

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.

By Stochastic Gradient Descent, Gradient Descent Rule, and hypothesis, we have

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

$$\theta^1 = \theta^{0+1} = \theta^0 - \alpha \nabla_{\theta} L(\theta; x_1, x_2, y) \text{ with } L(\theta; x_1, x_2, y) = (y - h_{\theta}(x_1, x_2))^2 \text{ and } \alpha > 0 \text{ called learning rate. - (*)}$$

$$\text{Let } z = b + w_1 x_1 + w_2 x_2, \text{ then } \nabla_{\theta} L = \frac{dL}{d\theta} = \frac{dL}{dz} \frac{dz}{dw_1} \frac{dw_1}{d\theta} \\ = -2(y - h_{\theta}(x_1, x_2)) \cdot \sigma'(z) \cdot \nabla_{\theta} z - (**)$$

$$\text{We have } \sigma(z) = \frac{1}{1+e^{-z}}, \sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2}, \text{ and } 1 - \sigma(z) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}.$$

$$\text{Thus } \sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z) \cdot (1 - \sigma(z)) - (***)$$

By (*), (**), and (***), we have

$$\theta^1 = \theta^0 - \alpha \cdot [-2(y - h_{\theta}(x_1, x_2)) \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}]$$

Substitute $\theta^0 = (4, 5, 6)$ and $(x_1, x_2) = (1, 2)$ into the update formula, we have

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

$$\text{Thus } \begin{pmatrix} b^1 \\ w_1^1 \\ w_2^1 \end{pmatrix} = \begin{pmatrix} 4 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 1 \\ 5 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 1 \\ 6 + 2 \cdot \alpha \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot 2 \end{pmatrix}. \quad \square$$

2. (a) Find the expression of $\frac{d^k}{dx^k} \sigma$ in terms of $\sigma(x)$ for $k = 1, \dots, 3$ where σ is the sigmoid function.

(b) Find the relation between sigmoid function and hyperbolic function.

$$\begin{aligned} \text{2. (a) For } k=1, \quad \frac{d\sigma(x)}{dx} &= \frac{d}{dx} \left(\frac{1}{1+e^x} \right) = \frac{d}{dx} [(1+e^x)^{-1}] = \frac{-(-e^x)}{(1+e^x)^2} = \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}} = \sigma(x) \cdot (1-\sigma(x)). \quad \square \end{aligned}$$

$$\begin{aligned} \text{For } k=2, \quad \frac{d^2\sigma(x)}{dx^2} &= \frac{d}{dx} \left(\frac{d\sigma(x)}{dx} \right) = \frac{d}{dx} [\sigma(x) \cdot (1-\sigma(x))] = \frac{d}{dx} [\sigma(x) - (\sigma(x))^2] \\ &= \frac{d}{dx} \sigma(x) - \frac{d}{dx} (\sigma(x))^2 \\ &= \sigma(x) \cdot (1-\sigma(x)) - 2\sigma(x) \cdot \frac{d}{dx} \sigma(x) \\ &= \sigma(x) \cdot (1-\sigma(x)) - 2\sigma(x) \cdot \sigma(x) \cdot (1-\sigma(x)) \\ &= \sigma(x) \cdot (1-\sigma(x)) \cdot (1-2\sigma(x)). \quad \square \end{aligned}$$

$$\begin{aligned} \text{For } k=3, \quad \frac{d^3\sigma(x)}{dx^3} &= \frac{d}{dx} \left[\frac{d^2}{dx^2} \sigma(x) \right] = \frac{d}{dx} [\sigma(x) \cdot (1-\sigma(x)) \cdot (1-2\sigma(x))] \\ &= \frac{d}{dx} [\sigma(x) - 2(\sigma(x))^2 - (\sigma(x))^2 + 2(\sigma(x))^3] \\ &= \frac{d}{dx} [\sigma(x) - 3(\sigma(x))^2 + 2(\sigma(x))^3] \\ &= \frac{d}{dx} \sigma(x) - 3 \frac{d}{dx} (\sigma(x))^2 + 2 \frac{d}{dx} (\sigma(x))^3 \\ &= \sigma(x) \cdot (1-\sigma(x)) - \\ &\quad 3 \cdot 2 \cdot \sigma(x) \cdot \sigma(x) \cdot (1-\sigma(x)) + \\ &\quad 2 \cdot 3 \cdot (\sigma(x))^2 \cdot \sigma(x) \cdot (1-\sigma(x)) \\ &= \sigma(x) \cdot (1-\sigma(x)) \cdot [1-6\sigma(x) + 6(\sigma(x))^2] \end{aligned}$$

$$(b) \text{ We know } \sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \text{ and } \tanh\left(\frac{x}{2}\right) = \frac{e^{x/2} - e^{-x/2}}{e^{x/2} + e^{-x/2}} = \frac{e^x - 1}{e^x + 1} \quad (*)$$

$$\text{By } (*), \text{ we have } \sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = \frac{e^x-1}{e^x+1} + \frac{1}{e^x+1} = \tanh\left(\frac{x}{2}\right) + \frac{1}{e^x+1}. \quad \square$$

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.

3. Although the lecture introduced the three main types of gradient descent methods, it didn't go into detail about how they differ in terms of computational efficiency or accuracy, or how to choose.

I know that the lecture mainly focuses on theory, I'd still like to learn about the theory's performance, such as when it is more appropriate to use Batch, SGD or Mini-Batch in different situations. ↪