DATE
PAGE

Part 1

1 Case 1 : when layer = output layer.

$$o = f(x) = x.$$

$$E(w) = \frac{1}{2} \sum_{k \in outputs} (t - o)^2$$

~~$\Delta w =$~~

$$\nabla E(w) = \frac{\partial E(w)}{\partial w} = \frac{1}{2} \sum_{k \in outputs} \frac{\partial (t-o)^2}{\partial w}$$

$$= \frac{1}{2} \sum_{k \in outputs} 2(t - o) (-x_k) = \sum_k (t-o)(-x_k)$$

$$\Rightarrow \Delta w_{ji} = \cancel{+} \eta \nabla E(w)$$

$$= \eta (t_i - o_i) (-x_{ji})$$

$$= - \eta (t - o) x_{ji}$$

$$\Rightarrow \delta_j = (t_j - o_j)$$

Case 2 : when layer = hidden layer.

$$o = \cancel{tah}\ tanh(x)$$

$$net = x$$

$$E(w) = \cancel{\frac{1}{2} \sum}$$

For a hidden unit h

$$\delta_h = \frac{\partial E_d(w)}{\partial net_h} = \sum_{k \in Downstream(h)} \frac{\partial E_d}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_h}$$

$$= \sum_{k \in Downstream(h)} -\delta_k \cdot \frac{\partial net_k}{\partial net_h}$$

$$= \sum_{k \in Downstream(h)} -\delta_k \cdot \frac{\partial net_k}{\partial o_h} \cdot \frac{\partial o_h}{\partial net_h}$$

$$= \sum_{k \in Downstream(h)} -\delta_k \cdot \frac{\partial x_{kh}}{\partial o_h} \cdot (1 - o_h^2)$$

$$= \sum_{k \in Downstream(h)} -\delta_k \cdot \cancel{\partial x_{kh}}\, w_{kh} \, (1 - o_h^2)$$

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

$$\Rightarrow \quad \delta_h = (1 - o_h^2) \sum_{k \in Downstream(h)} \delta_k \, w_{kh}$$

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

$$= \eta \, \delta_h \cdot x_{ji}$$

<u>Finally</u>:

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

For output layer:

$$\delta_j = (t_j - o_j)$$

For hidden layer:

$$\delta_j = (1 - o_j^2) \sum_{k \in downstream(j)} \delta_k \, w_{kj}$$

DATE

PAGE

2

Linear activation function.

$$E_d(w) = \cancel{\tfrac{1}{2}(t_d - O_d)^2}\ \frac{1}{2}\sum_{d\in D}(t_d - O_d)^2 \quad d \in D$$

D = training data

d = training instance

$$\Delta w_j = -\eta \; \nabla E_d(w)$$

$$= -\eta \; \frac{\partial E_d(w)}{\partial w}$$

$$= -\eta \; . \; \frac{1}{\cancel{2}}\sum_{d\in D}\cancel{2}(t_d - O_d) . \frac{\partial (t_d - O_d)}{\partial w_j}$$

$$= -\eta . \cancel{\tfrac{1}{2}}\sum_{d\in D}(t_d - O_d) . \left(-(x_j + x_j^2)\right)$$

$$\Rightarrow \Delta w_j = \sum_{d\in D} \eta (t_d - O_d)(x_j + x_j^2)$$

η = learning rate.

D = all training data

d = single training instance

t_d = target output for dth instance

O_d = observed output for dth instance

x_{id} = value of ith attribute for dth training example.

Final weight update rule :

$$w_j = w_j + \Delta w_j$$

Scanned by CamScanner

DATE

PAGE

3

(a)

$x_3 = w_{31} x_1 + w_{32} x_2$, $O_3 = x_3$

$x_4 = w_{41} x_1 + w_{42} x_2$, $O_4 = x_4$

$x_5 = w_{53} x_3 + w_{54} x_4$, $O_5 = y_5 = x_5$.

(b)

$$X_H = W^{(1)} \times X = \begin{pmatrix} x_3 \\ x_4 \end{pmatrix}$$

$$X_O = W^{(2)} \times X_H = (x_5)$$

$y_5 = x_5$.

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

$$h_1 + h_1 e^{-x} = 1$$
$$e^{-x} = \frac{1-h_1}{h_1}$$

$$h_2 = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} = \frac{1 - \left(\frac{1-h_1}{h_1}\right)^2}{1 + \left(\frac{1-h_1}{h_1}\right)^2}$$

$$= \frac{\cancel{h_1^2} - (1 + \cancel{h_1^2} - 2h_1)}{h_1^2 + 1 + h_1^2 - 2h_1}$$

$$= \frac{2h_1 - 1}{2h_1^2 + 1 - 2h_1}$$

$$h_1 = \frac{e^x}{1+e^x}$$

$$h_2 = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

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

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

$$= h_1(2x) - \frac{1}{1+e^{2x}}$$

$$= h_1(2x)\cancel{(1+e^{2x})} - \frac{1}{1+e^{2x}}$$

$$h_1 = \frac{1}{1+e^{-x}}$$

$$h_1(x) - \frac{1}{2} = \frac{1}{1+e^{-x}} - \frac{1}{2}$$

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

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

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

$$= \frac{1}{2} \cdot \tanh\left(\frac{x}{2}\right)$$

Or $\tanh\left(\frac{x}{2}\right) = 2\, h_1(x) - 1$

or $h_2(x) = 2\, h_1(2x) - 1$

Since $h_2(x) \propto h_1(2x) \Rightarrow$ linear relation b/w tanh and sigmoid $\Rightarrow$ Changing to tanh will not affect a lot.

4.

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

$$\Delta w_{ji} = -\eta\, \nabla E(w)$$

$$= -\eta \frac{\partial E(w)}{\partial w_{ji}}$$

$$\frac{\partial E(w)}{\partial w_{ji}} = \cancel{\frac{1}{2}}\sum_{d}\sum_{k}(t_{kd}-o_{kd})\cdot(-x_{ji}) + 2\gamma\sum_{i,j} w_{ji}$$

$$\Delta w_{ji} = -\eta\left\{\cancel{\frac{1}{2}}\sum_{d}\sum_{k}(t_{kd}-o_{kd})(-x_{ji}) + 2\gamma\sum_{i,j} w_{ji}\right\}$$

$$= \cancel{\frac{1}{2}}\,\eta\sum_{d}\cdot\sum_{k}(t_{kd}-o_{kd})\,x_{ji} - 2\eta\gamma\sum_{i,j} w_{ji}$$

$$\Delta w_{ji} = \eta\sum_{d\in D}\sum_{k\in outputs}(t_{kd}-o_{kd})\,x_{ji}$$

For backpropagation algorithm:

$$\cancel{\frac{\partial E(w)}{\partial w_{ji}} = \sum_{k\in outputs}(t_{kd}-o_{kd})\cdot}$$

$$\cancel{\sum_{d\in D}\delta_k(-x_{ji})}$$

$$\frac{\partial E(w)}{\partial w_{ji}} = \frac{\partial}{\partial w}\left\{\frac{1}{2}\sum_{d}\sum_{k}(t_{kd}-o_{kd})^2\right\} + 2\gamma\sum_{i,j} w_{ji}$$

$$= \delta_k \cdot x_{ji} + 2\gamma\sum_{i,j} w_{ji}$$

$$\Delta w_{ji} = \eta\,\delta_k\cdot x_{ji} + 2\eta\gamma\sum_{i,j} w_{ji}$$

⇒ For weight updation

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

$$= w_{ji} + \eta \delta_k x_{ji} + 2\eta\gamma \sum_{j,i} w_{ji}$$

So, while updating ~~the~~ a single weight this can be approximated to,

$$w_{ji} = w_{ji} + \eta \delta_k x_{ji} + 2\eta\gamma w_{ji}$$

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

$$w_{ji} = C w_{ji} + \eta \delta_k x_{ji}$$

⇒ In the traditional gradient descent, multiply the weight by a constant before updating.