Supervised learning

Four ingredients of supervised learning

- Input feature x
- Output target y
- Function hypothesis f
- Loss function

$y = f(x)$

ex: House price prediction

$$\left.\begin{array}{l} \underbrace{\text{size}}_{} = x_1 \\ \underbrace{\text{age}}_{\text{Input}} = x_2 \end{array}\right\} \rightarrow \underbrace{\text{price}}_{\text{Output}} = y$$

Data : $\{(x_1^i, x_2^i, y^i)\}_{i=1}^{N}$

Goal : To find h s.t. $y^i = h(x_1^i, x_2^i)$, $\forall i$.

Assume $h(x_1, x_2) = b + w_1 x_1 + w_2 x_2$, $b, w_1, w_2 \in \mathbb{R}$.

$$\Rightarrow h(x_1, x_2) = \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix} \begin{pmatrix} b \\ w_1 \\ w_2 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_2^1 \\ 1 & x_1^2 & x_2^2 \\ & \vdots & \\ 1 & x_1^N & x_2^N \end{pmatrix} \begin{pmatrix} b \\ w_1 \\ w_2 \end{pmatrix}$$

$$\underset{Y}{} \qquad \underset{X}{} \qquad \underset{\theta}{}$$

$Y \quad X \quad \theta$

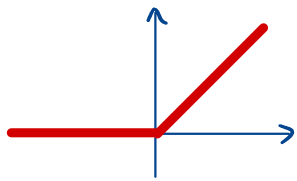
$$\min \| Y - X\theta \|_2^2$$

$$\Rightarrow \theta^* = (X^T X)^{-1} X^T Y$$

mean squared error loss (MSE) : $\frac{1}{N} \sum_{i=1}^{N} | y^i - h(x_1^i, x_2^i) |^2$

Relu activation function

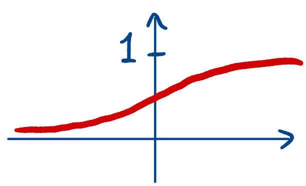
$$\sigma(x) = \begin{cases} x, & x > 0 \\ 0, & x \le 0 \end{cases}$$



hypothesis 2 : $h_2(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2)$

Sigmoid activation function

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



hypothesis 3 : $h_3(x_1, x_2) = w_3 \sigma(b + w_1 x_1 + w_2 x_2)$

Training / Validation / Test set

make sure training loss $\approx$ validation loss

then we except test loss to be in similar order

Runge phenamena

$$f(x) = \frac{1}{1+x^2}, \quad -5 \le x \le 5.$$

Input x

Output y

Hypothesis polynomial: $h(x) = a_0 + a_1 x + \cdots + a_n x^n$

Supervised learning

hypothesis: $h(x;\theta) = h_\theta(x) = h(x)$

$$\text{Loss}(\theta) = \frac{1}{N}\sum_{i=1}^{N} |y^i - h(x^i;\theta)|^2$$

$\theta \in \mathbb{R}^M$, $\text{Loss}: \mathbb{R}^M \to \mathbb{R}$

$$\theta^* = \arg\min_\theta \text{Loss}$$

Gradient descent method (GD)

Motivation: $f: \mathbb{R}^m \to \mathbb{R}$

$\nabla f(x_0)$ is the stepest ascent direction at x_0.

$-\nabla f(x_0)$ is the stepest descent direction at x_0.

GD algorithm

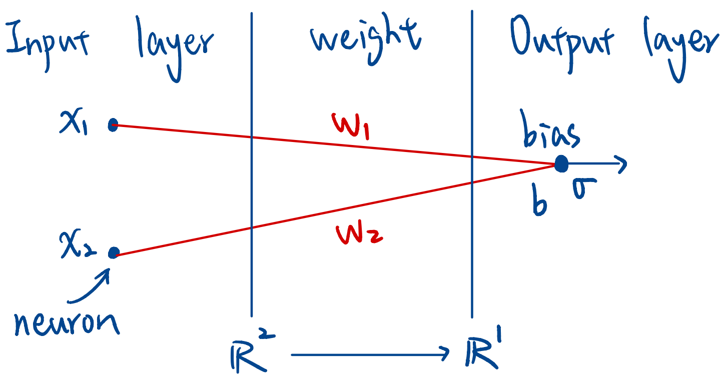
$$\theta^{n+1} = \theta^{n} - \alpha \nabla_\theta Loss \ , \quad \alpha > 0 : \text{learning rate}$$

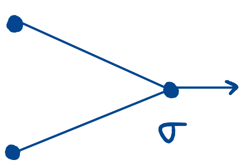
$$\theta^{n+1} = \theta^{n} + \frac{2\alpha}{m} \sum_{i=1}^{m} (y^i - h(x^i;\theta^n)) \cdot \nabla_\theta h(x^i;\theta^n)$$

- $m = N$ Batch gradient descent
- $m = 1$ stochastic gradient descent (SGD)
- $m < N$ mini-batch gradient descent

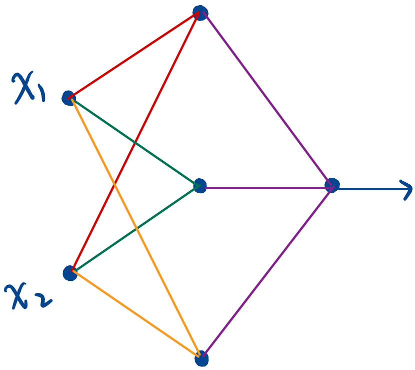
epoch : one full pass through the entire data set

$$h(x_1, x_2) = \sigma(b + w_1 + w_2)$$





$\sigma(b + w_1 + w_2)$



$\rightarrow \ : \ \sigma(w_{11}x_1 + w_{21}x_2 + b_1)$

$\rightarrow \ : \ \sigma(w_{21}x_1 + w_{22}x_2 + b_2)$

$\rightarrow \ : \ \sigma(w_{31}x_1 + w_{32}x_2 + b_3)$

$$\sigma\left(W^{[3]}\sigma\left(\underset{\begin{pmatrix} w_{11} & w_{21} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}}{\underset{\parallel}{W^{[2]}}}\vec{x} + \underset{\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}}{\underset{\parallel}{b^{[2]}}}\right) + b^{[3]}\right)$$