CS 4104
APPLIED MACHINE LEARNING

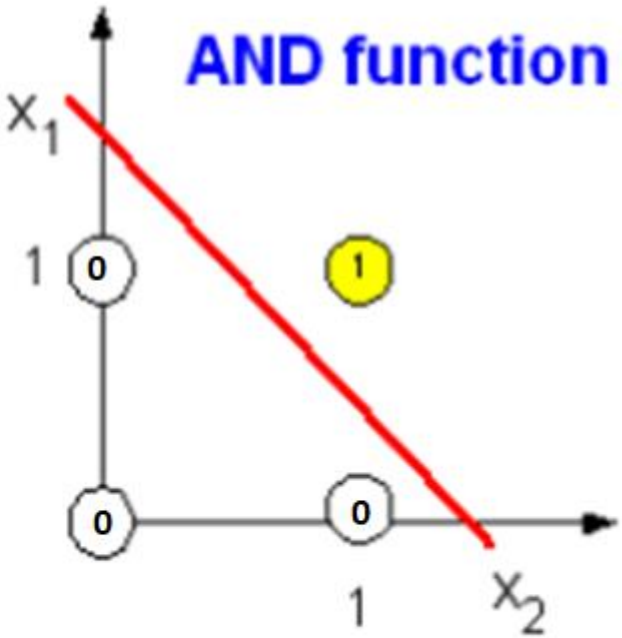
Dr. Hashim Yasin

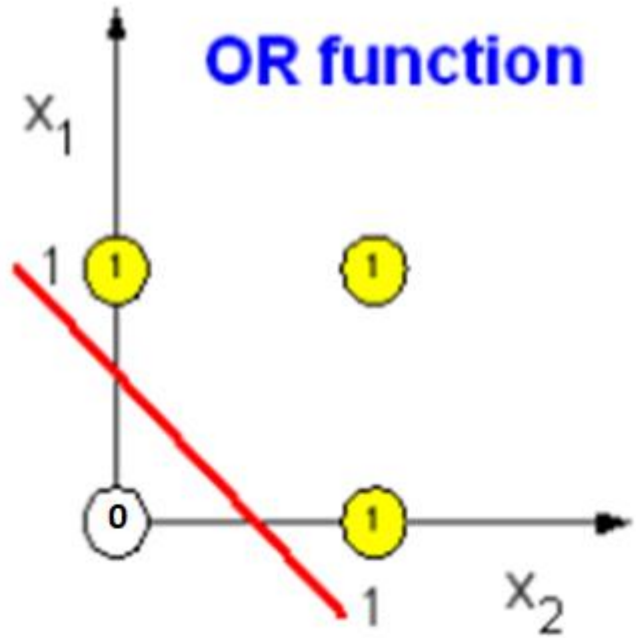
National University of Computer and Emerging Sciences,

Faisalabad, Pakistan.

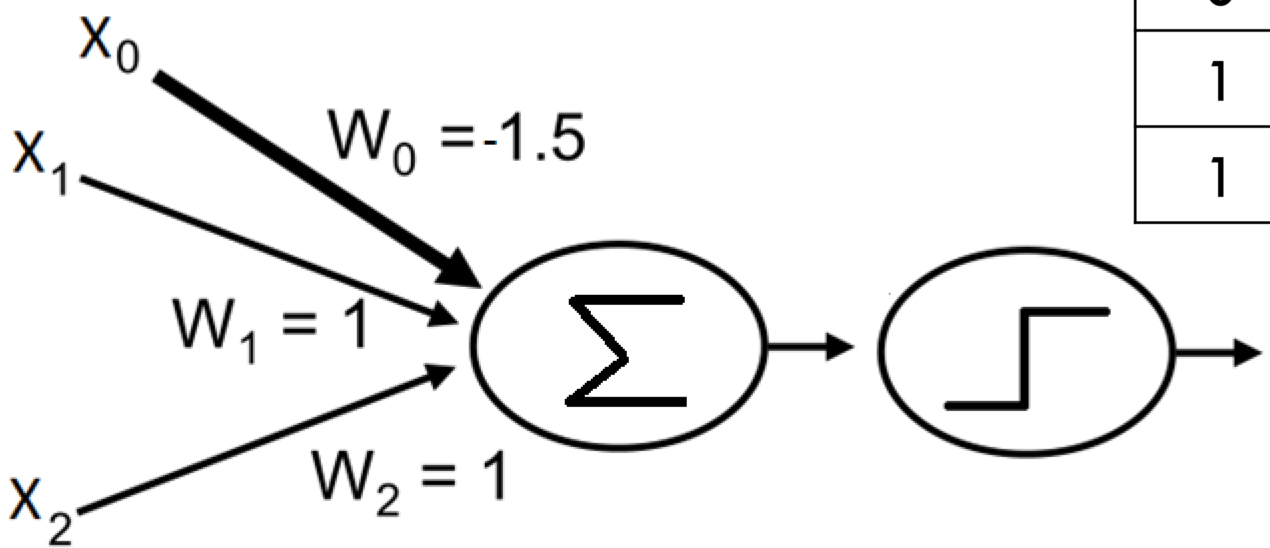
PERCEPTRON-RECAP

AND OR Functions





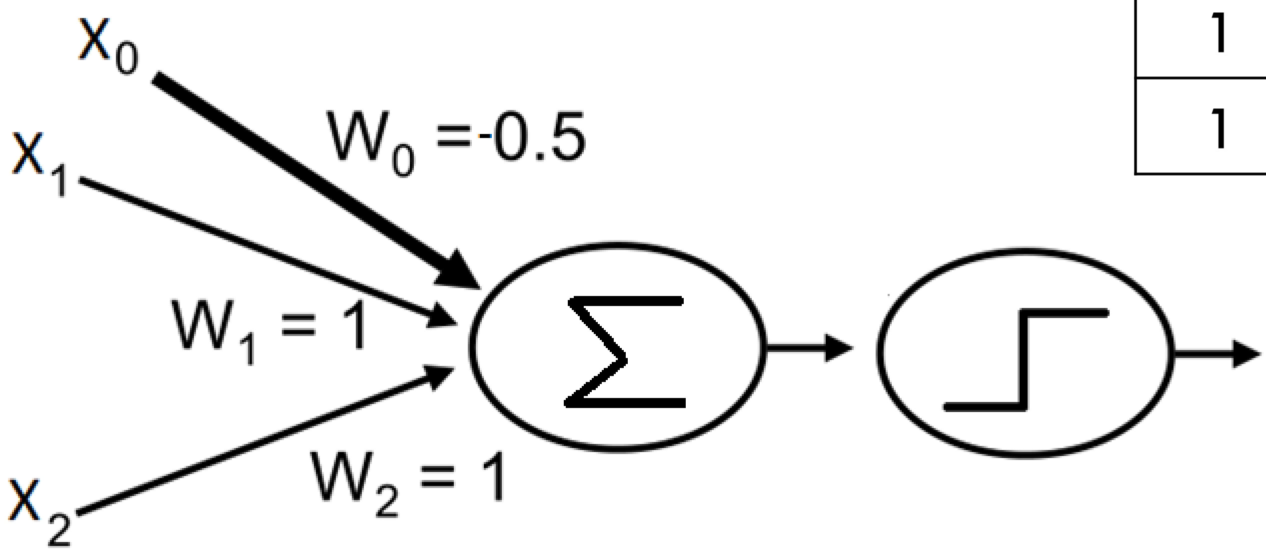
AND Function



X_1	X_2	Y
0	0	0
0	1	0
1	0	0
1	1	1

OR Function

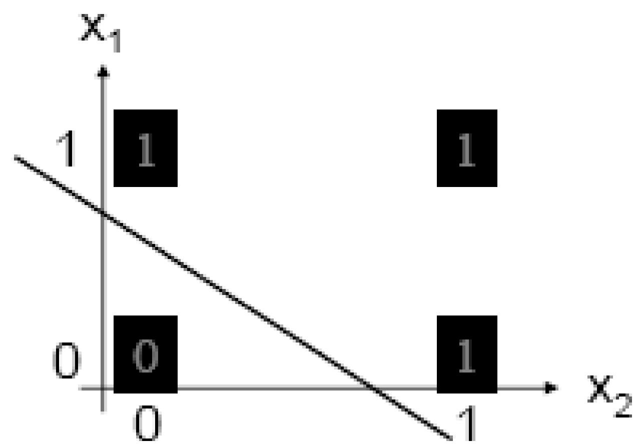
X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1



XOR Function

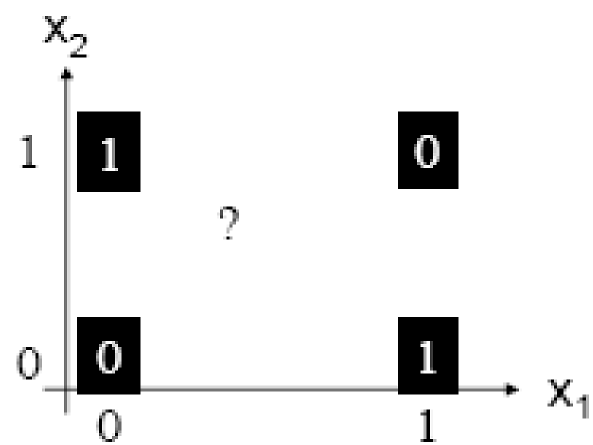
OR function

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



XOR function

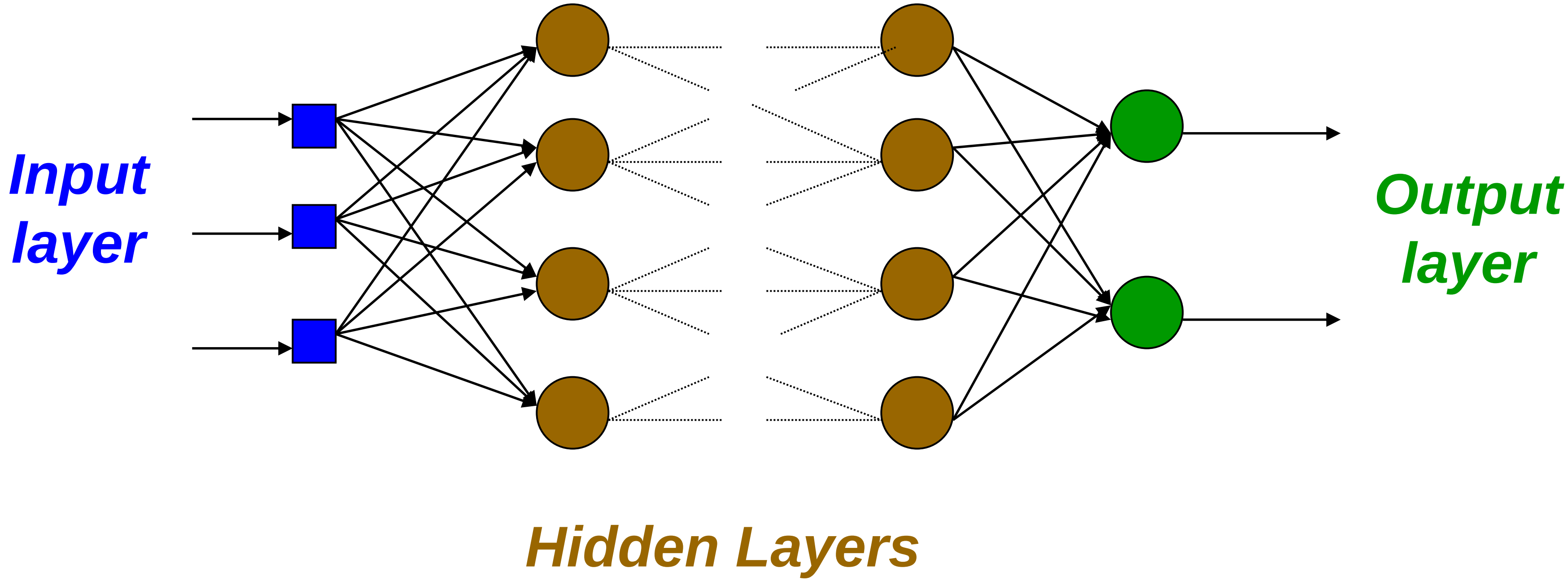
x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0



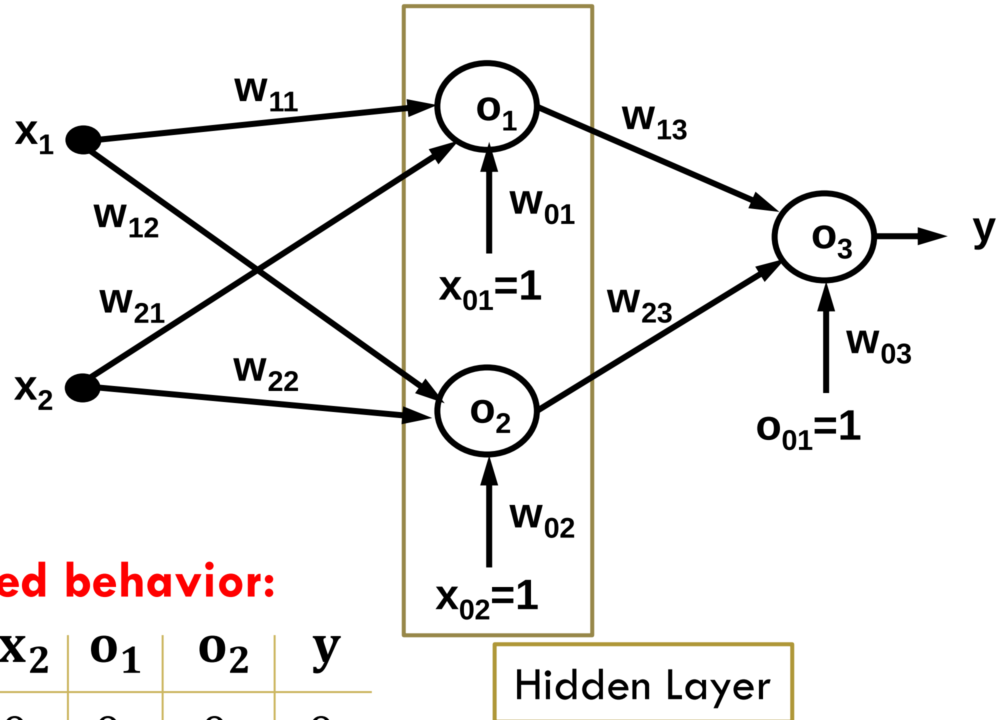
MULTILAYER PERCEPTRON

Multilayer Perceptron Architecture

MLP used to describe any general feedforward (no recurrent connections) network



Multilayer Networks



Network Topology:

2 hidden nodes

1 output

Weights:

$$w_{11} = w_{12} = 1$$
$$w_{21} = w_{22} = 1$$
$$w_{01} = -1.5$$
$$w_{02} = -0.5$$
$$w_{13} = -1$$
$$w_{23} = 1$$
$$w_{03} = -0.5$$

Desired behavior:

x_1	x_2	o_1	o_2	y
0	0	0	0	0
1	0	0	1	1
0	1	0	1	1
1	1	1	1	0

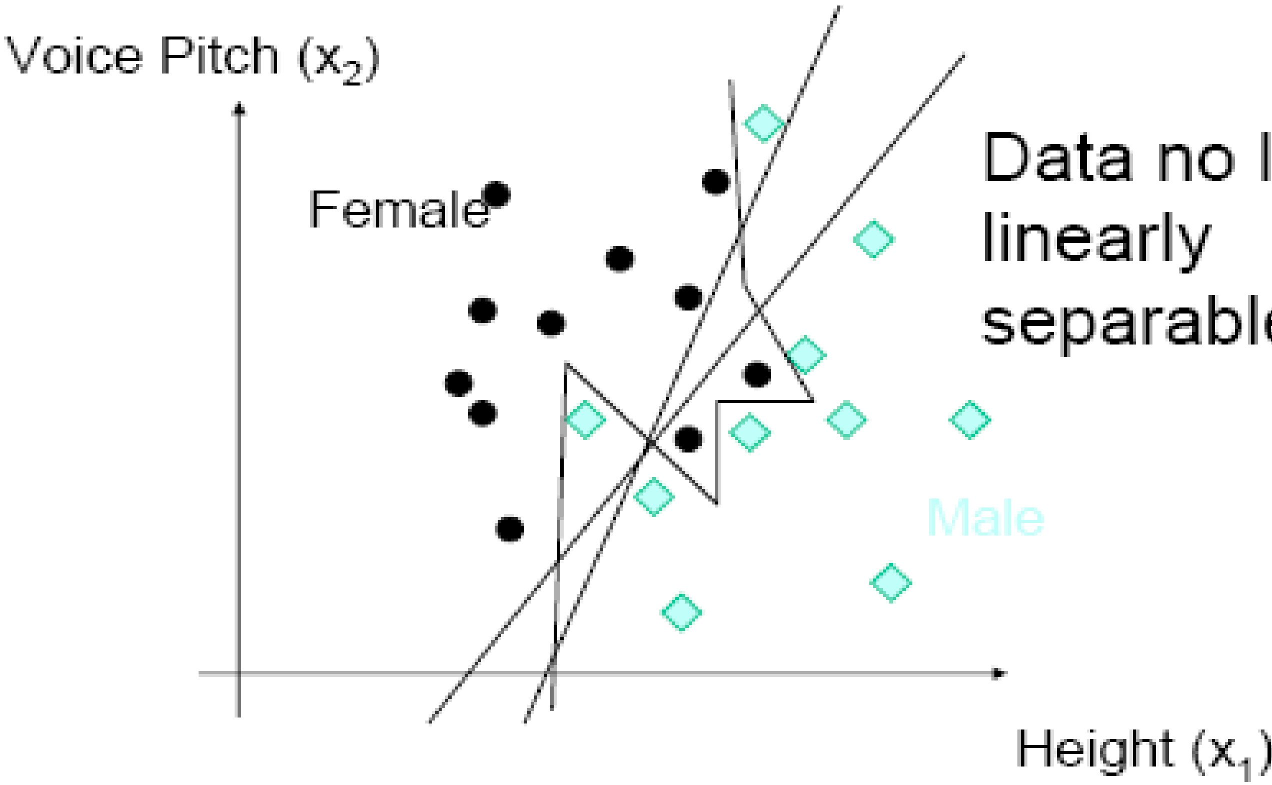
Piecewise linear classification using an MLP with threshold (perceptron) units

Multilayer Networks

- The single perceptron can only express **linear decision surfaces**.

- The kind of **multilayer networks** learned by the **back propagation** algorithm are capable of expressing a rich variety of **nonlinear decision surfaces**.

Multilayer Networks… Example



Data no longer linearly separable

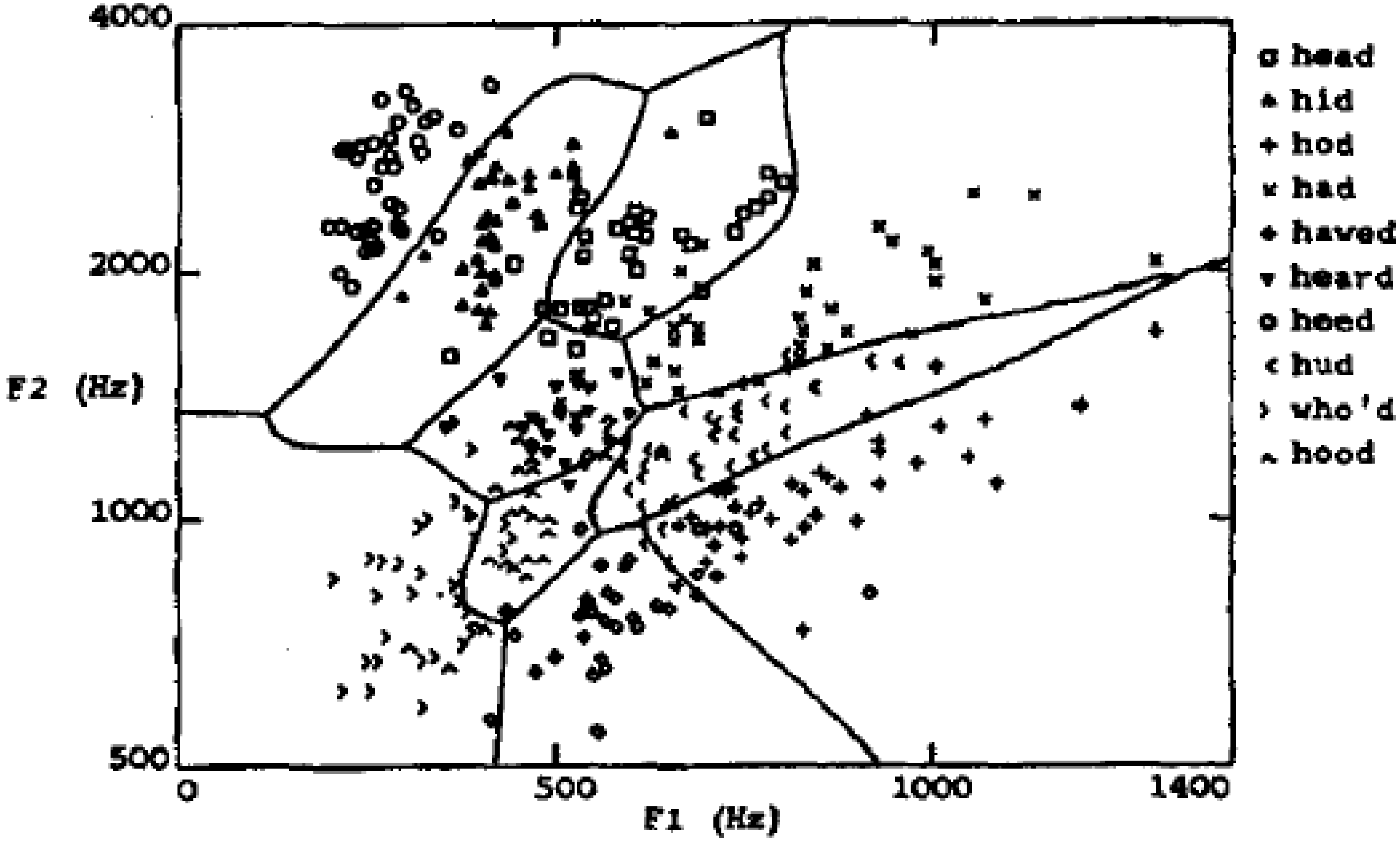
What is a good decision boundary ?

Multilayer Networks… Example

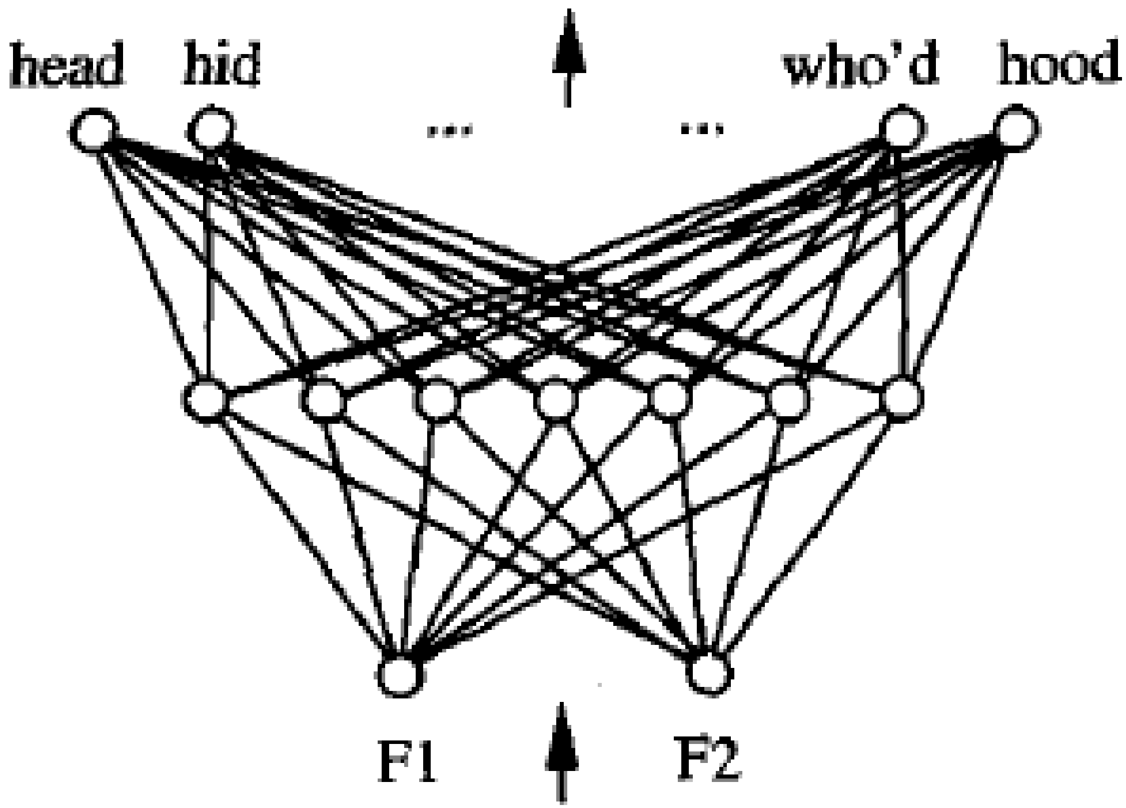
Example:

- The speech recognition task involves distinguishing among 10 possible vowels, all spoken in the context of "h-d" (i.e., "hid," "had," "head," "hood," etc.).

Multilayer Networks… Example



Multilayer Networks… Example



Multilayer Networks

- **What type of <u>unit</u> shall we use as the basis for constructing multilayer networks?**

- Can we use the delta/gradient descent learning rule?
  - **<u>multi-layers of linear units</u>**… multiple layers of cascaded linear units still produce only linear functions, and we prefer networks capable of representing highly nonlinear functions.

- The perceptron unit is another possible choice, is it?
  - its discontinuous threshold makes it undifferentiable and hence unsuitable for gradient descent.

Multilayer Networks

Solution:

- One solution is the **sigmoid unit:**
  - a unit very much like a perceptron, but based on a **smoothed, differentiable threshold function.**
- Like the perceptron, **the sigmoid unit**,
  - first computes a linear combination of its inputs,
  - then applies a threshold to the result. However, the threshold output is a continuous function of its input.

Multilayer Networks

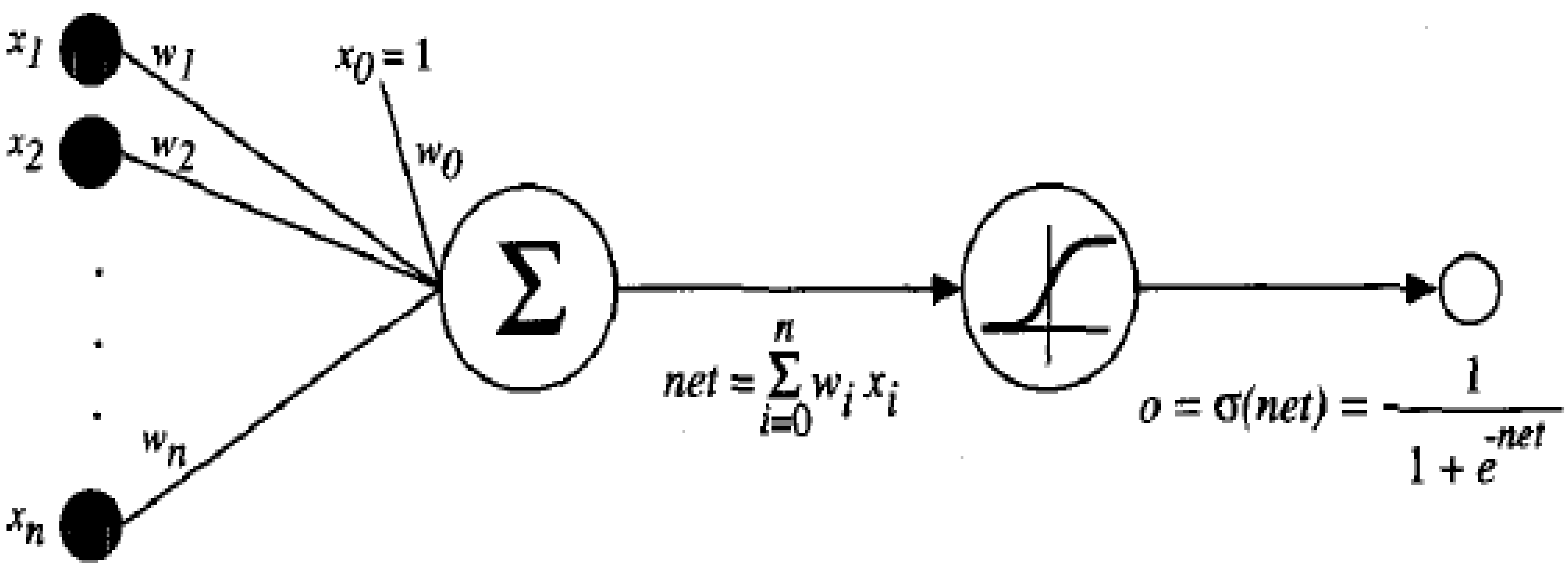
- In case of **sigmoid unit**, however, the **threshold output is a continuous function** of its input.
- More precisely, the sigmoid unit computes its output ***o*** as,

$$o = \sigma(\vec{w} \cdot \vec{x})$$

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

- σ is often called the sigmoid function or, alternatively, the logistic function.

Sigmoid Threshold Unit



Sigmoid Function

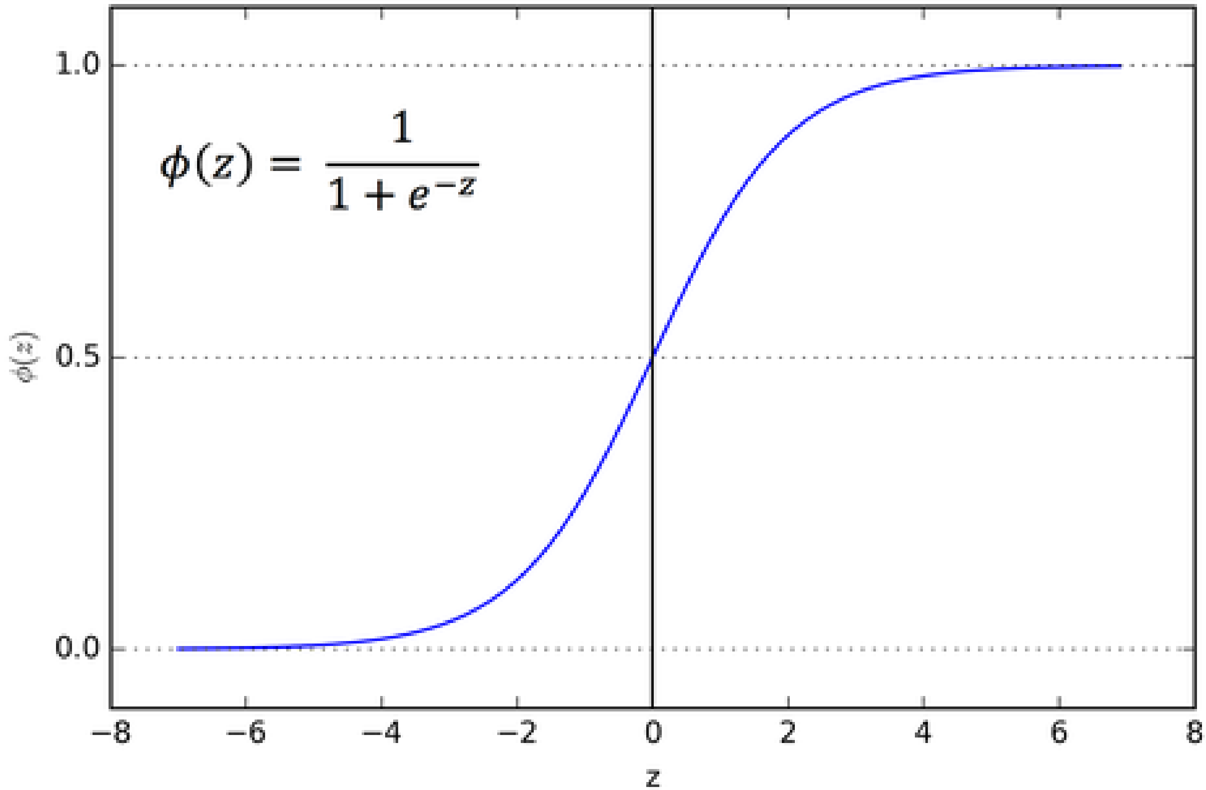
- Sigmoid function maps a very large input domain to a small range of outputs, it is often referred to as the ***squashing function*** of the unit.
- The sigmoid function has the **useful property** that **its derivative is easily expressed in terms of its output**.

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

$$\frac{d\sigma(y)}{dy} = \sigma(y) \cdot (1 - \sigma(y))]$$

Sigmoid Function

- Sigmoid function exists between 0 and 1.



Sigmoid Function

- The term e^{-y} in the sigmoid function definition is sometimes replaced by $e^{-k.y}$
  - where k is some positive constant that *determines the steepness*.
- The function ***tanh*** is also sometimes used in place of the sigmoid function.

Tangent Function

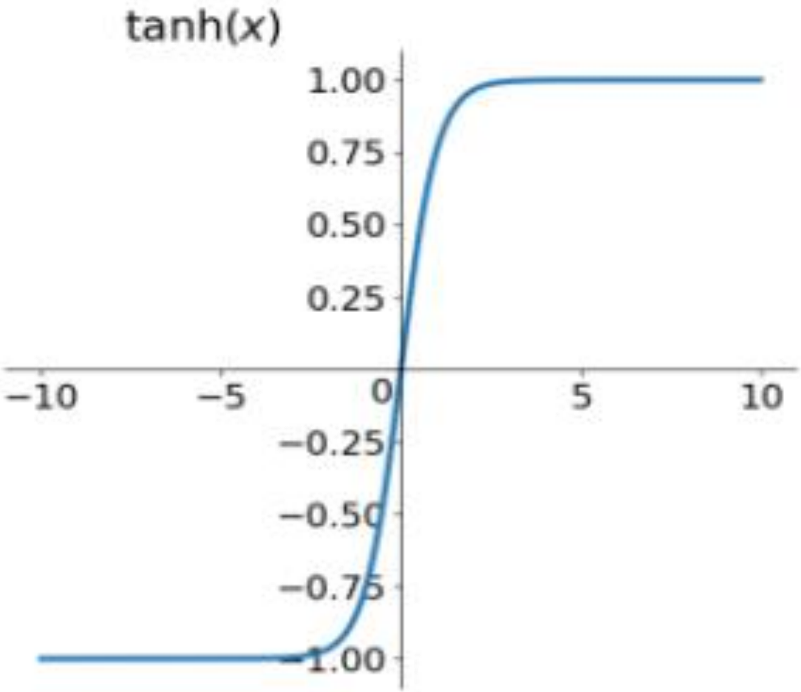
- The tangent function ***tanh*** is:

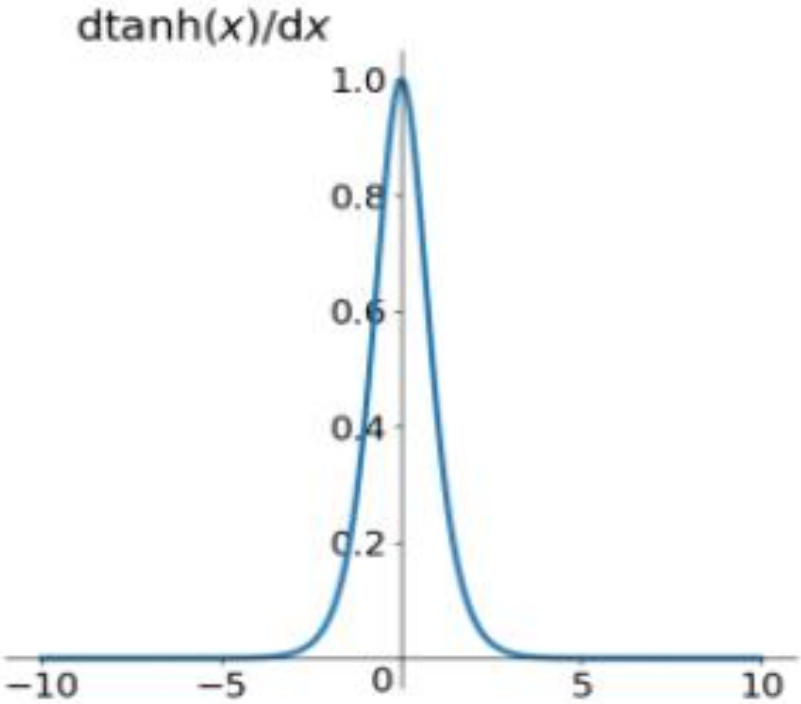
$$\tanh(x) = \frac{e^{x} + e^{-x}}{e^{x} - e^{-x}}$$

- Its derivative can also easily be expressed as;

$$\tanh'(x) = 1 - \tanh^2(x)$$

Tangent Function



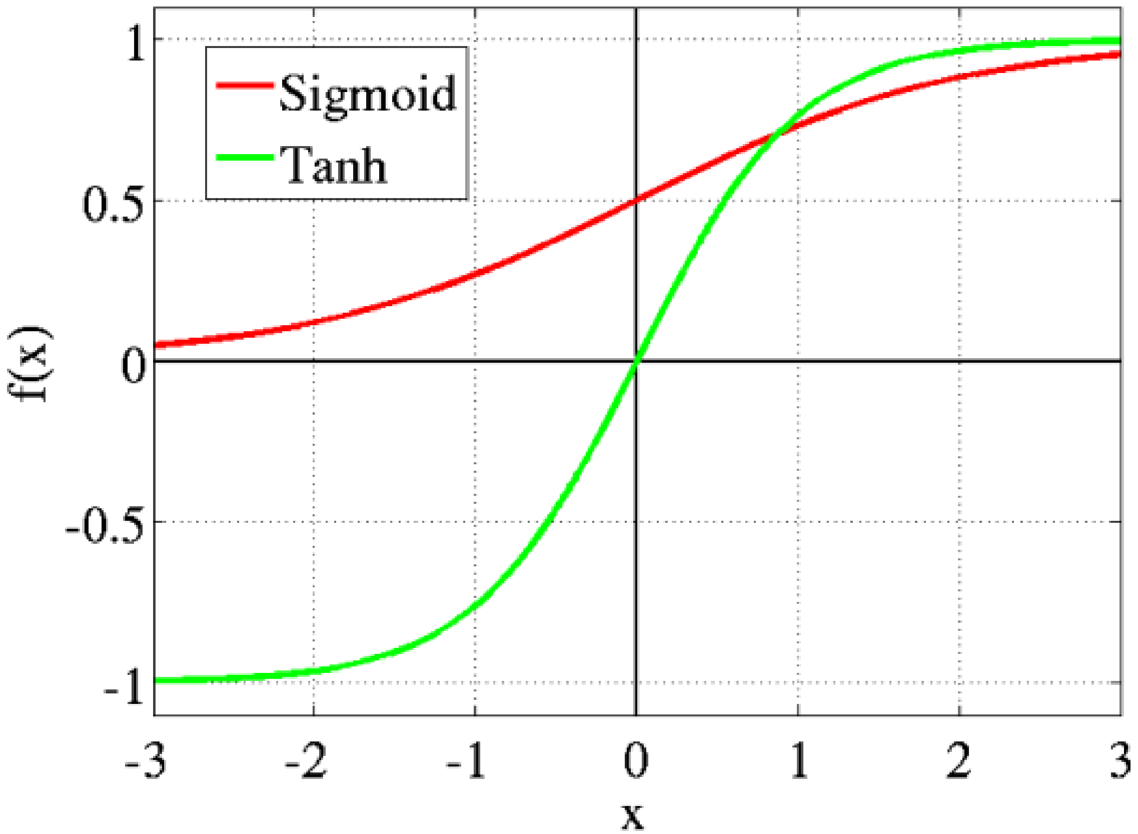


Sigmoid vs Tangent

- **Smooth gradient,** preventing "jumps" in output values.
- **Output values bound** between 0 and 1, normalizing the output of each neuron.

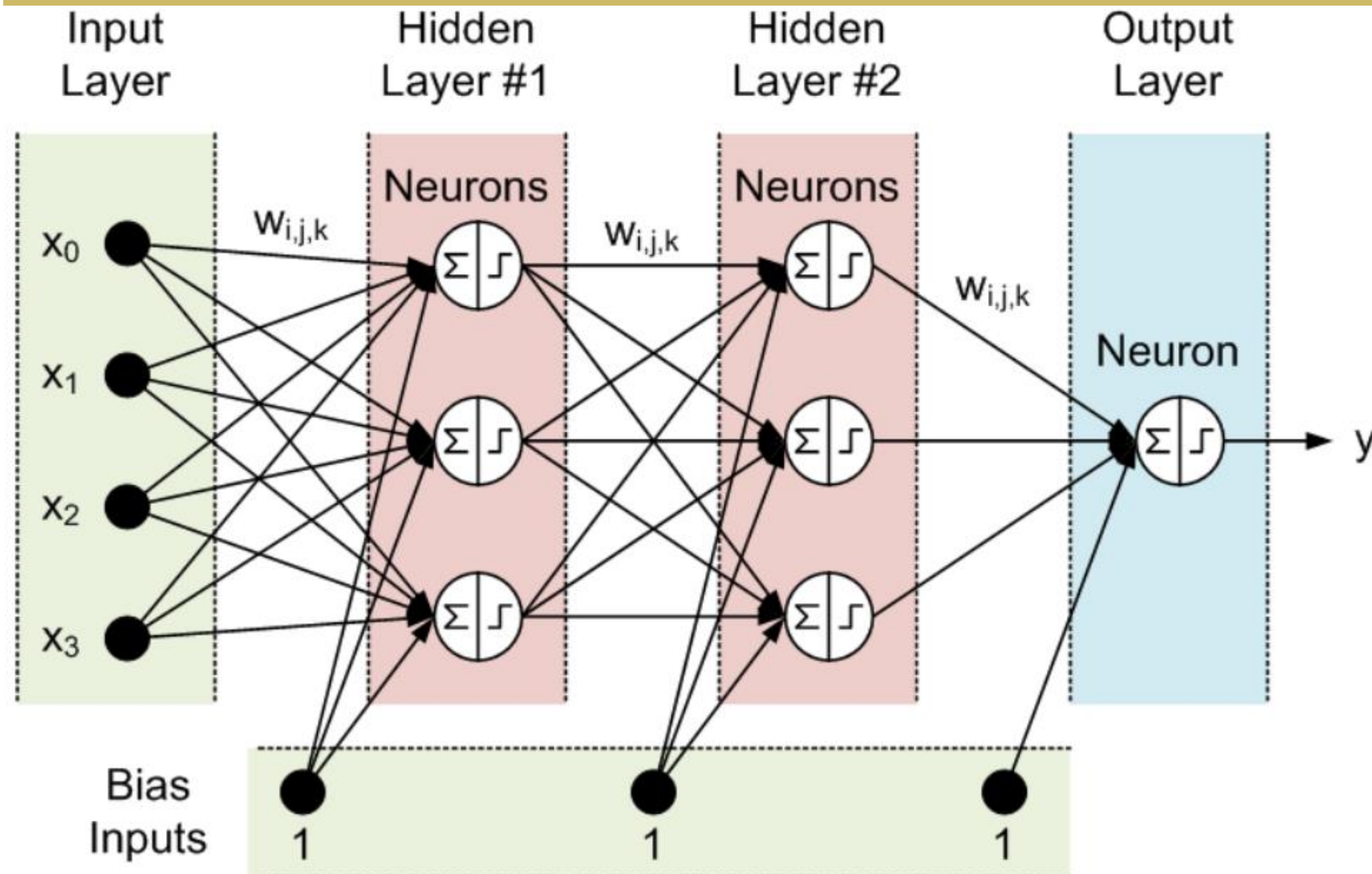
- **Zero centered**—making it easier to model inputs that have strongly negative, neutral, and strongly positive values.
- Its range is between -1 to 1.

Sigmoid vs Tangent

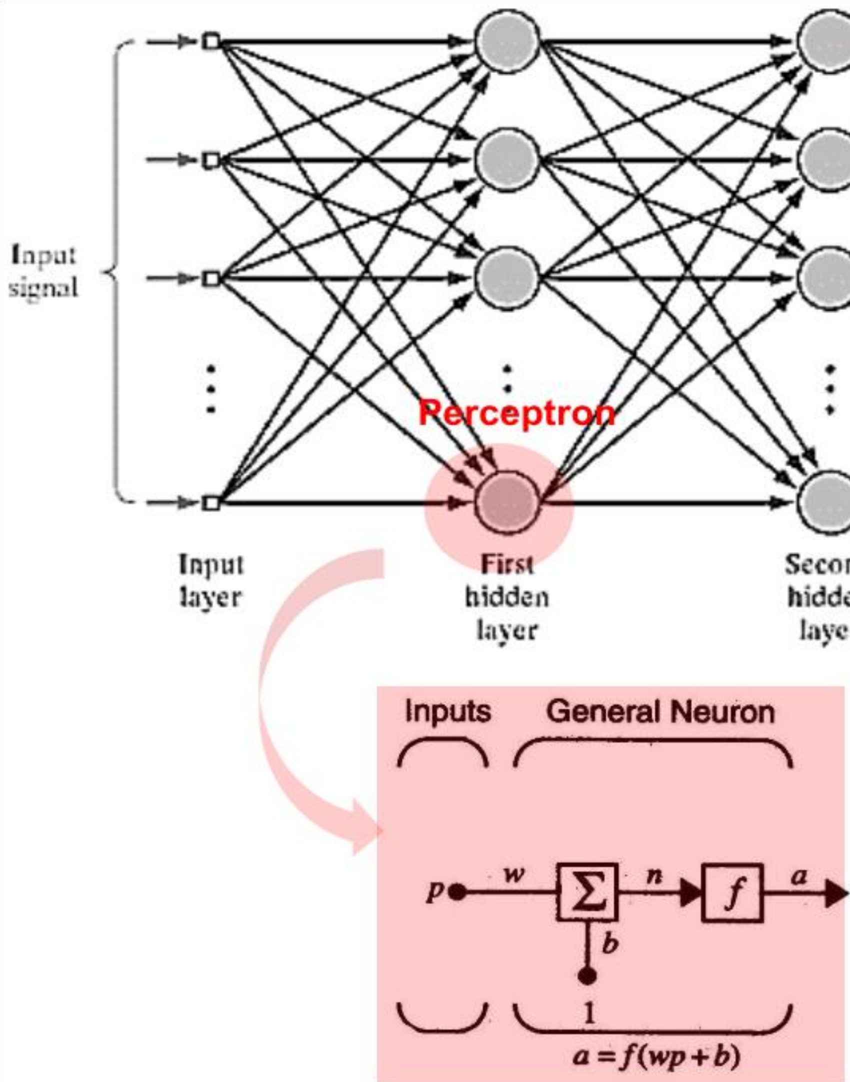


MULTILAYER NETWORKS

Multilayer Perceptron Architecture



Multilayer Network Architecture



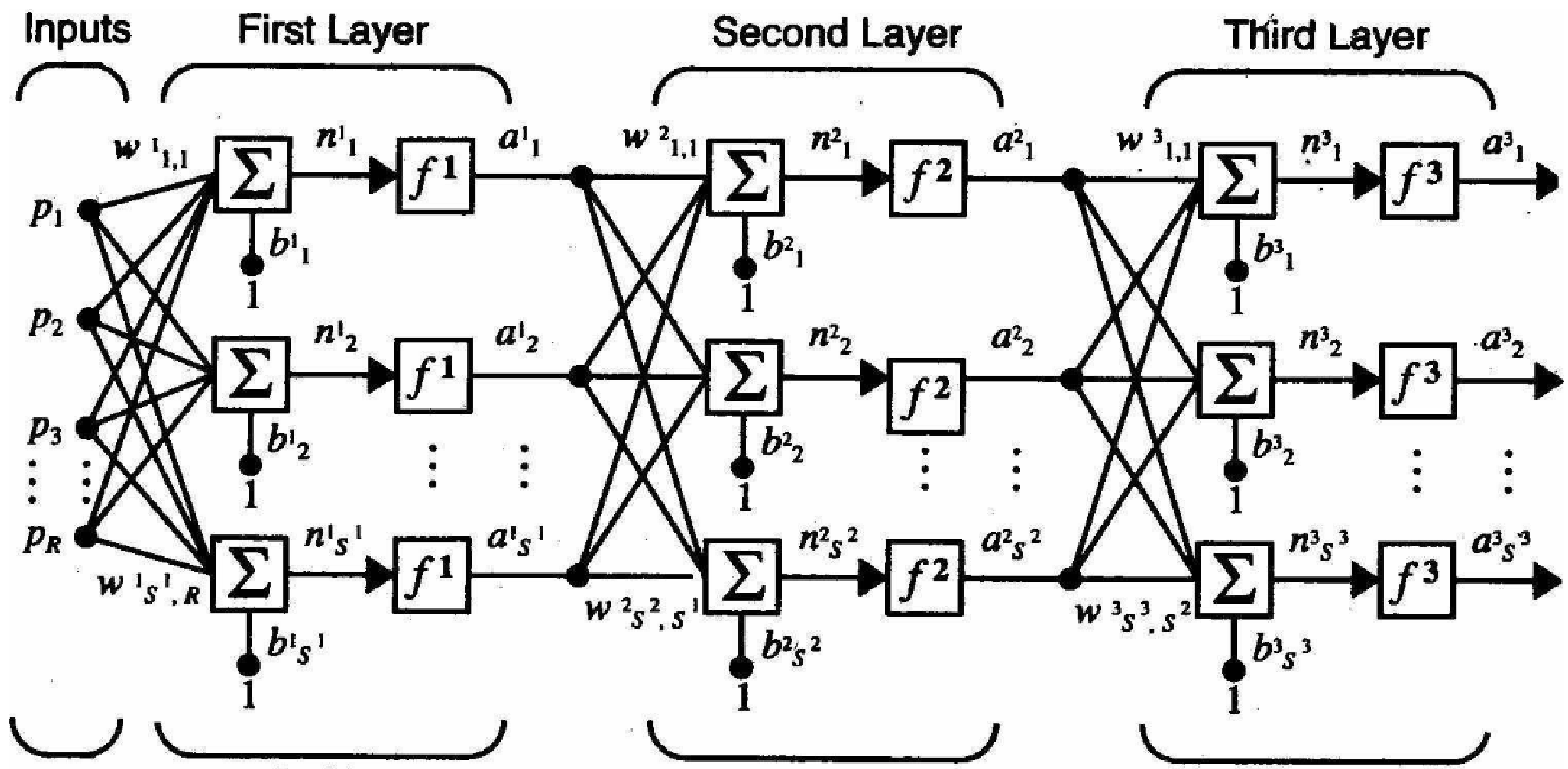
The neuron output is calculated as

$$a = f(wp + b) .$$

If, for instance, $w = 3$, $p = 2$ and $b = -1.5$, then

$$a = f(3(2) - 1.5) = f(4.5)$$

Multilayer Network Architecture



$$\mathbf{a}^1 = \mathbf{f}^1(\mathbf{W}^1\mathbf{p}+\mathbf{b}^1) \qquad \mathbf{a}^2 = \mathbf{f}^2(\mathbf{W}^2\mathbf{a}^1+\mathbf{b}^2) \qquad \mathbf{a}^3 = \mathbf{f}^3(\mathbf{W}^3\mathbf{a}^2+\mathbf{b}^3)$$

$$\mathbf{a}^3 = \mathbf{f}^3(\mathbf{W}^3\mathbf{f}^2(\mathbf{W}^2\mathbf{f}^1(\mathbf{W}^1\mathbf{p}+\mathbf{b}^1)+\mathbf{b}^2)+\mathbf{b}^3)$$

MLP – A static composite (nested) function

Reading Material

- **Artificial Intelligence,** A Modern Approach

  Stuart J. Russell and Peter Norvig
  - **Chapter 18.**
- **Machine Learning**

  Tom M. Mitchell
  - **Chapter 4.**