COSC 522 – Machine Learning

Backpropagation (BP) and Multi-Layer Perceptron (MLP)

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu

Questions

- Limitations of perceptron
- Why go deeper?

- MLP structure
- MLP cost function and optimization method (BP)
- The importance of the threshold function
- Relationship between BPNN and MPP

- Various aspects of practical improvements of BPNN

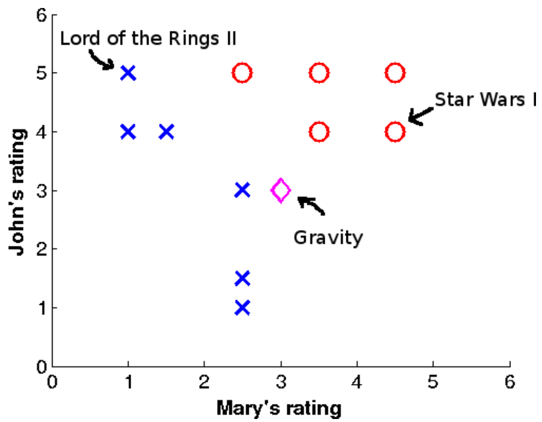
Limitations of Perceptron

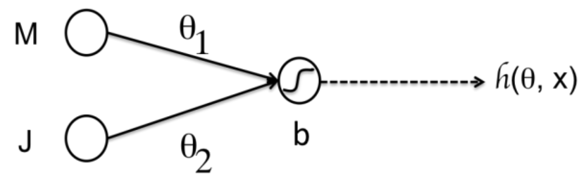
AICIP RESEARCH

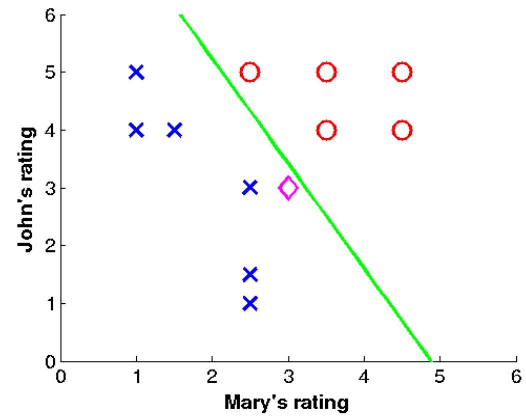
- The output only has two values (1 or 0)
- Can only classify samples which are linearly separable (straight line or straight plane)
- Single layer: can only train AND, OR, NOT
- Can’t train a network functions like XOR

Why deeper?

Movie name	Mary's rating	John's rating	I like?
Lord of the Rings II	1	5	No
...	...	...	...
Star Wars I	4.5	4	Yes
Gravity	3	3	?

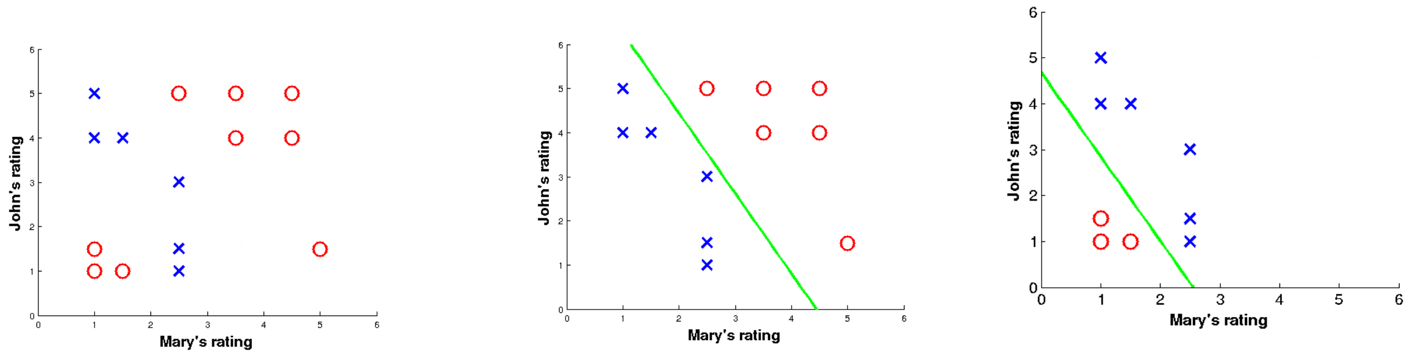




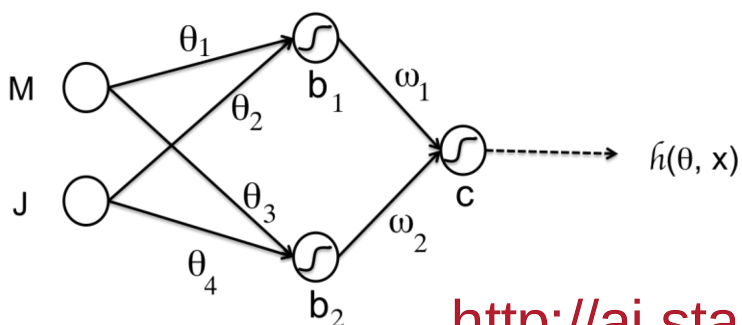


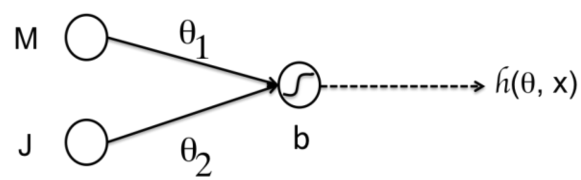
http://ai.stanford.edu/~quocle/tutorial2.pdf

Why deeper?



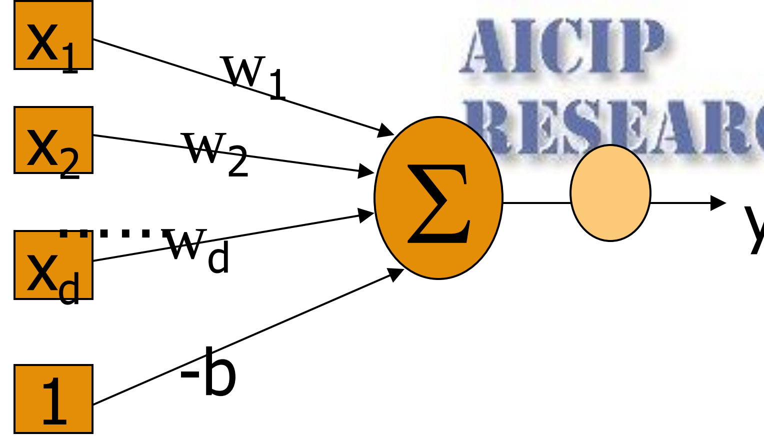
Movie name	Output by decision function h_1	Output by decision function h_2	Susan likes?
Lord of the Rings II	$h_1(x^{(1)})$	$h_2(x^{(2)})$	No
...	...	...	...
Star Wars I	$h_1(x^{(n)})$	$h_2(x^{(n)})$	Yes
Gravity	$h_1(x^{(n+1)})$	$h_2(x^{(n+1)})$	?

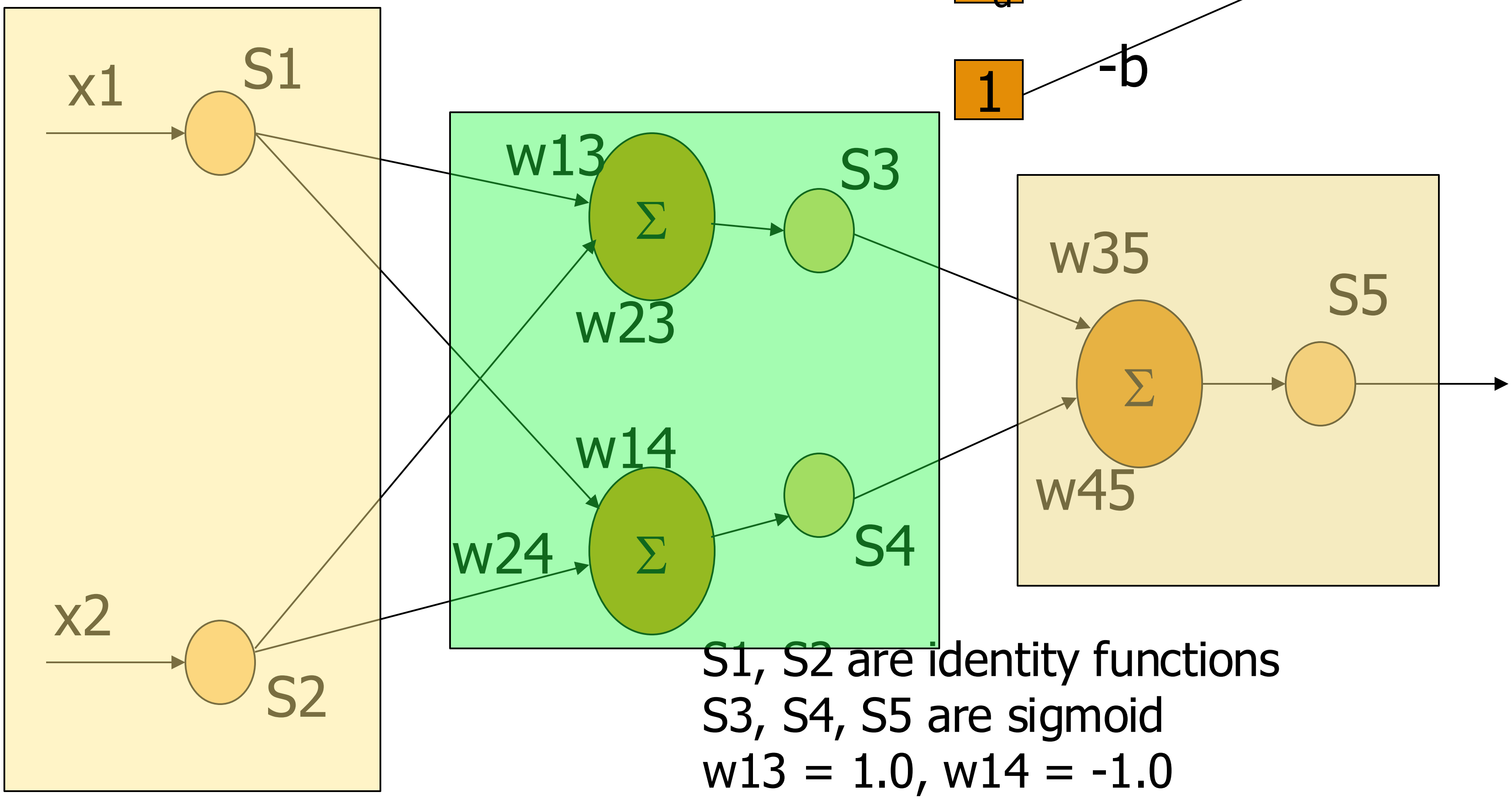




http://ai.stanford.edu/~quocle/tutorial2.pdf

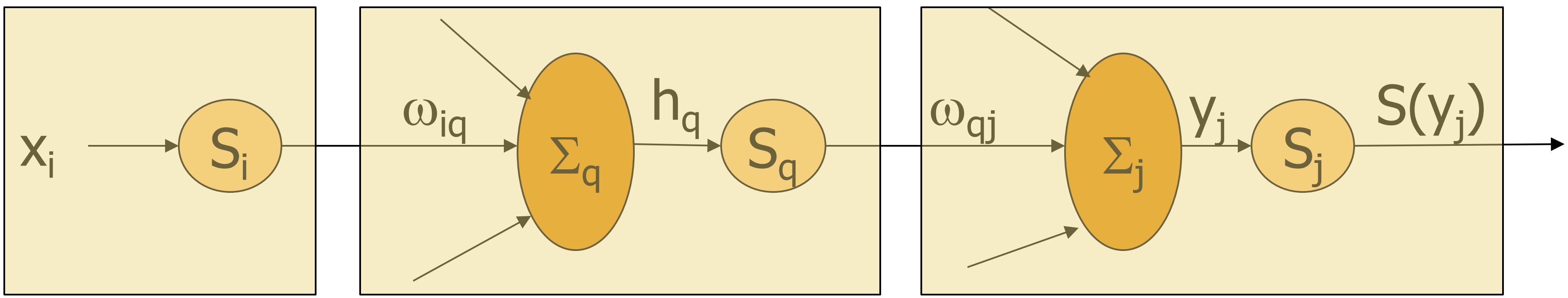
XOR (3-layer NN)





S1, S2 are identity functions
S3, S4, S5 are sigmoid
w13 = 1.0, w14 = -1.0
w24 = 1.0, w23 = -1.0
w35 = 0.11, w45 = -0.1
The input takes on only –1 and 1

MLP – 3-Layer Network



$$E = \frac{1}{2}\sum_j \left(T_j - S\left(y_j\right)\right)^2$$

Choose a set of initial w_{st}

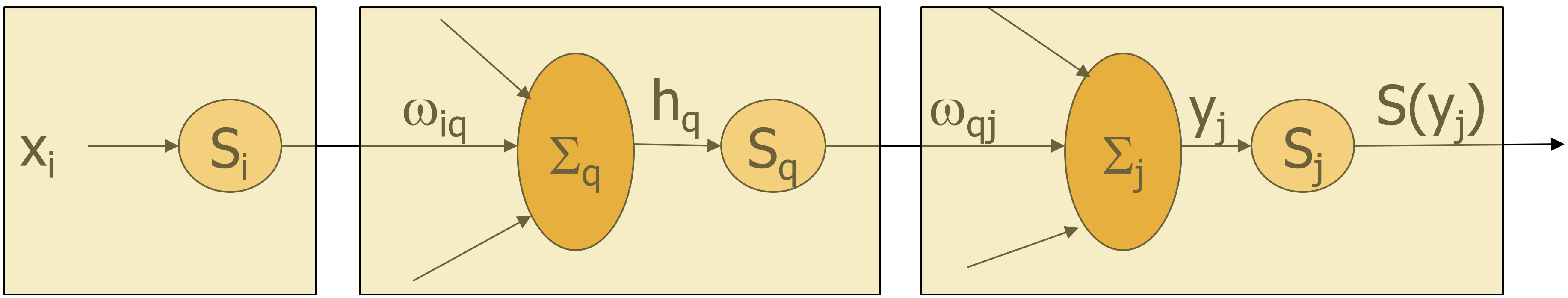
$$w_{st}^{k+1} = w_{st}^{k} - c^{k}\frac{\partial E^{k}}{\partial w_{st}^{k}}$$

ω_{st} is the weight connecting input *s* at neuron *t*

The problem is essentially "how to choose weight ω to minimize the error between the expected output and the actual output"

The basic idea behind BP is gradient descent

Exercise

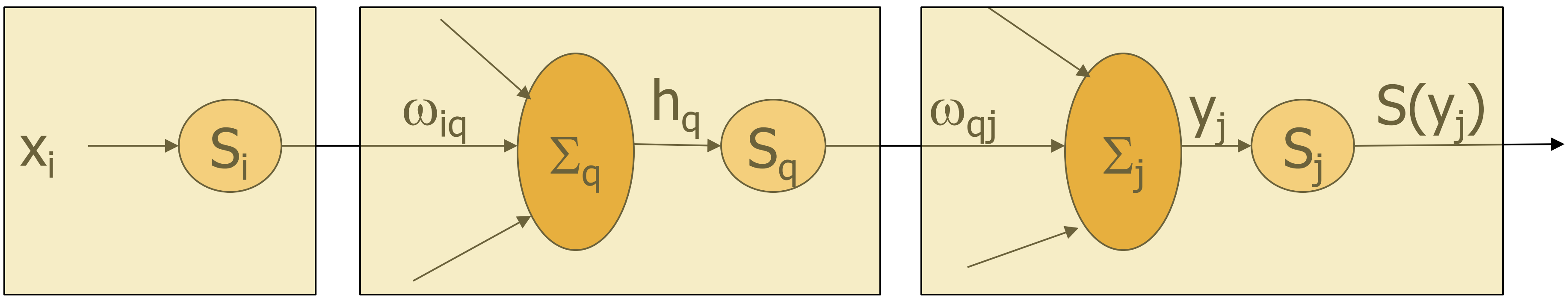


$$y_j = \sum_q S_q(h_q) w_{qj} \Rightarrow \frac{\partial y_j}{\partial S_q} = w_{qj} \quad \text{and} \quad \frac{\partial y_j}{\partial w_{qj}} = S_q(h_q)$$

$$h_q = \sum_i x_i w_{iq} \Rightarrow \frac{\partial h_q}{\partial x_i} = w_{iq} \quad \text{and} \quad \frac{\partial h_q}{\partial w_{iq}} = x_i$$

AICIP RESEARCH

The Derivative – Chain Rule



$$Dw_{qj} = -\frac{\partial E}{\partial w_{qj}} = -\frac{\partial E}{\partial S_j}\frac{\partial S_j}{\partial y_j}\frac{\partial y_j}{\partial w_{qj}}$$

$$= -\left(T_j - S_j\right)\left(S_j'\right)\left(S_q\left(h_q\right)\right)$$

$$Dw_{iq} = -\frac{\partial E}{\partial w_{iq}} = \left[\sum_j \frac{\partial E}{\partial S_j}\frac{\partial S_j}{\partial y_j}\frac{\partial y_j}{\partial S_q}\right]\frac{\partial S_q}{\partial h_q}\frac{\partial h_q}{\partial w_{iq}}$$

$$= \left[\sum_j \left(T_j - S_j\right)\left(S_j'\right)\left(w_{qj}\right)\right]\left(S_q'\right)\left(x_i\right)$$

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Threshold Function

AICIP RESEARCH

- Traditional threshold function as proposed by McCulloch-Pitts is binary function
- The importance of differentiable
- A threshold-like but differentiable form for S (25 years)
- The sigmoid

$$S(x) = \frac{1}{1 + \exp(-x)}$$

BP vs. MPP

$$E(\omega) = \sum_{\mathbf{x}} [g_k(\mathbf{x};\mathbf{w}) - T_k]^2 = \sum_{\mathbf{x}\in\omega_k} [g_k(\mathbf{x};\mathbf{w}) - 1]^2 + \sum_{\mathbf{x}\notin\omega_k} [g_k(\mathbf{x};\mathbf{w}) - 0]^2$$

$$= n\left\{ \frac{n_k}{n} \frac{1}{n_k} \sum_{\mathbf{x}\in\omega_k} [g_k(\mathbf{x};\mathbf{w}) - 1]^2 + \frac{n-n_k}{n} \frac{1}{n-n_k} \sum_{\mathbf{x}\notin\omega_k} [g_k(\mathbf{x};\mathbf{w})]^2 \right\}$$

$$\lim_{n\to\infty} \frac{1}{n} E(\mathbf{w}) = P(\omega_k) \int [g_k(\mathbf{x};\mathbf{w}) - 1]^2 p(\mathbf{x} \mid \omega_k) d\mathbf{x} + P(\omega_{i\neq k}) \int g_k^2(\mathbf{x};\mathbf{w}) p(\mathbf{x} \mid \mathbf{w}_{i\neq k}) d\mathbf{x}$$

$$= \int [g_k{}^2(\mathbf{x};\mathbf{w}) - 2g_k(\mathbf{x};\mathbf{w}) + 1] p(\mathbf{x},\omega_k) d\mathbf{x} + \int g_k^2(\mathbf{x};\mathbf{w}) p(\mathbf{x},\mathbf{w}_{i\neq k}) d\mathbf{x}$$

$$= \int g_k{}^2(\mathbf{x};\mathbf{w}) p(\mathbf{x}) d\mathbf{x} - 2\int g_k(\mathbf{x};\mathbf{w}) p(\mathbf{x},\omega_k) d\mathbf{x} + \int p(\mathbf{x},\omega_k) d\mathbf{x}$$

$$= \int [g_k(\mathbf{x};\mathbf{w}) - P(\omega_k \mid \mathbf{x})]^2 p(\mathbf{x}) d\mathbf{x} + C$$