Machine Learning HW1

Student ID: 111652015

Problem 1

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

$$\boldsymbol{\theta}^{(0)} = (b, w_1, w_2) = (4, 5, 6),$$

evaluate $\boldsymbol{\theta}^{(1)}$.

Just write the expression and substitute the numbers; no need to simplify or evaluate.

By the question, we have

$$h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2), \qquad (x_1, x_2, y) = (1, 2, 3),$$
$$\boldsymbol{\theta}^{(0)} = (b, w_1, w_2) = (4, 5, 6).$$

Loss function We define the loss function as the *Mean Squared Error (MSE)*:

$$L \;=\; \frac{1}{2}\big(y - h(x_1, x_2)\big)^2.$$

SGD Using SGD , the parameters are updated by:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \alpha \nabla_{\boldsymbol{\theta}} L.$$

First derivative of sigmoid function In part (a) of Problem 2, I proved the following formula:

$$\frac{d}{dx}\sigma(x) = \sigma(x)\big(1 - \sigma(x)\big).$$

Compute the partial derivatives

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial b} \\
&= -\big(y - h(x_1, x_2)\big)\, \sigma'(b + w_1 x_1 + w_2 x_2) \frac{\partial}{\partial b}(b + w_1 x_1 + w_2 x_2) \\
&= -\big(y - h(x_1, x_2)\big)\, \sigma(b + w_1 x_1 + w_2 x_2)\, \big[1 - \sigma(b + w_1 x_1 + w_2 x_2)\big] \cdot 1 \\
&= -\big(y - h(x_1, x_2)\big)\, h(x_1, x_2)\, \big[1 - h(x_1, x_2)\big].
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial L}{\partial w_1} &= -\big(y - h(x_1,x_2)\big)\,\frac{\partial}{\partial w_1}\,\sigma(b + w_1x_1 + w_2x_2) \\
&= -\big(y - h(x_1,x_2)\big)\,\sigma'(b + w_1x_1 + w_2x_2)\,\frac{\partial}{\partial w_1}(b + w_1x_1 + w_2x_2) \\
&= -\big(y - h(x_1,x_2)\big)\,\sigma'(b + w_1x_1 + w_2x_2)\,x_1 = x_1\,\frac{\partial L}{\partial b}, \\
\frac{\partial L}{\partial w_2} &= -\big(y - h(x_1,x_2)\big)\,\frac{\partial}{\partial w_2}\,\sigma(b + w_1x_1 + w_2x_2) \\
&= -\big(y - h(x_1,x_2)\big)\,\sigma'(b + w_1x_1 + w_2x_2)\,\frac{\partial}{\partial w_2}(b + w_1x_1 + w_2x_2) \\
&= -\big(y - h(x_1,x_2)\big)\,\sigma'(b + w_1x_1 + w_2x_2)\,x_2 = x_2\,\frac{\partial L}{\partial b}.
\end{aligned}
$$

Substitution Substituting the given data and parameters:

$$(x_1, x_2, y) = (1, 2, 3), \qquad (b, w_1, w_2) = (4, 5, 6),$$

we have

$$b + w_1x_1 + w_2x_2 = 4 + 5(1) + 6(2) = 21, \qquad h(x_1, x_2) = \sigma(21).$$

Based on the stochastic gradient descent, the parameters are updated as

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - \alpha\nabla_{\boldsymbol{\theta}}L = (\theta_b^1,\, \theta_{w_1}^1,\, \theta_{w_2}^1).$$

and

$$
\begin{cases}
\theta_b^1 = 4 - \alpha\Big[-(3 - \sigma(21))\,\sigma(21)\,[1 - \sigma(21)]\Big], \\
\theta_{w_1}^1 = 5 - \alpha\Big[-(3 - \sigma(21))\,\sigma(21)\,[1 - \sigma(21)]\cdot 1\Big], \\
\theta_{w_2}^1 = 6 - \alpha\Big[-(3 - \sigma(21))\,\sigma(21)\,[1 - \sigma(21)]\cdot 2\Big].
\end{cases}
$$

Problem 2

(a)Find the expression of

$$\frac{d^k}{dx^k}\sigma(x)$$

in terms of $\sigma(x)$ for $k = 1, \ldots, 3$, where σ is the sigmoid function.
(b)Find the relation between sigmoid function and hyperbolic function.

(a) Derivatives of the sigmoid function

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

Then its first derivative is

$$
\begin{aligned}
\frac{d}{dx}\,\sigma(x) &= \frac{d\,\sigma(x)}{d(1+e^{-x})} \;\cdot\; \frac{d(1+e^{-x})}{d\,e^{-x}} \;\cdot\; \frac{d\,e^{-x}}{dx} \\
&= -\,\frac{1}{(1+e^{-x})^2} \;\cdot\; 1 \;\cdot\; (-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}+1}{(1+e^{-x})^2} \;-\; \frac{1}{(1+e^{-x})^2} \\
&= \sigma(x) \;-\; \sigma^2(x) \\
&= \sigma(x)\big(1-\sigma(x)\big)
\end{aligned}
$$

For the second derivative,

$$
\begin{aligned}
\frac{d^2}{dx^2}\sigma(x) &= \frac{d}{dx}\big(\sigma(x)-\sigma^2(x)\big) \\
&= \sigma(x)-\sigma^2(x) \;-\; 2\sigma(x)\big(\sigma(x)-\sigma^2(x)\big) \\
&= 2\sigma^3(x) - 3\sigma^2(x) + \sigma(x).
\end{aligned}
$$

For the third derivative,

$$
\begin{aligned}
\frac{d^3}{dx^3}\sigma(x) &= \frac{d}{dx}\big(2\sigma^3(x) - 3\sigma^2(x) + \sigma(x)\big) \\
&= 6\sigma^2(x)\cdot\sigma'(x) - 6\sigma(x)\cdot\sigma'(x) + \sigma'(x) \\
&= \sigma(x)\big[1-\sigma(x)\big]\big[6\sigma^2(x) - 6\sigma(x) + 1\big].
\end{aligned}
$$

(b) Relation between the sigmoid and hyperbolic tangent. We know

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

Also,

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

Hence,

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

Problem 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.

---

1. Why can a small number of neurons approximate so well?
2. Can we estimate how many neurons are needed?