Lecture Series on

Intelligent Control

Lecture 7
Artificial Neural Networks
Backpropagation

Kwang Y. Lee
Professor of Electrical & Computer Engineering
Baylor University
Waco, TX 76706, USA
Kwang_Y_Lee@baylor.edu

1

Neural Network Terminology

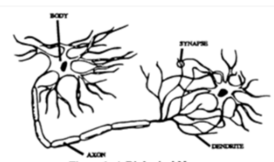


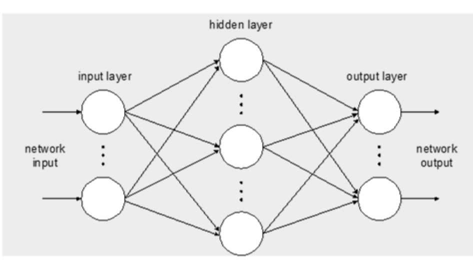
Figure 1: A Biological Neuron

A Biological Neuron Model:

- Each neuron in the brain is composed of *body, axon,* and a multitude of *dendrites*.
- Axon is a long tube which splits into branches terminating in endbulbs, almost touching the dendrites of other cells.
- The small gap between an endbulb and a dendrite is called *synapse*.

2

Neural Network Terminology

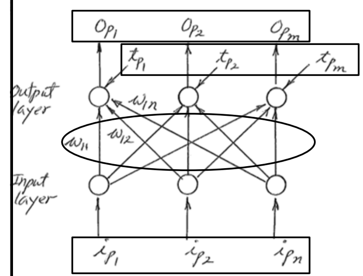


A Feedforward Multi-layer Network: each circle corresponds to a node and each arrow represents a weighted link.

3

Backpropagation

Delta Rule



p : pattern

i_p : input vector for pattern p

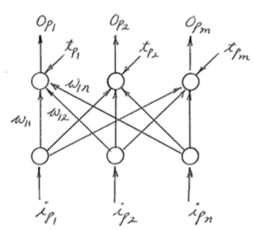
O_p : output vector for pattern p

t_p : target vector for pattern p

w_{ij} : weight from jth input to ith output component.

4

Delta Rule:



For a pattern p, the error is

$$E_p = \frac{1}{2} \sum_j (t_{p_j} - O_{p_j})^2 , \quad \epsilon = \sum_p \epsilon_p$$

The gradient is, by chain rule,

$$\frac{\partial \epsilon_p}{\partial w_{ji}} = \frac{\partial \epsilon_p}{\partial O_{p_j}} \frac{\partial O_{p_j}}{\partial w_{ji}}$$

$$\frac{\partial \epsilon_p}{\partial O_{p_j}} = -(t_{p_j} - O_{p_j}) \triangleq -\delta_{p_j}$$

Assuming linear units,

$$O_{p_j} = \sum_i w_{ji} i_{p_i} \Rightarrow \frac{\partial O_{p_j}}{\partial w_{ji}} = i_{p_i}$$

Therefore, the negative gradient is

$$-\frac{\partial \epsilon_p}{\partial w_{ji}} = \delta_{p_j} i_{p_i}$$

$\therefore$ The weight w_{ji} can be updated by

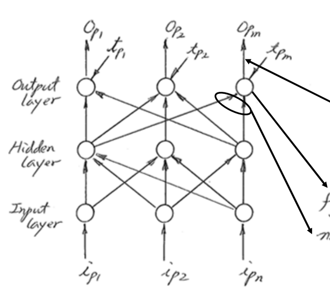
$$\Delta_p w_{ji} = \eta \, \delta_{p_j} i_{p_i} , \quad \underline{\eta : \text{learning rate}}$$

Since $\frac{\partial \epsilon}{\partial w_{ji}} = \sum_p \frac{\partial \epsilon_p}{\partial w_{ji}}$,

the Delta rule implements a gradient descent on E.

5

Generalized Delta Rule:



Now we add hidden layers (can be multiple).

Assume semilinear activation function:

$$O_{p_j} = f_j(net_{p_j}) \quad \text{--- (1)}$$

$$net_{p_j} = \sum_i w_{ji} O_{p_i} \quad \text{-- (2)}$$

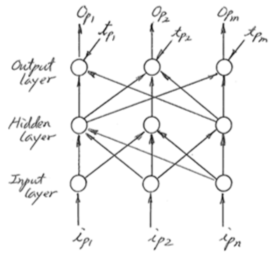
f_j : differentiable, nondecreasing

net_{p_j} : weighted sum of inputs to jth neuron for pattern p.

Note : inputs to jth neuron are outputs of ith neuron in the previous layer.

6

Generalized Delta Rule:



$$O_{p_j} = f_j(net_{p_j}) \quad \text{- - - (1)}$$

$$net_{p_j} = \sum_i w_{ji} O_{p_i} \quad \text{-- (2)}$$

Define the error function:

$$E_p = \tfrac{1}{2}(t_{p_j} - O_{p_j})^2, \qquad E = \sum_p E_p$$

Then the gradient is

$$\frac{\partial E_p}{\partial w_{ji}} = \underbrace{\frac{\partial E_p}{\partial net_{p_j}}} \; \underbrace{\frac{\partial net_{p_j}}{\partial w_{ji}}}_{O_{p_i} \text{ from (2)}}$$

The negative gradient is

$$\Rightarrow \; -\frac{\partial E_p}{\partial w_{ji}} = \delta_{p_j} O_{p_i} \quad \cdots (3)$$

7

Generalized Delta Rule:

$$E_p = \tfrac{1}{2}(t_{p_j} - O_{p_j})^2, \qquad E = \sum_p E_p$$

Then the gradient is

$$\frac{\partial E_p}{\partial w_{ji}} = \underbrace{\frac{\partial E_p}{\partial net_{p_j}}} \; \underbrace{\frac{\partial net_{p_j}}{\partial w_{ji}}}_{O_{p_i} \text{ from (2)}}$$

The negative gradient is

$$\Rightarrow \; -\frac{\partial E_p}{\partial w_{ji}} = \delta_{p_j} O_{p_i} \quad \cdots (3)$$

$$O_{p_j} = f_j(net_{p_j}) \quad \text{- - - (1)}$$

$$net_{p_j} = \sum_i w_{ji} O_{p_i} \quad \text{-- (2)}$$

Define

$$\delta_{p_j} \triangleq -\frac{\partial E_p}{\partial net_{p_j}}$$

$$= -\frac{\partial E_p}{\partial O_{p_j}} \underbrace{\frac{\partial O_{p_j}}{\partial net_{p_j}}}_{f_j'(net_{p_j}) \text{ from (1)}} \quad \text{- - - - - - - - - - (4)}$$

Now we have two different cases to compute the first term: $\frac{\partial E_p}{\partial O_{p_j}}$

8

Generalized Delta Rule:

Case I. Output layer

$$\frac{\partial E_p}{\partial O_{p_j}} = -(t_{p_j} - O_{p_j}) \qquad E_p = \tfrac{1}{2}(t_{p_j} - O_{p_j})^2,$$

Thus, from (4),

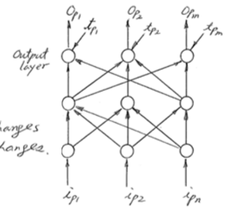
$$\underline{\delta_{p_j} = (t_{p_j} - O_{p_j}) f_j'(net_{p_j})} \qquad (5)$$

$$\delta_{p_j} \triangleq -\frac{\partial E_p}{\partial net_{p_j}}$$

$$= -\frac{\partial E_p}{\partial O_{p_j}} \underbrace{\frac{\partial O_{p_j}}{\partial net_{p_j}}}_{f_j'(net_{p_j}) \text{ from (1)}} \quad \text{- - - - - - - - - - (4)}$$

9

Generalized Delta Rule:



Case II. Hidden layers

Note that when the output O_{pj} of hidden layer changes the net inputs net_{pk} to the output layer changes.

$$net_{pk} = \sum_i w_{ki} O_{pi}$$

Thus,

$$\frac{\partial E_p}{\partial O_{pj}} = \sum_k \frac{\partial E_p}{\partial net_{pk}} \underbrace{\frac{\partial net_{pk}}{\partial O_{pj}}}$$

$$= \sum_k \underbrace{\frac{\partial E_p}{\partial net_{pk}}}_{-\delta_{pk}} w_{kj}$$

$-\delta_{pk}$: already calculated at the output layer.

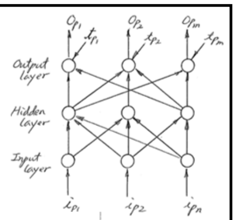
$$= -\sum_k \delta_{pk} w_{kj}$$

Thus, from (4),

$$\underline{\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj}}$$: error is propagated backward (6)

10

Generalized Delta Rule:



The negative gradient is

$$-\frac{\partial E_p}{\partial w_{ji}} = \delta_{pj} O_{pi} \cdots (3)$$

Thus, from (4),

$$\underline{\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj}}$$: error is propagated backward (6)

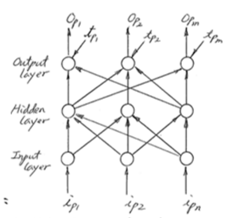
The weight adjustment is, therefore, in the direction of the negative gradient:

$$\underline{\Delta_p w_{ji} = \eta \, \delta_{pj} O_{pi}} \quad \cdots\cdots\cdots (7)$$

NOTE: δ_{pj} is propagated backward (to neuron j from the next layer).
O_{pi} is propagated forward (from neuron i in the previous layer)

11

Generalized Delta Rule:



Thus, from (4),

$$\underline{\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj}} \quad (6)$$

The weight adjustment is, therefore, in the direction of the negative gradient:

$$\underline{\Delta_p w_{ji} = \eta \, \delta_{pj} O_{pi}} \quad \cdots\cdots\cdots (7)$$

Note that the errors in the output layer (5) are propagated backward through output weights (weights between output layer and hiddenlayer) (6), and multiplied by inputs propagated forward (7) to compute the weight adjustment.

12

Generalized Delta Rule:

Thus, from (4),

$$\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj} \quad \text{: error is propagated backward} \quad (6)$$

Derivative of the activation function, $f_j'(net_{pj})$, can be computed directly without differentiating numerically. For example, for a sigmoidal function

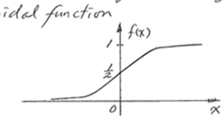
$$f(x) = \frac{1}{1+e^{-(x+T)}}$$

it can be shown that

$$f'(x) = f(x)\,[1 - f(x)]$$

Therefore, from (1),

$$f_j'(net_{pj}) = O_{pj}\,[1 - O_{pj}]$$



Activation function ($T = 0$)

$O_{pj} = f_j(net_{pj})$ - - - (1)

$net_{pj} = \sum_i w_{ji} O_{pi}$ - - (2)

13

Generalized Delta Rule:

Thus, from (4),

$$\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj} \quad \text{: error is propagated backward} \quad (6)$$

The weight adjustment is, therefore, in the direction of the negative gradient:

$$\Delta_p w_{ji} = \eta\, \delta_{pj}\, O_{pi} \quad \cdots\cdots\cdots (7)$$

In general, the momentum is added to avoid local minima:

$$\Delta_p w_{ji}^{new} = \eta\, \delta_{pj}\, O_{pi} + \underline{\alpha\, \Delta_p w_{ji}^{prev}}$$

where the second term on the right hand side is the <u>momentum term</u>. This term adds a portion of the most recent weight change when computing the new weight change. The momentum term is supposed to give the neuron momentum in weight space, enabling it to pass through local minima.

14