Hoang Nguyen

1.1

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = y \rightarrow \frac{dy}{dx} = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

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

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}, \quad E_d(w) = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

$$\frac{\partial E_d}{\partial w_{ij}} = \frac{\partial E_d}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} \qquad \frac{\partial net_j}{\partial w_{ji}} = x_{ji}$$

Case I: j is an output

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

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

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

$$\rightarrow \Delta w_{ji} = \eta (t_j - o_j)(1 - o_j^2) x_{ji}$$

Case II: j is a hidden unit

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

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

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

Hoang Nguyen

$$\frac{d}{dx} ReLU(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & x < 0 \\ \text{undefined} & x = 0 \end{cases} \rightarrow Relu'(x)$$

Case 1 output layers $\rightarrow \Delta w_{ji} = \eta (t_j - o_j) Relu'(x)$

Case 2 : hidden layers $\rightarrow \Delta w_{ji} = \eta\, Relu'(x) \sum_{k \in downstream} \delta_k w_{kj} \; x_{ij}$

with $\delta_k = (t_k - o_k) Relu(x_k)$

1.2 Gradient Descent $O = w_0 + w_1 (x_1 + x_1^2) + w_n (x_n + x_n^2)$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_d (t_d - o_d)^2 ,$$

Gradient $\nabla E[\vec{w}] =$

$$\left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{w_1}, \dots \frac{\partial E}{w_n} \right]$$

Training rule: $\Delta \vec{w} = -\eta \nabla E[\vec{w}]$

$$\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_d 2 (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d)$$

$$= \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \vec{w} \cdot (\vec{x}_d + \vec{x}_d^{\,2}))$$

$$= -\sum_d (t_d - o_d) (x_{i,d} + x_{i,d}^2) \rightarrow \Delta w_i = \eta \sum_d (t_d - o_d)(x_{id} + x_{id}^2)$$

the final result for weight update

$$w_i^{new} = w_i^{old} + \eta \sum_d (t_d - o_d)(x_{id} + x_{id}^2)$$

η is learning rate

D is the set of all training example

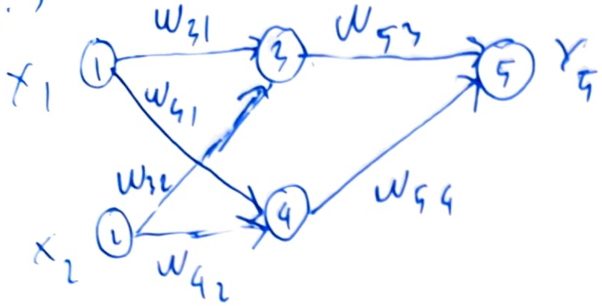
t_d is the target output for the d^{th} training example

o_d is the actual output for the d^{th} training example

x_{id} is the value of the i^{th} attribute for the d^{th} training example

1.3

a/



③ : $x_3 = h(w_{31} \cdot x_1 + w_{32} x_2)$

$x_4 = h(w_{41} x_1 + w_{42} \cdot x_2)$

$\rightarrow y_5 = h(w_{53} h(w_{31} x_1 + w_{32} x_2) + w_{54} h(w_{41} x_1 + w_{42} x_2))$

b/ for $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $w^{(1)} = \begin{pmatrix} w_{3,1} & w_{3,2} \\ w_{4,1} & w_{4,2} \end{pmatrix}$ $w^{(2)} = (w_{5,3}, w_{5,4})$

Vector format: $y_5 = h(w^{(2)} h(w^{(1)} \cdot X))$

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

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

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

$O_t(x) = w_0 + w_h h_t(x) = w_0 + w_h [h_s(2x) - h_s(-2x)]$

$= w_0 + w_h h_s(2x) - w_h h_s(-2x)$

$O_s(x) = w_0' + w_h h_s(x)$, assume $w_0 = 0 = w_0' - w_0'$

$\rightarrow O_t(x) = w_0' + w_h h_s(2x) - (w_0' + w_h h_s(-2x))$

$= O_s(2x) - O_s(-2x)$

$\rightarrow O_t(x) = O_s(2x) - O_s(-2x)$

$\rightarrow$ We can see that $O_t(x)$ is exactly the same as the expression $O_s(x)$ with weights and biases adjusted, differing only by linear transformation and constant