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

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

4 + 5 + 12

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.

11265011 鄭晨鈞

For the gradient descent algorithm with MSE loss, we have

$$\theta^{n+1} = \theta^n + 2\alpha\left[\frac{1}{m}\sum_{i=1}^{m}(y^i - h(x_1^i, x_2^i))\nabla_\theta h\right]$$

Let $n=0$, $m=1$ (stochastic)

4 + 5 + 12.

$$\theta^1 = \theta^0 + 2\alpha\left[(y^1 - h(x_1^1, x_2^1))\nabla_\theta h\right]$$
$$= \theta^0 + 2\alpha\left[(3 - h(1,2))\nabla_\theta h\right]$$
$$= \theta^0 + 2\alpha\left[(3 - \sigma(21))\nabla_\theta h\right]$$

$$\nabla_\theta h = \nabla_\theta h(1,2) = \nabla_\theta \sigma(b + w_1 + 2w_2)$$
$$= \langle \sigma'(b+w_1+2w_2), \sigma'(b+w_1+2w_2), 2\sigma'(b+w_1+2w_2)\rangle$$
$$= \langle \sigma'(21), \sigma'(21), 2\sigma'(21)\rangle$$

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

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

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

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

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

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

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.

隨著 gradient 一路被 update 到了不能再小，通常這時會認為已到達了局部極小值，但這不一定表示我們已經找到了全局最小值，有可能我們目前仍卡在山谷。

因此，該如何設計最佳的學習率策略？