314652021 羅瑋翔

Assignment 1

Written assignment

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.

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.

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) Recall : gradient decent algorithm.

$\theta^{n+1} = \theta^n - \alpha \nabla \text{Loss}$, α is learning rate and $\alpha > 0$.

For MSE loss

$$\Rightarrow \theta^1 = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + 2\alpha\left(3 - h_{(4,5,6)}(1,2)\right) \cdot h_{(4,5,6)}(1,2) \cdot \left(1 - h_{(4,5,6)}(1,2)\right) \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}. \#$$

(2) - a.

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

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

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

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

(ii) $\sigma''(x) = \sigma'(x) \cdot (1-\sigma(x)) + \sigma(x)(1-\sigma(x))'$

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

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

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

$= \sigma(x) \cdot (1-\sigma(x)) \cdot (1-2\sigma(x))$ #

(iii) $\sigma^{(3)}(x) = \sigma'(x)(1-\sigma(x))(1-2\sigma(x))$

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

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

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

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

$+ \sigma(x) \cdot (1-\sigma(x)) \cdot (-2) \cdot \sigma(x)(1-\sigma(x))$

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

(2) - b.

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

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

(3)

Q: Assume $h(x_1, x_2) = b + w_1 x_1 + w_2 x_2 = [1, x_1, x_2]\begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$.

$$\underbrace{\begin{bmatrix} y^1 \\ \vdots \\ y^N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_1^1 & x_2^1 \\ & \vdots & \\ 1 & x_1^N & x_2^N \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}}_{\theta}.$$

$$\min \|Y - X\theta\|_2^2 \Rightarrow \theta^* = (X^T X)^{-1} X^T Y \ ?$$

A: Goal: 找到 1個 向量 $X\theta - Y$ 垂直於 $col(X)$ 的平面上.

$$\therefore \quad \min \|Y - X\theta\|_2^2$$

$$\Rightarrow Y - X\theta \in col(X)^{\perp} = N(X^T).$$

$$\Rightarrow X^T(X\theta - Y) = 0.$$

Suppose X is full rank $\Rightarrow rank(X) = N$

$$\Rightarrow rank(X^T X) = rank(X) \Rightarrow rank(X^T X) = N.$$

故 X^TX is full rank, $\det(X^TX) \neq 0$,

$\Rightarrow$ X^TX is invertible.

Then, $\theta = (X^TX)^{-1}X^TY$ #

Q：上課中提到選取 learning rate 時，
有說可以在每一步選取一大一小的 α.
然後可以找到 $\underset{\theta}{\operatorname{argmin}}\ Loss(\theta)$, why?

A：應該是想要保有 ① 加速逼近最佳點
和 ② 保持穩定，避免跳過最佳點.

問 chatGPT

$\Rightarrow$ 使用這種方法在取大的 learning rate 時，
要讓它在穩定範圍內，
讓小的 learning rate 將它拉回來.