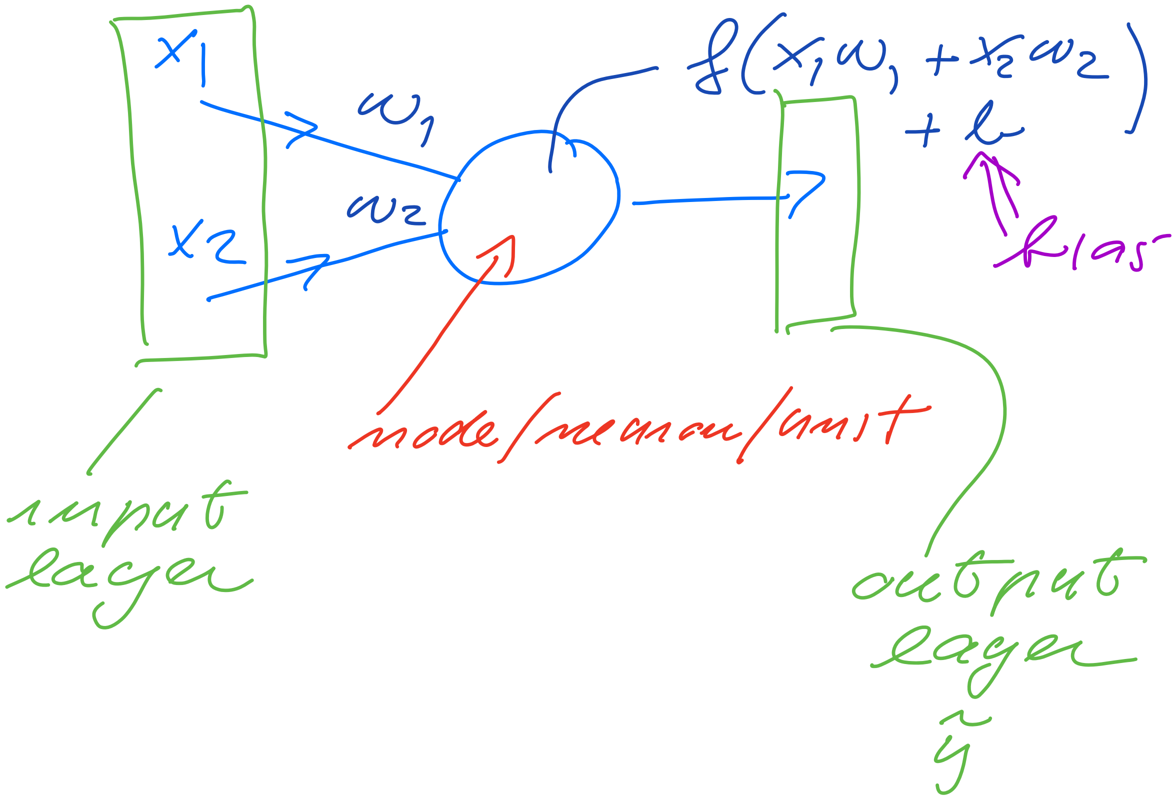
Lecture October 8

simple perception



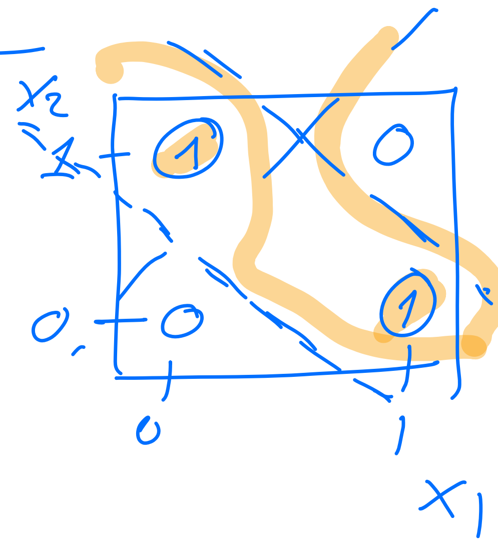
XOR, OR, AND

$$X = [x_1 \, x_2]^T = \{ [0,0]^T, [0,1]^T, [1,0]^T [1,1] \}$$

XOR gate

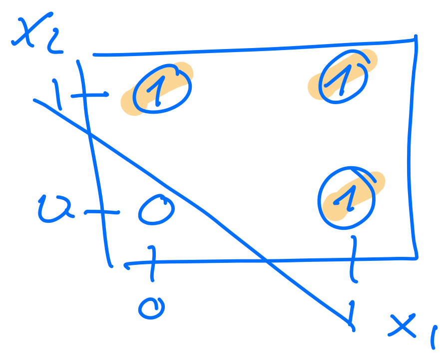
XOR-gate

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



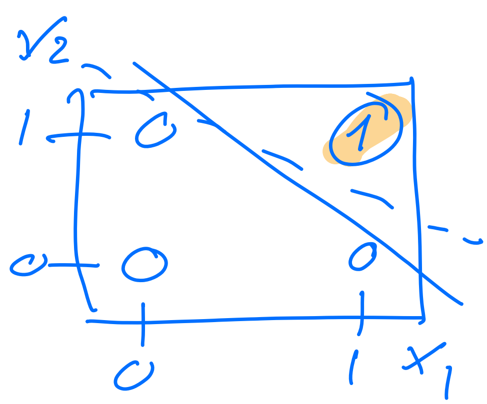
OR-gate

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



AND gate

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



$$\tilde{y} = x_1 w_1 + x_2 w_2 + b$$

$$= x^T w + b$$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

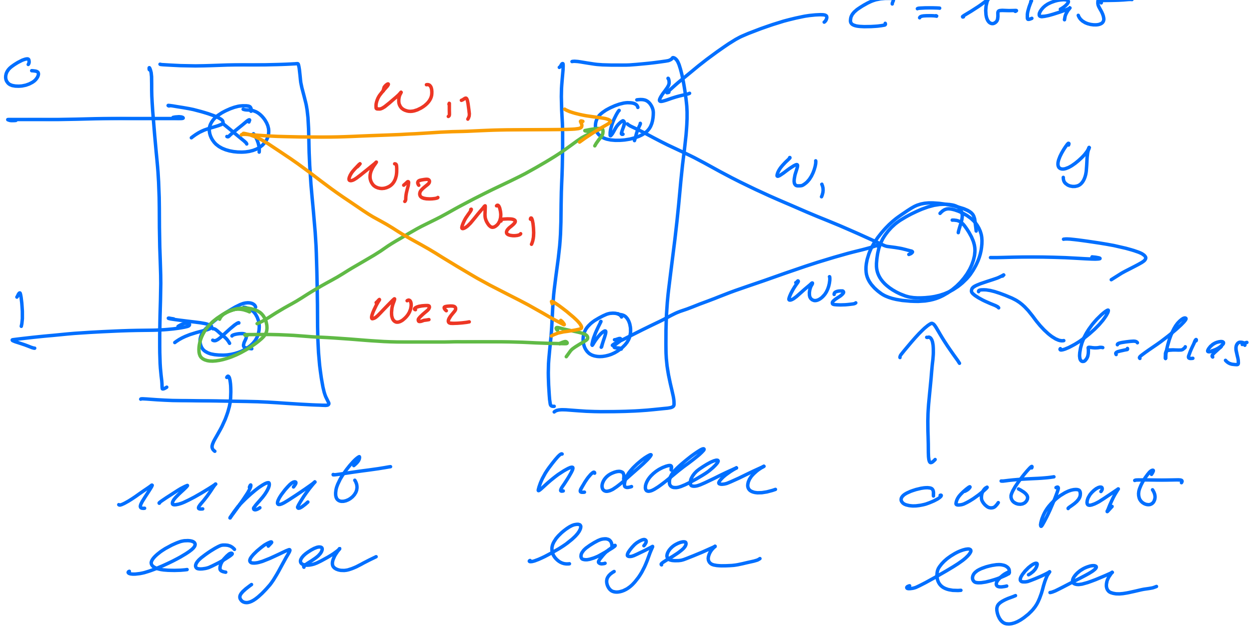
$$X^T X = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

$$\Theta = \{w, b\} =$$

$$(\beta) = \Theta = [\underbrace{\tfrac{1}{2}}_{= b}, 0, 0]^T$$

$$y\_predict = X\Theta = [\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}]^T$$

NN with one hidden layer



$$h = [h_1, h_2]^T \qquad W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$= \underline{f^{(1)}(x; W, c)}$$

$$\hat{y} = f^{(2)}(h; w, b) \qquad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= f^{(2)}(f^{(1)}; w, b)$$

$$W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$c = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$w = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \qquad b = 0$$

$$X = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \end{array}$$

$$z = XW + c$$

$$f^{(1)}(z) = h$$

$$XW + c = \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} = z$$

$$h = \boxed{f^{(1)}(z)}$$

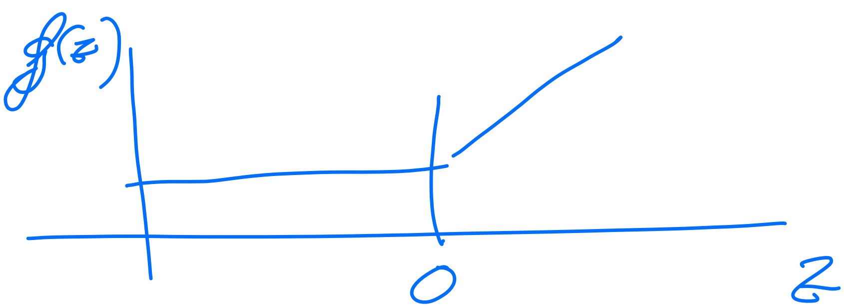
$$f^{(1)} = ?$$

$$XW = \begin{bmatrix} \boxed{\begin{matrix} 0 \\ 1 \\ 1 \\ 2 \end{matrix}} & \begin{matrix} 0 \\ 1 \\ 1 \\ 2 \end{matrix} \end{bmatrix}$$

$f^{(1)}$ = sigmoid, tanh, ReLU, eLU, ...

ReLU = Rectified Linear Unit

$$f(z) = \max\{0, z\}$$



$$h = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

multiply with $w = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$
and b ($b = 0$)

$\Rightarrow$ output =

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$ = correct output for XOR gate.

Feed Forward pass.

But do we train it?

Backpropagation algo:
(chain rule)

MSE

$$C(W, b) = C(\theta)$$ target

$$= \frac{1}{2} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

output from

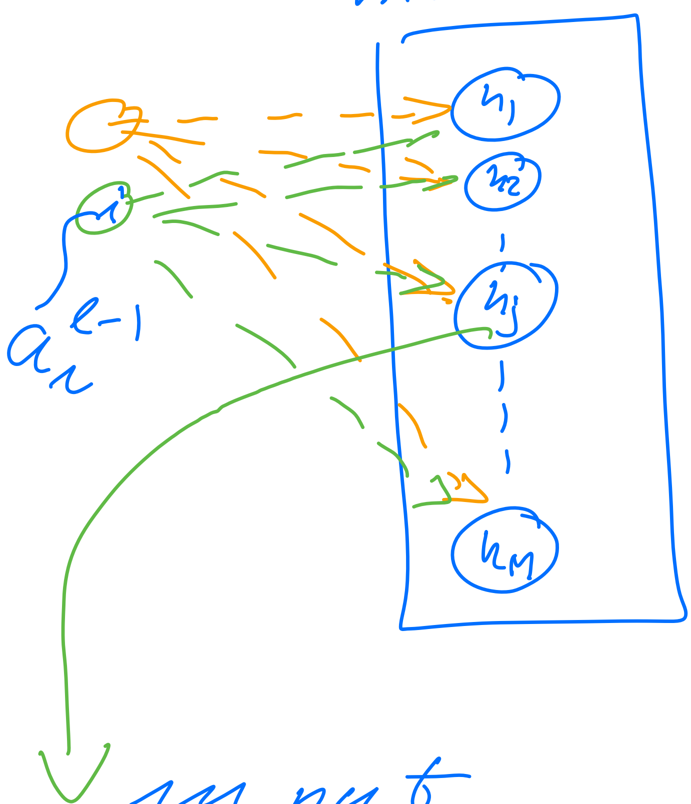
$$\{ \quad \frac{1}{2} \sum (t_i - y_i)^2$$

$$\left\{ \begin{aligned} &\tfrac{1}{2}\textstyle\sum (\ldots)\\ &= \tfrac{1}{2}\sum_i \left(a_j^L - y_i\right)^2 \end{aligned} \right.$$ your NN

Definitions:

Layer $-l-$ and node j

hidden layer



input

output from node i in layer $l-1$

$$z_j^l = \sum_{i=1}^{M_{l-1}} w_{ij}^l\, a_i^{l-1} + b_j^l$$

weights connect layer l with

layer $(l-1)$

output from node -j- in layer -l-

$$a_j^l = f(z_j^l)$$

↑ activation function

$$f(z_j^l) = \frac{1}{1+e^{-z_j^l}}$$

$$\frac{\partial C}{\partial w} = \, ? \quad , \quad \frac{\partial C}{\partial b}$$

intermediate steps:

$$\frac{\partial z_j^l}{\partial w_{ij}^l} = a_i^{l-1}$$

∂z_j^l $\quad$ l

$$\frac{\quad_j}{\partial a_k^{l-1}} = w_{jk}$$

$$\frac{\partial a_j^l}{\partial z_j^l} = f(z_j^l)\left(1 - f(z_j^l)\right)$$

Output layer $l = L$

$$\frac{\partial C}{\partial w_{jk}^l} = \underline{(a_j^L - \overset{\text{target}}{y_j})\,\frac{\partial a_j^L}{\partial w_{jk}^L}}$$

output a_j^L $(\hat{y}_j)$

$$\frac{\partial a_j^L}{\partial w_{jk}^L} = \frac{\partial a_j^L}{\partial z_j^L}\,\frac{\partial z_j^L}{\partial w_{jk}^L}$$

$$= a_j^L(1 - a_j^L)\,a_k^{L-1}$$

$$\Rightarrow$$

$$\frac{\partial C}{\partial w_{jk}^L} = (a_j^L - y_j)(1 - a_j^L) \times a_j^L \times a_k^{L-1}$$

<u>Define</u>

$$\delta_j^L = \underbrace{a_j^L(1 - a_j^L)}(a_j^L - y_j)$$

$$= f'(z_j^L)\,\frac{\partial C}{\partial a_j^L}$$

$$\boxed{\frac{\partial C}{\partial w_{jk}^L} = \delta_j^L\, a_k^{L-1}}$$

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\,\frac{\partial a_j^L}{\partial z_j^L}$$

$$\delta_j^L = \frac{\partial C}{\partial b_j^L}\,\frac{\partial b_j^L}{\partial z_j^L} = \frac{\partial C}{\partial b_j^L}$$

Backprop : $L \Rightarrow l$

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l}$$

$$z_j^{l+1} = \sum_i w_{ij}^{l+1} \underbrace{a_i^l}_{f(z_i^l)} + b_j^{l+1}$$

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{kj}^{l+1} f'(z_j^l)$$

$$\delta_j^l = \sum_k \delta_k^{l+1} w_{kj}^{l+1} f'(z_j^l)$$

$$w_{jk}^l \leftarrow w_{jk}^l - \eta\, \delta_j^l a_k^{l+1}$$

$$b_j^l \leftarrow b_j^l - \eta\, \delta_j^l$$