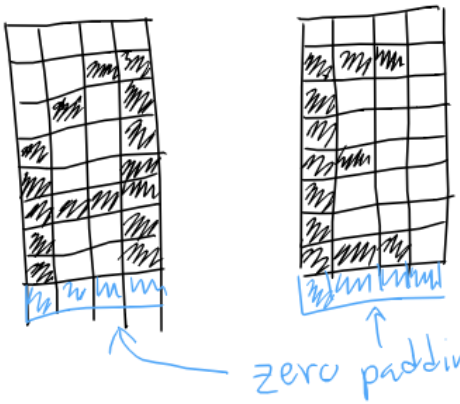
Task 1

- stride : $S_1 = 1$, $S_2 = 3$
- Output of convolution layer : V (3×2)
- After max pooling : m (2×1)

1) Calculate entries of V matrix.
2) Attempt : set all $w_{ij} = 1$
3) Get entries for m
4) Determine θ and $\bar{W}$



A

$$V_{11} = w_{23} + w_{32} - \theta$$

$$V_{12} = w_{22} + w_{23} + w_{31} + w_{33} - \theta$$

$$V_{21} = w_{11} + w_{21} + w_{31} + w_{32} + w_{33} - \theta$$

$$V_{22} = w_{13} + w_{23} + w_{31} + w_{32} + w_{33} - \theta$$

$$V_{31} = w_{11} + w_{21} - \theta$$

$$V_{32} = w_{13} + w_{23} - \theta$$

E

$$V_{11} = w_{21} + w_{22} + w_{23} + w_{31} - \theta$$

$$V_{12} = w_{21} + w_{22} - \theta$$

$$V_{21} = w_{11} + w_{21} + w_{22} + w_{31} - \theta$$

$$V_{22} = w_{21} - \theta$$

$$V_{31} = w_{11} + w_{21} + w_{22} + w_{23} - \theta$$

$$V_{32} = w_{21} + w_{22} - \theta$$

$$m(A) = (5 - \theta,\ 5 - \theta)^T \quad \rightarrow \quad \bar{W}_1(5 - \theta) + \bar{W}_2(5 - \theta) < 0$$

$$m(E) = (4 - \theta,\ 4 - \theta)^T \quad \rightarrow \quad \bar{W}_1(4 - \theta) + \bar{W}_2(4 - \theta) > 0$$

ANS With stride $S_1 = 1$, $S_2 = 3$: All $w_{ij} = 1$, $\theta = 4, 5$,
$\bar{W}_1 = -1$, $\bar{W}_2 = -1$

note: Infinite possible solutions.

(2.)(a) Starting with the KL divergence,

$$D_{KL} = \sum_{\mu=1}^{P} P_{data}(x^\mu) \log \frac{P_{data}(x^\mu)}{P_B(S = x^\mu)} = -\sum_{\mu=1}^{P} P_{data} \log \frac{P_B}{P_{data}}$$

use the inequality $\log(x) \leq x - 1$, where $\log(x) = x - 1$ only if $x = 1$.

We find

$$D_{KL} = -\sum_{\mu=1}^{P} P_{data} \log \frac{P_B}{P_{data}} \geq -\sum_{\mu=1}^{P} P_{data} \left(\frac{P_B}{P_{data}} - 1 \right)$$

$$= -\sum_{\mu=1}^{P} \left[P_B - P_{data}(x^\mu) \right]$$

$$= -\sum_{\mu=1}^{P} \left[P_B - 1 \right]$$

$$\geq 0$$

$$\boxed{D_{KL} \geq 0}$$

(b) The equivalence between maximising the log-likelihood and minimising the KL divergence can be seen as follows. The log-likelihood reads

$$\log \mathcal{L} = \sum_{\mu=1}^{p} \log P_B\left(s = x^{(\mu)}\right)$$

where, as stated in the lecture notes, the sum runs over a sequence of input patterns $x^{(\mu)}$, $\mu = 1, \dots, p$. Any pattern x may appear more than once in the sequence, with frequency proportional to $P_{data}(x)$. In the limit of a large number of samples, $p \gg 1$,

$$\log \mathcal{L} = p\left[\frac{1}{p}\sum_{\mu=1}^{p} \log P_B\left(s = x^{\mu}\right)\right]$$

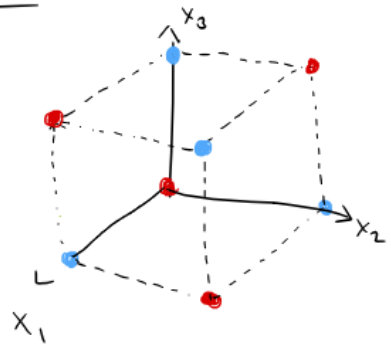
~~is interpreted as~~ a Using the law of large numbers, the quantity in the square brackets reads

$$\frac{1}{p}\sum_{\mu=1}^{p} \log P_B\left(s = x^{\mu}\right) \overset{p \gg 1}{=} \sum_{\mu=1}^{p} P_{data}(x^{\mu}) \log P_B(\cdot)$$

thus, $\log \mathcal{L} = p\left[\underbrace{\left(\sum P_{data} \log P_{data}\right)}_{\text{Constants}} - D_{KL}\right]$

In conclusion, maximising $\log \mathcal{L}$ is equivalent to minimising D_{KL}.

Task 3

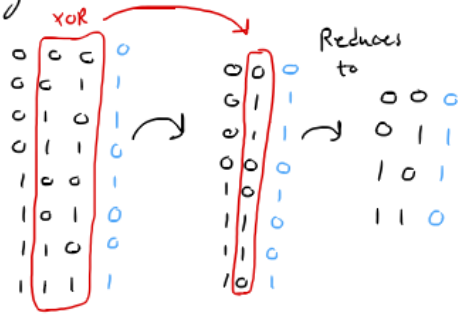


- 1
- 0

No way to construct a plane that separates the points.

A network which can separate the blue from the red can be found by first constructing an XOR network for the first two variables and then using the output of that network and the third variable as input to a second XOR network

This can be seen by looking at the logic table:



$$W^{(1)} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$W^{(2)} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$W^{(3)} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$$

$$W^{(4)} = \begin{pmatrix} 1 & 1 \end{pmatrix}$$

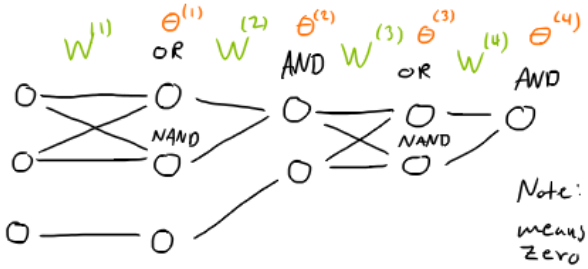
$$\theta^{(1)} = (0{,}5, -1{,}5, 0)^T$$

$$\theta^{(2)} = (1{,}5, 0)^T$$

$$\theta^{(3)} = (0{,}5, -1{,}5)^T$$

$$\theta^{(4)} = 1{,}5$$

$\therefore$



Note: No line means weight is zero

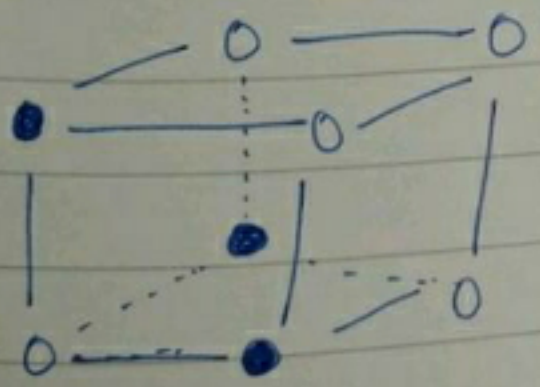
(4). Boolean function II.

Truth table.

x_1	x_2	x_3	t
0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	0

Empty circle = 0

Filled circle = 1



Not linearly separable = requires hidden layer.

A neural network representing these functions in N ($=3$ or 4) dimensions can be constructed using the winning neuron construction discussed in section 7.1 of the lecture notes

The hidden layer weights and thresholds need to be modified from those discussed in the lecture notes. The lecture notes define the hidden layer weights,

$$w_{jk} = \begin{cases} \delta & \text{if the } k^{\text{th}} \text{ digit of binary representation of } j \text{ is } 1 \\ -\delta & \text{otherwise.} \end{cases}$$

and thresholds $\theta_j = N(\delta - 1)$.

The above weights and thresholds are for ± 1 neurons. Let the weights and thresholds for $0,1$ neurons be $\tilde{w}_{ij}$ and $\tilde{\theta}_{ij}$. Then

$$\tilde{w}_{jk} = 2 w_{jk}$$

$$\tilde{\theta}_j = \sum_k w_{jk} + \theta_j .$$

Task 5

Back-propagation:

$$w^{(l)} \leftarrow w^{(l)} - \eta \frac{\partial H}{\partial w^{(l)}}$$

$$\frac{\partial H}{\partial w^{(l)}} = \frac{\partial}{\partial w^{(l)}} \frac{1}{2} \left(t - V^{(L)}\right)^2 = -\left(t - V^{(L)}\right) \frac{\partial V^{(L)}}{\partial w^{(l)}}$$

$$= -\left(t - V^{(L)}\right) g'\left(b^{(L)}\right) w^{(L)} \frac{\partial}{\partial w^{(l)}} V^{(L-1)}$$

$$= \boxed{-\left(t - V^{(L)}\right) \left[\prod_{i=0}^{L-l-1} g'\left(b^{(L-i)}\right) w^{(L-i)}\right] g'\left(b^{(l)}\right) V^{(l-1)}}$$

$$\frac{\partial V^{(L)}}{\partial V^{(l)}} = \frac{\partial}{\partial V^{(l)}} g\left(w^{(L)} V^{(L-1)} - \theta^{(L)}\right) = g'\left(b^{(L)}\right) w^{(L)} \frac{\partial V^{(L-1)}}{\partial V^{(l)}}$$

$$= \boxed{\prod_{i=0}^{L-l-1} g'\left(b^{(L-i)}\right) w^{(L-i)}}$$

$$\therefore \quad \boxed{\delta^{(l-1)} = \left(t - V^{(L)}\right) \frac{\partial V^{(L)}}{\partial V^{(l-1)}} g'\left(b^{(l-1)}\right)}$$

(6.)

(a) Start ~~with~~ by differentiating the average immediate reward w.r.t w_n,

$$\frac{\partial \langle r \rangle}{\partial w_n} = \sum_{y=\pm 1} \left\langle r(\bar{x}, y) \frac{\partial P(y, \bar{x})}{\partial w_n} \right\rangle$$

where,

$$P(y, \bar{x}) = \begin{cases} p(b) & \text{if } y = +1 \\ 1 - p(b) & \text{if } y = -1 \end{cases}$$

$$\Rightarrow \frac{\partial \langle r \rangle}{\partial w_n} = \left\langle r(\bar{x}, 1) \frac{\partial p(b)}{\partial w_n} \right\rangle + \cancel{\left\langle r(\bar{x}, -1) \frac{\partial}{\partial w_n} \right.}$$

$$+ \left\langle r(\bar{x}, -1) \frac{\partial [1 - p(b)]}{\partial w_n} \right\rangle$$

$$= \left\langle r(\bar{x}, 1)\, p(b)[1 - \tanh(b)] \right\rangle \frac{\partial b}{\partial w_n}$$

$$+ \left\langle r(\bar{x}, -1)[1 - p(b)][-1 - \tanh(b)] \right\rangle \frac{\partial b}{\partial w_n}$$

$$= \sum_{y=\pm 1} \left\langle r(\bar{x}, y)\, P(y, \bar{x})[y - \tanh(b)] \right\rangle x_n$$

(b) choosing $\delta w_n = -\eta\, r(\bar{x}, y)\,[y - \tanh(b)]\,x_n$

gives.

$$\langle \delta w_n \rangle = -\eta \langle r(\bar{x}, y)[y - \tanh(b)] \rangle x_n$$

$$= -\eta \frac{\partial \langle r \rangle}{\partial w_n}.$$

(c) An unbiased estimator is an estimator whose expected value equals the parameter being estimated. Suppose the X is an estimator for θ. Then X is an unbiased estimator if

$$\mathbb{E}[X] = \theta.$$