net id: rxs179030

Q1.1 Consider the error on the outputs of training example d

$$E_d(w) = \frac{1}{2} \sum_{k \in d} (t_k - o_k)^2 \quad \text{where } k \in \text{output layer}$$

Let w_{ji} be the weight between nodes i and j

So $\quad i \xrightarrow{w_{ji}} j$

and let x_{ji} be the input coming from i to j.

Let f_a be an activation function such as:

$$o_j = f_a(net_j) \quad \text{where } net_j = \sum_i w_{ji} x_{ji}$$

Consider the update rule $w_{ji}^{new} = w_{ji}^{old} + \Delta w_{ji}$

where $\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$. Let us compute $\frac{\partial E_d}{\partial w_{ji}}$.

By the chain rule of derivation we have:

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}}$$

we have that $\dfrac{\partial net_j}{\partial w_{ji}} = x_{ji}$

we have two cases for $\dfrac{\partial E_d}{\partial net_j}$

Case 1 j is an output unit.

By the chain rule we have that

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \frac{\partial o_j}{\partial net_j}$$

for $\dfrac{\partial E_d}{\partial o_j}$ we have $\dfrac{\partial E_d}{\partial o_j} = \dfrac{\partial}{\partial o_j}\left(\frac{1}{2}\sum_k (t_k - o_k)^2\right) = -(t_j - o_j)$

now $\dfrac{\partial o_j}{\partial net_j} = \dfrac{\partial f_a(net_j)}{\partial net_j} = f_a'(net_j)$

let $f_a(x) = \tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

now by the division rule of derivation $\left(\frac{p}{q}\right)' = \frac{p'q - q'p}{q^2}$

we have:
$$\tanh'(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)' = \frac{(e^x+e^{-x})(e^x+e^{-x}) - (e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2}$$

$$= \frac{(e^x+e^{-x})^2 - (e^x-e^{-x})^2}{(e^x+e^{-x})^2} = 1 - \left(\frac{e^x-e^{-x}}{e^x+e^{-x}}\right)^2 = 1 - \tanh^2(x)$$

So
$$\frac{\partial o_j}{\partial net_j} = \frac{\partial f_a(net_j)}{\partial net_j} = f_a'(net_j) = 1 - \tanh^2(net_j)$$
$$= 1 - o_j^2$$

Therefore if $f_a = \tanh$ and j is an output layer:

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

and
$$\frac{\partial E_d}{\partial w_{ji}} = -(t_j - o_j)(1 - o_j^2)\, x_{ji}$$

So $\Delta w_{ji} = \eta\,(t_j - o_j)(1 - o_j^2)\, x_{ji}$ for j an output layer and $f_a = \tanh$.

now if $f_a(x) = \text{Relu}(x) = \max(0, x)$

we can write $\text{Relu}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$

then $\text{Relu}'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases} = \underset{(x>0)}{1}(x)$ where $\underset{x>0}{1}(x)$ is the 0-1 indicator function over the set $x > 0$

so if $f_a = \text{Relu}$ and j is an output layer.

then $\dfrac{\partial o_j}{\partial net_j} = \dfrac{\partial f(net_j)}{\partial net_j} = f_a'(net_j) = \text{Relu}'(net_j)$

$= \underset{(x>0)}{1}(net_j)$

So if $f_a = \text{Relu}$ and j is an output layer

$$\frac{\partial E_d}{\partial net_j} = -(t_j - o_j) \cdot \underset{(x>0)}{1}(net_j)$$

and $$\frac{\partial E_d}{\partial w_{ji}} = -(t_j - o_j) \cdot \underset{(x>0)}{1}(net_j) \cdot x_{ji}$$

So $\Delta w_{ji} = \eta\,(t_j - o_j) \cdot \underset{x>0}{1}(net_j)\, x_{ji}$ for j an output layer and $f_a = \text{Relu}$.

Case 2: j is a hidden unit

Let's denote $-\delta_j = -(t_j - o_j)\, f_a'(net_j) = \dfrac{\partial E_d}{\partial net_j}$

if j is a hidden unit then

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

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

Now $\dfrac{\partial o_j}{\partial net_j} = f_a'(net_j)$

and $\dfrac{\partial net_k}{\partial o_j} = \dfrac{\partial(\sum w_{kj} o_{kj})}{\partial o_j} = w_{kj}$

So $\dfrac{\partial E_d}{\partial net_j} = f_a'(net_j) \sum -\delta_k w_{kj}$

So in the case $f_a = \tanh$, $f_a'(net_j) = 1 - o_j^2$

we have $$\frac{dE_d}{dnet_j} = (1-o_j^2) \sum_{k \in Downstream(j)} -\delta_k \cdot w_{kj}$$

So if j is a hidden layer and $f_a = \tanh$

$$\Delta w_{ji} = \eta \, (1-o_j^2) \Big(\sum_{k \in Downstream(j)} \delta_k w_{kj} \Big) x_{ji}$$

In the case $f_a = \text{Relu}$, $f_a'(net_j) = \mathbb{1}_{(x>0)}(net_j)$

we have $$\frac{dE_d}{dnet_j} = \mathbb{1}_{x>0}(net_j) \sum_{k \in Downstream(j)} -\delta_k w_{kj}$$

So if j is hidden layer and $f_a = \text{Relu}$

$$\Delta w_{ji} = \eta \cdot \mathbb{1}_{x>0}(net_j) \Big(\sum_{k \in Downstream(j)} \delta_k w_{kj} \Big) x_{ji}$$

Note: $\delta_k = (t_k - o_k)(1 - o_k^2)$ if $f_a = \tanh$

and $\delta_k = (t_k - o_k) \, \mathbb{1}_{x>0}(net_k)$ if $f_a = \text{Relu}$

Q 1.2

Let $o = w_0 + w_1(x_1 + x_1^2) + \cdots + w_n(x_n + x_n^2)$

where w_0 is the bias weight and $\{x_i\}_{i=1,\ldots,n}$, $\{w_i\}_{i=1}^{n}$ are the inputs and weights respectively.

Consider the identity activation function.

Let the Error be $E = \frac{1}{2} \sum_{k \in D} (t_k - o_k)^2$ where D is the set of training examples.

and the update rule $w_i^{new} = w_i^{old} + \Delta w_i$

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

Let us find $\frac{\partial E}{\partial w_i}$

$$\frac{\partial E}{\partial w_i} = \frac{\partial \left(\frac{1}{2} \sum_k (t_k - o_k)^2\right)}{\partial w_i} = \frac{1}{2} \sum_k \frac{\partial (t_k - o_k)^2}{\partial w_i}$$

applying derivative of a finite sum

now

$$\frac{\partial (t_k - o_k)^2}{\partial w_i} = \frac{\partial (t_k - o_k)^2}{\partial (t_k - o_k)} \cdot \frac{\partial (t_k - o_k)}{\partial w_i}$$

by the chain rule

$$\frac{\partial (t_K - o_K)^2}{\partial (t_K - o_K)} = t_K - o_K$$

and $$\frac{\partial (t_K - o_K)}{\partial w_i} = \frac{\partial \left[t_K - \sum_j w_j (x_{jK} + x_{jK}^2)\right]}{\partial w_i}$$

$$= -(x_{iK} + x_{iK}^2)$$

Here x_{jK} denotes the j^{th} component of the K^{th} training example.

So $$\frac{\partial E}{\partial w_i} = -\frac{1}{2} \cdot 2 \sum_{K \in D} (t_K - o_K) \cdot (x_{iK} + x_{iK}^2)$$

$$= -\sum_{K \in D} (t_K - o_K)(x_{iK} + x_{iK}^2)$$

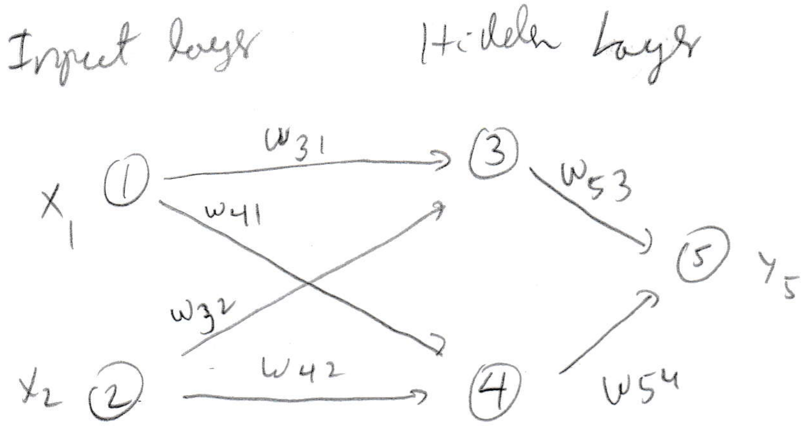
Therefore the gradient descent update rule would be

$$w_i^{new} = w_i^{old} + \Delta w_i$$

$$= w_i^{old} + \eta \sum_{k \in D} (t_K - o_k)(x_{iK} + x_{iK}^2)$$

Q1.3

Consider



a) Let z_3 and z_4 be:

$$z_3 = w_{31}x_1 + w_{32}x_2$$

$$z_4 = w_{41}x_1 + w_{42}x_2$$

then $x_3 = h(z_3)$

$x_4 = h(z_4)$

So $z_5 = w_{53}x_3 + w_{54}x_4$

and $y_5 = h(z_5)$

Putting all together we obtain

$$y_5 = h\left[w_{53}h(w_{31}x_1 + w_{32}x_2) + w_{54}h(w_{41}x_1 + w_{42}x_2)\right]$$

b) Let $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\quad W^1 = \begin{pmatrix} w_{31} & w_{32} \\ w_{41} & w_{42} \end{pmatrix}$ 10

$W^2 = \begin{pmatrix} w_{53} & w_{54} \end{pmatrix}$

Following the notation from a)

we have:

$$\begin{pmatrix} z_3 \\ z_4 \end{pmatrix} = W^1 X$$

$$\begin{pmatrix} x_3 \\ x_4 \end{pmatrix} = h\begin{pmatrix} z_3 \\ z_4 \end{pmatrix}$$

then $z_5 = W^2 \begin{pmatrix} x_3 \\ x_4 \end{pmatrix}$

$Y_5 = h(z_5)$

Thus: $Y_5 = h\left[W^2 h(W^1 X) \right]$

Note: We used the simplified notation $h(v)$ to represent the vector $\begin{pmatrix} h(v_1) \\ h(v_2) \\ \vdots \\ h(v_n) \end{pmatrix}$ where $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$

c) Let

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

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

Let us try to find a relationship between h_s an h_t

$$h_t(x) = \frac{e^{x} - e^{-x}}{e^{x}+e^{-x}} \cdot \frac{e^{-x}}{e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{1+1-1-e^{-2x}}{1+e^{-2x}}$$

$$= \frac{2-1-e^{-2x}}{1+e^{-2x}} = \frac{2-(1+e^{-2x})}{1+e^{-2x}}$$

$$= 2 \cdot \frac{1}{1+e^{-2x}} - 1$$

we can notice that $h_s(2x) = \frac{1}{1+e^{-2x}}$

So $h_t(x) = 2h_s(2x) - 1$

Therefore h_s and h_t generate functions with parameters just differing by linear transformations and constants.

Q1.4 Consider the error of the network as

$$E(w) = \frac{1}{2} \sum_{d \in D} \sum_{k \in outputs} (t_{kd} - o_{kd})^2 + \gamma \sum_{ij} w_{ji}^2$$

Consider the error on the outputs of the training example d.

$$E_d(w) = \frac{1}{2} \sum_{k \in d} (t_k - o_k)^2 + \gamma \sum_{ij} w_{ji}^2$$

Let us derived the update rule:

$$w_{ji}^{new} = w_{ji}^{old} + \Delta w_{ji}$$

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

Let us find $\frac{\partial E_d}{\partial w_{ji}}$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial \frac{1}{2} \sum_{k \in d} (t_k - o_k)^2}{\partial w_{ji}} + \frac{\partial \gamma \sum w_{ji}^2}{\partial w_{ji}}$$

$$\frac{\partial r \sum_{ij} w_{ji}^2}{\partial w_{ji}} = 2\gamma \cdot w_{ji}$$

Now $\dfrac{\partial \frac{1}{2} \sum_{k \in d} (t_k - o_k)^2}{\partial w_{ji}}$ was obtained on Q 1.1

for the two cases j an output layer and j a hidden layer.

So if j is an output layer.

$$\frac{\partial \frac{1}{2} \sum (t_k - o_k)^2}{\partial w_{ji}} = -(t_j - o_j) f_a'(net_j) x_{ij}$$

So $w_{ji}^{new} = w_{ji} - \eta \dfrac{\partial E_d}{\partial w_{ji}}$

$$= w_{ji} - \eta \left(-(t_j - o_j) f_a'(net_j) + 2\gamma \cdot w_{ji}\right)$$

$$= w_{ji} - 2\gamma\eta w_{ji} + \eta (t_j - o_j) f'(net_j)$$

$$= (1 - 2\gamma\eta) w_{ji} + \eta (t_j - o_j) f'(net_j)$$

Now if j is a hidden layer we have that

$$\frac{\partial \frac{1}{2}\sum_{k}(t_k-o_k)^2}{\partial w_{ji}} = f_a'(net_j)\Big(\sum_{k\in Downstream(j)} -(t_j-o_j)\, f_a'(net_k)\cdot w_{kj}\Big)\, X_{ji}$$

So $$w_{ji}^{new} = w_{ji} - \eta \frac{\partial E_d}{\partial w_{ji}}$$

$$= w_{ji} - \eta\Big[-f_a'(net_j)\Big(\sum_{k}(t_j-o_j)\cdot f_a'(net_k)\, w_{kj}\Big) X_{ji} + 2\gamma w_{ji}\Big]$$

$$= w_{ji} - 2\gamma\eta + \eta f_a'(net_j)\cdot\Big(\sum_{k}(t_j-o_j)\, f_a'(net_k)\, w_{kj}\Big) X_{ji}$$

$$= (1-2\gamma\eta)\, w_{ji} + \eta f_a'(net_j)\Big(\sum_{k}(t_j-o_j)\, f_a'(net_k)\, w_{kj}\Big) X_{ji}$$

We can notice in either case j an output or hidden layer we have obtained a weight update rule identical to the original backpropagation rule except that each weight is multiplied by the constant $1-2\gamma\eta$