Machine Learning

313704071 陳安定 Assignment 1

1.

Since SGD, m=1

$$\theta^{n+1} = \theta^n - \alpha\nabla_\theta Loss = \theta^n + 2\alpha\left[\frac{1}{m}\sum_{i=1}^{m}\left(y^i - h\left(x_1^i, x_2^i\right)\right)\nabla_\theta h\right]$$

$$\begin{aligned}\theta^1 &= \theta^0 + 2\alpha[(3 - h(1,2))\nabla_\theta h] \\ &= (4,5,6) + 2\alpha[(3 - \sigma(b + w_1 + 2w_2))\nabla_\theta \sigma(b + w_1 + 2w_2)]\end{aligned}$$

2(a).

$$\sigma(x) = sigmoid\ function$$

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

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

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

$$\begin{aligned}s'' &= \left(s(1-s)\right)' = s'(1-s) + s(-s') \\ &= s'(1-s-s) = s'(1-2s) \\ &= s(1-s)(1-2s)\end{aligned}$$

$$\begin{aligned}s''' &= \left(s'(1-2s)\right)' = s''(1-2s) + s'(-2s') \\ &= s'(1-2s)^2 - 2s'(s') \\ &= s'(1-4s+4s^2-2s') \\ &= s(1-s)(1-6s+6s^2)\end{aligned}$$

2(b).

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

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

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

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

3.

The sigmoid saturates for large-magnitude inputs: if $x \gg 0$, $\sigma(x) = \frac{1}{1+e^{-x}} \approx 1$; if $x \ll 0$, $\sigma(x) \approx 0$. Since $\sigma'(x) = \sigma(x)(1-\sigma(x))$, the previous situation will make $\sigma'(x)$ become near zero. During backpropogation, $\delta^{[i]} = \sigma'\left(z^{[i]}\right) \circ \left(W^{[i+1]}\right)^T \delta^{[i+1]}$

Hence saturation makes $\delta^{[i]}$ tiny, leading to vanishing gradients and slow or stalled learning in lower layers.