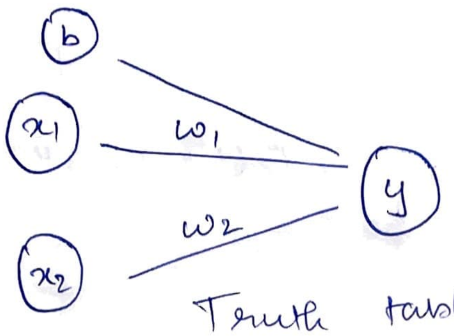
HOMEWORK - 2 CS 541 DEEP LEARNING.

1. XOR Problem.



Truth table of XOR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Input x_1, x_2
Output y.
weights w
bias b

$\Big\}\rightarrow$ for the network

$\Rightarrow X = (x_1^{(i)}, x_2^{(i)}) = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$

1×2 2×1 +b

$y = [y^{(i)}] = [0 \quad 1 \quad 1 \quad 0]^T$

$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ & $b =$

Loss function as given in equation 6.1 of Deep Learning book.

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

where θ ~~a~~ is ~~the~~ w_1, w_2 & b in our case

& ~~y^*~~ $f^*(y)$ = true value of XOR from truth table.

$$J(w_1, w_2, b) = \frac{1}{4} \sum_{x \in X} \left[y^{(i)} - \left(x^{(i)T} w + b\right)\right]^2$$

$$= \frac{1}{4}\left[[0 \quad 1 \quad 1 \quad 0]^T - \left(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}}_{4\times 2} \underbrace{\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}}_{2\times 1} + b\right)\right]^2$$

$$= \frac{1}{4}\left[(0 \quad 1 \quad 1 \quad 0)^T - \left(\begin{bmatrix} 0 \\ w_2 \\ w_1 \\ w_1 + w_2 \end{bmatrix} + b\right)\right]^2$$

$$= \frac{1}{4}\left[[0 \quad 1 \quad 1 \quad 0]^T - [b \quad w_2+b \quad w_1+b \quad w_1+w_2+b]^T\right]^2$$

$$\cancel{= \frac{1}{4}\left[[b \quad 1+w_2+b \quad 1+w_1+b \quad w_1+w_2\right.}$$

$$= \frac{1}{4}\left[[-b \quad 1-(w_2+b) \quad (1-(w_1+b)) \quad -(w_1+w_2+b)]^T\right]^2$$

So $\left(\begin{bmatrix} -b & 1-(w_2+b) & 1-(w_1+b) & -(w_1+w_2+b) \end{bmatrix}^T\right)^2$

$$= \begin{bmatrix} -b & 1-(w_2+b) & 1-(w_1+b) & -(w_1+w_2+b) \end{bmatrix} \begin{bmatrix} -b \\ 1-(w_2+b) \\ 1-(w_1+b) \\ -(w_1+w_2+b) \end{bmatrix}$$

$$= \left[b^2 + (1-(w_2+b))^2 + (1-(w_1+b))^2 + (w_1+w_2+b)^2 \right]$$

$$= \left[b^2 + 1 + (w_2+b)^2 - 2(w_2+b) + 1 + (w_1+b)^2 - 2(w_1+b) + (w_1+w_2)^2 + 2(w_1+w_2)b + b^2 \right]$$

$$= \left[b^2 + 2 + w_2^2 + b^2 + 2w_2 b - 2(w_1+w_2+2b) + w_1^2 + b^2 + 2w_1 b + w_1^2 + w_2^2 + 2w_1 w_2 + 2w_1 b + 2w_2 b + b^2 \right]$$

$$= \left[4b^2 + 2 + 2(w_1^2 + w_2^2) + 4b(w_1+w_2) - 2(w_1+w_2+2b) + 2w_1 w_2 \right]$$

$J(w, \theta) =$

$$\left[4b^2 + 2(w_1^2 + w_2^2) + 4b(w_1+w_2) + 2w_1 w_2 + 2 - 2(w_1+w_2+2b) \right]$$

$\nabla J(w_1) = [0 + 4w_1 + 4b + 2w_2 - 2] = 0 \rightarrow (1)$

$\nabla J(w_2) = [0 + 4w_2 + 4b + 2w_1 - 2] = 0 \rightarrow (2)$

$\nabla J(b) = [8b + 4(w_1 + w_2) - 4] = 0 \rightarrow (3)$

$4w_1 + 2w_2 + 4b = 2 \rightarrow (4)$

$4w_2 + 2w_1 + 4b = 2 \rightarrow (5)$

$8b + 4w_1 + 4w_2 = 4 \rightarrow (6)$

(4) - (5)

$2w_1 - 2w_2 = 0$

$w_1 = w_2 \rightarrow (7)$

Substituting (7) in (6) & (5)

$8b + 8w_1 = 4 \quad \times 3$

$6w_1 + 4b = 2 \quad \times 4$

we get

$24b + \cancel{24w_1} = 12$

$(-)\ \cancel{24w_1} \overset{(-)}{+} 16b \overset{(-)}{=} 8$

$8b = 4$

$\boxed{b = 1/2} \rightarrow (8)$

Substituting $b = 1/2$ in (5)

$4w_1 + 2w_1 + 4 \times \frac{1}{2} = 2$

$6w_1 + \cancel{2} = \cancel{2}$

$\boxed{w_1 = 0} \rightarrow (9)$

Since $w_1 = w_2$

$\Rightarrow w_2 = 0$

∴ The values of w_1, w_2 & b to minimize the function $J(w, b) = \frac{1}{4} \sum_{x \in X} (f^*(y) - f(x; w, b))^2$

is

$$\boxed{\begin{aligned} w_1 &= 0 \\ w_2 &= 0 \\ b &= 1/2 \end{aligned}}$$

Q3.

L_2 regularization term

be $= \frac{\alpha}{2n} w^T w$.

Given input $x = [x_1 \quad x_2]$

Let weights be $w = [w_1 \quad w_2]^T$

Let a matrix S be included in the L_2 regularization term.

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

$\Rightarrow$ L_2 cost $= J_{L_2} = \frac{\alpha}{2n} w^T S w$

$$= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} S_{11} w_1 + S_{12} w_2 \\ S_{21} w_1 + S_{22} w_2 \end{bmatrix}$$

$$= w_1^2 S_{11} + w_1 w_2 S_{12} + w_1 w_2 S_{12} + w_2^2 S_{22}$$

$$J_{L2} = w_1^2 S_{11} + w_1 w_2 (S_{12} + S_{21}) + w_2^2 S_{22}.$$

$J_{L2} = 0$ when w_1 approximately equal to w_2, such that we need not to penalise the cost function as the weights are reflective over the middle column $\boxed{\text{LHS} \mid \text{RHS}}$ ~~LHS=RHS~~

$\Rightarrow w_1 - w_2 \approx 0$

$\Rightarrow (w_1 - w_2)^2 \approx 0.$

$$\Rightarrow (w_1 - w_2)^2 = w_1^2 S_{11} + w_1 w_2 (S_{12} + S_{21}) + w_2^2 S_{22}$$

$$w_1^2 + w_2^2 - 2 w_1 w_2 = w_1^2 S_{11} + w_2^2 S_{22} + w_1 w_2 (S_{12} + S_{21})$$

comparing coefficients of w_1^2, w_2^2 & $w_1 w_2$.

$S_{11} = 1$, $S_{22} = 1$, $S_{12} + S_{21} = -2$.

$S_{12} = S_{21} = -1$ (assume).

To prove $S_{12} \stackrel{?}{=} S_{21} = -1$.

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} w_1 - w_2 \\ w_2 - w_1 \end{bmatrix}$$

$$= w_1(w_1 - w_2) + w_2(w_2 - w_1) \rightarrow (1)$$

Equation (1) $= 0$.

$$w_1^2 - w_1 w_2 + w_2^2 - w_1 w_2 = 0.$$

$$w_1^2 - 2 w_1 w_2 + w_2^2 = 0$$

$$(w_1 - w_2)^2 = 0$$

is the condition when the weights are symmetric & this can be derived only when $\boxed{S_{12} = S_{21} = -1}$

Explanation for matrix S.

when weights w_1 & w_2 are not symmetric

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

$$= w_1(w_1 - w_2) + w_2(w_2 - w_1)$$

$$= \frac{\alpha}{2n}\left[(w_1 - w_2)(w_1 - w_2)\right]$$

$$= \frac{\alpha}{2n}\left[(w_1 - w_2)^2\right].$$

which penalises the cost function when $w_1 = w_2$, ie the weights are symmetric (in the sense reflective along the middle column) (not $X^T = X$)

$$J_{L_2} = \frac{\alpha}{2n}\left[(w_1 - w_2)\right]^2$$

$$= \frac{\alpha}{2n} \times 0$$

$$= 0$$

$$\left[\begin{array}{c|c} LHS & RHS \end{array}\right]$$

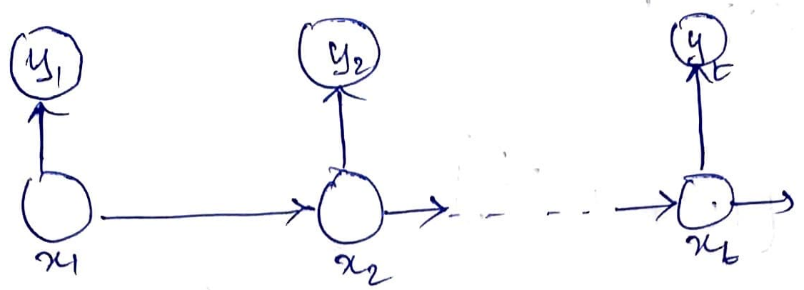
~~LHS = 255 - RH~~

Since the weights are already symmetric it does not penalise the cost function

Hence L_2 ~~can be~~ regularization can be used to discourage asymmetric weights ~~xx~~ if the datas are symmetric in nature.

$$\boxed{\therefore S = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}}$$

4.



$y_1, y_2, y_3 \dots y_t$ are the observable random variables, ~~here~~ here the student's behaviour which is observable ~~which~~ is determined by that student's current state which is thought process inside his brain which is not observable given by ~~$x_t, x_{t-1} \dots x_1$~~ $x_1, x_2 \dots x_t$ (corresponding y_i, $i \to 1$ to t)

Hence $x_1 \dots x_t$ are known as hidden units & $y_1 \dots y_t$ are known as observable units.

Given:

$$P(x_t \mid x_1, \dots, x_{t-1}) = P(x_t \mid x_{t-1})$$

markov property.

$$P(y_t \mid x_t, y_1, \dots, y_{t-1}) = P(y_t \mid x_t)$$

The Goal of the teacher is to estimate the current state ie x_t given the observations $y_1, \dots, y_t$ and update her belief about the student.

ie; to ~~find~~ prove
$$\boxed{P(x_t \mid y_1, \dots, y_{t-1}, y_t) \propto P(y_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid y_1, \dots y_{t-1})}$$

Applying Bayes' theorem:

$$P(x_t \mid y_t, y_{t-1}, \dots, y_1) = \frac{P(y_t \mid x_t, y_1, \dots y_{t-1}) P(x_t \mid y_1, \dots y_{t-1})}{P(y_t \mid y_1, \dots, y_{t-1})} \rightarrow ①$$

Since the denominator of ① does not involve x_t equation ① can be re-written as

$$P(x_t \mid y_1, \dots, y_t) \propto P(y_t \mid x_t, y_1, \dots y_{t-1}) \, P(x_t \mid y_1, \dots y_{t-1})$$

$$P(y_t \mid x_t, y_1, \dots y_{t-1}) = P(y_t \mid x_t) \rightarrow$$ Given & can be inferred from the chain.

$$\Rightarrow P(x_t \mid y_1, \dots, y_t) \propto P(y_t \mid x_t) P(x_t \mid y_1, \dots, y_{t-1}) \rightarrow ②.$$

x_t depends on x_{t-1} as given in the chain and x_{t-1} can be any state of mind of the student, hence according to law of total probability, $P(x_t) = \sum_{x_{t-1}} P(x_t, x_{t-1})$

Applying law of probability to equation (2) we get.

$$P(x_t|y_1,\dots,y_t) \propto P(y_t|x_t) \sum_{x_{t-1}} P(x_t, x_{t-1}|y_1,\dots,y_{t-1})$$

$\hookrightarrow$ (3)

Applying conditional probability distribution concept to equation (3) we get.

$$P(x_t|y_1,\dots y_t) \propto P(y_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}|y_1,\dots,y_{t-1})$$

where $P(x_{t-1}|y_1,\dots,y_{t-1})$ is the teacher's belief from the time stamp of $t-1$ &

the summation is over the all possible values of x_{t-1} i.e; all the thought process in the student's mind.

$$\therefore \boxed{P(x_t|y_t,\dots,y_1) \propto P(y_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}|y_1,\dots,y_{t-1})}$$

Hence Proved

~~5~~

~~Cost fn.~~

5.

$$P(y \mid x, w, \sigma^2) = \mathcal{N}(y; x^T w, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right).$$

$$P(\mathcal{D} \mid w, \sigma^2) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, w; \sigma^2) \quad \to (1)$$

where $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

Taking log on both sides of eqn. (1).

$$\log P(\mathcal{D} \mid w, \sigma^2) = \log \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log P(y^{(i)} \mid x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - x^{(i)T} w)^2}{2\sigma^2}\right)\right)$$

$$= \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log\left(\exp\left(-\frac{(y^{(i)} - x^{(i)T} w)^2}{2\sigma^2}\right)\right)\right]$$

$$= \sum_{i=1}^{n}\left[\log\frac{1}{\sqrt{2\pi}} + \log\frac{1}{\sqrt{\sigma^2}} + \left(-\frac{(y^{(i)} - x^{(i)T}w)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{n}\log\frac{1}{\sqrt{\sigma^2}} - \sum_{i=1}^{n}\frac{(y^{(i)} - x^{(i)T}w)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y^{(i)} - x^{(i)T}w)^2}{2\sigma^2}$$

$$= -\frac{n}{2}(\log 2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y^{(i)} - x^{(i)T}w)^2}{2\sigma^2} \rightarrow (2)$$

Differentiating eqn. (2) wrt σ^2 & equating it to zero

$$0 = 0 - \frac{n}{2}\times\frac{1}{\sigma^2} - (-1)\sum_{i=1}^{n}\frac{(y^{(i)} - x^{(i)T}w)^2}{2\sigma^4}$$

$$\frac{n}{\cancel{2\sigma^2}} = \sum_{i=1}^{n}\frac{(y^{(i)} - x^{(i)T}w)^2}{\cancel{2}\sigma^{\cancel{4}2}}$$

$$\therefore \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - x^{(i)T}w)^2$$

Taking (-1) common from $(y^{(i)} - x^{(i)T}w)^2$

$$\Rightarrow \boxed{\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)T}w - y^{(i)})^2}$$

Differentiating equation (2) w.r.t w and equating it to zero.

$$0 = 0 + 0 - 2\sum_{i=1}^{n} \left(y^{(i)} - x^{(i)T} w\right) \times \left(-x^{(i)T}\right)$$

since $\boxed{\nabla \left(x^{(i)T} w\right) = x^{(i)}}$

$$0 = -\sum_{i=1}^{n} x^{(i)} y^{(i)} + \sum_{i=1}^{n} x^{(i)} x^{(i)T} w$$

$$\sum_{i=1}^{n} x^{(i)} x^{(i)T} w = \sum_{i=1}^{n} x^{(i)} y^{(i)}$$

$$w \sum_{i=1}^{n} x^{(i)} x^{(i)T} = \sum_{i=1}^{n} x^{(i)} y^{(i)}$$

$$\therefore \boxed{w = \left(\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right)^{-1} \sum_{i=1}^{n} x^{(i)} y^{(i)}}$$

& $\boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)T} w - y^{(i)}\right)^2}$

Q 2.

Set of learning Rates = [0.001, 0.005, 0.01, 0.0005]

Set of no. of epochs = [50, 100, 200, 400]

Set of batch sizes = [128, 256, 512, 1024]

Set of regularization constant = [0.01, 0.2, 0.5, 0.005]

Result

Best Epochs = 400

Best Learning Rate = 0.001

Best batch size = 512

Best regularization constant = 0.5

Best Validation Loss = 118.10845242

~~Best~~
TEST MSE LOSS = 120.87438984