Assignment 1.

1. Consider stochastic gradient descent method to learn the house price model

$$h(x_1, x_2) = \sigma(b + w_1x_1 + w_2x_2),$$

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.

By SGD, $\theta' = \theta^0 - \alpha \nabla_\theta \text{Loss}$, $\alpha > 0$.

Using MSE, $\text{Loss} = \frac{1}{2}(h-y)^2 = L$.

$h(x_1, x_2) = \sigma(z)$

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial \theta} = (h-y)\sigma'(z)\frac{\partial z}{\partial \theta} = (h-y)\sigma(z)(1-\sigma(z))\frac{\partial z}{\partial \theta}$$

$$\Rightarrow \frac{\partial L}{\partial b} = (h-y)\sigma(z)(1-\sigma(z))\cdot 1$$

$$\frac{\partial L}{\partial w_1} = (h-y)\sigma(z)(1-\sigma(z))\cdot x_1$$

$$\frac{\partial L}{\partial w_2} = (h-y)\sigma(z)(1-\sigma(z))\cdot x_2$$

$$\Rightarrow b' = b - \alpha(h-y)\sigma(z)(1-\sigma(z))$$

$$w_1' = w_1 - \alpha(h-y)\sigma(z)(1-\sigma(z))\, x_1$$

$$w_2' = w_2 - \alpha(h-y)\sigma(x)(1-\sigma(x))\, x_2$$

$$z = b + w_1 x_1 + w_2 x_2 = 4 + 5 \cdot 1 + 6 \cdot 2 = 21$$

$$\Rightarrow h(x_1, x_2) = \sigma(z) = \sigma(21)$$

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

$$w_1' = 5 - \alpha(\sigma(21) - 3)\sigma(21)(1 - \sigma(21))$$

$$w_2^2 = 6 - 2\alpha(\sigma(21) - 3)\sigma(21)(1 - \sigma(21))$$

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.

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

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

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

$$= \sigma(x)(1-\sigma(x))(1-\sigma(x)) - \sigma(x)\sigma(x)(1-\sigma(x))$$

$$= \sigma(x)(1-\sigma(x))(1-2\sigma(x)).$$

$$\frac{d^3\sigma}{dx^3} = \sigma'(x)(1-\sigma(x))(1-2\sigma(x)) - \sigma(x)\sigma'(x)(1-2\sigma(x))$$
$$-2\sigma(x)(1-\sigma(x))\sigma'(x).$$

$$= \sigma(x)(1-\sigma(x))^2(1-2\sigma(x)) - \sigma(x)^2(1-\sigma(x))(1-2\sigma(x))$$
$$-2\sigma(x)^2(1-\sigma(x))^2$$

$$= \sigma(x)(1-\sigma(x))\left[(1-\sigma(x))(1-2\sigma(x)) - \sigma(x)(1-2\sigma(x))\right.$$
$$\left.-2\sigma(x)(1-\sigma(x))\right]$$

$$= \sigma(x)(1-\sigma(x))(1-6\sigma(x)+6\sigma^2(x))$$

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

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

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

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.

Convergence of SGD:

SGD: $\theta^{t+1} = \theta^t - \alpha^t g^t$, where $g^t = \nabla_\theta L(x^t)$.

Assume $\mathbb{E}[g^t | \theta^t] = \nabla f(\theta^t)$ (unbiasedness).

Let θ^* be the optimal solution, i.e. $\theta^* = \arg\min_\theta f(\theta)$.

Then $\| \theta^{t+1} - \theta^* \|^2 = \| \theta^t - \alpha^t g^t - \theta^* \|^2$

$$= \| \theta^t - \theta^* \|^2 - 2\alpha^t g^{tT} (\theta^t - \theta^*) + \alpha^{t2} \| g^t \|^2$$

$\mathbb{E}[\| \theta^{t+1} - \theta^* \|^2] \le \mathbb{E}[\| \theta^t - \theta^* \|^2] - 2\alpha^t \mathbb{E}[f(\theta^t) - f(\theta^*)] + \alpha^{t^2} G^2$,

here we assume $\| g^t \|^2 \le G^2$.

Then $\sum_{t=1}^{N} \alpha^t \mathbb{E}[f(\theta^t) - f(\theta^*)] \le \frac{1}{2} \| \theta^1 - \theta^* \|^2 + \frac{1}{2} G^2 \sum_{t=1}^{N} \alpha^{t2}$

Define the weighted average iterate $\bar{\theta}_N = \frac{\sum_{t=1}^{N} \alpha^t \theta^t}{\sum_{t=1}^{N} \alpha^t}$.

We can bound the error:

$$\mathbb{E}[f(\bar{\theta}_N)] - f(\theta^*) \le \frac{\| \theta^1 - \theta^* \|^2 + G^2 \sum_{t=1}^{N} \alpha^{t2}}{2 \sum_{t=1}^{N} \alpha^t}$$

- If $\alpha^t = O(\frac{1}{\sqrt{t}})$, the error $= O(\frac{1}{\sqrt{N}})$
- If f is strongly convex and $\alpha^t = O(\frac{1}{t})$, the error $= O(\frac{1}{N})$