Assignment - 3 [Theory]

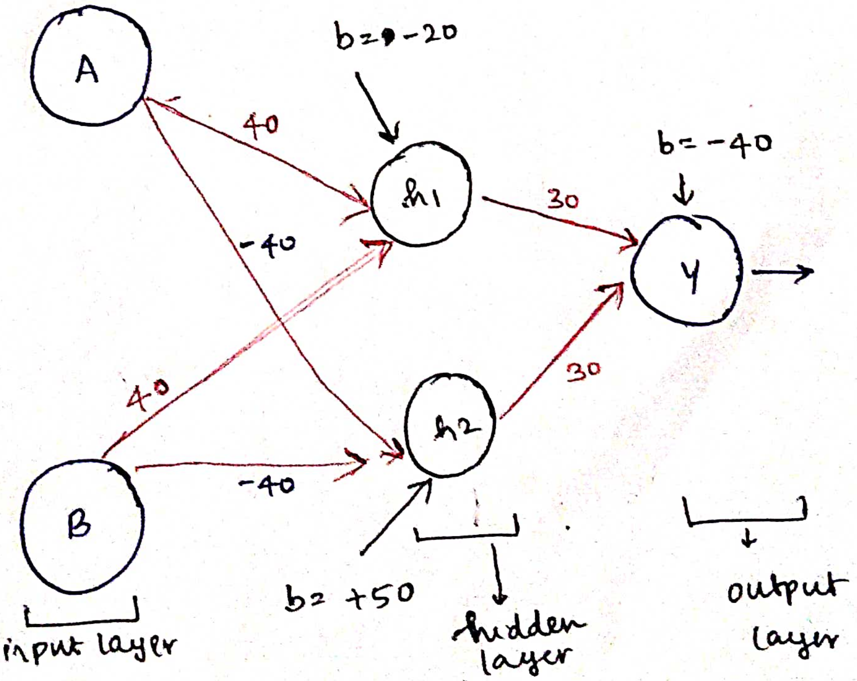
1. Neural Networks.

a) The truth table of XOR is

A	B	Y = $A \oplus B$
0	0	0
0	1	1
1	0	1
1	1	0

__To prove:__ A two layer perceptron can solve the XOR problem.

__Solution:__ Consider the two layer perceptron as



Let

$$h_1 = \sigma(40A + 40B - 20)$$

$$h_2 = \sigma(-40A - 40B + 50)$$

σ → sigmoid function

and $y = \sigma(30h_1 + 30h_2 - 40)$.

(d)

case (i) when $A = 0$, $B = 0$

$h_1 = \sigma(40(0) + 40(0) - 20) = \sigma(-20) = 0$

$h_2 = \sigma(-40(0) - 40(0) + 50) = \sigma(50) = 1$

$y = \sigma(30(0) + 30(1) - 40) = \sigma(-10) = 0$

$\therefore$ when $A = 0$, $B = 0 \Rightarrow y = 0$ ✓.

case (ii) when $A = 0$, $B = 1$

$h_1 = \sigma(40(0) + 40(1) - 20) = \sigma(20) = 1$

$h_2 = \sigma(-40(0) - 40(1) + 50) = \sigma(10) \approx 1$

$y = \sigma(30(1) + 30(1) - 40) = \sigma(20) = 1$

$\therefore$ when $A = 0$, $B = 1 \Rightarrow y = 1$ ✓

case (iii) when $A = 1$, $B = 0$

$h_1 = \sigma(40(1) + 40(0) - 20) = \sigma(20) = 1$

$h_2 = \sigma(-40(1) - 40(0) + 50) = \sigma(10) \approx 1$

$y = \sigma(30(1) + 30(1) - 40) = \sigma(20) \approx 1$

$\therefore$ when $A = 1$, $B = 0 \Rightarrow y = 1$ ✓

case (iv) when $A = 1$, $B = 1$

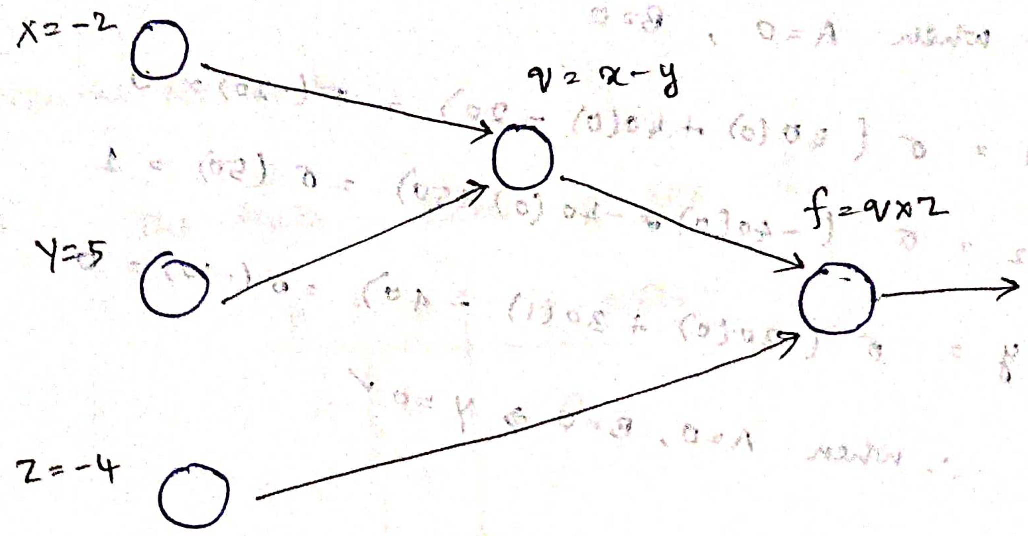
$h_1 = \sigma(40(1) + 40(1) - 20) = \sigma(60) = 1$

$h_2 = \sigma(-40(1) - 40(1) + 50) = \sigma(-30) \approx 0$

$y = \sigma(30(1) + 30(0) - 40) = \sigma(-10) \approx 0$

$\therefore$ when $A = 1$, $B = 1$, $y = 0$ ✓ Thus XOR truth table is verified

(b)



Gradient f with respect to x = $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x}$

$= z(1)$

$= -4.$

Gradient f with respect to y = $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial y}$

$= z(-1) = -z$

$= -(-4) = 4$

Gradient f with respect to z = $\frac{\partial f}{\partial z} = q = x-y = -7$

$$\boxed{\therefore \frac{\partial f}{\partial x} = -4, \quad \frac{\partial f}{\partial y} = 4, \quad \frac{\partial f}{\partial z} = -7}$$

2)

Given: cross entropy error function is.

$$E(w) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{kn} \ln y_k(x_n, W) \rightarrow (1)$$

N → No. of data samples, K → number of classes

$$y(x_n, w) = p(t_k = 1 / x) = \frac{\exp(a_k(x, w))}{\sum_j \exp(a_k(x, w))}$$

$0 \le y \le 1$ and $\sum_k y_k = 1$; $a_k \rightarrow$ pre softmax activation of output layer neurons

We know that by chain rule

$$\frac{\partial E}{\partial w_{ki}} = \sum_j \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial a_k} \cdot \frac{\partial a_k}{\partial w_{ki}} \rightarrow (2)$$

$E(w) =$ Partial derivative of (1) with respect to y_j for a particular value of j is

$$\frac{\partial E}{\partial y_j} = -\frac{t_j}{y_j} \rightarrow (3)$$

Let us find $\frac{\partial y_j}{\partial a_k}$ Two cases: $j = k$, $j \neq k$.

when $j = k$

$$\frac{\partial y_j}{\partial a_k} = \frac{\sum \exp(a_i) \exp(a_j) - \exp(a_j) \exp(a_j)}{(\sum \exp(a_i))^2}$$

$$\frac{\partial y_j}{\partial a_k} = y_i - y_i^2 = y_j(1-y_j) \quad \text{④}$$

when $j \neq k$,

$$\frac{\partial y_j}{\partial a_k} = -\frac{\exp(a_j)\exp(a_k)}{(\sum \exp(a_i))^2} = -y_j y_k. \quad \text{⑤}$$

④ & ⑤ ⇒ generalised as $\boxed{\frac{\partial y_j}{\partial a_k} = y_j(\delta_{jk} - y_k)}$ ⑥

δ_{jk} ⇒ Kronecker delta defined as

$\delta_{jk} = 1$ when $j = k$

$\delta_{jk} = 0$ when $j \neq k$.

substituting ⑥, ③ in ②

$$\frac{\partial E}{\partial w_{ki}} = \sum_j -\frac{t_j}{y_j} \cdot y_j(\delta_{jk} - y_k) \cdot \frac{\partial a_k}{\partial w_{ki}}$$

$$= \sum_j t_j(y_k - \delta_{jk}) \cdot \frac{\partial a_k}{\partial w_{ki}}$$

given in question $t_n = [0, 0, 1, 0 \ldots, 0]$

$$\sum_j t_j = 1$$

for an one given input,

$$\therefore \frac{\partial E}{\partial w_{ki}} = (y_k - \delta_{jk}) \cdot \frac{\partial a_k}{\partial w_{ki}}$$

$$\Rightarrow \cancel{\frac{\partial E}{\partial w_{ki}}} \quad \frac{\partial E}{\partial a_k} \cdot \cancel{\frac{\partial a_k}{\partial w_{ki}}} = (y_k - \delta_{jk}) \cdot \cancel{\frac{\partial a_k}{\partial w_{ki}}}$$

$\delta_{jk} \rightarrow t_k$ (target)

$$\Rightarrow \boxed{\frac{\partial E}{\partial a_k} = (y_k - t_k)}$$

3) $f(x) = x^2$

$$E_{AV} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_x[(y_m(x) - f(x))^2]$$

$$E_{ENS} = \mathbb{E}_x\left[\left(\frac{1}{M} \sum_{m=1}^{M} y_m(x) - f(x)\right)^2\right]$$

To prove: $E_{ENS} \leq E_{AV}$

given

$$E_{ENS} = \mathbb{E}_x\left[\left(\frac{1}{M} \sum_{m=1}^{M} y_m(x) - f(x)\right)^2\right]$$

$$= \mathbb{E}_x\left[\frac{1}{M^2}\left(\sum_{m=1}^{M} y_m(x) - f(x)\right)^2\right]$$

$$= \mathbb{E}_x\left[\frac{1}{M^2}\left[(y_1(x) - f(x)) + (y_2(x) - f(x)) + \dots + (y_m(x) - f(x))\right]^2\right]$$

$$= \mathbb{E}_x\left[\frac{1}{M^2}\left(\sum_{m=1}^{M} \epsilon_m\right)^2\right]$$

$$= E_x\left[\left(\frac{\sum_{m=1}^{M} \epsilon_m}{M}\right)^2\right]$$

By Jensen Inequality,

$$f(E(x)) \leq E(f(x))$$

$$E_{ENS} = E_x\left[\left(\frac{\epsilon_1 + \epsilon_2 + \epsilon_3 + \dots \epsilon_m}{M}\right)^2\right]$$

$$\leq \frac{1}{M} E_x(\epsilon_1^2) + \frac{E_x}{M}(\epsilon_2^2) + \dots + \frac{E_x(\epsilon_M^2)}{M}$$

$$\leq \frac{1}{M}\sum_{m=1}^{M} E_x(\epsilon_m^2)$$

$$\Rightarrow \boxed{E_{ENS} \leq E_{AV}}$$

The Jensen's inequality mainly cares about the convexity of the function rather than error function $E(y)$. Hence the result stands true for any error function