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.

1.

Define $Loss = MSE = (h(x_1,x_2) - y)^2$

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

$\because (x_1, x_2, y) = (1,2,3)$

$(b, w_1, w_2) = (4,5,6)$

$\therefore h = \sigma(4 + 5\cdot 1 + 6 \cdot 2) = \sigma(21)$

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

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

$b' = b - \alpha\left(\frac{dL}{db}\right)$

$= b - \alpha\left[2(\sigma(21) - y)\cdot \sigma'(21)\right]$

$= 4 - \alpha \cdot 2(\sigma(21) - 3)\sigma(21)(1-\sigma(21))$

where α is learning rate

$w_1' = w_1 - \alpha\left(\frac{dL}{dw_1}\right)$

$= 5 - \alpha\left[2(\sigma(21) - y)\sigma(21)(1-\sigma(21))\cdot 1\right.$

where α is learning rate

$w_2' = w_2 - \alpha\left(\frac{dL}{dw_2}\right)$

$= w_2 - \alpha \cdot 2(\sigma(21) - y)\sigma(21)(1-\sigma(21))\cdot x_2$

$= 6 - 4\alpha(\sigma(21) - 3)\sigma(21)(1-\sigma(21))$

where α is learning rate

$\therefore \theta' = (b', w_1', w_2')$

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.

2.(a)

$K=1$

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

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

$K=2$

$$\frac{d^2\sigma(x)}{dX^2} = \frac{d\,\sigma(x)-\sigma^2(x)}{dX} = \sigma'(X) - 2(\sigma(X))\sigma'(X)$$

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

$$= \underline{[\sigma(X)\,(1-\sigma(X))](1-2\,\sigma(X))} \quad \#$$

$K=3$

$$\frac{d^3\sigma(x)}{dX^3} = \frac{d\,[\sigma(x)(1-\sigma(x))]\,(1-2\sigma(x))}{dX}$$

$$= \frac{d\;\sigma'(X)\,(1-2\sigma(X))}{dX}$$

$$= \sigma''(X) - 2[\sigma''(x)\sigma(X) + (\sigma'(x))^2]$$

$$= [\sigma(X)(1-\sigma(X))](1-2\sigma(X)) - 2[\sigma^2(X)(1-\sigma(X))](1-2\sigma(X)) + 2\,(1-\sigma(X))^2\sigma^2(X)$$

$$= \underline{\sigma(X)\,(1-\sigma(X))\,(1-6\sigma(x)+6\sigma^2(X))} \quad \#$$

2.(b)

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

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

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

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

$$\underline{=2\sigma(2x)-1}\ \#$$

$\because \sigma(x)=\frac{1}{2}\tanh(\frac{x}{2})+\frac{1}{2}$

$$\therefore\ \sigma(x)=\frac{\left(\frac{\sinh(\frac{x}{2})}{\cosh(\frac{x}{2})}+1\right)}{2}$$

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

$$\cosh(x)=\underline{\frac{e^x}{2\sigma(2x)}}\ \#$$

$$\sinh(x)=\frac{e^x}{2\sigma(2x)}\cdot(2\sigma(2x)-1)=\underline{\frac{e^x\sigma(2x)-1}{\sigma(2x)}}\ \#$$

$$\coth(x)=\underline{\frac{1}{2\sigma(2x)-1}}\ \#$$

$$\operatorname{sech}(x)=\underline{\frac{2\sigma(2x)}{e^x}}\ \#$$

$$\operatorname{csch}(x)=\underline{\frac{\sigma(2x)}{e^x\sigma(2x)-1}}\ \#$$

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.

梯度下降可能卡在 local minima 而非 Global minima

在高維度時，saddle point 和 Plateau 導致學習緩慢

如何找到適合的 Learning rate？太大可能 Divergence 太小又

收斂慢或卡在 Plateaus

可以使用 Scheduling 或 Optimizers