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.

By Stochastic Gradient Descent, Gradient Descent Rule, and hypothesis, we have

$\theta^0 = (b, w_1, w_2) = (4, 5, 6)$ and

$\theta^1 = \theta^{0+1} = \theta^0 - \alpha \nabla_\theta L(\theta; x_1, x_2, y)$ with $L(\theta; x_1, x_2, y) = (y - h_\theta(x_1, x_2))^2$ and $\alpha > 0$ called learning rate. $-(*)$

Let $z = b + w_1 x_1 + w_2 x_2$, then $\nabla_\theta L = \frac{dL}{d\theta} = \frac{dL}{dh}\frac{dh}{dz}\frac{dz}{d\theta}$

$= -2(y - h_\theta(x_1, x_2)) \cdot \sigma'(z) \cdot \nabla_\theta z \quad -(**)$

We have $\sigma(z) = \frac{1}{1+e^{-z}}$, $\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2}$, and $1 - \sigma(z) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$.

Thus $\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z) \cdot (1 - \sigma(z)) \quad -(***)$

By $(*)$, $(**)$ and $(***)$, we have

$$\theta^1 = \theta^0 - \alpha \cdot \left[-2(y - h_\theta(x_1, x_2)) \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}\right.$$

Substitute $\theta^0 = (4, 5, 6)$ and $(x_1, x_2) = (1, 2)$ into the update formula, we have

$$\theta^1 = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + 2 \cdot \alpha \cdot (3 - h_{(4,5,6)}(1, 2)) \cdot \sigma(4 + 5 \times 1 + 6 \times 2) \cdot (1 - \sigma(4 + 5 \times 1 + 6 \times 2)) \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

Thus

$$\begin{pmatrix} b^1 \\ w_1^1 \\ w_2^1 \end{pmatrix} = \begin{pmatrix} 4 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 1 \\ 5 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 1 \\ 6 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 2 \end{pmatrix}. \quad \square$$

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.

2. (a) For $k=1$, $\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = \frac{-(-e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$

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

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

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

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

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

$$= \sigma(x)\cdot(1-\sigma(x))\cdot(1-2\sigma(x)). \square$$

For $k=3$, $\frac{d^3\sigma(x)}{dx^3} = \frac{d}{dx}\left[\frac{d^2}{dx^2}\sigma(x)\right] = \frac{d}{dx}\left[\sigma(x)\cdot(1-\sigma(x))\cdot(1-2\sigma(x))\right]$

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

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

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

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

$$3\cdot 2\cdot \sigma(x)\cdot\sigma(x)\cdot(1-\sigma(x)) +$$

$$2\cdot 3\cdot(\sigma(x))^2\cdot\sigma(x)\cdot(1-\sigma(x))$$

$$= \sigma(x)\cdot(1-\sigma(x))\cdot\left[1 - 6\sigma(x) + 6(\sigma(x))^2\right]$$

(b) We know $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$ and $\tanh\left(\frac{x}{2}\right) = \frac{e^{x/2}-e^{-x/2}}{e^{x/2}+e^{-x/2}} = \frac{e^x-1}{e^x+1}$ $-(*)$

By $(*)$, we have $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = \frac{e^x-1}{e^x+1} + \frac{1}{e^x+1} = \tanh\left(\frac{x}{2}\right) + \frac{1}{e^x+1}. \square$

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.

3、課堂中有提及3種 Gradient Descent 方法，但沒有詳細說明這3種方法在應用上的優劣，或者是在不同的需求中，是否有方法判斷何種方法最為適合。

雖然課程最初教授就有提到，這堂課會專注在理論上，但我還是希望能稍微有關於應用的說明，或是像此次 assignment 的第1題，實際自己計算推導一次也讓我對於理論有更深刻地理解。