

hw1 11270502

- sigmoid- $\beta$ 皆導數

$$b'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \quad b'(x) = b(x)(1-b(x)) \quad L(\theta) = \frac{1}{2}(h_{\theta}(x) - y)^2$$

- 鏈連式法則拆解

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial b}$$

- 積次正梯度結果

$$\frac{\partial L}{\partial b} = (h-y) b'(z)$$

$$\frac{\partial L}{\partial w_1} = (h-y) b'(z) x_1$$

$$\frac{\partial L}{\partial w_2} = (h-y) b'(z) x_2$$

- SGD  $\rightarrow$  sample平均誤差

- 更新式

$$b' = b^0 - \eta(h-y) b'(z)$$

$$w_1' = w_1^0 - \eta(h-y) b'(z) x_1$$

$$w_2' = w_2^0 - \eta(h-y) b'(z) x_2$$

- evaluate  $\theta'$

$$z = 4 + 5 \cdot 1 + 6 \cdot 2 = 21 \quad h = b(21) \quad \text{learning rate}$$

$$b' = b^0 - \eta(b(21) - 3) b(21) (1 - b(21))$$

$$w_1' = w_1^0 - \eta(b(21) - 3) b(21) (1 - b(21)) \cdot 1$$

$$w_2' = w_2^0 - \eta(b(21) - 3) b(21) (1 - b(21)) \cdot 2$$

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

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

$$b''(x) = b'(x)(1-b(x)) - b(x)b'(x) = b'(x)(1-2b(x))$$

$$= b(x)(1-b(x))(1-2b(x)) \#$$

$$b'''(x) = b''(x)(1-2b(x)) + b'(x)(-2b'(x))$$

$$= b''(x)(1-2b(x)) - 2b'(x)^2$$

$$= b'(x)(1-2b(x))^2 - 2b'(x)^2$$

$$= b(x)(1-b(x))(1-2b(x)) \geq 2[b(x)(1-b(x))]^2$$

$$b(x)(1-b(x))(b(x)-b(x)+1) \#$$

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

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

$$\tanh(x) = 2b(2x) - 1 \#$$

31 為什麼 SGD 會收斂，是否收斂到同一個地方？

- SGD 每次更新都往負梯度方向走，在 learning rate 適當及 loss function 光滑情況下，遞歸會 keep 下降收斂。
- 若 loss function 是 convex，SGD 一定會收斂到唯一全局或局部極值  
若非 convex，則可能收斂到不同局部極值點或鞍點。

• 反正申反思

1) learning rate 太大  $\rightarrow$  震盪，太小  $\rightarrow$  收斂太慢，理論上 SGD 會收斂，  
但在實務上不能「有效收斂」depends on learning rate 要參看周來文。

2) SGD 的隨機性有時能幫助跳出局部極值，但非在 non-convex  
的 loss function 中，若用 batch GD 可能被卡在局部極值，但 SGD 更新方向  
有隨機性，反而有機會跳出找到更好的解。