CS 4700: Foundations of Artificial Intelligence

ONLINE

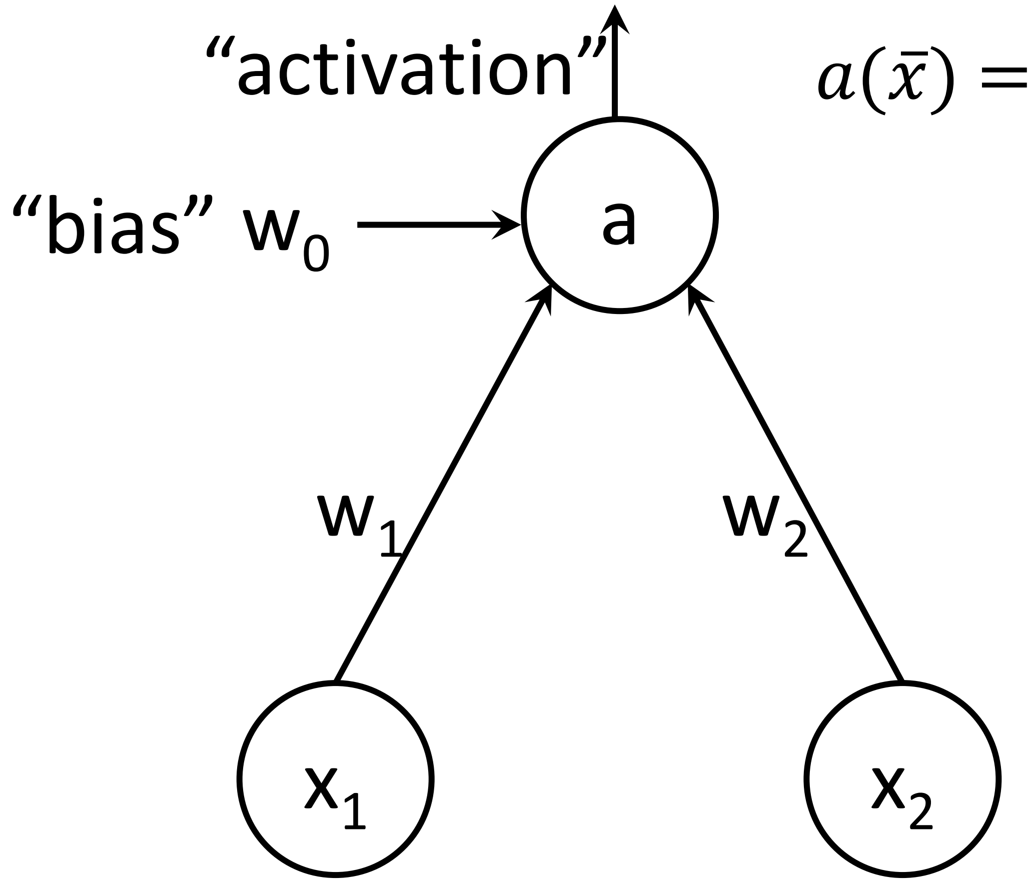
Spring 2020
Prof. Haym Hirsh

Lecture 28
April 20, 2020

Backup plans:

If this Zoom meeting ends prematurely,
five-minute break, check Piazza

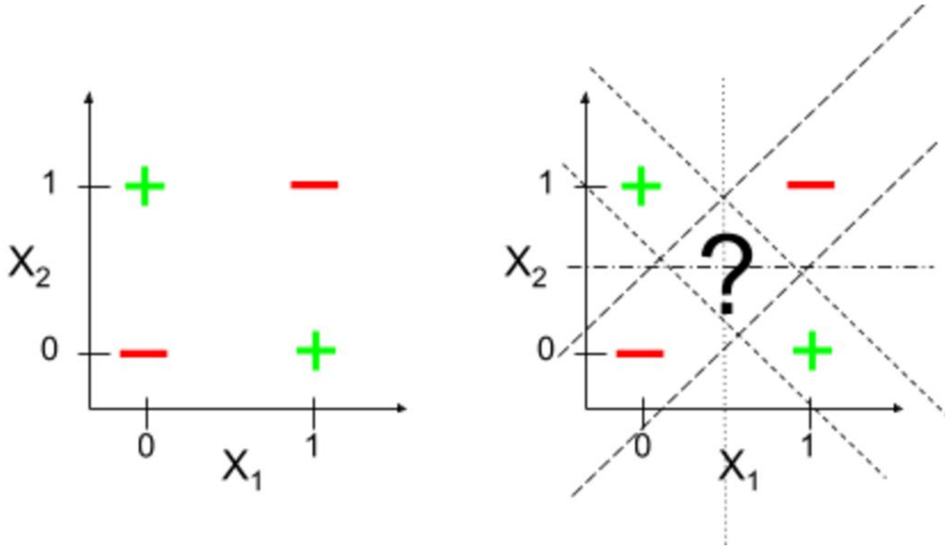
Abstract Representation of a Perceptron



$a(\bar{x}) = +1$ if $\sum_{i=0}^{n} w_i x_i = \bar{w} \cdot \bar{x} \geq 0$

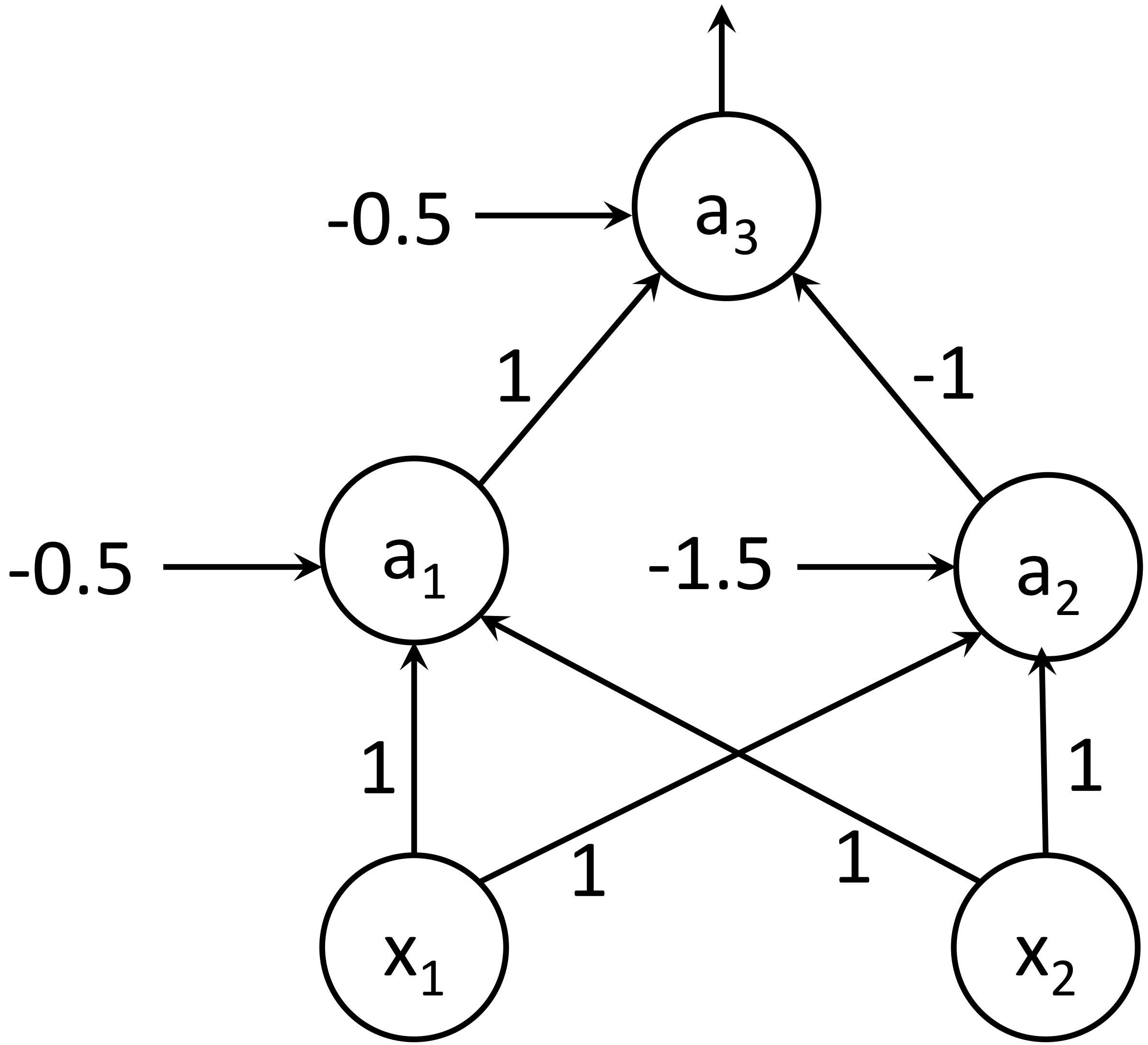
0 otherwise

Perceptrons: The Problem with XOR



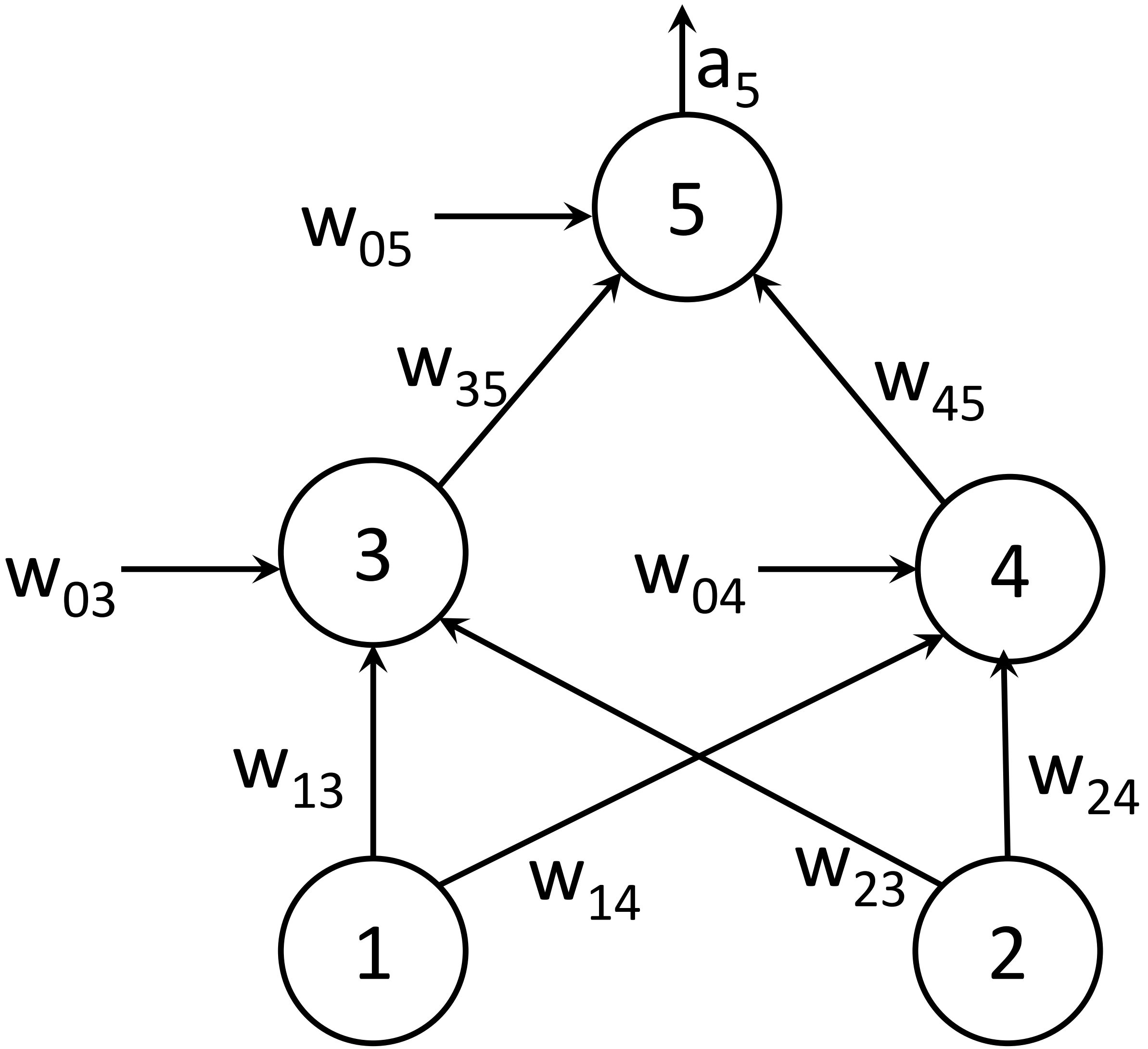
(https://medium.com/@claude.coulombe/the-revenge-of-perceptron-learning-xor-with-tensorflow-eb52cbdf6c60)

Perceptron: The Solution for XOR



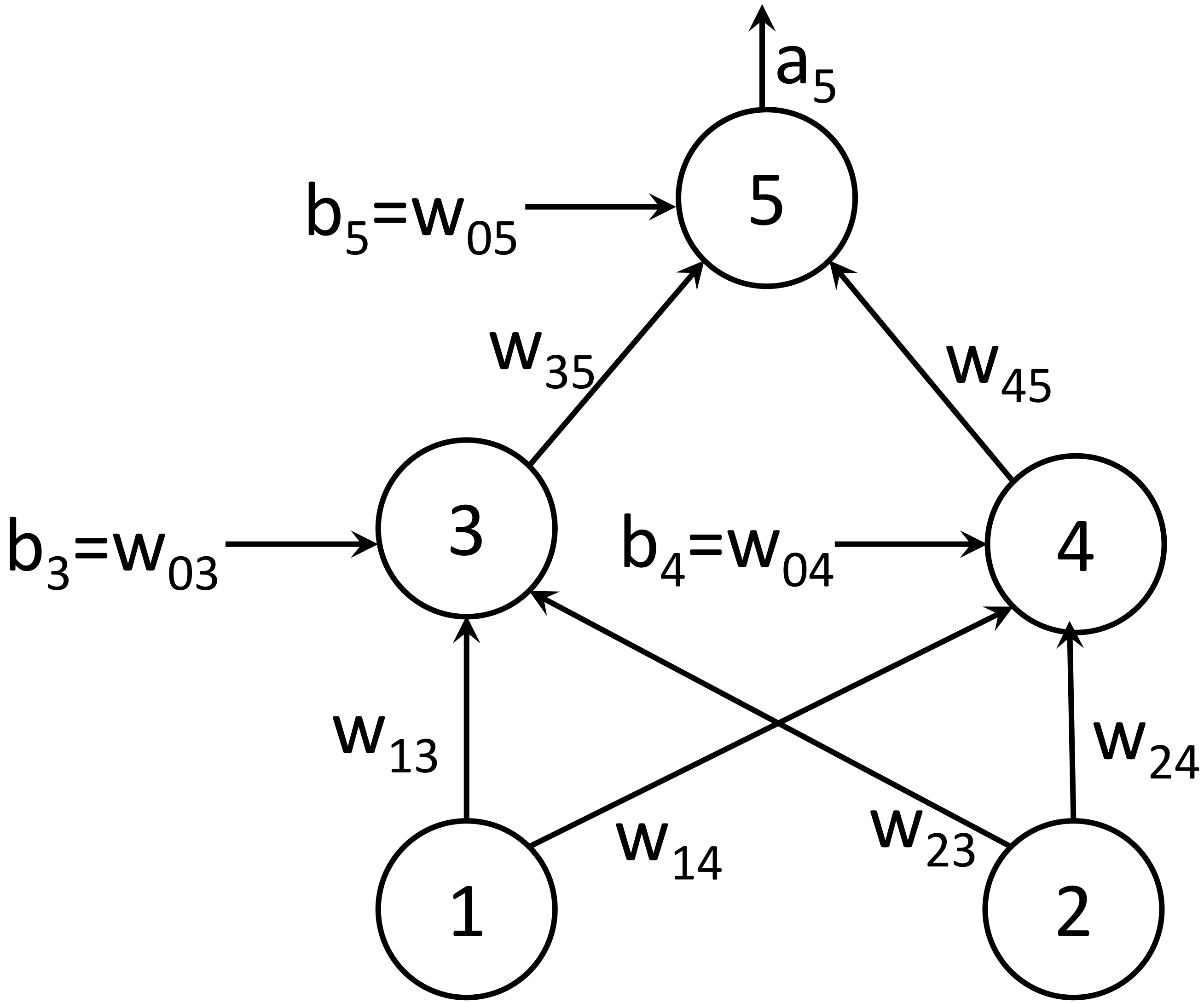
Multi-layer
Neural Network

Multi-Layer Perceptron



Multi-layer
Neural Network

Multi-Layer Perceptron

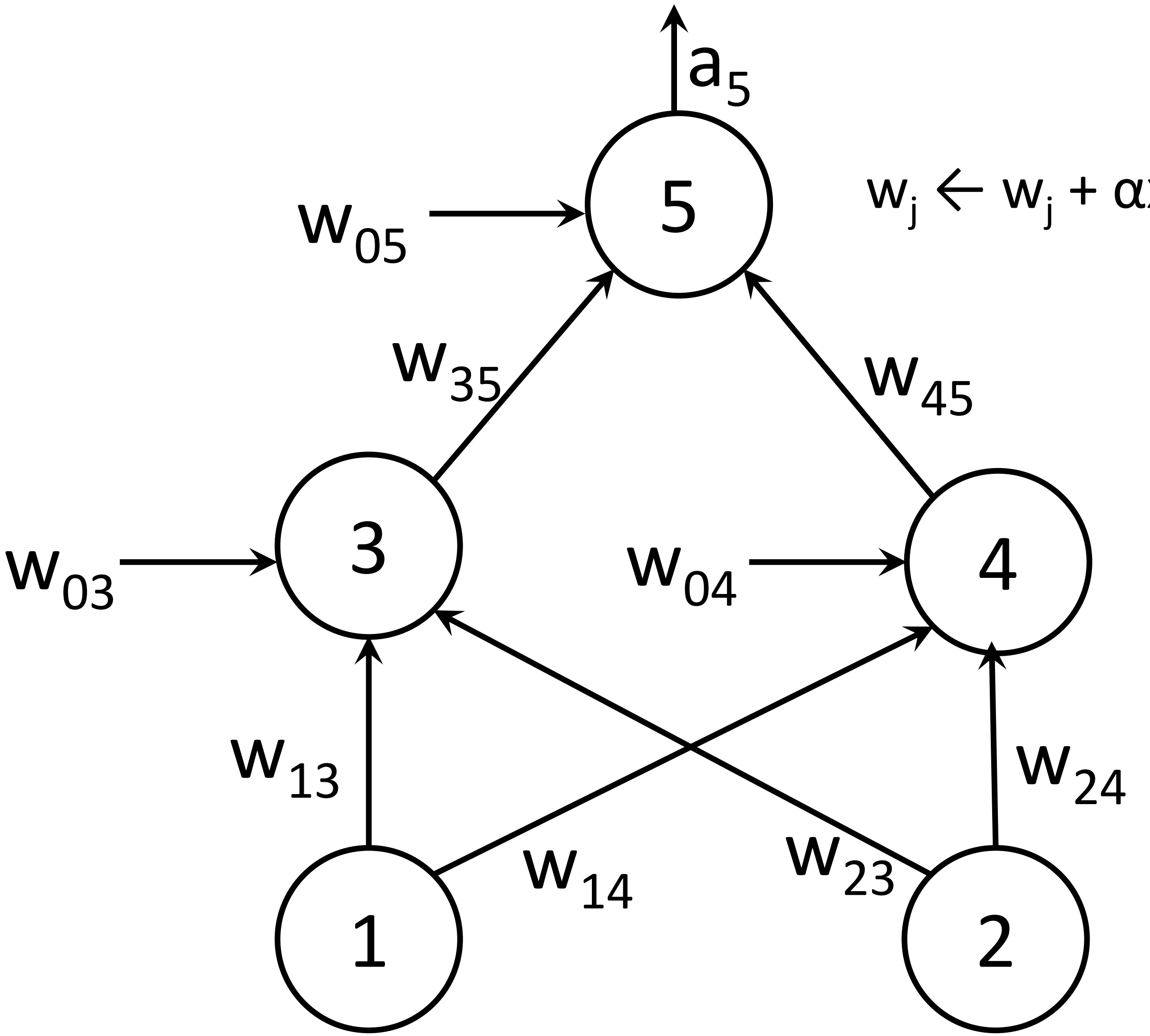


Multi-layer
Neural Network

Multi-Layer Neural Network

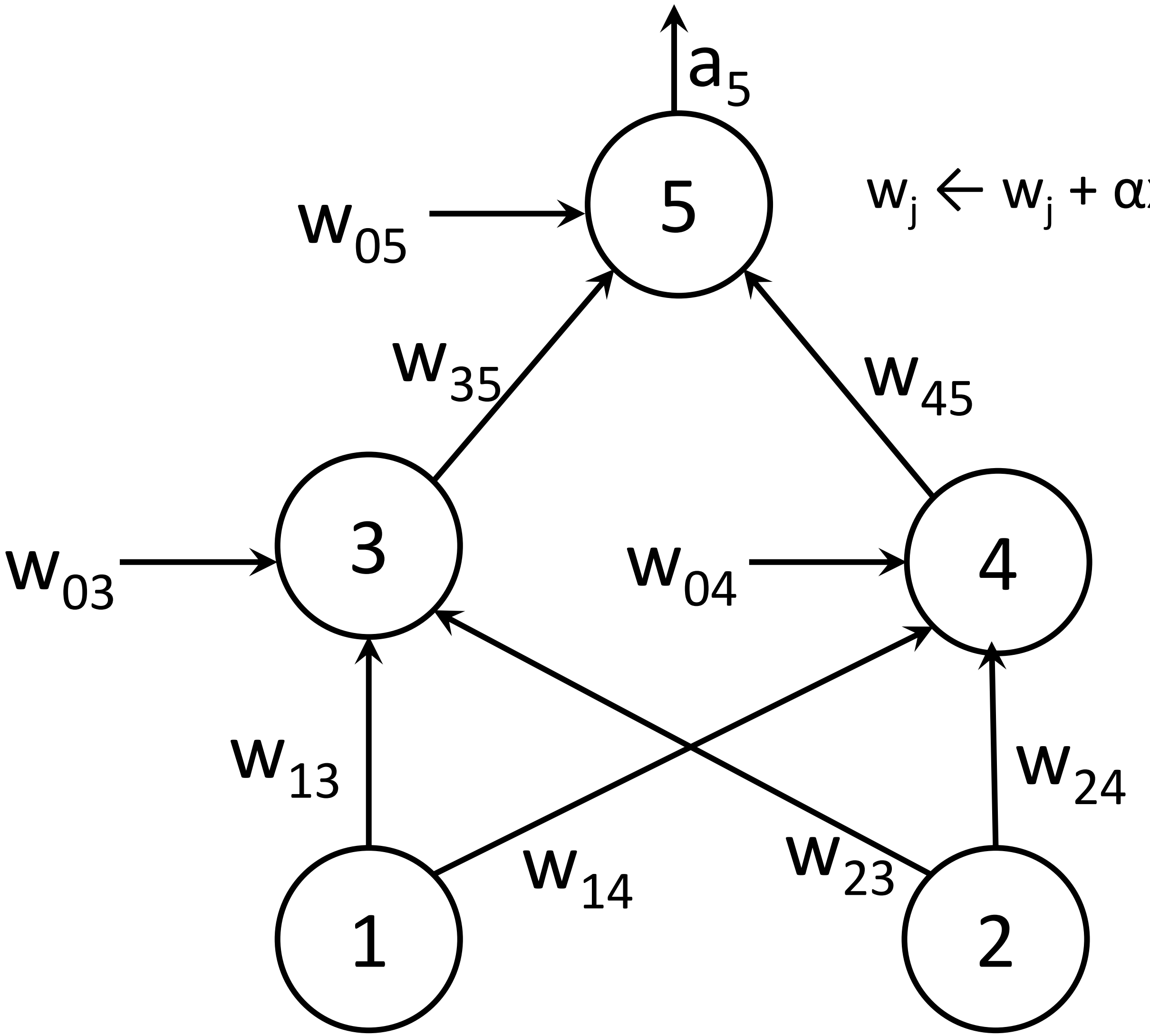
- Increases our “representational power”
- Learning the weights requires something different

Multi-Layer Perceptron



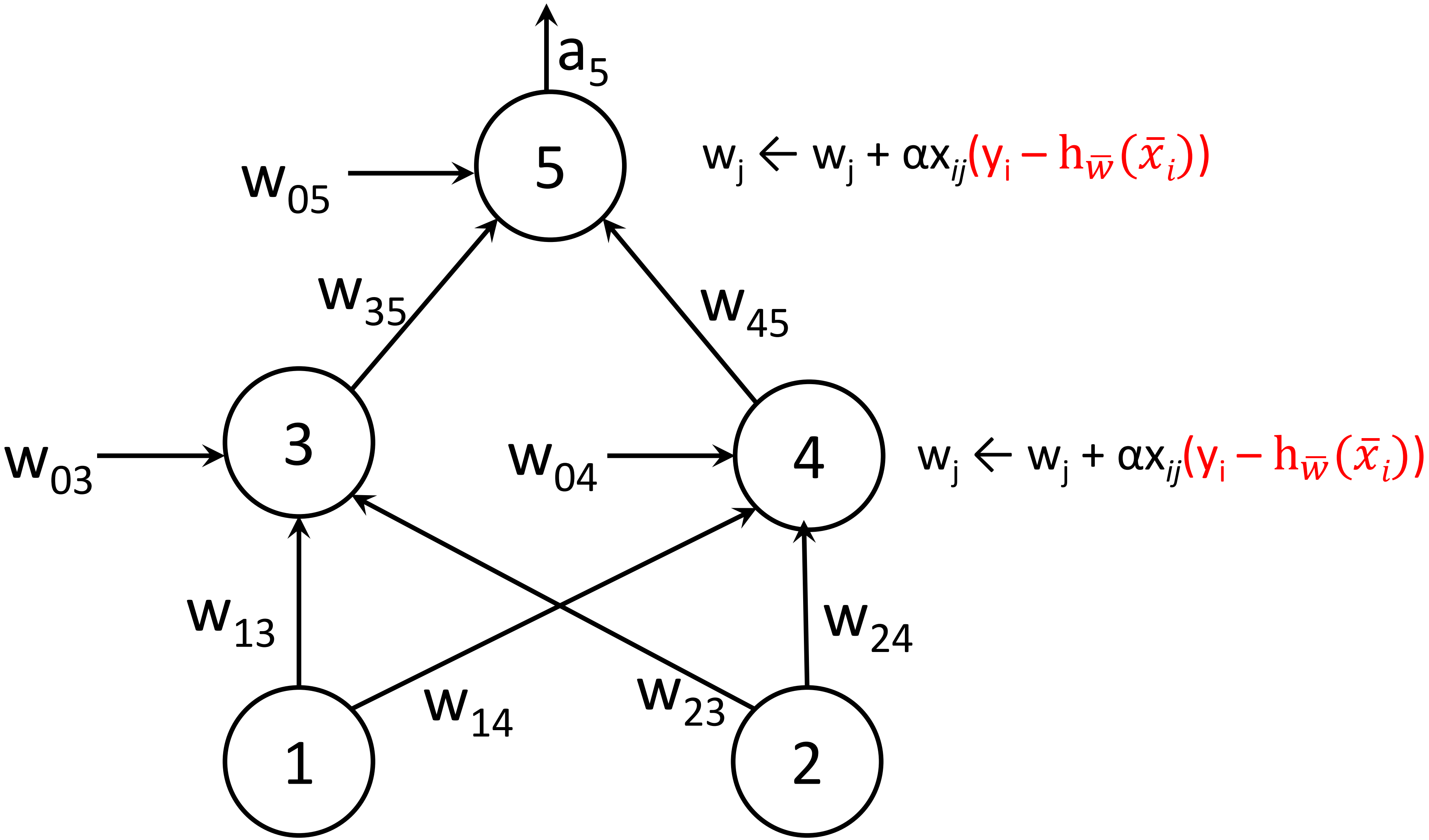
$w_j \leftarrow w_j + \alpha x_{ij}(y_i - h_{\bar{w}}(\bar{x}_i))$

Multi-Layer Perceptron

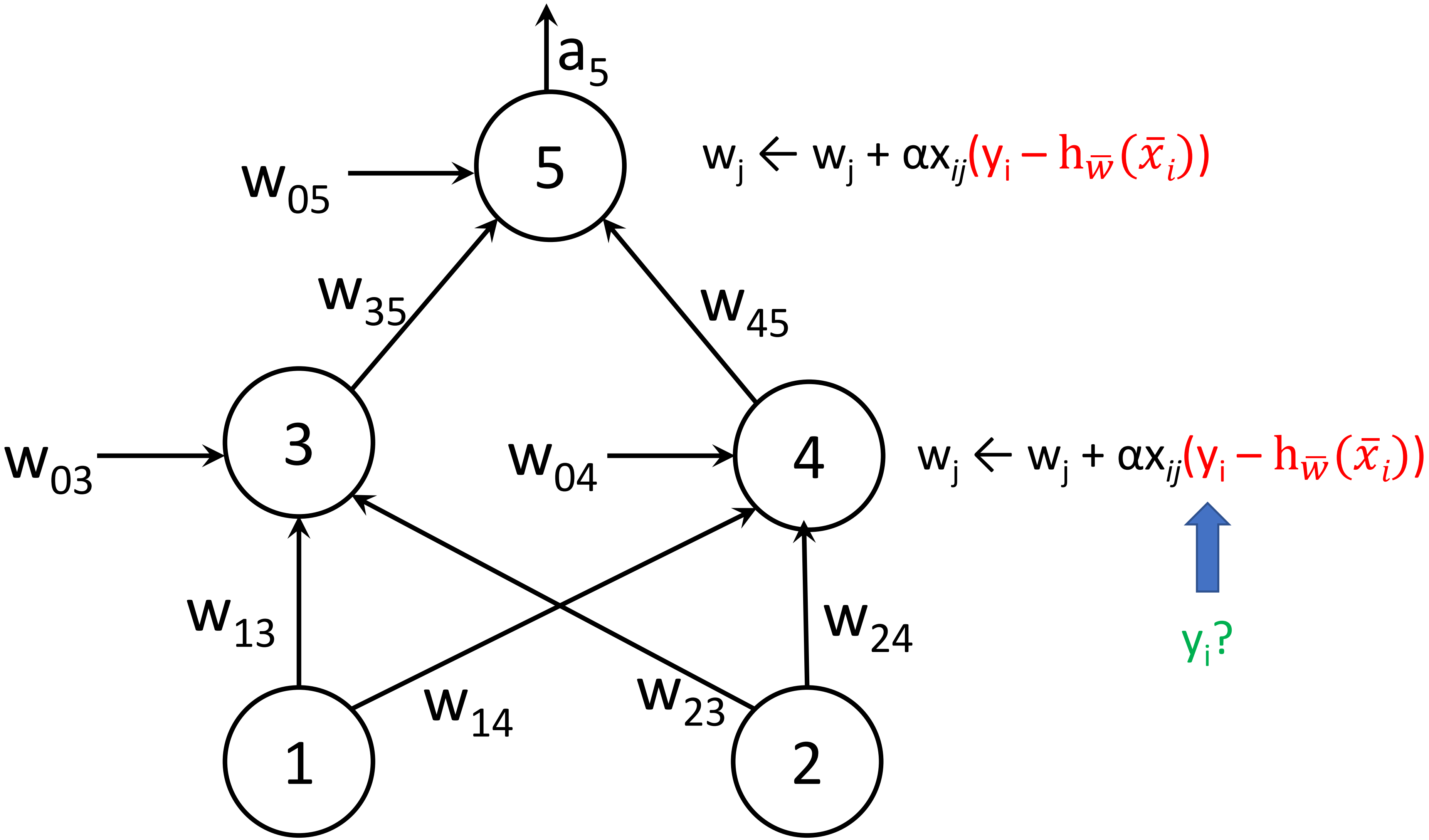


$w_j \leftarrow w_j + \alpha x_{ij}(y_i - h_{\bar{w}}(\bar{x}_i))$

Multi-Layer Perceptron



Multi-Layer Perceptron

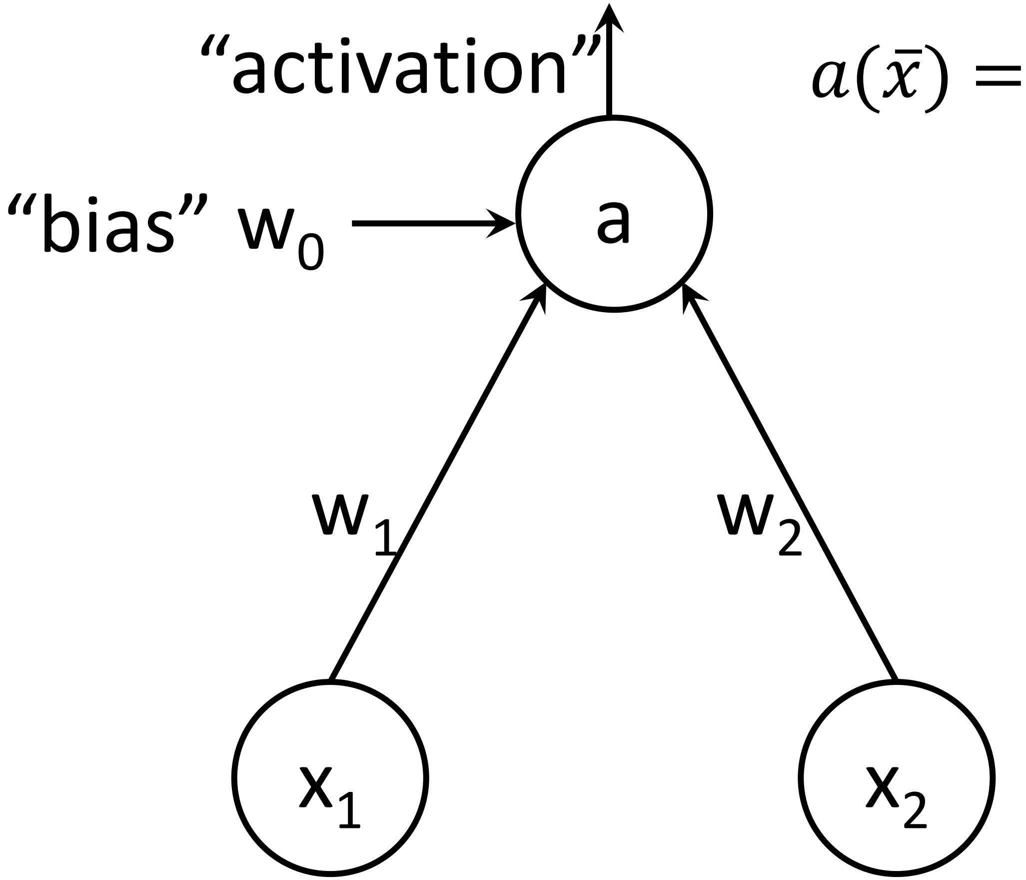


Solution

Replace the hard threshold with a continuous function

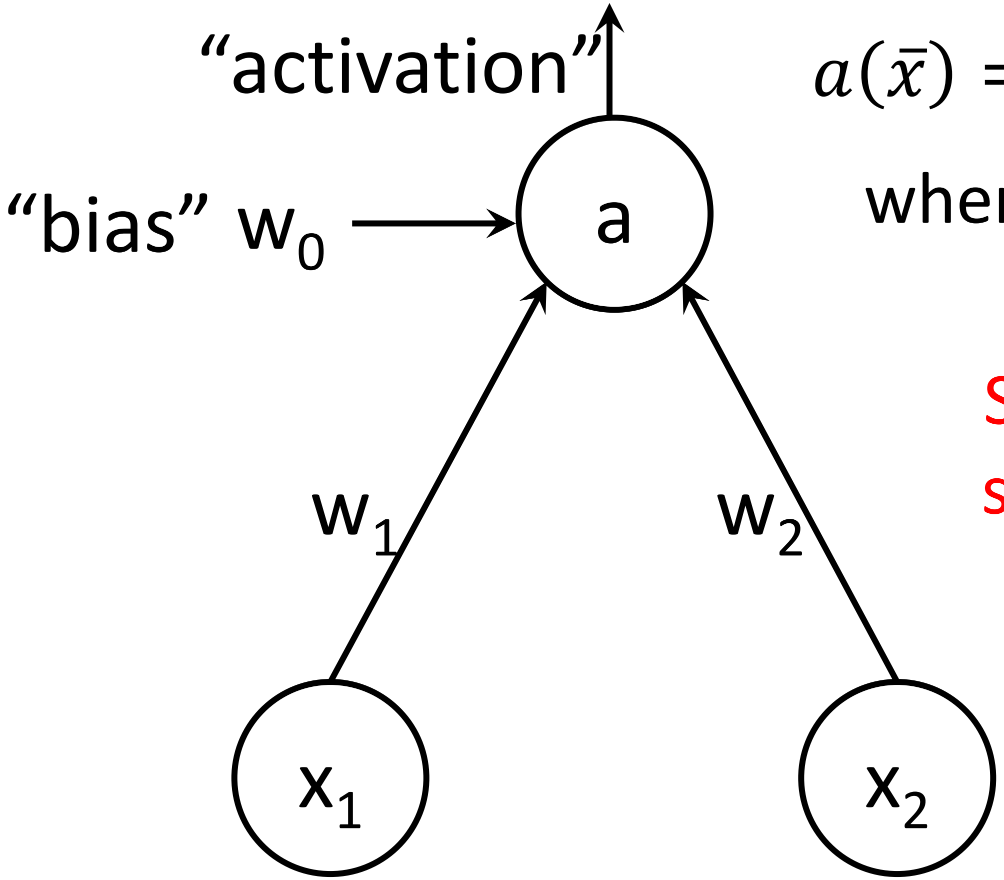
(and use calculus)

Abstract Representation of a Perceptron



$$a(\bar{x}) = \begin{cases} +1 & \text{if } \bar{w} \cdot \bar{x} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Abstract Representation of a Perceptron



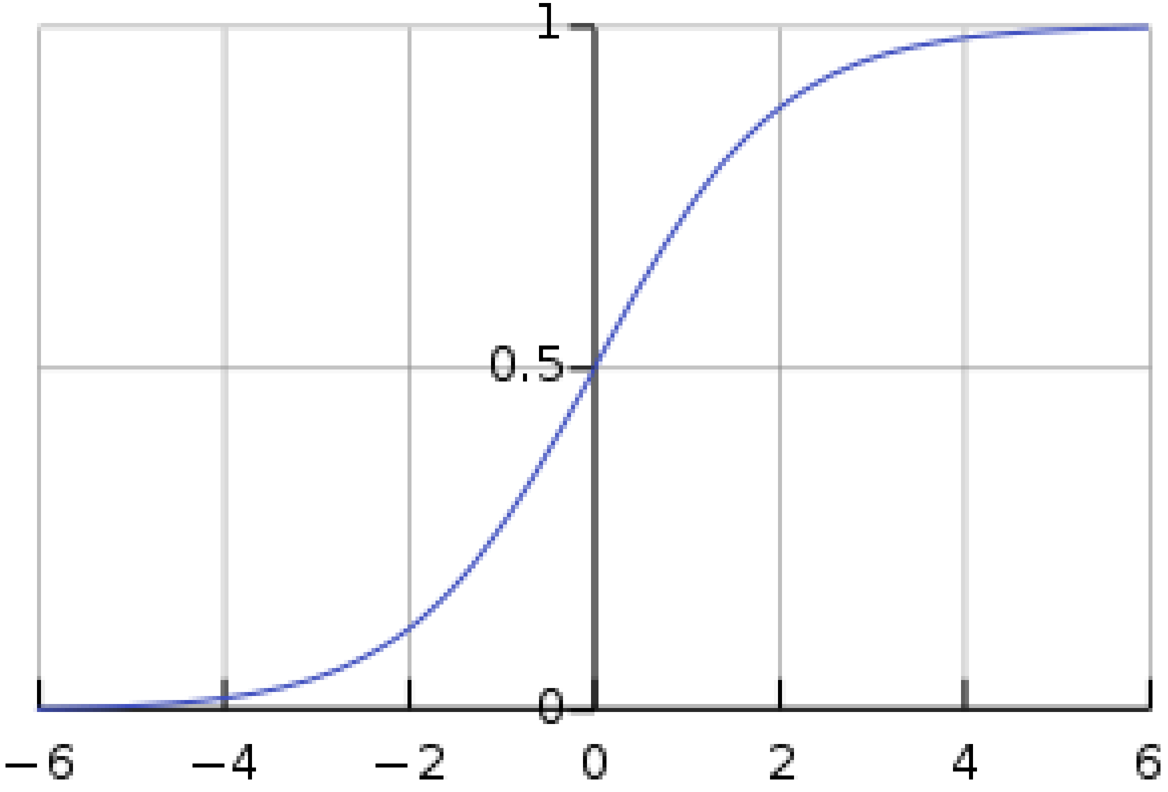
$$a(\bar{x}) = g(\bar{w} \cdot \bar{x})$$

where $g(z) = 1$ if $z \geq 0$

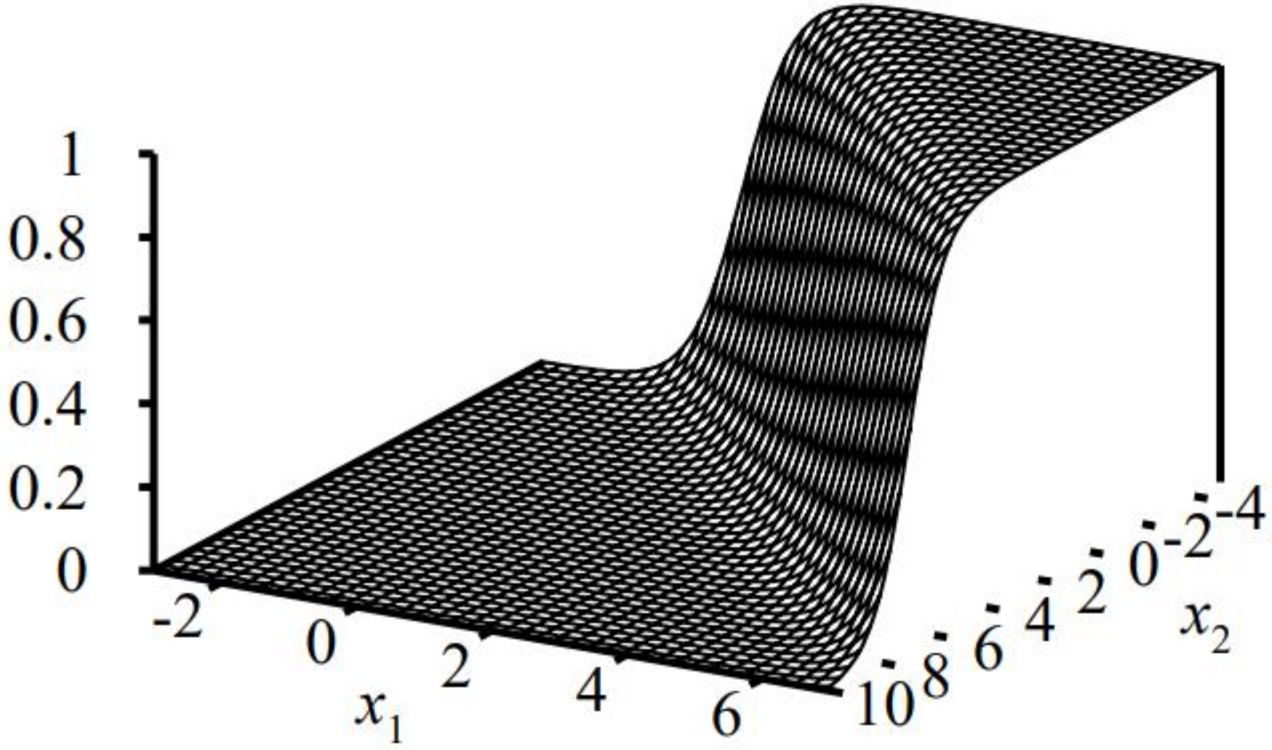
0 otherwise

Separates weighted sum from threshold

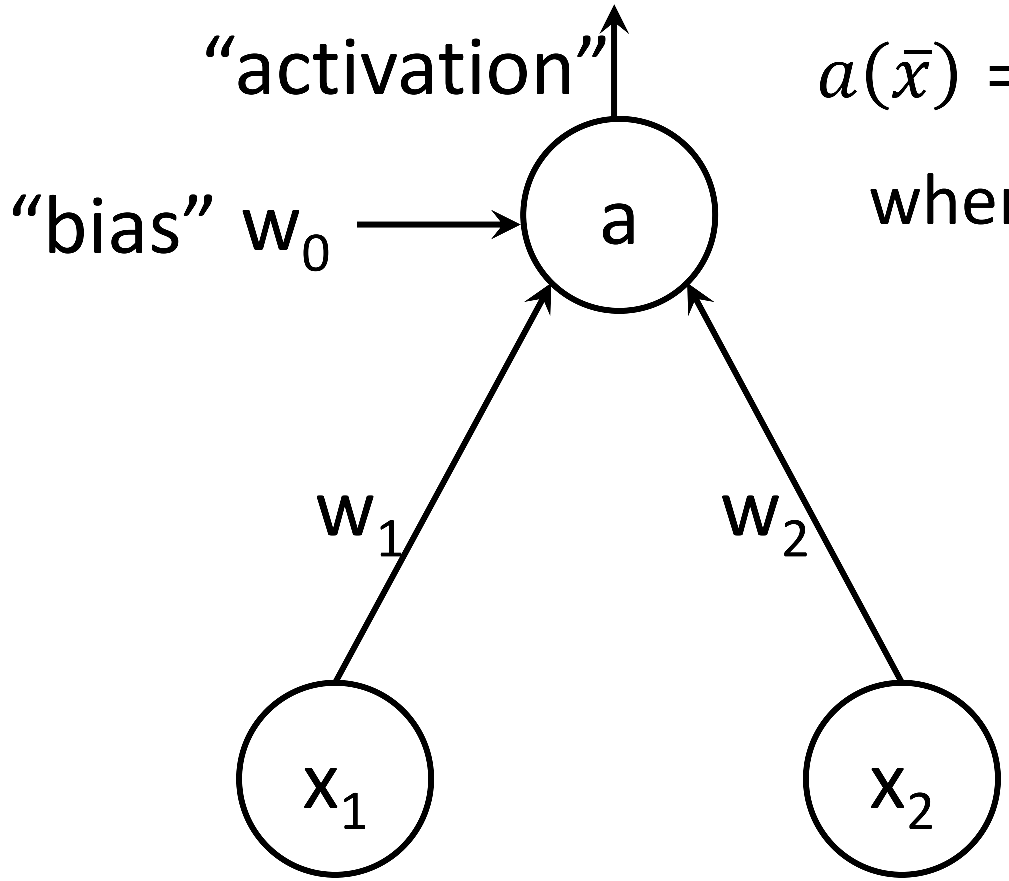
$$g(z) = \frac{1}{1 + e^{-z}}$$ "Logistic Function"



$$\boldsymbol{h}_{\overline{w}}(\bar{x}) = \boldsymbol{g}(\overline{w} \cdot \bar{x}) = \frac{1}{1 + e^{-\overline{w} \cdot \bar{x}}}$$ "Logistic Function"



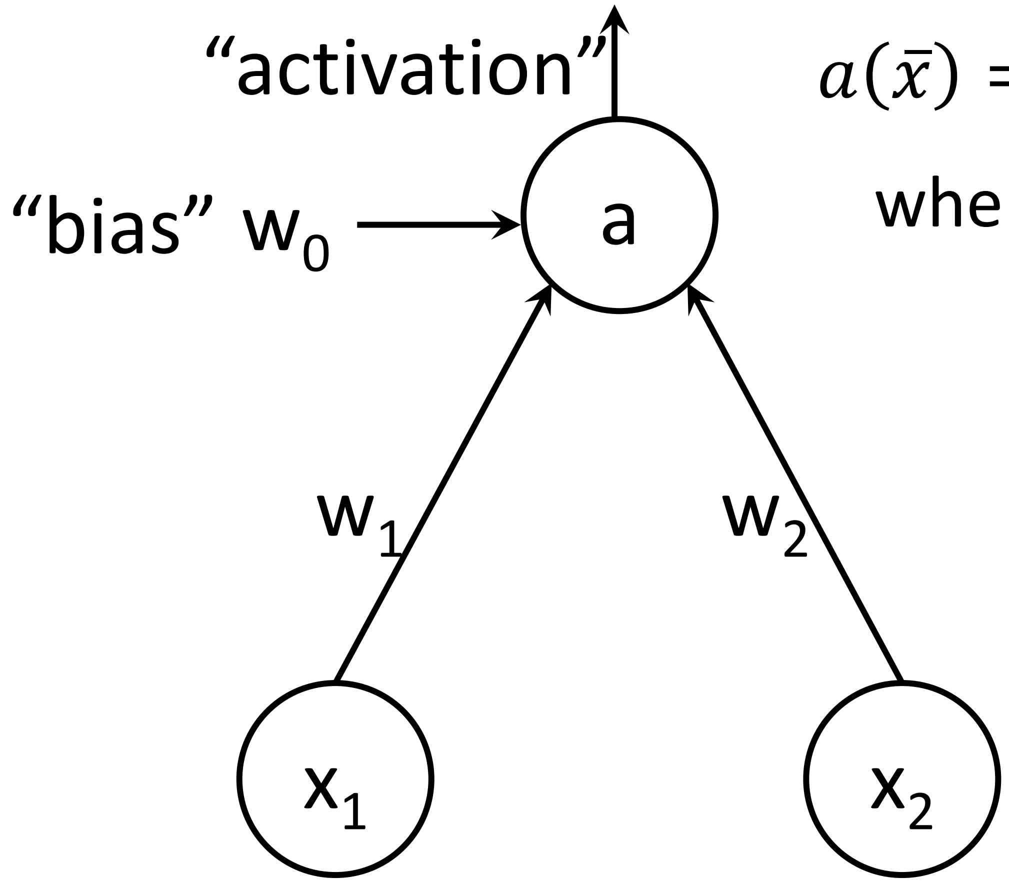
Abstract Representation of a Perceptron



$a(\bar{x}) = g(\bar{w} \cdot \bar{x})$

where $g(z) = 1$ if $z \geq 0$
0 otherwise

Abstract Representation of a Logistic Neuron



$$a(\bar{x}) = g(\bar{w} \cdot \bar{x})$$

where $g(z) = \frac{1}{1+e^{-z}}$

Learning

- Consider $\text{Error}_D(\overline{w})$:

$$\sum_{i=1}^{N}(y_i - h_{\overline{w}}(\bar{x}_i))^2$$

Learning

- Consider $\text{Error}_D(\overline{w})$:

$$\sum_{i=1}^{N}(y_i - h_{\overline{w}}(\bar{x}_i))^2$$

- For a given set of data D, the $\bar{x}_i$'s andy_i's are fixed
- Different $\overline{w}$ give different values for $\text{Error}_D(\overline{w})$

Learning

- Consider $\text{Error}_D(\overline{w})$:

$$\sum_{i=1}^{N}(y_i - h_{\overline{w}}(\bar{x}_i))^2$$

- For a given set of data D, the $\bar{x}_i$'s andy_i's are fixed
- Different $\overline{w}$ give different values for $\text{Error}_D(\overline{w})$
- We want to minimize $\text{Error}_\text{D}(\overline{w})$ by changing the $\overline{w}_i$'s

Learning

- Consider $\text{Error}_D(\overline{w})$:

$$\sum_{i=1}^{N}(y_i - h_{\overline{w}}(\bar{x}_i))^2$$

- For a given set of data D, the $\bar{x}_i$'s andy_i's are fixed
- Different $\overline{w}$ give different values for $\text{Error}_D(\overline{w})$
- We want to minimize $\text{Error}_\text{D}(\overline{w})$ by changing the $\overline{w}_i$'s
- We'll do this incrementally by following the derivative of $\text{Error}_\text{D}(\overline{w})$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\text{Error}_D(\overline{w}) = \sum_{(\bar{x}_i, y_i)\in D} (y_i - h_{\overline{w}}(\bar{x}_i))^2 = \sum_{(\bar{x}_i, y_i)\in D} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\text{Error}_D(\overline{w}) = \sum_{(\bar{x}_i, y_i) \in D} (y_i - h_{\overline{w}}(\bar{x}_i))^2 = \sum_{(\bar{x}_i, y_i) \in D} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$$\frac{\partial}{\partial w_j} \text{Error}_D(\overline{w}) = \frac{\partial}{\partial w_j} \sum_{(\bar{x}_i, y_i) \in D} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\text{Error}_D(\overline{w}) = \sum_{(\bar{x}_i, y_i) \in D} (y_i - h_{\overline{w}}(\bar{x}_i))^2 = \sum_{(\bar{x}_i, y_i) \in D} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$$\frac{\partial}{\partial w_j} \text{Error}_D(\overline{w}) = \frac{\partial}{\partial w_j} \sum_{i=1}^{N} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial w_j}\text{Error}_D(\overline{w}) = \frac{\partial}{\partial w_j}\sum_{i=1}^{N}(y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial w_j}\text{Error}_D(\overline{w}) = \frac{\partial}{\partial w_j}\sum_{i=1}^{N}(y_i - g(\overline{w}\cdot\bar{x}_i))^2$$

$$= \sum_{i=1}^{N}\frac{\partial}{\partial w_j}(y_i - g(\overline{w}\cdot\bar{x}_i))^2$$

$\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial w_j} \text{Error}_D(\overline{w}) = \frac{\partial}{\partial w_j} \sum_{i=1}^{N} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$$= \sum_{i=1}^{N} \frac{\partial}{\partial w_j} (y_i - g(\overline{w} \cdot \bar{x}_i))^2$$

$$= -2 \sum_{i=1}^{N} g'(\overline{w} \cdot \bar{x}_i)(y_i - g(\overline{w} \cdot \bar{x}_i)) x_{ij}$$

Logistic Regression $\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial wj}\text{Error}_{\text{D}}(\overline{w}) = -2\sum_{i=1}^{N} g'(\overline{w}\cdot\bar{x}_i)(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

$$g'(z) = g(z)(1 - g(z))$$

for $g(z) =$ logistic function

$$\frac{\partial}{\partial w_j}\text{Error}_{\text{D}}(\overline{w}) = -2\sum_{i=1}^{N} g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Logistic Regression $\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial wj}\text{Error}_\text{D}(\overline{w}) = -2\sum_{i=1}^{N} g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Logistic Regression $\text{Error}_D(\overline{w})$ and Its Gradient

$$\frac{\partial}{\partial wj}\text{Error}_\text{D}(\overline{w}) = -2\sum_{i=1}^{N} g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Focusing on a single example $(\bar{x}_i, y_i)$

$$\frac{\partial}{\partial w_j}\text{Error}_{\{(\bar{x}_i, y_i)\}}(\overline{w}) = -2g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Logistic Regression Update Rule

$$\frac{\partial}{\partial w_j}\text{Error}_{\{(\bar{x}_i, y_i)\}}(\overline{w}) = -2g(\overline{w}\cdot\bar{x}_i)(1-g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Logistic Regression Update Rule

$$\frac{\partial}{\partial w_j}\text{Error}_{\{(\bar{x}_i,y_i)\}}(\overline{w}) = -2g(\overline{w}\cdot\bar{x}_i)(1-g(\overline{w}\cdot\bar{x}_i))(y_i-g(\overline{w}\cdot\bar{x}_i))xi_j$$

$$w_j \leftarrow w_j + \alpha\, g(\overline{w}\cdot\bar{x}_i)(1-g(\overline{w}\cdot\bar{x}_i))(y_i-g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Logistic Regression Update Rule

$$\frac{\partial}{\partial w_j} \text{Error}_{\{(\bar{x}_i, y_i)\}}(\overline{w}) = -2g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))xi_j$$

$$\mathrm{w_j} \leftarrow \mathrm{w_j} + \alpha\, g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))x_{ij}$$

The -2 is captured in α

Logistic Regression Update Rule

$$\frac{\partial}{\partial w_j}\text{Error}_{\{(\bar{x}_i, y_i)\}}(\overline{w}) = -2g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))xi_j$$

$$w_j \leftarrow w_j + \alpha\, g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))x_{ij}$$

Logistic Regression Update Rule

$$\frac{\partial}{\partial w_j}\text{Error}_{\{(\bar{x}_i, y_i)\}}(\overline{w}) = -2g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

$$w_j \leftarrow w_j + \alpha\, g(\overline{w}\cdot\bar{x}_i)(1 - g(\overline{w}\cdot\bar{x}_i))(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

For other activation functions

$$w_j \leftarrow w_j + \alpha\, g'(\overline{w}\cdot\bar{x}_i)(y_i - g(\overline{w}\cdot\bar{x}_i))x_{ij}$$

Perceptron Learning Algorithm

Current hypothesis: $\mathrm{h}_{\bar{w}}(\bar{x})$

Initialize $w_0, w_1, w_2, \ldots, w_n$

Repeat

For i = 1 to N [for each example]

$h \leftarrow \mathrm{h}_{\bar{w}}(\overline{\mathrm{x}_i})$

For j = 0 to n [for each feature]

$w_j \leftarrow w_j + \alpha x_{ij}(y_i - h)$

Until $\mathrm{h}_{\bar{w}}(\bar{x})$ gets all data correct

Gradient Descent (Logistic Regression)

Current hypothesis: $\mathrm{h}_{\overline{w}}(\bar{x})$

Initialize $w_0, w_1, w_2, \ldots, w_n$

Repeat

For i = 1 to N [for each example]

$h \leftarrow \mathrm{h}_{\overline{w}}(\overline{\mathrm{x}_i})$

For j = 0 to n [for each feature]

$w_j \leftarrow w_j + \alpha\, g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))x_{ij}$

Until <stopping condition>

Comparing Update Rules

$$w_j \leftarrow w_j + \alpha x_{ij}(y_i - h)$$

$$w_j \leftarrow w_j + \alpha\, g(\overline{w} \cdot \bar{x}_i)(1 - g(\overline{w} \cdot \bar{x}_i))(y_i - g(\overline{w} \cdot \bar{x}_i))x_{ij}$$