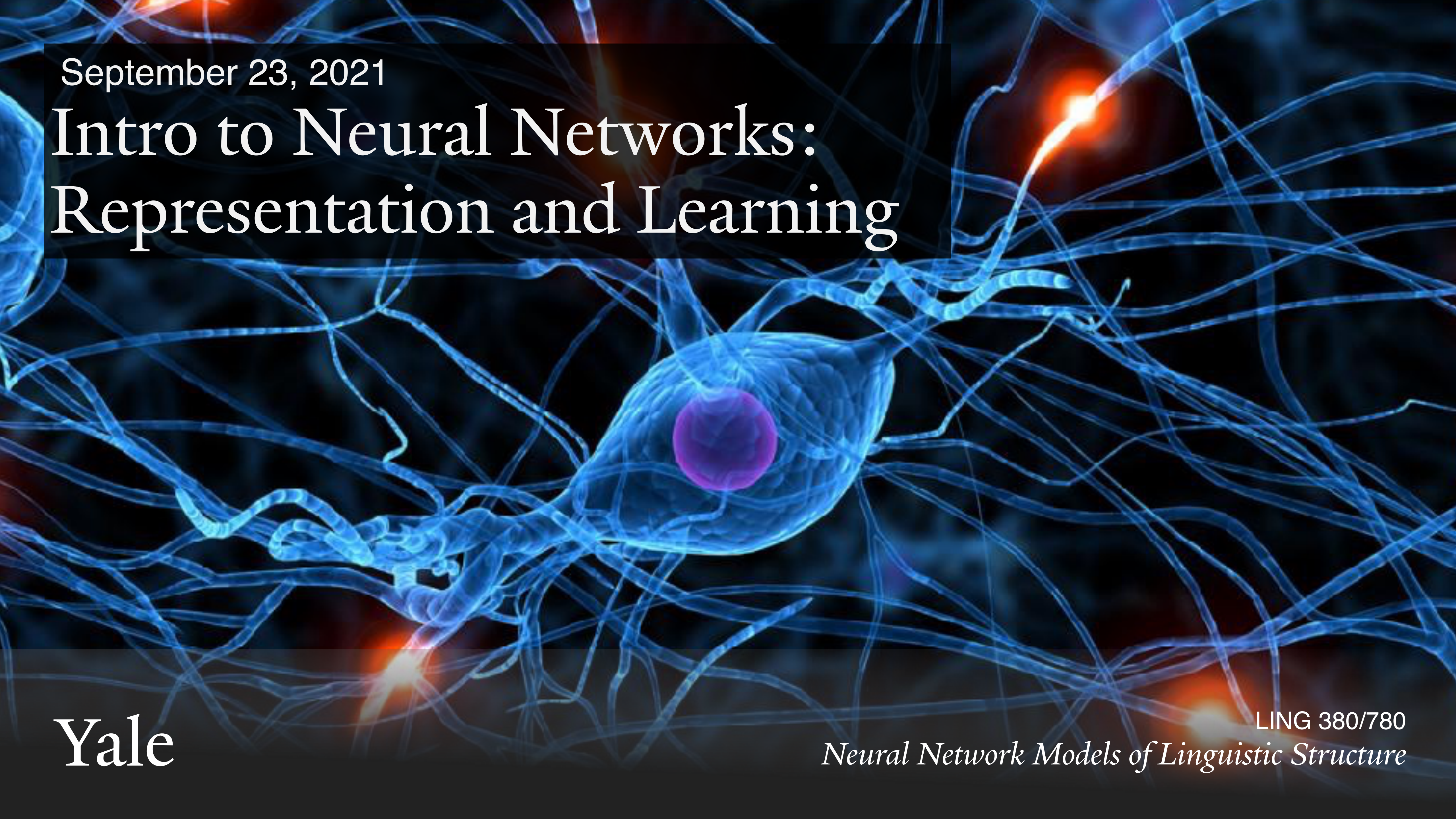


September 23, 2021

Intro to Neural Networks: Representation and Learning

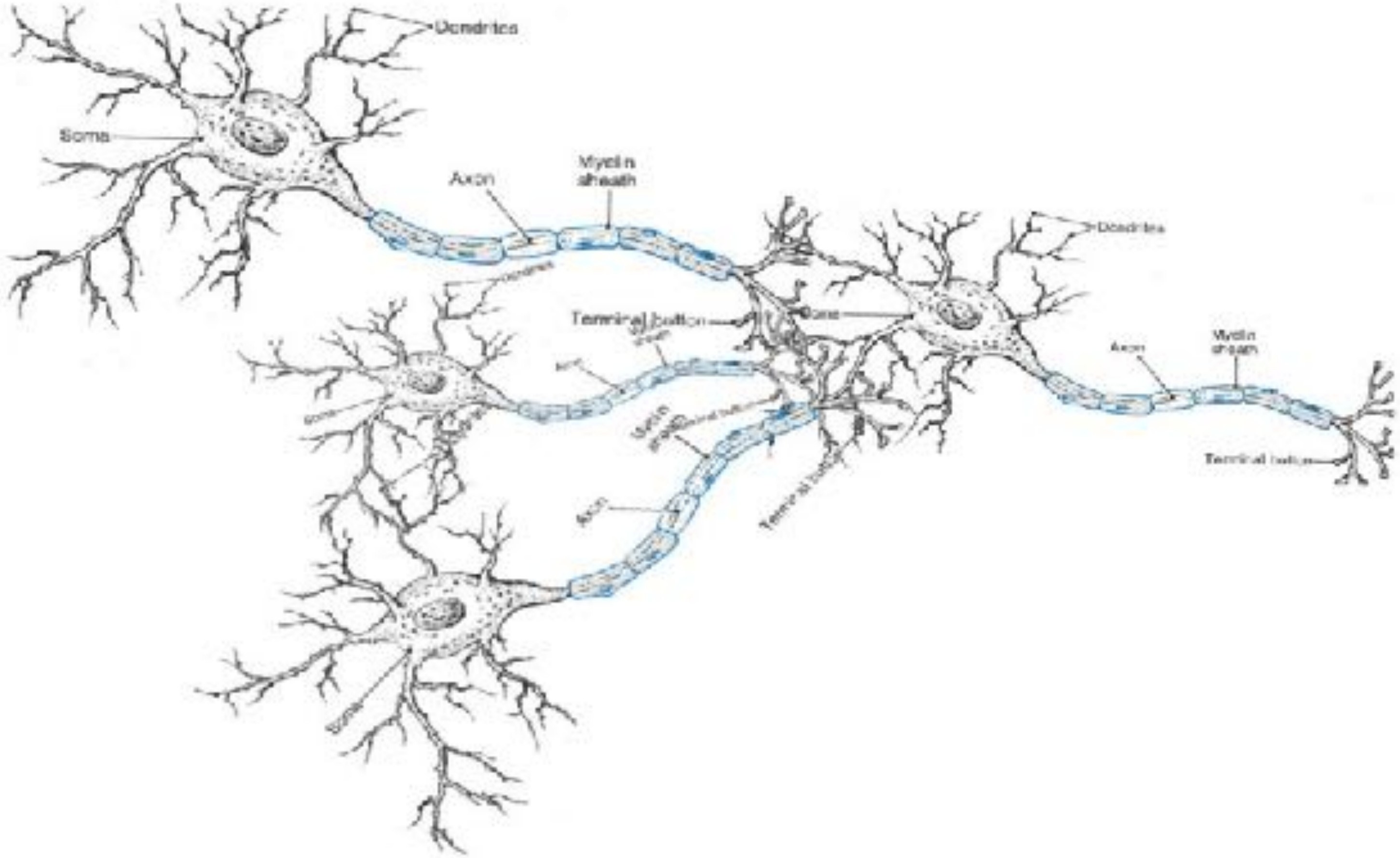
Yale

LING 380/780

Neural Network Models of Linguistic Structure

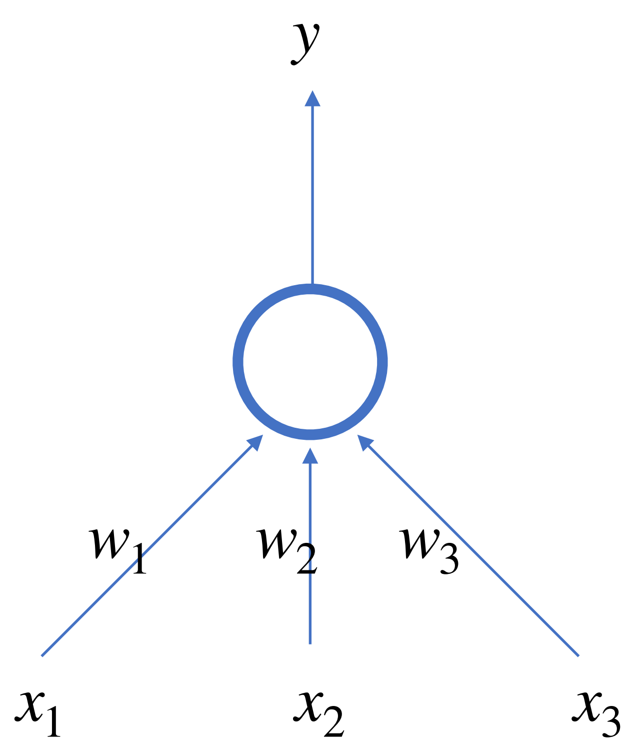
Perceptrons (Rosenblatt 1957)

- A model of neural activitation



Perceptrons (Rosenblatt 1957)

- A model of neural activation



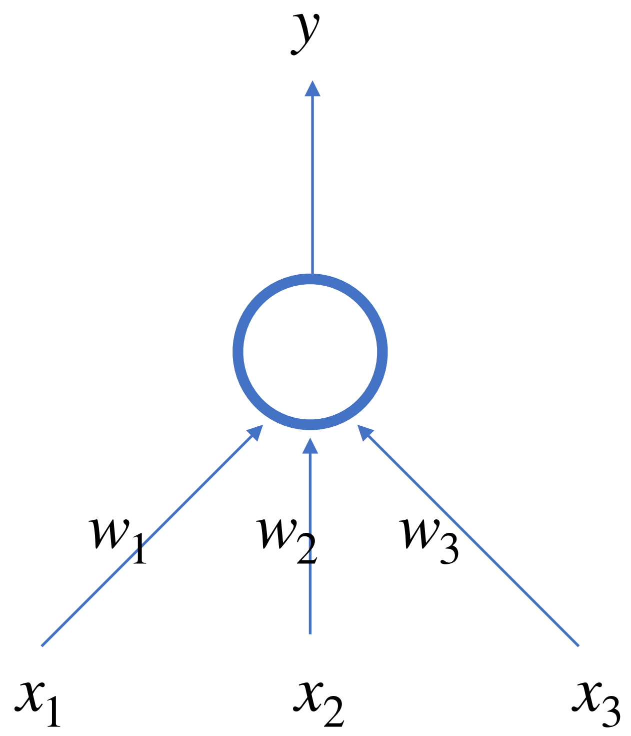
$$y = f\left(\sum w_i x_i\right)$$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$y = f(w^\top x)$$

Perceptrons (Rosenblatt 1957)

- A model of neural activation



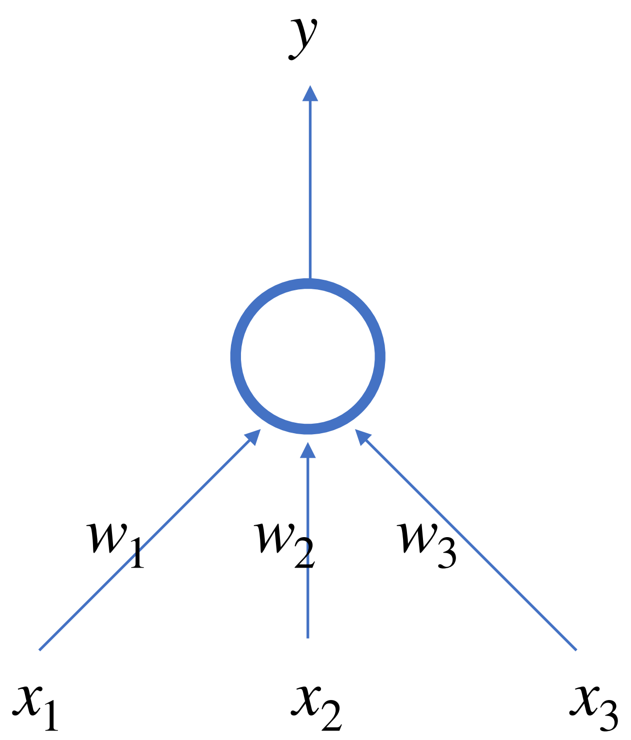
With identity activation function:

$$f(x) = x$$

we get linear regression model

Perceptrons (Rosenblatt 1957)

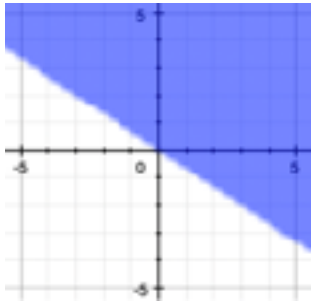
- A model of neural activation



Biologically-inspired activation function:

$$f(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{o.w.} \end{cases}$$

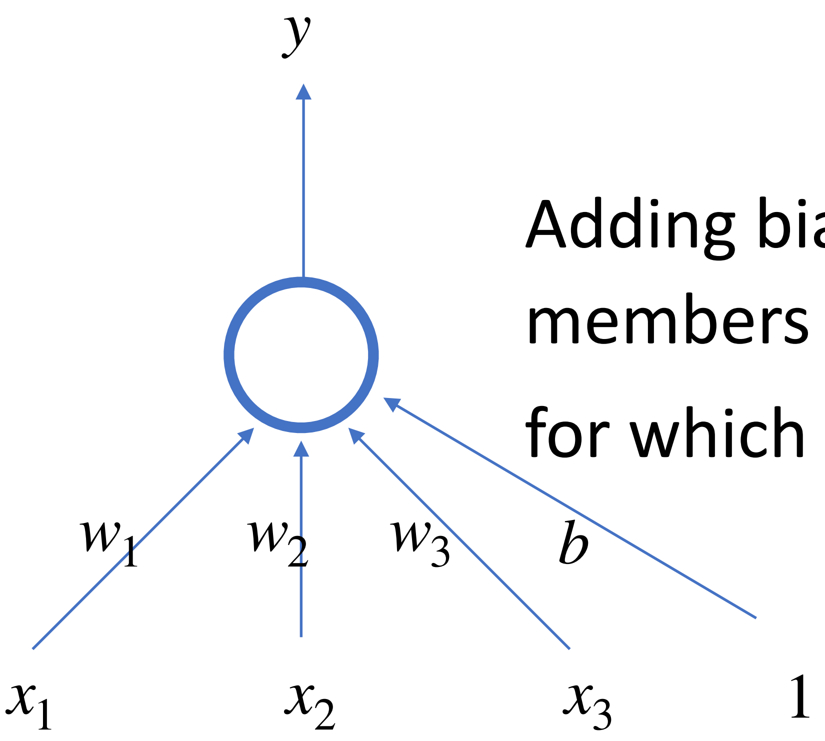
This yields a binary classifier: members of the class are those x for which $w^\top x > 0$.



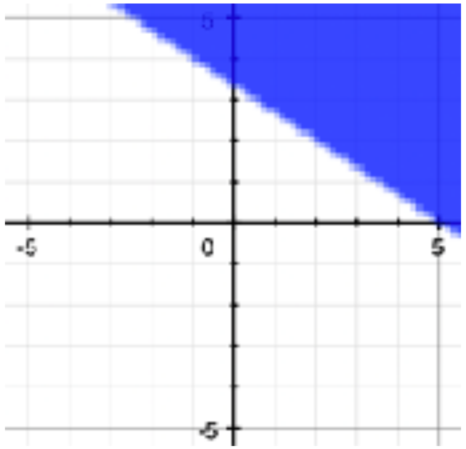
Perceptrons (Rosenblatt 1957)

- A model of neural activation

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ b \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix}$$

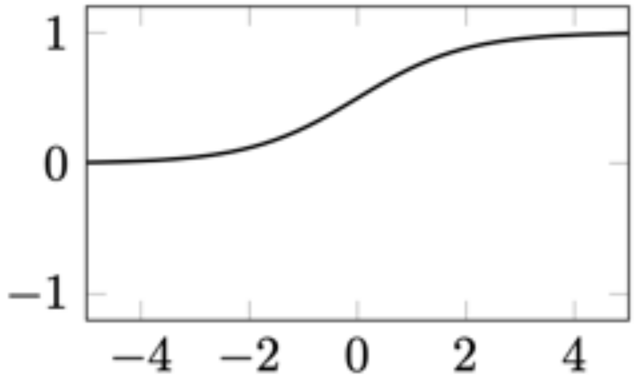


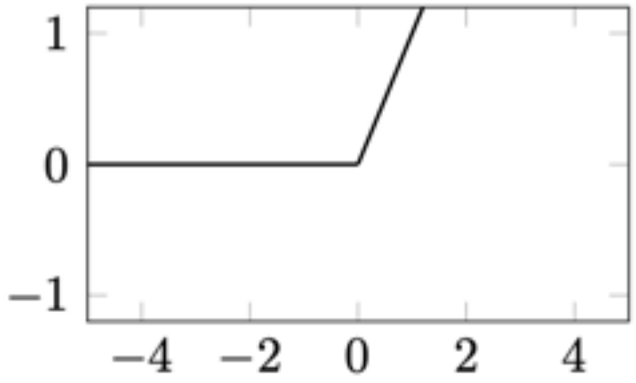
Adding bias yields a binary classifier: members of the class are those x for which $w^\top x > -b$.



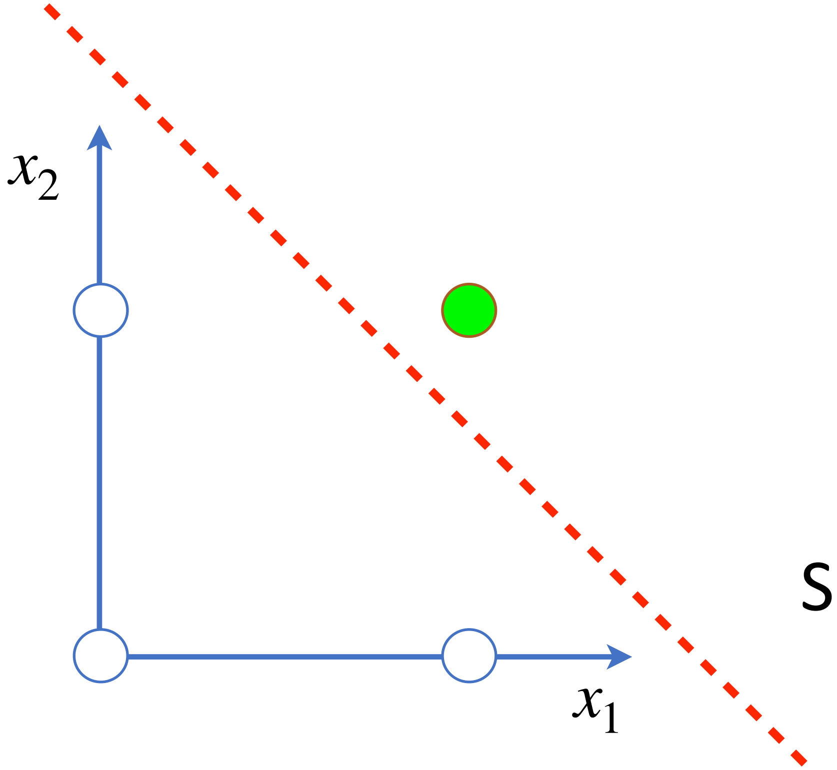
Perceptrons (Rosenblatt 1957)

- Other activation functions
  - **Sigmoid**
    (continuous and differentiable version of step function)
  - **ReLU**(x) = max(x,0)
    (rectified linear unit)





Perceptron classification



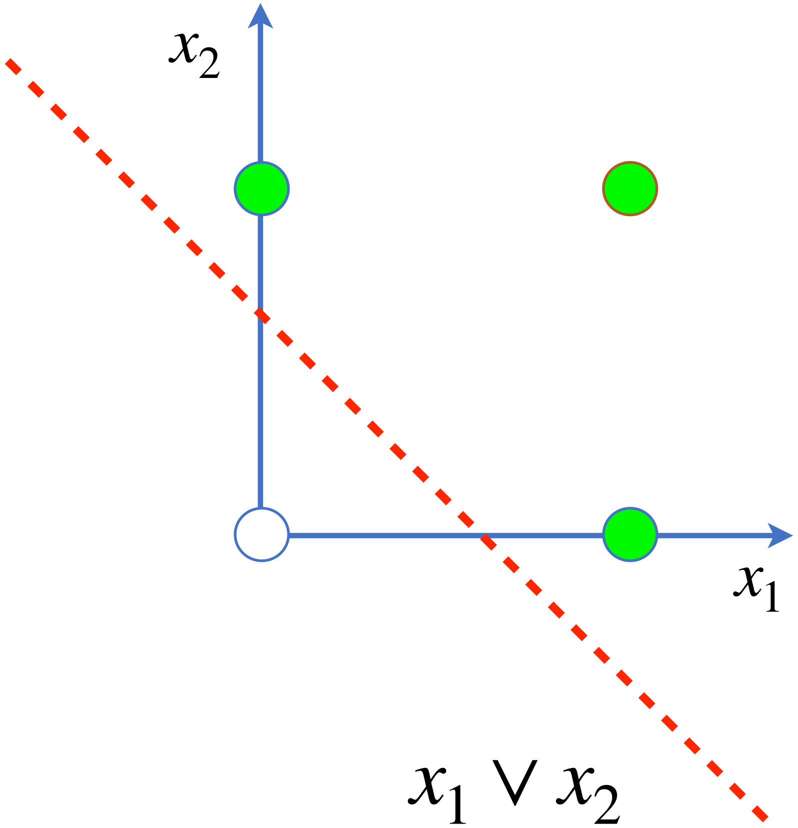
$$x_1 \wedge x_2$$

$$\theta = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1.5 \end{bmatrix}$$

Separating hyperplane

$$x_1 + x_2 > 1.5$$

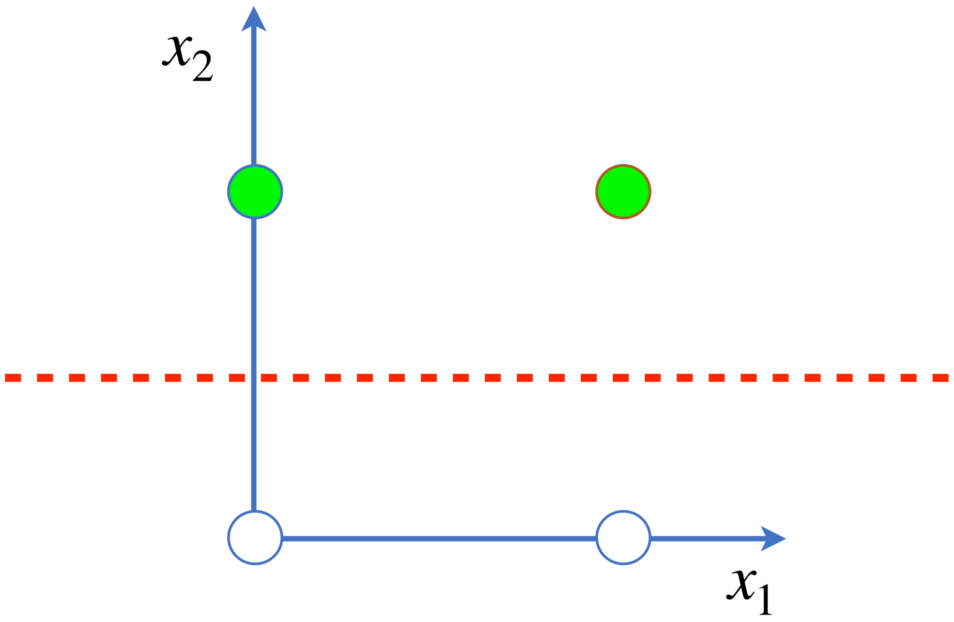
Perceptron classification



$$\theta = \begin{bmatrix} 1 \\ 1 \\ -.5 \end{bmatrix}$$

$$x_1 + x_2 > .5$$

Perceptron classification

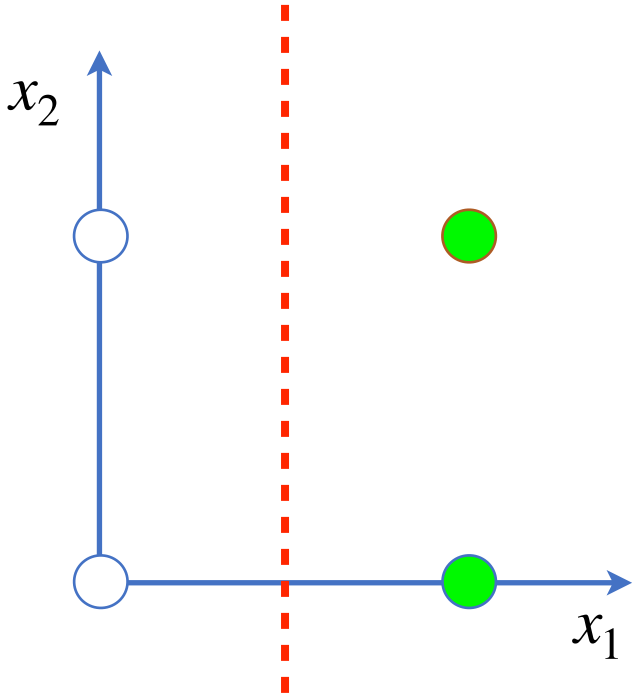


$$\theta = \begin{bmatrix} 0 \\ 1 \\ -.5 \end{bmatrix}$$

$x_2 > .5$

x_2

Perceptron classification

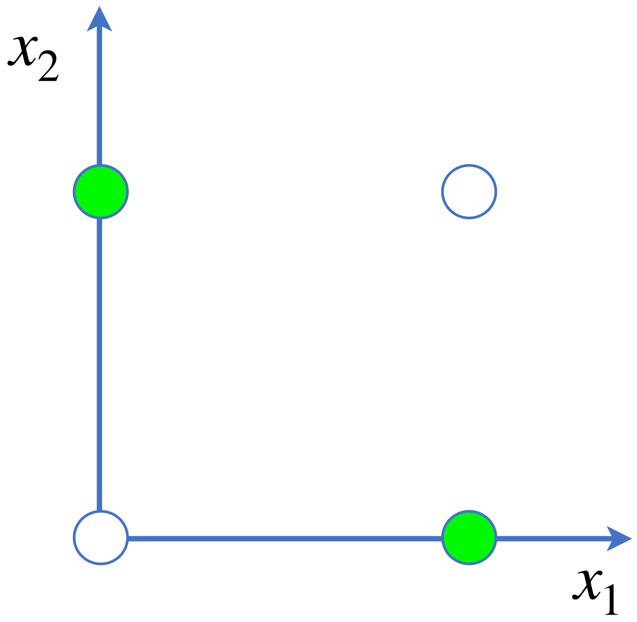


$$\theta = \begin{bmatrix} 1 \\ 0 \\ -.5 \end{bmatrix}$$

x_1

$x_1 > .5$

Perceptron classification

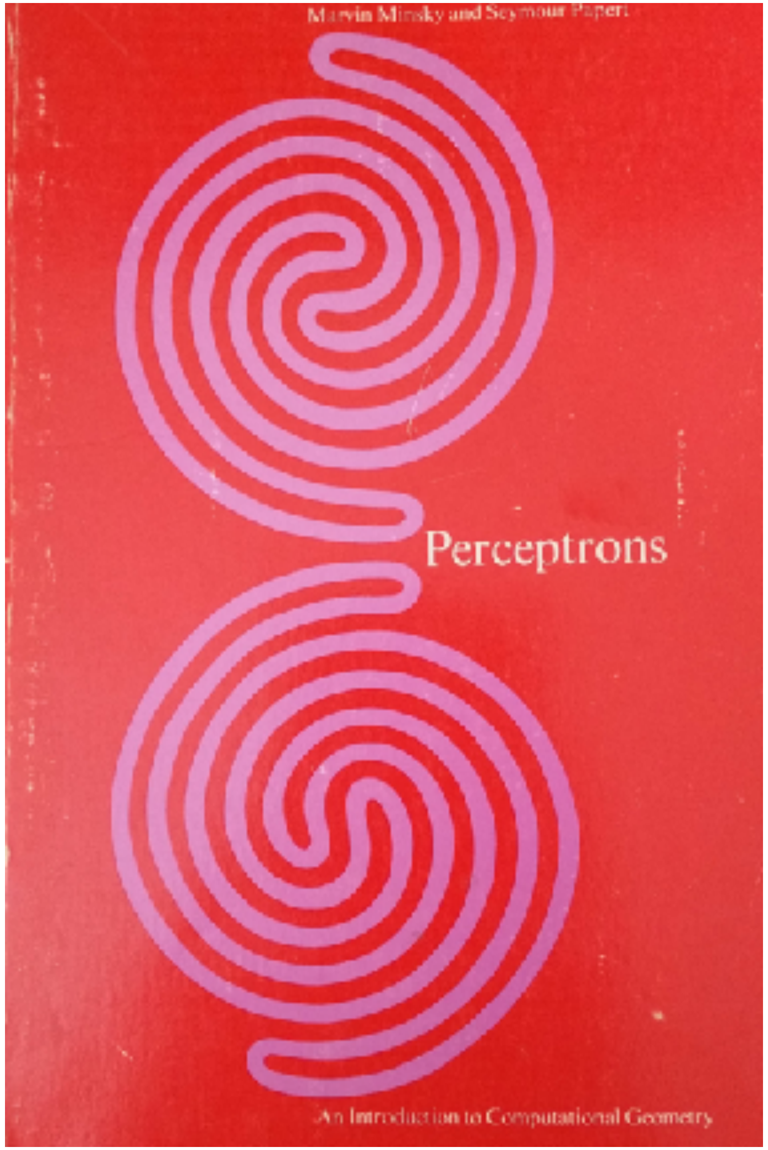


$x_1 \oplus x_2$

- $(0,0) \notin C$, so $b < 0$
- $(0,1) \in C$ and $(1,0) \in C$, so $w_1 x_1 > -b$ and $w_2 x_2 > -b$
- This means $w_1 x_1 + w_2 x_2 > -2b$
- Since $b < 0$, $w_1 x_1 + w_2 x_2 > -2b > -b$
- So, $(1,1) \in C$ 😫

Generalizing the XOR Problem

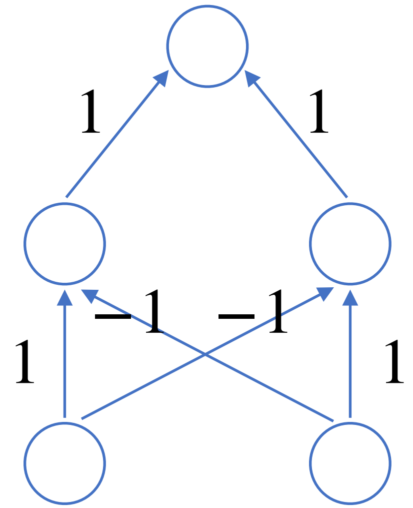
- **Minsky and Papert (1969):** Only linearly separable concepts can be represented by a perceptron (with any monotonic activation function)



Solving the XOR problem

- Multilayer perceptrons

hidden layer



$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Weight matrix dimensions:
rows = outputs
columns = inputs

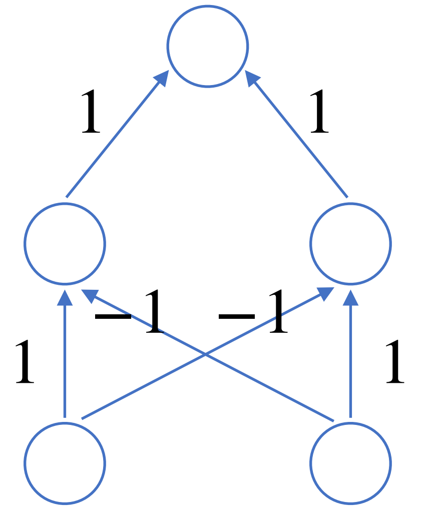
$$W_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad W_2 = [1 \quad 1]$$

$$h = \text{thresh}(W_1 x) \qquad \hat{y} = \text{thresh}(W_2 h)$$

Solving the XOR problem

- Multilayer perceptrons

hidden layer



$$W_2 = [1 \quad 1] \qquad \hat{y} = \text{thresh}(W_2 h)$$

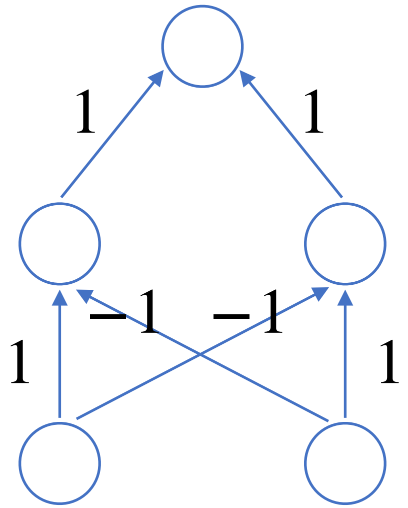
$$W_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad h = \text{thresh}(W_1 x)$$

$$h_{(1,0)} = \text{thresh}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) = \text{thresh}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Solving the XOR problem

- Multilayer perceptrons

hidden layer



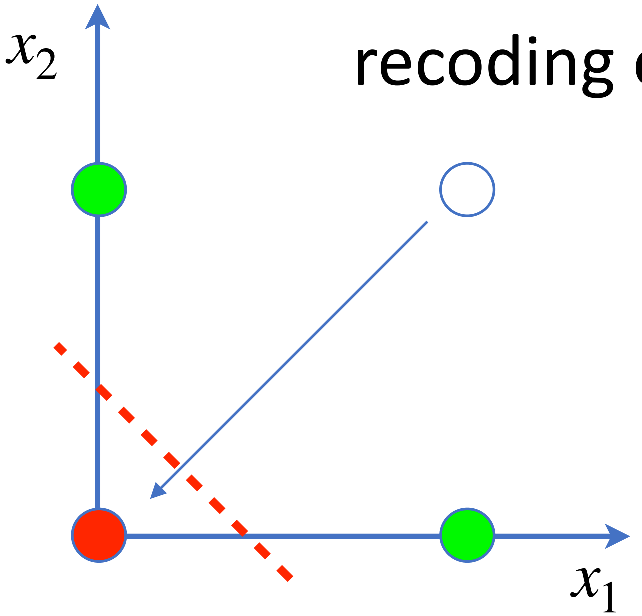
$$W_2 = [1 \quad 1] \qquad \hat{y} = \text{thresh}(W_2 h)$$

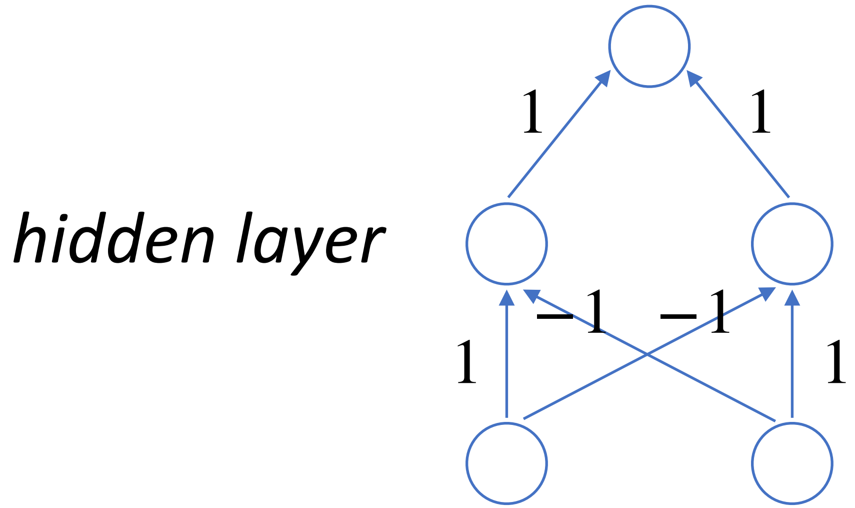
$$W_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad h = \text{thresh}(W_1 x)$$

$$\hat{y}_{(1,0)} = \text{thresh}\left([1 \quad 1] \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = 1$$

Solving the XOR problem

- Multilayer perceptrons

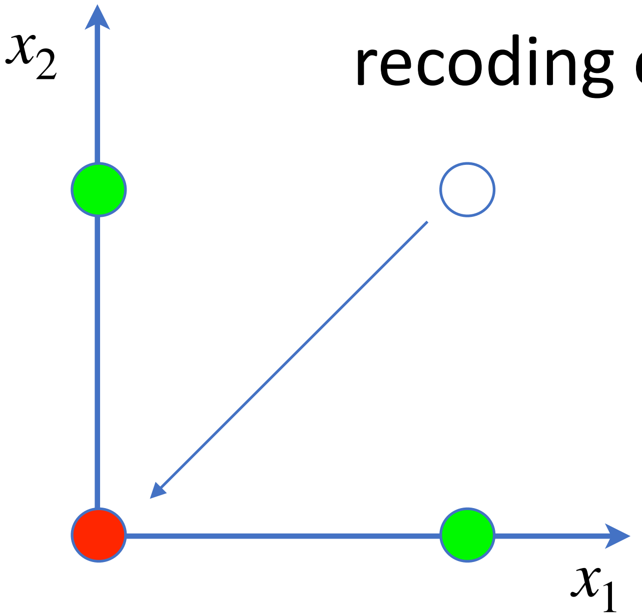




$$W_2 = [1 \quad 1] \qquad \hat{y} = \text{thresh}(W_2 h)$$

$$W_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad h = \text{thresh}(W_1 x)$$

$$h_{(1,1)} = \text{thresh}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \hat{y}_{(1,0)} = \text{thresh}\left([1 \quad 1] \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = 0$$



Batch computation

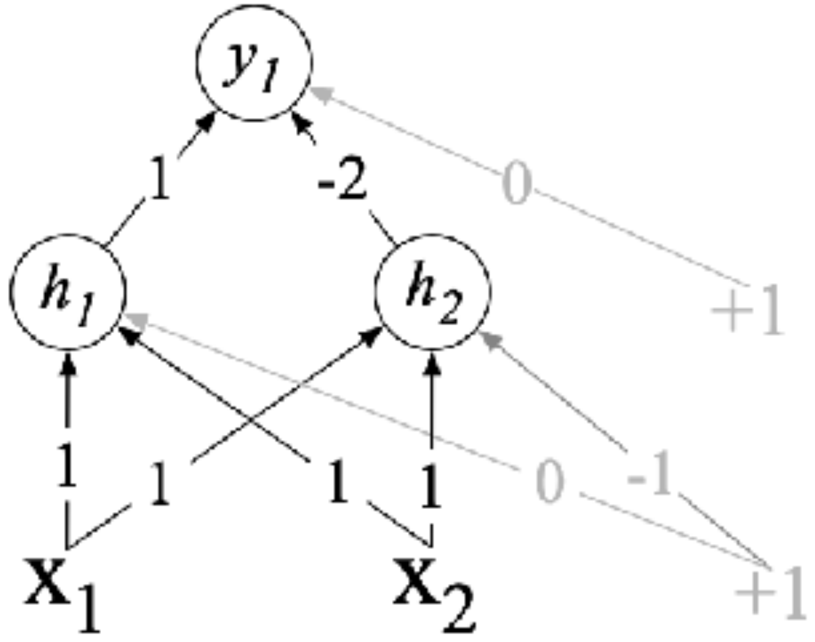
$$W_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad h = \text{thresh}(W_1 x)$$

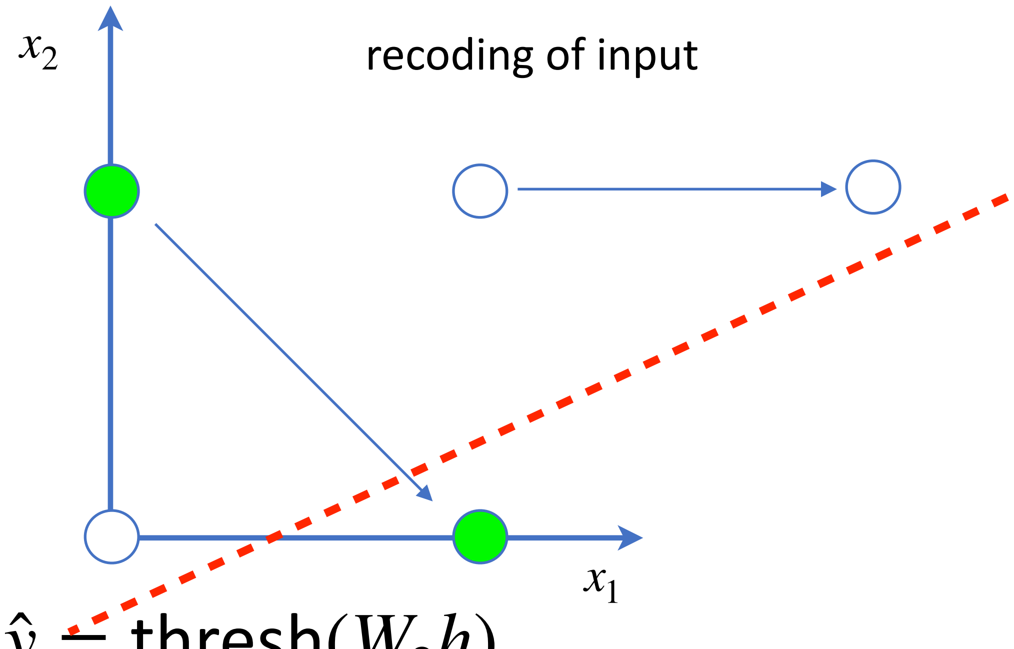
$$W_2 = [1 \quad 1] \qquad \hat{y} = \text{thresh}(W_2 h)$$

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

$$h = \text{thresh}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \right) = \text{thresh}\left(\begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Solving the XOR problem



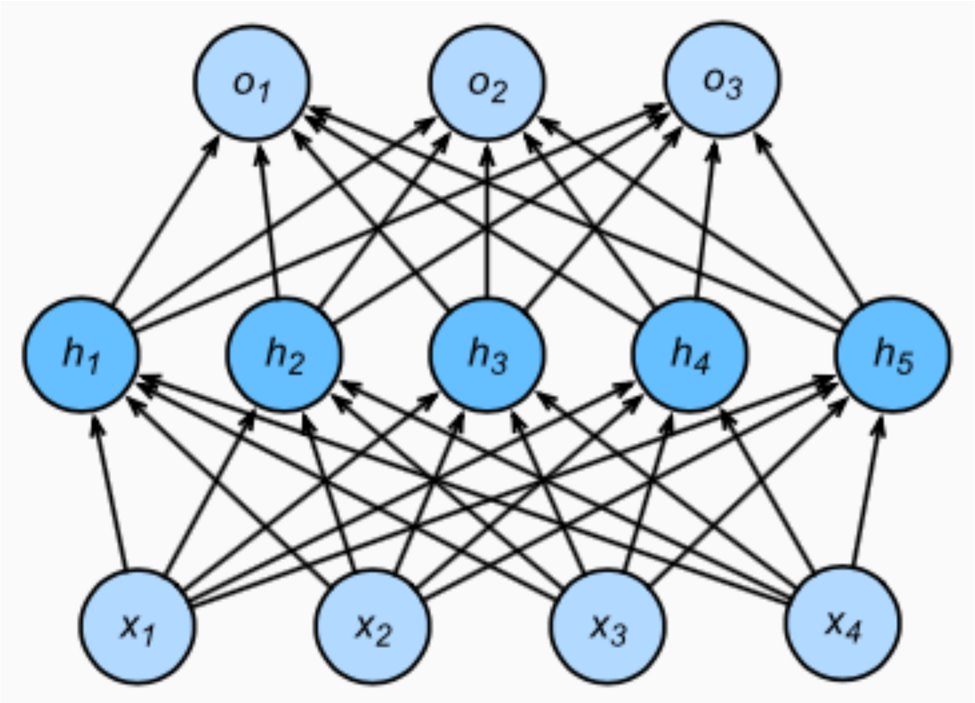


$$W_2 = [1 \quad -2] \qquad \hat{y} = \text{thresh}(W_2 h)$$

$$W_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \; b = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \qquad h = \text{ReLU}(W_1 x + b)$$

$$h = \text{ReLU}\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) = \text{ReLU}\left(\begin{bmatrix} 0 & 1 & 1 & 2 \\ -1 & 0 & 0 & 1 \end{bmatrix}\right) = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Multi-Layer Perceptrons (MLPs)



$$|y| = p \qquad y = f_2(W_2 h + b_2)$$

$$|h| = n \qquad h = f_1(W_1 x + b_1)$$

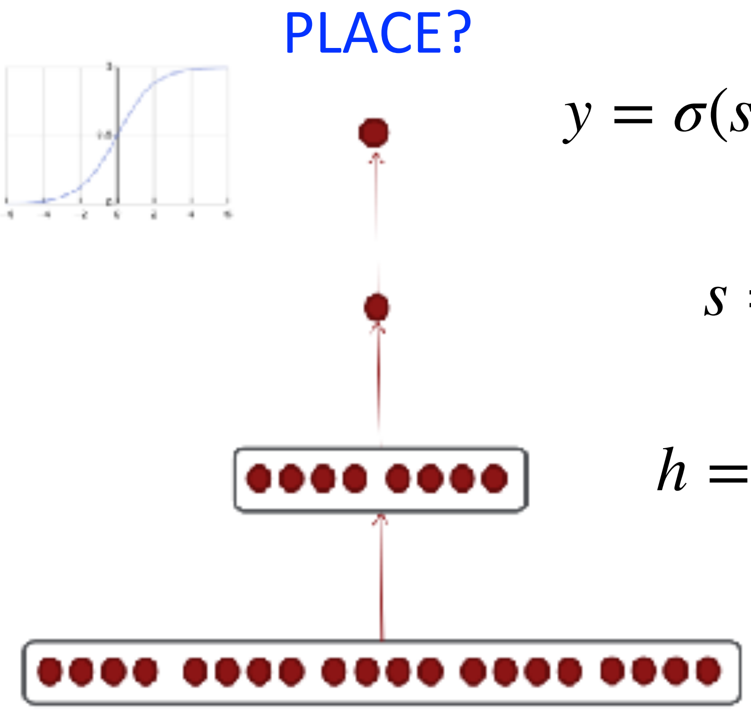
$$|x| = m$$

An Application: Named Entity Recognition

Brazil ’s health minister has tested positive for the coronavirus while in New York for the United Nations General Assembly , where President Jair Bolsonaro spoke on Tuesday .

- Problem: find and label named entities
- Approach: classify each word w on the basis of the words in a window around w

An Application: Named Entity Recognition



x = [while in **New** York for]

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

$$s = u^T h$$

$$h = \sigma(Wx + b)$$

Brazil ’s health minister has tested positive for the coronavirus while in New York for the United Nations General Assembly , where President Jair Bolsonaro spoke on Tuesday .

Other applications

- POS tagging
  - input: sequence of word embeddings of surrounding context
  - output: predicted part of speech (softmax)
- Language modeling
  - input: sequence of word embeddings of preceding context
  - output: predicted next word (softmax)
- Text classification
  - input: sum of word embeddings of text
  - output: predicted class (softmax)