I.V.N.S. Rohith Reddy
Amandeep Singh

(1.1)

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times \frac{\partial net_j}{\partial w_{ji}} \quad \text{where } \frac{\partial net_j}{\partial w_{ji}} = x_{ji}$$

$$= \frac{\partial E_d}{\partial net_j} \times x_{ji} \quad -(1) \qquad \text{and } net_j = \sum w_{ji} x_{ji}$$

x_{ji} is the ith input to unit j

Two cases - (1) unit j is an output unit for the network

(2) unit j is a hidden layer unit

Case 1 : Training rule for output unit weights

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \times \frac{\partial o_j}{\partial net_j} \quad -(2)$$

→ where o_j is the output computed by unit j

Consider the first part of eq (2)

$$\frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j} \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

$$\frac{\partial}{\partial o_j} (t_k - o_k)^2 = 0 \quad \forall k \neq j$$

$$\therefore \frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j} \frac{1}{2} (t_j - o_j)^2 \qquad k = j$$

$$= \frac{1}{2}(2)(t_j - o_j) \frac{\partial}{\partial o_j}(t_j - o_j)$$

$$\frac{\partial E_d}{\partial o_j} = -(t_j - o_j) \quad -(3)$$

Now let the activation function used be tanh

$o_j = \tanh(net_j)$

Therefore $\dfrac{\partial o_j}{\partial net_j} = \dfrac{\partial}{\partial net_j}\left[\tanh(net_j)\right]$

$$\boxed{\frac{d}{d\theta}\tanh\theta = 1-\tanh^2\theta}$$

$$= 1 - \tanh^2(net_j)$$

$$= 1 - o_j^2 \quad — (4)$$

From (2), (3), (4)

$$\frac{\partial E_d}{\partial net_j} = -(t_j - o_j)(1 - o_j^2) \quad — (5)$$

Gradient descent rule for output units

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$$

from (1) and (5), $\boxed{\Delta w_{ji} = \eta (t_j - o_j)(1 - o_j^2)\, x_{ji}}$ — (6)

Case 2 : Training rule for hidden unit weights

Downstream(j) refers to set of all units immediately downstream of unit j

$$\therefore \frac{\partial E_d}{\partial net_j} = \sum_{k \in down(j)} \frac{\partial E_d}{\partial net_k} \times \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in downst(j)} -\delta_k \cdot \frac{\partial net_k}{\partial net_j}$$

$$= \sum -\delta_k \times \frac{\partial net_k}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

$$= \sum_{k \in downstr(j)} -\delta_k\, w_{kj} \frac{\partial o_j}{\partial net_j}$$

from ④ $\frac{\partial O_j}{\partial net_j} = 1 - O_j^2$

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in downst(j)} -\delta_k w_{kj} (1 - O_j)^2 \quad - ⑦$$

Let $\delta_j = -\frac{\partial E_d}{\partial net_j}$

then $$\delta_j = (1 - O_j^2) \sum_{k \in downst(j)} \delta_k w_{kj} \quad - ⑧$$

$$\Delta w_{ji} = \eta \delta_j x_{ji} \quad - ⑨$$

Equations ⑧ and ⑨ ~~depict~~ shows the use of $\tanh(x)$ activation function to calculate Δw_{ji}

<u>Using Relu as activation</u>

Output from unit $j = O_j$

$$O_j = Relu(net_j)$$

$$O_j = \begin{cases} 0 & \text{for } net_j < 0 \\ net_j & \text{for } net_j \geq 0 \end{cases}$$

$$\frac{\partial O_j}{\partial net_j} = \begin{cases} 0 & \text{for } net_j < 0 \\ 1 & \text{for } net_j \geq 0 \end{cases}$$

Let $\frac{\partial O_j}{\partial net_j} = O_j'$

Two cases again

.

Case 1 : Training rule for output unit weights

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$$

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial net_j} \times x_{ji}$$

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial O_j} \times \frac{\partial O_j}{\partial net_j} \quad \text{from } 2$$

$$\frac{\partial E_d}{\partial O_j} = -(t_j - O_j)$$

Now $\dfrac{\partial O_j}{\partial net_j} = \begin{cases} 0, & net_j < 0 \\ 1, & net_j \geq 0 \end{cases}$

$$\frac{\partial O_j}{\partial net_j} = O_j'$$

$$\frac{\partial E_d}{\partial net_j} = -(t_j - O_j) O_j'$$

$\therefore \Delta w_{ji} = \eta (t_j - O_j) O_j' x_{ji}$

if $net_j < 0 \Rightarrow \Delta w_{ji} = 0$ as $O_j' = 0$

if $net_j \geq 0 \Rightarrow \Delta w_{ji} = \eta (t_j - O_j) x_{ji}$ (as $O_j' = 1$)

Case 2 : for hidden layer

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in down(j)} -\delta_k w_{kj} \times \frac{\partial O_j}{\partial net_j}$$

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in downstream(j)} -\delta_k w_{kj} O_j'$$

$$\delta_j = -\frac{\partial E_d}{\partial net_j} = O_j' \sum_{k \in down(j)} \delta_k \cdot w_{kj}$$

$\therefore \Delta w_{ji} = \eta \delta_j x_{ji}$

if $net_j < 0$ (as $o_j' = 0$)

$\Delta w_{ji} = 0$

if $net_j \geq 0$ (as $o_j' = 1$)

$$\Delta w_{ji} = \eta \left(\sum_{k \in downst(j)} \delta_k \cdot w_{kj} \right) x_{ji}$$

I.V.N.S. Rohith Reddy
Amandeep Singh

(2) $O = w_0 + w_1(x_1 + x_1^2) + w_2(x_2^2 + x_2) + \ldots \ldots + w_n(x_n + x_n^2)$

Error function $E(\vec{w}) = \frac{1}{2} \sum_{d \in D} (t_d - O_d)^2$

$w_i := w_i + \Delta w_i$ where $\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$

for term w_0

$$\frac{\partial E}{\partial w_0} = \frac{\partial}{\partial w_0} \frac{1}{2} \sum_{d \in D} (t_d - O_d)^2 = \frac{1}{2} \sum_{d \in D} \frac{\partial (t_d - O_d)^2}{\partial w_0}$$

$$= - \sum_{d \in D} (t_d - O_d)$$

Hence $\Delta w_0 = \eta \sum_{d \in D} (t_d - O_d)$

for $w_1, w_2 \ldots w_n$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - O_d)^2 = \frac{1}{2} \sum_{d \in D} \frac{\partial}{\partial w_i} (t_d - O_d)^2$$

$$= \frac{1}{2} \sum_{d \in D} 2(t_d - O_d) \frac{\partial}{\partial w_i} (t_d - O_d) = \sum_{d \in D} (t_d - O_d)(-(x_{id} + x_{id}^2))$$

$$\Delta w_i = \eta \cdot \sum_{d \in D} (t_d - O_d)(x_{id} + x_{id}^2)$$

③

Node	Net value	Output Value
1	x_1	x_1
2	x_2	x_2
3	$net_3 : \omega_{31}x_1 + \omega_{32}x_2$	$x_3 : h(net_3)$
4	$net_4 = \omega_{41}x_1 + \omega_{42}x_2$	$x_4 = h(net_4)$
5	$net_5 = \omega_{53}x_3 + \omega_{54}x_4$	$x_5 : h(net_5)$

(a) $y_5 : h(net_5)$

$= h(net_5) : h(\omega_{53}x_3 + \omega_{54}x_4)$

$: h\Big(\omega_{53}\big(h(\omega_{31}x_1 + \omega_{32}x_2)\big) + \omega_{54}\big(h(\omega_{41}x_1 + \omega_{42}x_2)\big)\Big)$

(b) Output $:$ $h\left[\omega^2 . h(\omega' x)\right]$

(c) $h_1(x) = \dfrac{1}{1+e^{-x}}$ $\qquad h_2(x) = \dfrac{e^{x} - e^{-x}}{e^{x}+e^{-x}}$

$h_1 = \dfrac{e^{x}}{e^{x}+1}$ $\qquad h_2(x) : \dfrac{e^{2x}-1}{e^{2x}+1}$

$2h_1(2x) : \dfrac{2e^{2x}}{e^{2x}+1}$ $\Rightarrow 2h_1(2x) - 1 : \dfrac{2e^{2x}}{e^{2x}+1} - 1$

$2h_1(2x) - 1 : \dfrac{e^{2x}-1}{e^{2x}+1} : h_2(x)$

$2h_1(2x) - 1 = h_2(x)$

$\Rightarrow$ The output of both the functions are same, the only difference is that to get similar output as $h_2(x)$, $h_1(x)$ have to linear transform by multiplying it with 2 and subtracting 1 from result

=) It shows that the $h_1(x)$ & $h_2(x)$ will generate same function

(4) Activation function $f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$

$f' = \sigma(1-\sigma)$

Error function $E(\vec{w}) = \underbrace{\frac{1}{2}\sum_{d\in D}\sum_{k\in outputs}(t_{kd}-O_{kd})^2}_{\text{Part 1}} + \underbrace{\gamma \sum w_{ji}^2}_{\text{part 2}}$

We know $\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$

$$\Delta w_{ji} = -\eta \frac{\partial \frac{1}{2}\sum_{d\in D}\sum_{k\in outputs}(t_{kd}-O_{kd})^2}{\partial w_{ji}} - \frac{\eta\, \partial \gamma \sum w_{ji}^2}{\partial w_{ji}}$$

Calculate part 1 — When j is output layer ... Case 1
— When j is hidden layer ... Case 2

Case 1 : $\frac{\partial part 1}{\partial net_j} = \frac{\partial part 1}{\partial O_j} \times \frac{\partial O_j}{\partial net_j}$

$= -(t_j - O_j)\, O_j\, (1-O_j)$

$= -\delta_j$ where $\delta_j = (t_j - O_j)\, O_j\, (1-O_j)$

Case 2 : $\frac{\partial part 1}{\partial net_j} = \sum_{k\in downstream(j)} \frac{\partial part 1}{\partial net_k} \frac{\partial net_k}{\partial net_j}$

$= -\sum_{k\in downstr(j)} \delta_k \frac{\partial net_k}{\partial net_j}$

$$\frac{\partial Part\,1}{\partial net_i} = O_i(1-O_i) \sum_{k \in downst(i)} \delta_k w_{ki}$$

For part 2 $\Rightarrow$ $\frac{\partial Part\,2}{\partial w_{ji}} = 2\gamma \sum_{i,j} w_{ji}$

$$\Delta w_{ji} = -\eta \delta_i x_{ji} - \eta 2\gamma \sum_{i,j} w_{ji}$$

where j = output layer then $\delta_i = (t_j - O_j) O_j (1-O_j)$

and j = input layer then $\delta_j = O_j(1-O_j) \sum_{k \in down} \delta_k w_{kj}$

$$w_{ji} := w_{ji} + \Delta w_{ji}$$

$$w_{ji} = w_{ji} - \eta \delta_j x_{ji} - \eta 2\gamma \sum_{i,j} w_{ji}$$

$$= (1 - 2\eta\gamma) w_{ji} + \eta \delta_j x_{ji}$$

$\Rightarrow$ $w_{ji} = \beta w_{ji} + \eta \delta_i x_{ji}$ where $\beta = (1-2\eta\gamma)$

This shows we have to multiply w_{ji} with constant β before performing Gradient descent