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.

* data $\{(x_1^i, x_2^i, y^i)\}_{i=1}^{N}$

* Loss function :

$$\text{Loss}(b, w_1, w_2) = \frac{1}{M} \sum_{i=1}^{M} (y^i - h(x_1^i, x_2^i))^2$$

* Gradient descent algorithm

$$\theta^{n+1} = \theta^n - \alpha \nabla_\theta \text{Loss}$$

1° $h^0(x_1, x_2) = \sigma(4 + 5 \cdot 1 + 6 \cdot 2)$

$$= \sigma(21)$$

2° $\text{Loss} = \frac{1}{2}(h^0 - 3)^2$

3° let $b + \omega_1 x_1 + \omega_2 x_2 = \ell$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial \ell} \frac{\partial \ell}{\partial b}$$

$$= (h^{0} - y) \cdot h^{0}(1 - h^{0}) \cdot 1$$

$$= (h^{0} - 3) h^{0} (1 - h^{0})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial \ell} \frac{\partial \ell}{\partial w_1}$$

$$= (h^{0} - y) \cdot h^{0}(1 - h^{0}) \cdot x_1$$

$$= (h^{0} - 3) h^{0} (1 - h^{0})$$

$$\frac{\partial L}{\partial \omega_2} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial \ell} \frac{\partial \ell}{\partial \omega_2}$$

$$= (h^{0} - y) \cdot h^{0}(1 - h^{0}) \cdot x_2$$

$$= (h^{0} - 3) h^{0} (1 - h^{0}) \cdot 2$$

4°

$$b^1 = 4 - \alpha(h^0 - 3)\, h^0 (1 - h^0)$$

$$w_1^1 = 5 - \alpha(h^0 - 3)\, h^0 (1 - h^0)$$

$$w_2^1 = 6 - 2\alpha(h^0 - 3)\, h^0 (1 - h^0)$$

learning rate: α, $h^0 = 6(21)$

$$\Theta^1 = (b^1, w_1^1, w_2^1) \quad \#1$$

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)

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

$$\frac{d}{dx}\sigma(x) = \sigma'(x) = \frac{0\cdot(1+e^{-x}) - 1\cdot(0-e^{-x})}{(1+e^{-x})^2}$$

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

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

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

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

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

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

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

$$\frac{d^3}{dx^3}\sigma = \sigma'''(x)$$

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

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

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

$$\left((1-\sigma(x))(1-2\sigma(x)) - \sigma(x)(1-2\sigma(x))\right.$$
$$\left.-2\sigma(x)(1-\sigma(x))\right)$$

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

E1

(b)

1° $\sigma(x) = \dfrac{1}{1+e^{-x}}$

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

$$= \frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}} + e^{-\frac{x}{2}}} \cdot \left(2 \cdot \frac{1}{2}\right)$$

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

$$= \frac{1}{2} \cdot \frac{e^{\frac{x}{2}} + e^{\frac{x}{2}} + \left(e^{-\frac{x}{2}} - e^{-\frac{x}{2}}\right)}{e^{\frac{x}{2}} + e^{-\frac{x}{2}}}$$

$$= \frac{1}{2} \cdot \left(1 + \frac{e^{\frac{x}{2}} - e^{-\frac{x}{2}}}{e^{\frac{x}{2}} + e^{-\frac{x}{2}}}\right)$$

$$= \frac{1}{2}\left(1 + \tanh\left(\frac{x}{2}\right)\right)$$

$$= \frac{1+\tanh\left(\frac{x}{2}\right)}{2} \quad \#$$

$$\Rightarrow \tanh\left(\frac{x}{2}\right) = 2\sigma(x) - 1 \quad \#$$

2°

$$\sigma'(x) = \sigma(x)\left(1 - \sigma(x)\right) \quad \text{by } 2(a)$$

$$= \frac{1+\tanh\left(\frac{x}{2}\right)}{2} \cdot \frac{1-\tanh\left(\frac{x}{2}\right)}{2}$$

$$= \frac{1}{4} \cdot \left(1 - \tanh^2\left(\frac{x}{2}\right)\right)$$

$$= \frac{1}{4} \operatorname{sech}^2\left(\frac{x}{2}\right) \quad \#$$

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.

In "derivation of MSE loss" and "gradient descent", the sample index often starts from $i=1 \cdots M$, while in implementations often $i=0 \cdots M-1$, reminds me of statistic, where degree of freedom often reduced 1. Are these differences mainly notation conventions, or do they reflect mathematical or practical implications?

sorry for didn't notice assignment earlier :c