Homogeneous representation

$$a_j^{(1)} = \sum_{i=0}^{D} \omega_{ji}^{(1)} x_i = \underline{w}_j^{(1)T} \underline{x}$$

$$z_j = h\left(a_j^{(1)}\right)$$

"activation"

"activation function"

usually logistic sigmoid, or tanh

Homogeneous representation

$$a_k^{(2)} = \sum_{j=0}^{M} \omega_{kj}^{(2)} z_j = \underline{w}_k^{(2)T} \underline{z}$$

$$y_k = \sigma\left(a_k^{(2)}\right)$$

"activation"

"activation function"

Usually determined by the nature of the problem/ problem specifications

- nature of the data
- assumed distribution of target variable s.

→ Regression: Identity function

$$y_k = a_k$$

→ Classification:

$y_k = \sigma(a_k)$ logistic sigmoid, for binary classification;

softmax: for multi-class classification

Nature of these activation functions

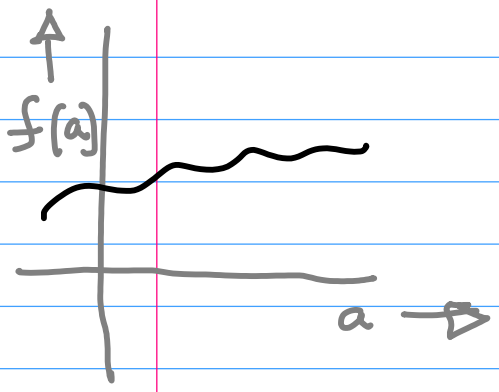
logistic sigmoid $\sigma(a) \triangleq \frac{1}{1+\exp(-a)}$

$\tanh\left(\frac{a}{2}\right) \triangleq 2\sigma(a) - 1$

softmax: relative exponential: $\exp / \sum \exp$

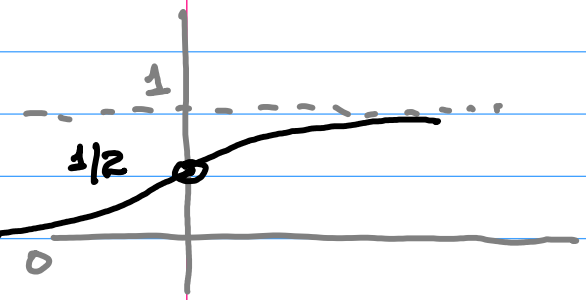
'alphabet' $\uparrow$ a/A, b/B
$\alpha\ \beta\ \gamma\ \delta \dots$
alpha, beta

$\sigma / \Sigma \rightarrow$ sigma
$\hookrightarrow s/S$
$\Sigma \rightarrow$ "sum"
$\int \rightarrow$ "S-shaped"

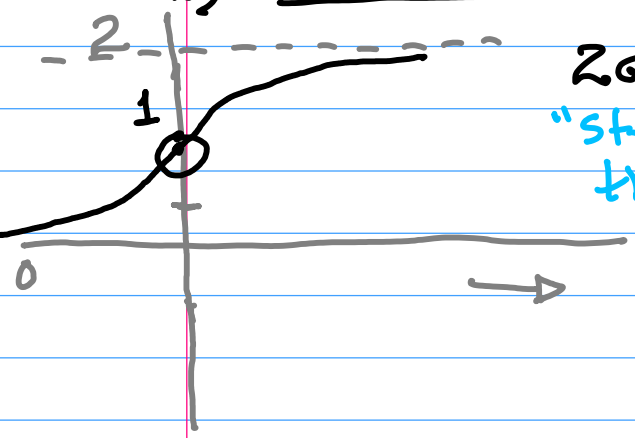


(1) $\sigma(a) = \dfrac{1}{1+\exp(-a)}$

when $a \to -\infty$, $\sigma(a) \to 0$
$a \to +\infty$, $\sigma(a) \to 1$
$a = 0$, $\sigma(a) = 1/2$

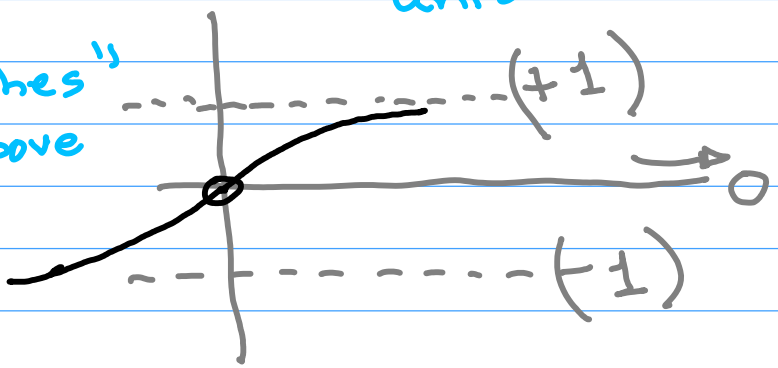


(2) $\tanh(\cdot)$



$2\sigma(a)$ "stretches" the above

$2\sigma(a) - 1$ bring the curve down by 1 unit



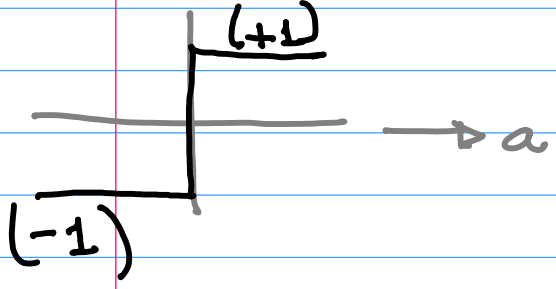
$\tanh\left(\frac{a}{2}\right) \triangleq 2\sigma(a) - 1$

why? what is $2\sigma(a) - 1$

$= \dfrac{2}{1+e^{-a}} - 1 = \dfrac{2 - 1 - e^{-a}}{1+e^{-a}} = \dfrac{(1-e^{-a})\,e^{a/2}}{(1+e^{-a})\,e^{a/2}}$

$= \dfrac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} = \tanh\left(\frac{a}{2}\right)$ by definition

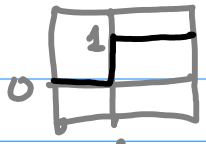
Signum function



$$f(a) = -1, a < 0$$
$$= +1, a \geq 0$$

"severe" form of the tanh( ) function.

Note on Activation Functions



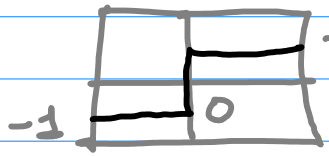
① Sigmoid ($0 \leftrightarrow 1$): smooth approximation of a unit step function

② tanh ($-1 \leftrightarrow 1$): smooth approximation to the signum function (e.g., Perceptron)

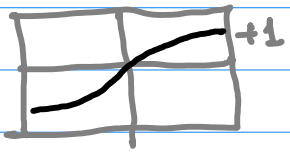
(+) smooth & differentiable

(−) gradient → 0 as the curve saturates

(−) Difficult computations with exponentials



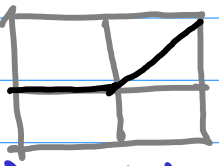
signum



tanh

↳ "vanishing gradient"

③ The ReLU function



(+) gradient = +1 for positive inputs
⇒ no vanishing gradients, as no saturation

(+) trivial to compute

(−) Doesn't work for negative inputs ⇒ no gradient-related feedback signal to help it escape from this parameter setting
↳ "Dying ReLU"

Two "simple" fixes: Leaky ReLU (LReLU) & ELU

(1) Leaky ReLU (LReLU)

$$\text{LReLU}(a, \alpha) \triangleq \max(\alpha a, a) \; ; \alpha \in (0, 1)$$

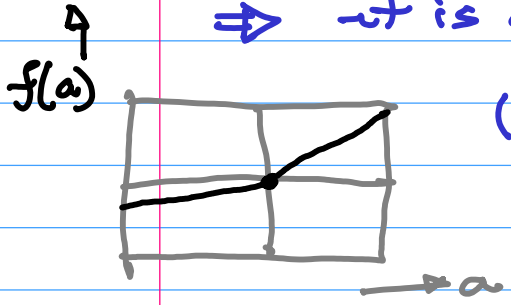
When $a \geq 0$, $\max(\alpha a, a) = a$ → ReLU

why? α: fraction $\in (0, 1)$

When $a < 0$, a: -ve and αa is a fraction of a "more negative" quantity a

$\alpha a > a$

$\Rightarrow$ it is a line with slope < 1



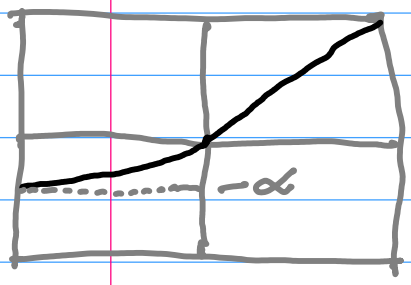
(-) continuous, but not smooth (not differentiable: $a = 0$)

"kink"

(2) ELU

$$ELU(a, \alpha) \triangleq \begin{cases} a, & a > 0 \\ \alpha(e^a - 1), & a \leq 0 \end{cases}$$

(same as ReLU / LReLU)



when $a = 0$, $ELU = \alpha \times 0 = 0$

a is "very negative"

$ELU = \alpha(\frac{1}{e^{large}} - 1)$

$\rightarrow -\alpha$

(if $\alpha = 1$, then here, $ELU \rightarrow -1$)

(-) exponential: 'more' computations

(*) Role of α: allows us to control the degree of saturation

(?)

X-OR problem (contd.) Formulate this as a regression problem and use MSE loss (Mean Square Error)

Training Set $\underline{x} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$ $X = [\underline{x}_{(1)}\ \underline{x}_{(2)}\ \underline{x}_{(3)}\ \underline{x}_{(4)}] = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$

$\underline{t} = [t_{(1)}\ t_{(2)}\ t_{(3)}\ t_{(4)}] = [0\ 1\ 1\ 0]$

MSE loss $J(\underline{w}) \triangleq \frac{1}{4}\sum_{n=1}^{4} [y(\underline{x}_{(n)}, \underline{w}) - t_{(n)}]^2$ → parameters

linear model

$$y(\underline{x}) = y(\underline{x}, \underline{w}) = \underbrace{\underline{w}^T \underline{x}}_{\text{homogeneous}} = \underbrace{\underline{w}^T \underline{x} + b}_{\text{non-homogeneous}} \quad \text{or } w_0$$

$$(*)\ \frac{\partial J(\underline{w})}{\partial \underline{w}} = \frac{1}{4} \cdot 2 \sum_{(n)} \{ y(\underline{x}_{(n)}, \underline{w}) - t_{(n)} \} \frac{\partial y}{\partial \underline{w}} (\underline{x}_{(n)}, \underline{w}) = 0$$

$$\Rightarrow \frac{\partial (\underline{w}^T \underline{x})}{\partial \underline{w}} = \underline{x} = \frac{\partial (\underline{x}^T \underline{w})}{\partial \underline{w}} \quad \text{(Rule)}$$

$$\Rightarrow \boxed{\frac{1}{N} \sum_{(n)} \{ y(\underline{x}_{(n)}, \underline{w}) - t_{(n)} \} \underline{x}_{(n)} = \underline{0}} \quad (1)$$

$$(*)\ \frac{\partial J(\underline{w})}{\partial b} = \frac{1}{4} \cdot 2 \sum_{(n)} \{ y(\underline{x}_{(n)}, \underline{w}) - t_{(n)} \} \cdot 1 = 0$$

$$\frac{\partial}{\partial b} (\underline{w}^T \underline{x} + b) = 1$$

$$\Rightarrow \boxed{\sum_{(n)} \{ \underline{w}^T \underline{x}_{(n)} + b - t_{(n)} \} = 0} \quad (2)$$

(1) $\Rightarrow \{w_2(0)+w_1(0)+b-0\}\begin{bmatrix}0\\0\end{bmatrix}$

$+\{w_2(0)+w_1(1)+b-1\}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$

$+\{w_2(1)+w_1(0)+b-1\}\begin{bmatrix}1\\0\end{bmatrix}$

$+\{w_2(1)+w_1(1)+b-0\}\begin{bmatrix}1\\1\end{bmatrix}$

$\Rightarrow \begin{bmatrix} w_2+b-1+w_2+w_1+b \\ w_1+b-1+w_2+w_1+b \end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$

$\Rightarrow \begin{bmatrix} 2w_2+w_1+2b \\ 2w_1+w_2+2b \end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix} \Rightarrow \begin{cases} 2w_2+w_1+2b=1 \\ 2w_1+w_2+2b=1 \end{cases}$ — (1)

(2) $\Rightarrow w_2(0)+w_1(0)+b-0$

$+w_2(0)+w_1(1)+b-1$

$+w_2(1)+w_1(0)+b-1$

$+w_2(1)+w_1(1)+b-0=0$

$\Rightarrow 4b+2(w_2+w_1)=2 \Rightarrow 2b+(w_2+w_1)=1$ — (2)

Put (2) in (1) to get

$2w_2+w_1+1-(w_2+w_1)=1$

$2w_1+w_2+1-(w_2+w_1)=1$

$\Rightarrow w_2=0,\ w_1=0,\ b=0.5$

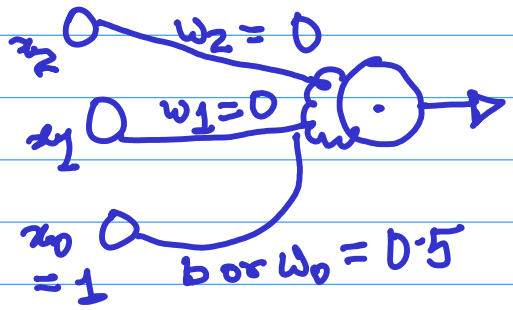
$\underline{w}=\underline{0}$, $b=0.5$

$y_{(n)}=\underline{w}^T\underline{x}_{(n)}+0.5$

whatever is $\underline{x}$ → the output should be 0.5

What did we try?

$y = \sigma(a)$ = a unit function



$y = \underline{w}^T \underline{x} + b$ (non-homogeneous)

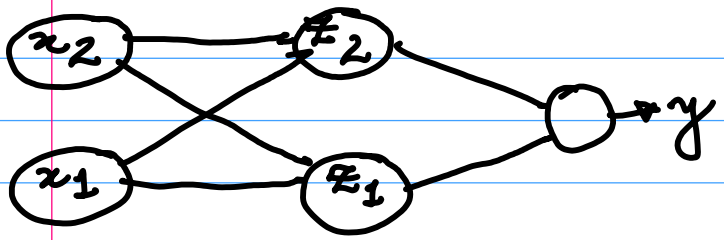
$= w_2 x_2 + w_1 x_1 + b$

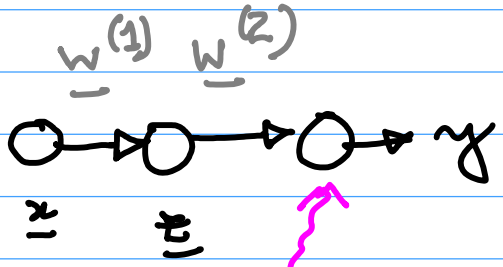
$= 0.5$ always.

Some points: →

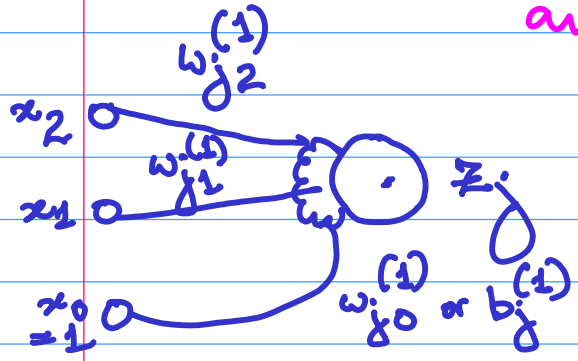
- This is a simple example with regression (though the problem is actually one of classification)
- This is one of the myriad possible solutions
- Unnecessarily trying to build a classifier out of a regressor (that too, for binary inputs)
- Handcrafted

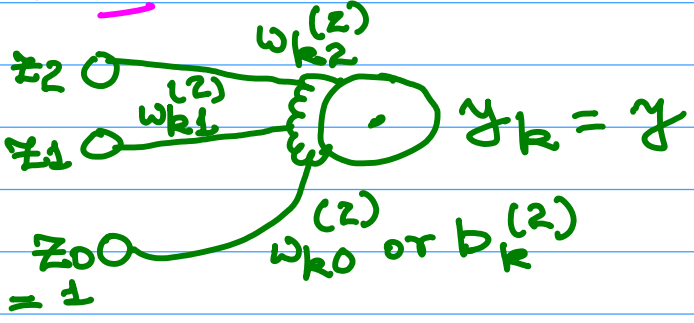
ONE possible solution: –





regressor applied to $\underline{z} = \underline{\phi}(\underline{x})$, and not $\underline{x}$ itself





$z_j = h(a_j)$

choose $h(\cdot)$ to be a ReLU

$a_j = \underline{w}_{j\cdot}^{(1)T} \underline{x} = \underline{w}_{j\cdot}^{(1)T} \underline{x} + b_j^{(1)}$ or $w_{j0}^{(1)}$

homogeneous

non-homogeneous

$y_k = y = \sigma(a_k)$

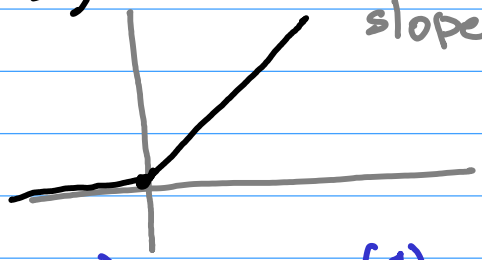
choose σ to be the unit function

$y_k = y = \underline{w}_k^{(2)T} \underline{z} + b_k^{(2)}$

non-homogeneous

ReLU as an activation function

(often the choice in most deep feedforward networks)



$h(a) = \max\{0, a\}$

piecewise linear

$a_j = \underline{w}_{j\cdot}^{(1)} \underline{x} = b_j^{(1)}$ or $w_{j0}^{(1)}$

$z_j = h(a_j) = \max\{0, a_j\} = \max\{0, \underline{w}_{j\cdot}^{(1)T} \underline{x} + b_j^{(1)}\}$

Handcrafted example:— (all these values are given!)

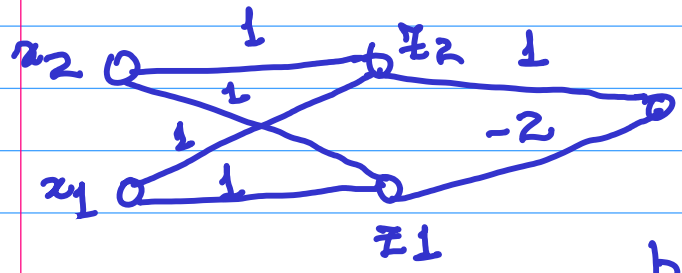
$\underline{w}_{j\cdot}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\underline{b}^{(1)} = \begin{bmatrix} b_2^{(1)} \\ b_1^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

$\underline{w}^{(2)} = \begin{bmatrix} w_2^{(2)} \\ w_1^{(2)} \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$; $b^{(2)} = 0$

X = matrix of all possible inputs

$$\underline{x} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\underline{x}_{(1)} \quad \underline{x}_{(2)} \quad \underline{x}_{(3)} \quad \underline{x}_{(4)}$$



$$\underline{b}^{(1)} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

(given)

$$b^{(2)} = 0$$

$$a_j = \underline{w}_j^{(1)T} \underline{x} + b_j^{(1)} \quad \text{or } w_{j0}^{(1)}$$

$$a_2 = \underline{w}_2^{(1)T} \underline{x} + b_2^{(1)}$$

$$a_1 = \underline{w}_1^{(1)T} \underline{x} + b_1^{(1)}$$

$$\underline{a}_{(i)} = \begin{bmatrix} \underline{w}_2^{(1)T} \\ \underline{w}_1^{(1)T} \end{bmatrix} \underline{x}_{(i)} + \begin{bmatrix} b_2^{(1)} \\ b_1^{(1)} \end{bmatrix}$$

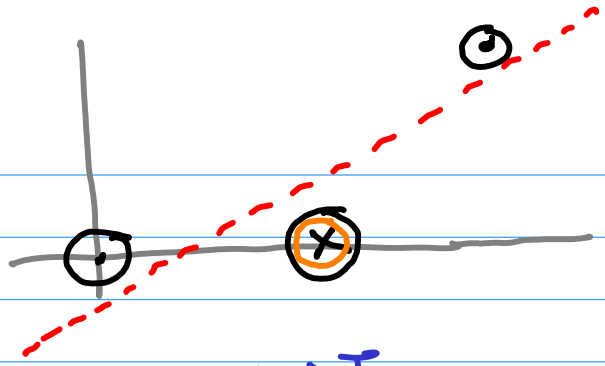
$$\Rightarrow \underline{a}_{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad \underline{z}_{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\underline{a}_{(2)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \underline{z}_{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\underline{a}_{(3)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \underline{z}_{(3)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\underline{a}_{(4)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \underline{z}_{(4)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$z_j = h(a_j) \Rightarrow \underline{z} = \begin{bmatrix} z_2 \\ z_1 \end{bmatrix} = \begin{bmatrix} h(a_2^{(1)}) \\ h(a_1^{(1)}) \end{bmatrix} = \begin{bmatrix} \max(a_2, 0) \\ \max(a_1, 0) \end{bmatrix}$$



$$a_k^{(2)} = \underbrace{\underline{w}_k^{(2)T} \underline{z}}_{\text{homogeneous}} = \underbrace{\underline{w}_k^{(2)T} \underline{z} + b_k^{(2)}}_{\text{non-homogeneous}}$$

given

$$a_{(n)}^{(2)} = \underbrace{\begin{bmatrix} 1 & -2 \end{bmatrix} \underline{z}_{(n)}}_{\underline{w}_k^{(2)}} + \underbrace{b^{(2)}}_{=0}$$

$$\underline{w}_k^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \quad b_k^{(2)} = 0$$

$$a_{(1)}^{(2)} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$$

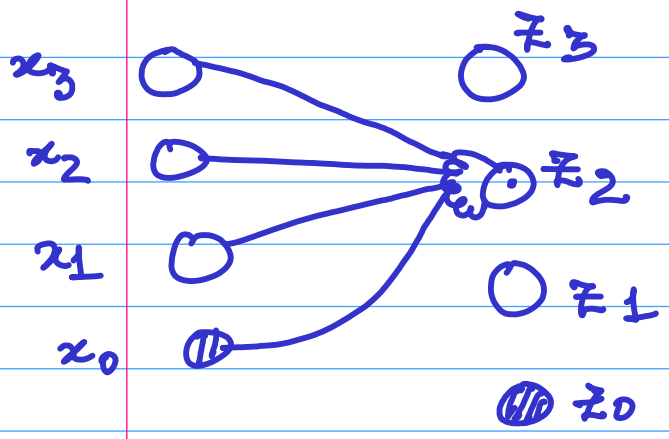
$$a_{(2)}^{(2)} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

$$a_{(3)}^{(2)} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

$$a_{(4)}^{(2)} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 0$$

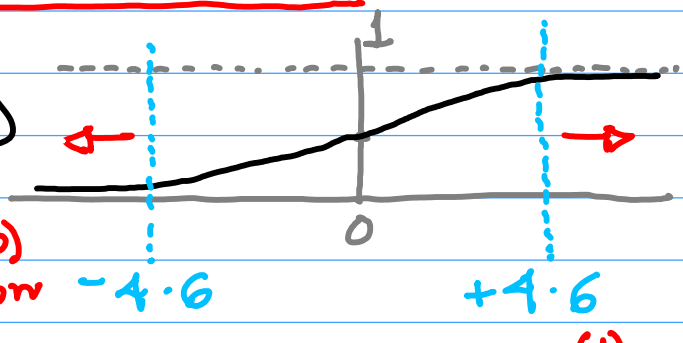
this is what we wanted as the XOR outputs

Notation (Vector-Matrix Notation)



SOME HANDCRAFTED NN EXAMPLES

$$h(a) = \frac{1}{1+e^{-a}} \quad \text{(sigmoid)}$$



$e^{4.6} \approx 100$

When $a = -4.6$, $h(a) = \frac{1}{1+e^{4.6}} \approx \frac{1}{1+100} \approx 0.01$

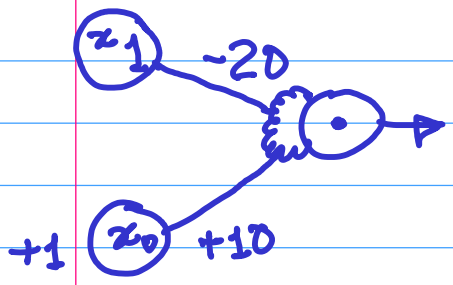
When $a = +4.6$, $h(a) = \frac{1}{1+e^{-4.6}} \approx \frac{1}{1+1/100} = \frac{100}{101} \approx 0.99$

Example 1 : NOT

$a = \underline{w}^T \underline{x} + b$

$= -20x_1 + 10$

x_1	$\bar{x}_1$
0	1
1	0



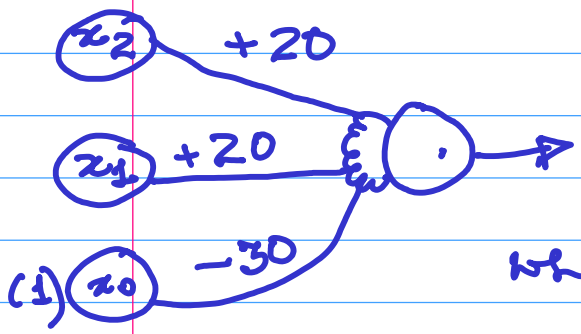
$x_1 = 0$, $a = 10$, $h(+10) \approx 1$

$x_1 = 1$, $a = -10$, $h(-10) \approx 0$

Example 2 : AND

x_2	x_1	$x_2 \cdot x_1$
0	0	0
0	1	0
1	0	0
1	1	1

$a = \underline{w}^T \underline{x} + b$



$a = 20x_2 + 20x_1 - 30$

When $x_2 = 0$, $x_1 = 0$: $a = -30$

$h(-30) \approx 0$

When $x_2 = 0$, $x_1 = 1$, $a = 20(0) + 20(1) - 30 = -10$, $h(-10) \approx 0$