frac{e^{-x}+1-1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \times \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x)\left(1-\sigma(x)\right) \quad \#$$

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

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

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

$$\sigma'''(x) = (\sigma(x)'')'$$
$$= \left(\sigma'(x)\,(1-2\sigma(x))\right)'$$
$$= \sigma''(x)(1-2\sigma(x)) + \sigma'(x)(1-2\sigma(x))'$$
$$= \sigma''(x)(1-2\sigma(x)) - 2(\sigma'(x))^2$$
$$= \sigma(x)(1-\sigma(x))(1-2\sigma(x))^2 - 2\sigma(x)^2(1-\sigma(x))^2$$
$$= \sigma(x)(1-\sigma(x))\left[1-4\sigma(x)+4\sigma^2(x)-2\sigma(x)(1-\sigma(x))\right]$$
$$= \sigma(x)(1-\sigma(x))\left[6\sigma^2(x)-6\sigma(x)+1\right]$$
※

(b) Since $\sigma(x) = \dfrac{1}{1+e^{-x}}$, we have $\sigma(x) = \dfrac{e^x}{e^x+1}$.

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

For convenience, we write $\sigma(x)$ as σ.

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$
$$= \frac{\dfrac{\sigma}{1-\sigma} - \dfrac{1-\sigma}{\sigma}}{2} = \frac{2\sigma-1}{2\sigma(1-\sigma)}$$
※

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

$$= \frac{\frac{\sigma}{1-\sigma} + \frac{1-\sigma}{\sigma}}{2} = \frac{2\sigma^2 - 2\sigma + 1}{2\sigma(1-\sigma)} \quad ※$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{2\sigma - 1}{2\sigma^2 - 2\sigma + 1} \quad ※$$

$$\coth(x) = \frac{\cosh(x)}{\sinh(x)} = \frac{2\sigma^2 - 2\sigma + 1}{2\sigma - 1} \quad ※$$

$$\operatorname{sech}(x) = \frac{1}{\cosh(x)} = \frac{2\sigma(1-\sigma)}{2\sigma^2 - 2\sigma + 1} \quad ※$$

$$\operatorname{csch}(x) = \frac{1}{\sinh(x)} = \frac{2\sigma(1-\sigma)}{2\sigma - 1} \quad ※$$

As mentioned in the course, maybe we can have a more in-depth discussion about the relationship between $\sigma(x)$ and $\tanh(x)$.

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

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

Therefore, $\tanh(x) = 2\sigma(2x) - 1$ ※ or equivalently

$$\sigma(x) = \frac{\tanh(\frac{x}{2}) + 1}{2}.$$ ※

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.

在做 gradient descent 時，不同的初始參數 θ° 的選擇是否都會收斂？都會收斂到相同的 $\tilde{\theta}$？ learning rate α 該如何挑選？