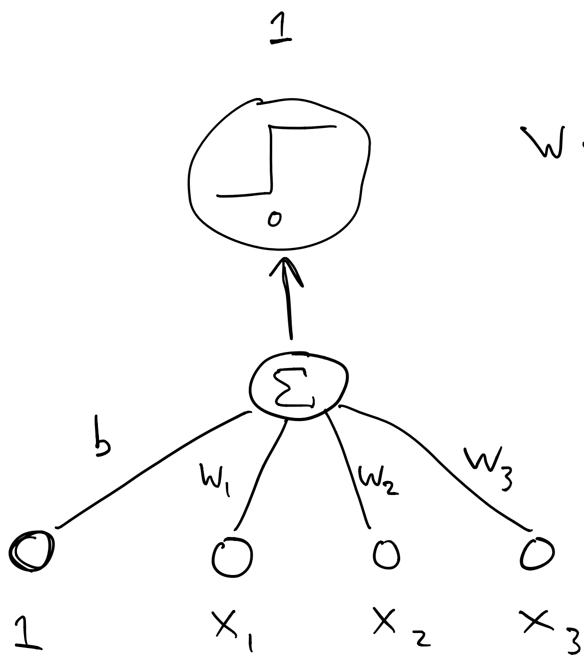
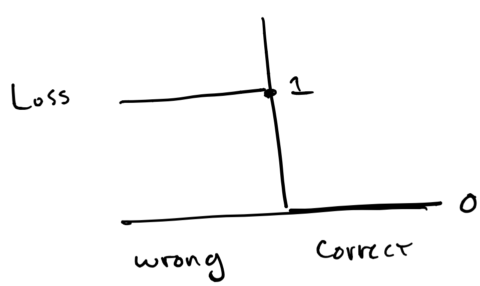
Perceptron

$$\hat{y} = \begin{cases} 1 & \text{if } w \cdot x > \emptyset \\ -1 & \text{otherwise} \end{cases}$$

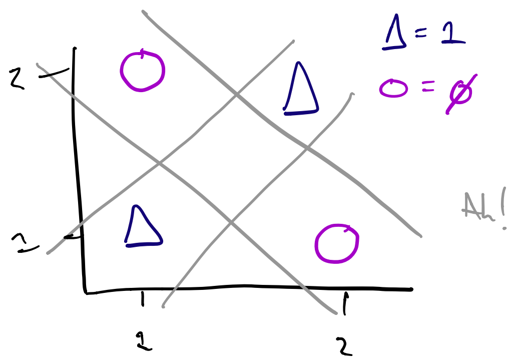


$$w \cdot x = 1 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

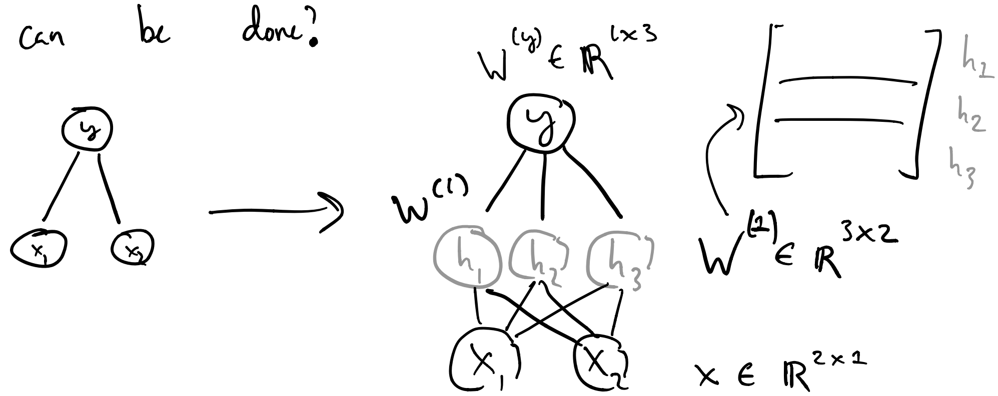
Problems with 0/1 loss?



Problems with Linear Models



What can be done?



$$h = W^{(1)} \cdot x \in \mathbb{R}^{3 \times 1}$$

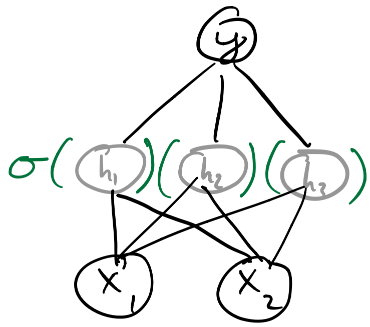
$$y = W^{(y)} \cdot h \in \mathbb{R}^{1}$$

But this is still a linear model!
Just weirdly parametrized.

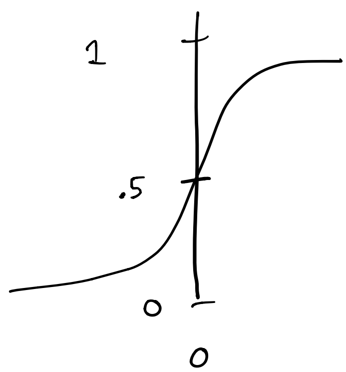
So. How do we learn non-linear models?

Activations.

Introduce non linearity $\sigma(\ )$ activation function



What to use for σ?



$\boxed{\text{Sigmoid}}$ is a common choice

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)\,(1 - \sigma(x)) \quad \star$$

$$\star \quad (\underbrace{1+e^{-x}}_{u})^{-1}$$

$\boxed{\text{Rectified Linear Units}}$ are popular

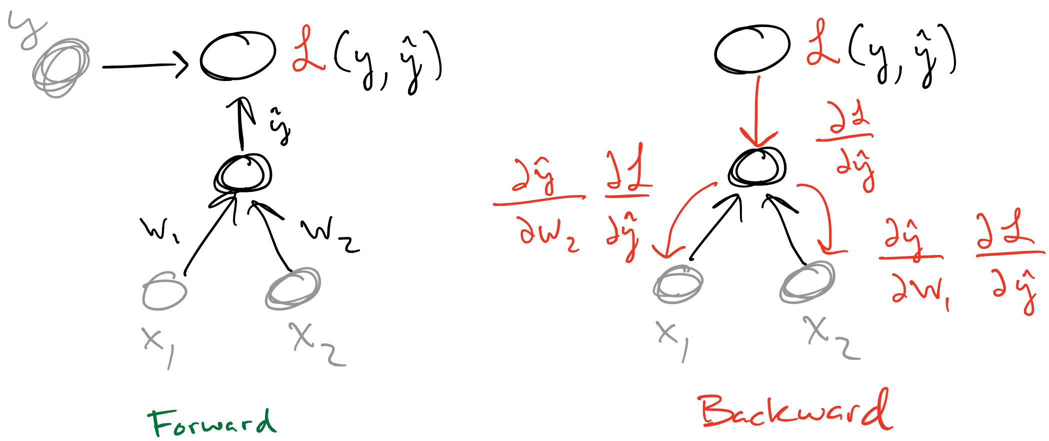
(ReLUs)

$$\sigma(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d}{dx}\text{ReLu}(x) = \ ? \quad \begin{array}{l} 1 \text{ if } x>0 \\ \text{undef @ } 0 \end{array}$$

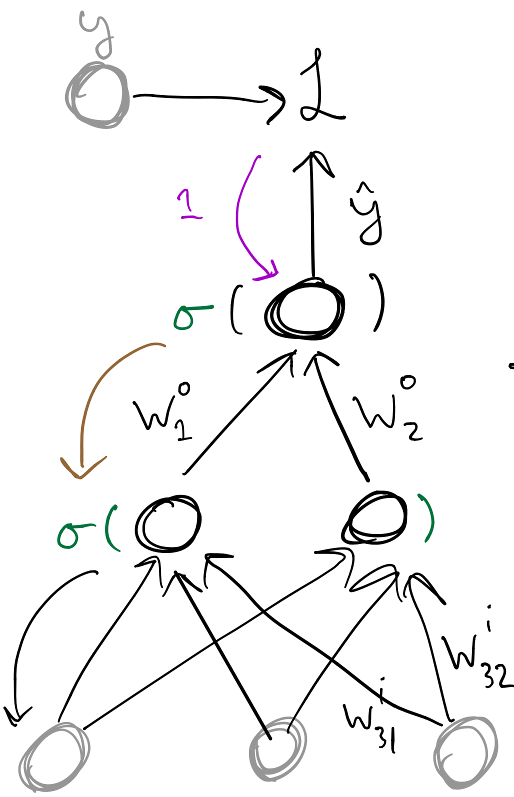
$\emptyset$ elsewhere.

BACKPROP



$$\frac{\partial}{\partial \hat{y}} \mathcal{L} = \frac{d}{d\hat{y}} -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

$$= \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$$



BCE:

$$-y \lg \hat{y} - (1-y)\lg(1-\hat{y})$$

$$\hat{y} = \sigma(z^o)$$

$$z^o = h \underset{(2\times 1)}{W^o}$$

$$h = \sigma(z^h)$$

$$z^h = \underset{(1\times 3)}{x}\,\underset{(3\times 2)}{W^i}$$

1 $$\frac{\partial}{\partial \hat{y}}\mathcal{L} = \frac{\partial}{\partial \hat{y}}\left(-y\lg\hat{y} - (1-y)\lg(1-\hat{y})\right)$$

1a $$\frac{\partial}{\partial z^o}\mathcal{L} = \frac{\partial \hat{y}}{\partial z^o}\frac{\partial}{\partial \hat{y}}\mathcal{L} = \frac{\partial}{\partial z^o}\sigma(z^o)\,\frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$= \sigma(z^o)(1-\sigma(z^o))\,\frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$= \cancel{\hat{y}(1-\hat{y})}\,\frac{\hat{y}-y}{\cancel{\hat{y}(1-\hat{y})}} = \delta^o$$

$$2 \quad \frac{\partial}{\partial W^{(o)}} \mathcal{L} = \frac{\partial z^o}{\partial W^{(o)}} \frac{\partial}{\partial z^o} \mathcal{L} = h(\hat{y} - y) = h\delta^o$$

$$2a \quad \frac{\partial}{\partial h} \mathcal{L} = \frac{\partial z^o}{\partial h} \frac{\partial}{\partial z^o} \mathcal{L} = W^{(o)}(\hat{y} - y)$$

$$\delta^h = \frac{\partial}{\partial z^h} h \frac{\partial}{\partial h} \mathcal{L} = \sigma(z^h)(1 - \sigma(z^h))W^{(o)}(\hat{y} - y)$$

$$3 \quad \frac{\partial \mathcal{L}}{\partial W^{(i)}} = \frac{\partial}{\partial W^{(i)}} z^h \cdot \delta^h$$

$$= x \cdot \underbrace{\sigma(z^h)(1 - \sigma(z^h))(\hat{y} - y)}_{\delta^h}$$