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 $L = \frac{1}{2} \| h(x_1, x_2) - y \|^2$, $z = b + w_1 x_1 + w_2 x_2$

SGD : $\theta' = \theta^\circ - \alpha \nabla_\theta L(\theta^\circ)$, α is learning rate

Note $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Then, $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial z} \frac{\partial z}{\partial b} = (h - y)\, \sigma'(z) \cdot 1$

$= (\sigma(z) - y)\, \sigma(z)(1 - \sigma(z))$

$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial z} \frac{\partial z}{\partial w_1} = (\sigma(z) - y)\, \sigma(z)(1 - \sigma(z))\, x_1$

$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial z} \frac{\partial z}{\partial w_2} = (\sigma(z) - y)\, \sigma(z)(1 - \sigma(z))\, x_2$

$h(x_1, x_2) = \sigma(4 + 5 \times 1 + 6 \times 2) = \sigma(21)$

$b : b' = b^\circ - \alpha \frac{\partial L}{\partial b}\Big|_{\theta^0} = 4 - \alpha(\sigma(21) - 3) \cdot \sigma(21)(1 - \sigma(21))$

$w_1 : w_1' = w_1^\circ - \alpha \frac{\partial L}{\partial w_1}\Big|_{\theta^0} = 5 - \alpha(\sigma(21) - 3) \cdot \sigma(21)(1 - \sigma(21))$

$w_2 : w_2' = 6 - \alpha(\sigma(21) - 3) \cdot \sigma(21)(1 - \sigma(21)) \cdot 2$

$\Rightarrow \theta' = (b', w_1', w_2')$ #

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

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

$\Rightarrow \sigma'(x)(1+e^{-x}) - \sigma(x)e^{-x} = 0$

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

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

$k=2$: $\sigma'(x) = \sigma(x) - \sigma^2(x)$

$\Rightarrow \sigma''(x) = \sigma'(x) - 2\sigma(x)\sigma'(x)$

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

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

$k=3$: $\sigma''(x) = \sigma'(x)(1-2\sigma(x))$

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

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

$= \sigma(x)(1-\sigma(x))(1-2\sigma(x))(1-2\sigma(x)) -$
$2(\sigma(x)(1-\sigma(x)))^2$ $\qquad 1-4\sigma(x)+4\sigma^2(x)$

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

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

(b) Note $\tanh x = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

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

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

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

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

3. ML Question

(a) If the learning rate is too large, it will fluctuate; If it's too small, it will converge slowly. How should we choose or adjust it in practice? Is it necessary to use a dynamic learning rate?

(b) In which supervised learning application is model interpretability as important as model accuracy?