DEEP LEARNING

by Ian Goodfellow.
Yosua Bengio.
Aaron Courville.

$$J(\theta) = \frac{1}{4} \sum \left[f^*(x) - f(x, \theta) \right]^2$$

$$f(x;\theta) = f(x; w, b) = x^T w + b$$

$$J(\theta) = \frac{1}{4} \sum_{x \in X} \left[f^*(x) - (x^T w + b) \right]^2$$

$$\frac{J(\theta)}{\partial w} = -\frac{1}{4} 2 (x^T w + b) \cdot x^T = -\frac{1}{2} x^T (x^T w + b)$$

$$\frac{J(\theta)}{\partial b} = -\frac{1}{4} 2 (x^T w + b) \cdot = -\frac{1}{2} (x^T w + b)$$

MINIMISE (MSE)

$$MSE_{test} = \frac{1}{m} \sum_i \left(\hat{y}^{(test)} - y^{(test)} \right)_i^2$$, with m = number example

we can see that $MSE_{test} = 0$ when $\hat{y}^{(test)} = y^{(test)}$

we can see also that

$$MSE_{test} = \frac{1}{m} \left\| \hat{y}^{(test)} - y^{(test)} \right\|_2^2$$

$$\|x\|_2 = \|x\| = x^T x.$$

Let Minimize MSE_{train}, $\Rightarrow$ solve where $\nabla_w MSE_{train} = 0$

$$\nabla_w MSE_{train} = 0$$

$$\Rightarrow \nabla_w \left[\frac{1}{m} \left\| \hat{y}^{(train)} - y^{(train)} \right\|_2^2 \right] = 0$$

$$\nabla_w \left\| \hat{y}^{(train)} - y^{(train)} \right\|_2^2 = 0$$

$$\nabla_w \left\| X^{(train)} w - y^{(train)} \right\|_2^2 = 0$$ with $\hat{y}^{(train)} = X^{(train)} w$

$$\nabla_w (X^{(train)} w - y^{(train)})^T (X^{(train)} w - y^{(train)}) = 0$$

$$\nabla_w \left[w^T X^{(train)T} X^{(train)} w - \underbrace{w^T X^{(train)T} y^{(train)} - y^{(train)T} X^{(train)} w}_{-2 w^T X^{(train)T} y^{(train)}} + y^{(train)T} y^{(train)} \right] = 0$$

$$\nabla_W \left(W^T X^{(train)T} X^{(train)} W - 2 W^T X^{(train)T} y^{(train)} + y^{(train)T} y^{(train)} \right) = 0$$

$$2 X^{(train)T} X^{(train)} W - 2 X^{(train)T} y^{(train)} = 0$$

$$\Rightarrow \boxed{W = \left(X^{(train)T} X^{(train)} \right)^{-1} X^{(train)} y^{(train)}}$$ known as normal equations

X	Y	XOR
0	0	0
1	0	1
0	1	1
1	1	0

$$y = f^*(x)$$

$J(\theta) = \frac{1}{4} \sum_{x \in X} \left(f^*(x) - f(x;\theta) \right)^2$

$$f(x;\theta) = f(x; W, b) = x^T W + b$$

$$\nabla \left[\frac{1}{4} \sum \left(f^*(x) - (x^T W + b) \right)^2 \right] = 0$$

$$\Rightarrow \nabla \| f^*(x) - (x^T W + b) \|_2^2 = 0$$

$$\Rightarrow \nabla \left[f^*(x) - (x^T W + b) \right]^T \left[f^*(x) - (x^T W + b) \right] = 0$$

$$\Rightarrow \nabla \left[f^*(x)^T f^*(x) - f^*(x) (x^T W + b) - (x^T W + b)^T f^*(x) + (x^T W + b)^T (x^T W + b) \right] = 0$$

---

suppose:

$$f^{(1)}(x) = W^T x$$

$$f^{(2)}(h) = h^T \omega$$

$h = f^{(1)}(x; W, c)$ and $y = \underbrace{f^{(2)}(h; \omega, b)}_{\text{output of the network}}$

$$\Rightarrow \underbrace{f(x)}_{f(x; W, c, \omega, b)} = f^{(2)}\left(f^{(1)}(x; W, c) \right) = f^{(2)}(W^T x) = (W^T x)^T \omega = \boxed{x^T W \omega} = (x^T W \omega)^T = \boxed{\omega^T W^T x}$$

$$= x^T \omega' \text{ where } \omega' = W\omega$$

putting an activation function "g" = ReLU $\Rightarrow \boxed{g(z) = \max\{0, z\}}$

Considering the bias term $f(x)$ can be written as:

$$f(x; W^1, b^1, W^2, b^2) = f^{(2)}\left(f^{(1)}(x; W^1, b^1) \right) = f^{(2)}\left[f^{(1)}\left((W^1)^T x + b^1 \right) + b^2 \right]$$

$$= f^{(2)}\left(\max\{0, (W^1)^T x + b^1\} + b^2 \right)$$

$$= g\left[(W^2)^T \max\{0, (W^1)^T x + b^1\} + b^2 \right]$$

$$= \max\left\{ 0, \max\left\{ (W^2)^T \max\{0, (W^1)^T x + b^1\} + b^2 \right\} \right.$$