Homework week 1

Problem 1 Note that

$$\ell(\theta) = \frac{1}{2}\left(y - h(x_1, x_2, \theta)\right)^2, \quad h(x_1, x_2, \theta) = \sigma(b + w_1 x_2 + w_2 x_2), \quad \theta = (b, w_1, w_2)^T, \tag{0.1}$$

and the stochastic gradient descent method:

$$\theta^{n+1} = \theta^n - \alpha \nabla \ell(\theta), \quad n \geq 0. \tag{0.2}$$

Since

$$\frac{\partial}{\partial b} h = \sigma', \quad \frac{\partial}{\partial w_1} h = \sigma' x_1, \quad \frac{\partial}{\partial w_2} h = \sigma' x_2, \tag{0.3}$$

it follows that

$$\theta^1 = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - \alpha \begin{pmatrix} \partial_1 \ell(\theta^0) \\ \partial_2 \ell(\theta^0) \\ \partial_3 \ell(\theta^0) \end{pmatrix}_{\theta^0 = (4,5,6)^T} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \alpha(y - \sigma(21)) \begin{pmatrix} \sigma(21)(1-\sigma(21)) \\ \sigma(21)(1-\sigma(21)) \\ 2\sigma(21)(1-\sigma(21)) \end{pmatrix}, \tag{0.4}$$

where sigmoid function satisfies $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

Problem 2

In the case of (a)**.** Recall that

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \qquad \sigma'(x) = \sigma(x)(1 - \sigma(x)). \tag{0.5}$$

For the second derivative of $\sigma(x)$, we have

$$\begin{aligned} \sigma''(x) &= \sigma'(x)(1 - \sigma(x)) - \sigma(x)\sigma'(x) \\ &= \sigma(x)(1 - \sigma(x))(1 - 2\sigma(x)). \end{aligned} \tag{0.6}$$

Next, we find

$$\begin{aligned} \sigma'''(x) &= \sigma''(x)(1 - 2\sigma(x)) - 2(\sigma'(x))^2 \\ &= \sigma(x)(1 - \sigma(x))(1 - 6\sigma(x) + 6\sigma^2(x)) \end{aligned} \tag{0.7}$$

In the case of (b)**.** Note that

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \qquad 1 - \sigma(x) = \frac{e^{-x}}{1 + e^{-x}} \tag{0.8}$$

Then

$$\tanh\left(\frac{x}{2}\right) = \frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} - \frac{e^{-\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} = \sigma(x) - (1-\sigma(x)), \tag{0.9}$$

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

Problem 3

Question 1 The issue of convergence in the stochastic gradient descent method.

Question 2 For practical problems, how should one determine the number of hidden layers and the number of neurons in each layer?