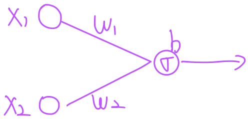
ML 第一週筆記整理

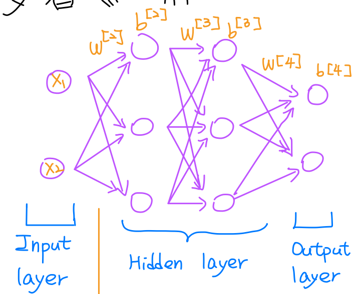
Affine Linear Transformation (仿射線性變換)



$$\sigma(\underbrace{b + w_1x_1 + w_2x_2}) = \sigma(Wx + b)$$

$$= [b, w_1, w_2]\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

多層結構



$h(x_1, x_2): \mathbb{R}^2 \longrightarrow \mathbb{R}^2$

$$h(x) = W^{[4]}\sigma(W^{[3]}\sigma(W^{[2]}x + b^{[2]}) + b^{[3]}) + b^{[4]}$$

$$W^{[2]} = \begin{bmatrix} W_{11} & W_{21} \\ W_{12} & W_{22} \\ W_{13} & W_{23} \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

→ 要確保 訓練損失≈驗證損失，並預期測試損失在同一量級

資料集 $\{(x^i, y^i)\}_{i=1}^{N}$，損失函數：$L = \frac{1}{N}\sum_{i=1}^{N}\|y^i - u(x^i)\|^2$

Gradient Descent 梯度下降 → 用全部訓練集

參數更新規則：$\theta^{n+1} = \theta^n - \alpha \nabla_\theta \text{Loss}$

ex: $w^{n+1} = w^n - \alpha\frac{\partial L}{\partial w}$, $b^{n+1} = b^n - \alpha\frac{\partial L}{\partial b}$

$\alpha > 0$，學習率 (Learning rate)

MSE Loss：$\theta^{n+1} = \theta^n - 2\alpha\left[\frac{1}{m}\sum_{i=1}^{m}(y^i - h(x^i))\nabla_\theta h\right]$

- Stochastic Gradient Descent (SGD) 隨機梯度下降 → 一次跑一個樣本
- Mini-Batch Stochastic Gradient Descent → 一次跑幾個樣本

※ Runge's phenomenon：等距節點做高次多項式插值，易在邊緣劇烈振盪

Backpropagation 反向傳播

MSE Loss of SGD : $C = \frac{1}{2}\|y - h(x)\|^2$

兩層隱藏層 : $h(x) = \underbrace{W^{[4]}\sigma(\underbrace{W^{[3]}\sigma(\underbrace{W^{[2]}x + b^{[2]}}_{z^{[2]}}) + b^{[3]}}_{z^{[3]}}) + b^{[4]}}_{z^{[4]}}$

$\Rightarrow \delta^{[4]} = \frac{\partial C}{\partial z^{[4]}} = h(x) - y$

Hadamard product

$\delta^{[3]} = \frac{\partial C}{\partial z^{[3]}} = \sigma'(z^{[3]}) \circ [(W^{[4]})^T \delta^{[4]}]$

$\delta^{[2]} = \frac{\partial C}{\partial z^{[2]}} = \sigma'(z^{[2]}) \circ [(W^{[3]})^T \delta^{[3]}]$

and $\frac{\partial C}{\partial b^{[4]}} = \delta^{[4]}$ $\qquad \frac{\partial C}{\partial b^{[3]}} = \delta^{[3]}$

$\frac{\partial C}{\partial W^{[4]}} = \delta^{[4]} \sigma(z^{[3]})^T$ $\qquad \frac{\partial C}{\partial W^{[3]}} = \delta^{[3]} \sigma(z^{[2]})^T$