assigment 1. (week 1)

1. Given one single data point $(x_1, x_2, y) = (1,2,3)$,
and assume that the current parameter is $\theta^0 = (b, w_1, w_2) = (4,5,6)$.

Consider SGD methods,

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

Then, given $h_{(4,5,6)}(1,2) = \sigma(4 + 5(1) + 6(2)) = \sigma(21)$ — (1)

By the gradient descent algorithm, since $\theta^0 = (4,5,6)$

we have $\theta^1 = \theta^0 - \alpha \nabla_\theta L(\theta^0; (1,2))$ where α the learning rate.

and $L(\theta) = (y - h_\theta(x_1, x_2))^2$.

Then $\nabla_\theta L = -2(y - h_\theta(x_1, x_2))\, h_\theta(x_1, x_2)(1 - h_\theta(x_1, x_2)) \cdot \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$.

Therefore, $\theta^1 = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + 2 \cdot \alpha \cdot \left(3 - h_{\theta^0}(x_1, x_2)\right) \cdot h_{\theta^0}(x_1, x_2) \cdot \left(1 - h_{\theta^0}(x_1, x_2)\right) \cdot \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

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

Given $\theta^1 = (b^*, w_1^*, w_2^*)$

we get $\begin{pmatrix} b^* \\ w_1^* \\ w_2^* \end{pmatrix} = \begin{pmatrix} 4 + 2\alpha(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \\ 4 + 2\alpha(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \\ 4 + 4\sigma(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \end{pmatrix}$

2. (a)

For $k=1$,

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

For $k=2$,

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

For $k=3$,

we have
$$\frac{d^3\sigma(x)}{dx^3} = \frac{d}{dx}\left(\sigma(x)(1-\sigma(x))(1-2\sigma(x))\right)$$
$$= \frac{d}{dx}\left\{\sigma(x)\left[1-3\sigma(x)+2\sigma^2(x)\right]\right\}$$
$$= \frac{d\sigma(x)}{dx} - \frac{d}{dx}(3\sigma^2(x)) + \frac{d}{dx}(2\sigma^3(x))$$
$$= \sigma(x)(1-\sigma(x)) - 6\sigma(x)\sigma(x)(1-\sigma(x)) + 6\sigma^2(x)\sigma(x)(1-\sigma(x))$$
$$= \sigma(x)(1-\sigma(x)) + 6\sigma^2(x)(1-\sigma(x))(\sigma(x)-1)$$
$$= \sigma(x)(1-\sigma(x))(1+6\cdot\sigma(x)\cdot(\sigma(x)-1))$$

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

(b) Discussion with roommate

Since $\sigma(x) = \dfrac{1}{1+e^{-x}}$,

we can rewrite that $\sigma(x) = \dfrac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} \cdot \dfrac{e^{-\frac{x}{2}}}{e^{-\frac{x}{2}}}$.

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

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

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

Hence $\tanh\left(\frac{x}{2}\right) = 2\sigma(x) - 1$.

3.

In the begining of this cource, for me, compared to the math problems, what confuses me more is that I seill cannot fully connect these theories and applications to how they actually work.