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.

$$\mathrm{Loss}(\theta) = \frac{1}{N}\sum_{i=1}^{N} |y^i - h(x^i; \theta)|$$

$$= |y^1 - h(x^1; \theta)| \qquad (N=1)$$

$$= 3 - h(1, 2; \theta)$$

$$= 3 - \sigma(b + w_1 + 2w_2)$$

$$\theta^1 = \theta^0 - \alpha \nabla_\theta \mathrm{Loss}(\theta^0) \quad , \quad \alpha > 0 .$$

$$= \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial \mathrm{Loss}}{\partial b}(4, 5, 6) \\ \frac{\partial \mathrm{Loss}}{\partial w_1}(4, 5, 6) \\ \frac{\partial \mathrm{Loss}}{\partial w_2}(4, 5, 6) \end{bmatrix}$$

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) $k=1$:

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

$k=2$:

$$\frac{d^2}{dx^2}\sigma(x) = \frac{d}{dx}\sigma' = \frac{d}{dx}[\sigma(1-\sigma)] = \sigma'(1-\sigma)+\sigma(-\sigma') = \sigma(1-\sigma)(1-\sigma)-\sigma\cdot\sigma(1-\sigma)$$

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

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

$k=3$:

$$\frac{d^3}{dx^3}\sigma(x) = \frac{d}{dx}\left(\frac{d^2}{dx^2}\sigma\right) = \frac{d}{dx}[\sigma(1-\sigma)(1-2\sigma)] = \sigma'(1-\sigma)(1-2\sigma)+\sigma(-\sigma')(1-2\sigma)+\sigma(1-\sigma)(-2\sigma')$$

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

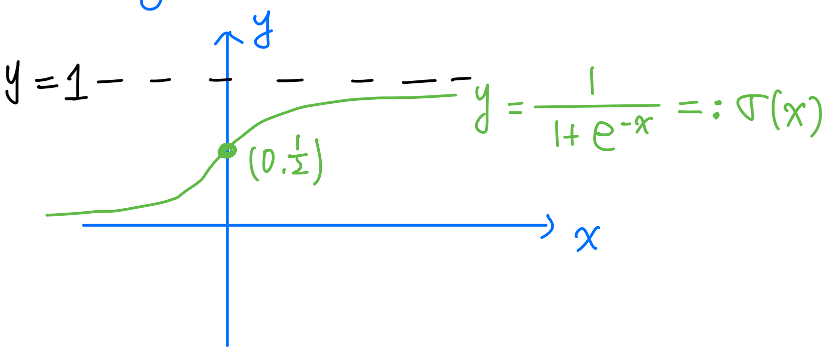
$$= \sigma(1-\sigma)[(1-\sigma)(1-2\sigma) - \sigma(1-2\sigma) - 2\sigma(1-\sigma)]$$

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

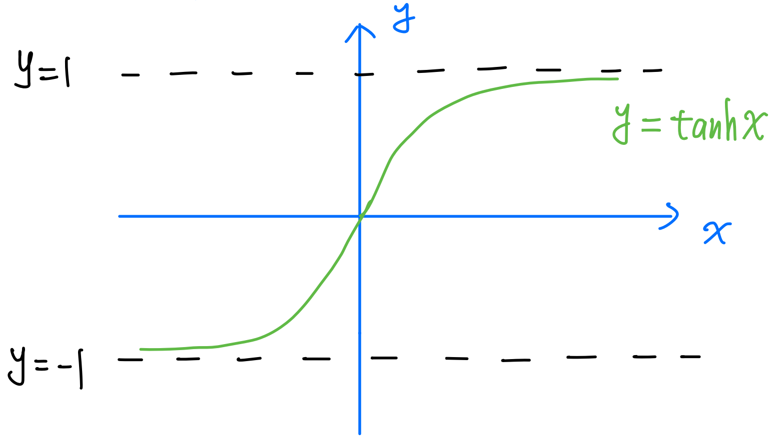
$$= \sigma(1-\sigma)(6\sigma^2-6\sigma+1)$$

(b)

Sigmoid function



hyperbolic function



Let $\tanh x = c_1\sigma(c_2x+c_3)+c_4$. $c_i\in\mathbb{R}$, $i=1,\cdots,4$

根據觀察圖形, guess $c_3=0$

Let $x=0$, then $0=c_1\sigma(c_2\cdot 0+0)+c_4 = c_1\sigma(0)+c_4 = \frac{1}{2}c_1+c_4 \Rightarrow c_4=-\frac{1}{2}c_1$

∴ Let $\tanh x = c_1\sigma(c_2x)-\frac{1}{2}c_1 = \frac{c_1}{1+e^{-c_2x}}-\frac{1}{2}c_1$

$$\overset{\frac{d}{dx}}{\Rightarrow} 1-\tanh^2 x = \frac{c_1c_2e^{-c_2x}}{(1+e^{-c_2x})^2}.$$

$$\Rightarrow 1-\left[\frac{c_1^2}{(1+e^{-c_2x})^2}-\frac{c_1^2}{1+e^{-c_2x}}+\frac{1}{4}c_1^2\right]=\frac{c_1c_2e^{-c_2x}}{(1+e^{-c_2x})^2}$$

$$\Rightarrow 1-\frac{1}{4}c_1^2+\frac{-c_1^2+c_1^2+c_1^2e^{-c_2x}}{(1+e^{-c_2x})^2}=\frac{c_1c_2e^{-c_2x}}{(1+e^{-c_2x})^2} \Rightarrow c_1=\pm 2,\ c_2=c_1=\pm 2$$

$(c_1=c_2=2)$

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

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

$(c_1=c_2=-2)$

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.

Gradient descent algorithm : $\theta^{n+1} = \theta^{n} - \alpha \nabla_{\theta} Loss$. $\alpha > 0$: learning rate

為什麼會收斂？

會收斂到同個值嗎？

收斂的值是最大/最小值嗎？

α 要怎麼選？