ML - AS 1 111652042 顏友君

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.

The Loss function of SGD (m = 1) is

$$Loss(\theta) = (y - h(x_1, x_2))^2 = (y - \sigma(b + \omega_1 x_1 + \omega_2 x_2))^2, \text{ which } \theta = (b, \omega_1, \omega_2)$$

And, the derivative of sigmoid function is

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

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

So, the gradient of $Loss(\theta)$ with respect to θ is

$$\nabla_\theta Loss = \begin{bmatrix} \frac{\partial Loss}{\partial b} \\ \frac{\partial Loss}{\partial \omega_1} \\ \frac{\partial Loss}{\partial \omega_2} \end{bmatrix} = \begin{bmatrix} -2(y - \sigma(b + \omega_1 x_1 + \omega_2 x_2)) \cdot \sigma'(b + \omega_1 x_1 + \omega_2 x_2) \\ -2(y - \sigma(b + \omega_1 x_1 + \omega_2 x_2)) \cdot \sigma'(b + \omega_1 x_1 + \omega_2 x_2) \cdot x_1 \\ -2(y - \sigma(b + \omega_1 x_1 + \omega_2 x_2)) \cdot \sigma'(b + \omega_1 x_1 + \omega_2 x_2) \cdot x_2 \end{bmatrix}$$

substitute $(x_1, x_2, y) = (1, 2, 3)$ and $\theta^0 = (b, \omega_1, \omega_2) = (4, 5, 6)$

$$= \begin{bmatrix} -2 \cdot (3 - \sigma(21)) \cdot \sigma'(21) \\ -2 \cdot (3 - \sigma(21)) \cdot \sigma'(21) \cdot 1 \\ -2 \cdot (3 - \sigma(21)) \cdot \sigma'(21) \cdot 2 \end{bmatrix} = \begin{bmatrix} -2 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \\ -2 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \\ -4 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \end{bmatrix}$$

By Gradient descent algorithm,

$$\theta' = \theta^0 - \alpha \nabla_\theta Loss, \text{ where } \alpha > 0 \text{ is the learning rate}$$

$$= \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \alpha \begin{bmatrix} -2 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \\ -2 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \\ -4 \cdot (3 - \sigma(21)) \cdot \sigma(21) \cdot (1 - \sigma(21)) \end{bmatrix}$$

$$= \begin{bmatrix} 4 + 2\alpha \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot (3 - \sigma(21)) \\ 5 + 2\alpha \cdot \sigma(21) \cdot (1 - \sigma(21)) \cdot (3 - \sigma(21)) \\ 6 + 4\alpha \cdot \sigma(21)(1 - \sigma(21)) \cdot (3 - \sigma(21)) \end{bmatrix}$$

#

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) The sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$

For $k=1$, $\frac{d}{dx}\sigma = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = -(1+e^{-x})^{-2} \cdot e^{-x} \cdot (-1)$

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

For $k=2$, $\frac{d^2}{dx^2}\sigma = \frac{d}{dx}\left(\frac{d}{dx}\sigma\right) = \frac{d}{dx}\left(\sigma(x)(1-\sigma(x))\right) = \sigma'(x)(1-\sigma(x)) + \sigma(x)\cdot(-\sigma'(x))$

$$= \sigma'(x)\left((1-\sigma(x)) - \sigma(x)\right) = \sigma(x)(1-\sigma(x))(1-2\sigma(x))$$

For $k=3$, $\frac{d^3}{dx^3}\sigma = \frac{d}{dx}\left(\frac{d^2}{dx^2}\sigma\right) = \frac{d}{dx}\left(\sigma(x)(1-\sigma(x))(1-2\sigma(x))\right)$

$$= \sigma'(x)(1-\sigma(x))(1-2\sigma(x)) + \sigma(x)(-\sigma'(x))(1-2\sigma(x)) + \sigma(x)(1-\sigma(x))(-2\sigma'(x))$$

$$= \sigma'(x)\left[1-3\sigma(x)+2(\sigma(x))^2 - \sigma(x) + 2(\sigma(x))^2 - 2\sigma(x) + 2(\sigma(x))^2\right]$$

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

(b) Observe the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ and hyperbolic function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1}$.

both are defined in terms of e^x, $\sigma : \mathbb{R} \to (0,1)$ and $\tanh : \mathbb{R} \to (-1,1)$.

So, the sigmoid function can be obtained by scaling and shifting the hyperbolic function as: $\sigma(x) = \frac{1+\tanh\left(\frac{x}{2}\right)}{2}$

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.

我覺得在 gradient descent algorithm 中，如果使用固定的 learning rate 可能會導致：

1. 如果 learning rate 設定太小了，收斂的速度就會很慢；

2. 如果 learning rate 設定太大了，當接近最低點時就有可能會跑過頭，導致 θ 在最低點旁振盪，就到不了最佳解。

所以為了避免這些情況，可以剛開始先設定一個較大的 learning rate，再設定一個像是 0.99 的參數，每跑完一次就讓 learning rate × 0.99，就可以同時避免 learning rate 太大或太小的缺點。