CS 6375
ASSIGNMENT _____Neural Network_______

Names of students in your group:
Pallavi Pandey- PXP17009
Swapna Chintapalli- SXC180048

Number of free late days used: __0__________________

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

①

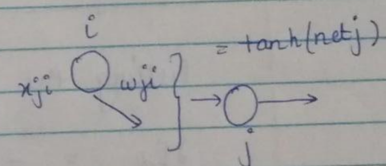
Q1. $\tanh(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \dfrac{\sinh(x)}{\cosh(x)}$

$$\frac{\partial(\tanh(x))}{\partial x} = \frac{\partial}{\partial x}\left(\frac{\sinh(x)}{\cosh(x)}\right)$$

$$= \frac{\frac{\partial}{\partial x}\sinh(x) \times \cosh(x) - \frac{\partial}{\partial x}\cosh(x) \times \sinh(x)}{\cosh^2(x)}$$

$$= \frac{\cos^2 h(x) - \sin^2 h(x)}{\cos^2 h(x)} = \frac{\cos^2 h(x)}{\cos^2 h(x)} - \frac{\sin^2 h(x)}{\cos^2 h(x)}$$

$$\boxed{\frac{\partial(\tanh(x))}{\partial x} = 1 - \tan^2 h(x)} \quad — ①$$



$$E_d = \frac{1}{2} \sum_{k \in output} (t_k - O_k)^2$$

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

η is learning rate

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times \underbrace{\frac{\partial net_j}{\partial w_{ji}}}_{x_{ji}}$$

$$\boxed{\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} . x_{ji}} \quad — ②$$

Case 1 :- When j is an output unit

$$\boxed{\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial O_j} . \frac{\partial O_j}{\partial net_j}} \rightarrow ⓐ$$

2

$$\frac{\partial E_d}{\partial O_j} = \frac{\partial}{\partial O_j}\left(\frac{1}{2}\sum_{k \in outputs}(t_k - O_k)^2\right) = \frac{\partial}{\partial O_j}\left[\frac{1}{2}(t_j - O_j)^2\right]$$

$$\boxed{\frac{\partial E_d}{\partial O_j} = -(t_j - O_j)} \quad — \ 3$$

$$\boxed{\frac{\partial O_j}{\partial net_j} = 1 - O_j^2} \quad — \ 4$$

Putting the values of 3 and 4 in x

$$\boxed{\frac{\partial E_d}{\partial net_j} = -(t_j - O_j)(1 - O_j^2)} \rightarrow 5$$

Putting the value of 5 in 2

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

$$\Delta w_{ji} = -\eta(-(t_j - O_j)(1 - O_j^2)x_{ji})$$
$$= \eta \underbrace{(t_j - O_j)(1 - O_j^2)}_{\delta_j} x_{ji}$$

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

Summary:

When j is O/P unit : $\delta_j = (t_j - O_j)(1 - O_j^2)$

When j is hidden unit $\rightarrow \delta_j = (1 - O_j^2) \sum \delta_k w_{kj}$

W_{ij} new $= W_{ij}$ old $+ \Delta w_{ij}$

$$\boxed{\Delta w_{ij} = \eta \delta_j x_{ij}}$$

3

b) $Relu(x) = \max(0, x)$

$$= \begin{cases} 1 & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\boxed{E_d = \frac{1}{2} \sum_{k \in output} (t_k - o_k)^2} \quad - \ 1$$

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}, \quad \eta - \text{learning rate}$$

As derived in part a)

$$\boxed{\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times x_{ji}} \quad - \ 2$$

Case 1: When j is o/p unit

$$\boxed{\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}} \quad - \ 3$$

$$\cancel{\frac{\partial E_d}{\partial net_j}} \quad \frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j} \left[\frac{1}{2} \sum_{k \in output} (t_k - o_k)^2 \right]$$

$$\boxed{\frac{\partial E_d}{\partial o_j} = -(t_j - o_j)} \quad - \ 4$$

$$\left. \begin{aligned} \frac{\partial o_j}{\partial net_j} &= 1 \quad \text{when } net_j > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \right\} \quad - \ 5$$

$$\begin{aligned} \frac{\partial E_d}{\partial net_j} &= -(t_j - o_j) \quad \text{when } net_j > 0 \\ &= 0 \quad \text{when } net_j \le 0 \end{aligned}$$

4

$$\frac{\partial E_d}{\partial w_{ji}} = 0 \quad \text{when } net_j \le 0$$

$$= -(t_j - o_j) x_{ji} \quad \text{when } net_j > 0$$

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

$$\Delta w_{ji} = \eta \underbrace{(t_j - o_j)}_{\delta_j} x_{ji} \quad \text{when } net_j > 0$$

$$\Delta w_{ji} = 0 \quad \text{for } net_j \le 0$$

Case II when j is hidden unit

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in downstream} \frac{\partial E_d}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in downstream_j} -\delta_k \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in downstream_j} -\delta_k \cdot \frac{\partial net_k}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

Using eq^n 5

$$\frac{\partial E_d}{\partial net_j} = \sum -\delta_k w_{kj} \quad \text{for } net_j \ge 0$$

$$= 0 \quad \text{Otherwise.}$$

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

$$= -\eta \underbrace{\left[\sum -\delta_k w_{kj}\right]}_{\delta_j} x_{ji} \quad \text{for } net_j > 0$$

5

$= 0 \quad \text{for } net_j \leq 0$

Summary

When j is O/P unit

$\delta_j = (t_j - o_j), \; net_j > 0$

$\delta_j = 0, \; net_j \leq 0$

When j is hidden layer

$\delta_j = \sum \delta_k W_{kj}, \; net_j > 0$

$\delta_j = 0, \; net_j \leq 0$

$$\boxed{\Delta W_{ij} = \eta \, \delta_i \, x_{ij}}$$

Q1.2

$$O = \underbrace{w_0}_{bias} + w_1(x_1 + x_1^2) + \dots \; w_n(x_n + x_n^2)$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$E_d = \frac{1}{2} \sum_{k \in output} (t_k - O_k)^2$$

$\Delta w_{ji} = \eta \frac{\partial E_d}{\partial w_{ji}}$, where η is the learning rate.

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times \underbrace{\frac{\partial net_j}{\partial w_{ji}}}_{(x_j + x_j^2)}$$

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial O_j} \times \underbrace{\frac{\partial O_j}{\partial net_j}}_{1}$$ using identity activation function

Pallavi Pandey(Net ID: PXP17009)
Swapna Chintapalli (Net ID: SXC180048)

6

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

$$\Delta w_{ji} = -\eta \times \left[-(t_j - o_j)(x_j + x_j^2)\right]$$

$$\boxed{\Delta w_{ji} = \eta (t_j - o_j)(x_j + x_j^2)}$$

$$\Delta w_{ji} + w_{j\,old} = w_{j\,new}.$$

Q1.3 output 1 = x_1
Output 2 = x_2

$net_3 = x_1 w_{31} + x_2 w_{32}$, $O/P_3 = h(x_1 w_{31} + x_2 w_{32})$

$net_4 = x_1 w_{41} + x_2 w_{42}$, $O/P_4 = h(x_1 w_{41} + x_2 w_{42})$

$net\ 5 = w_{53} \times h(x_1 w_{31} + x_2 w_{32}) + w_{54} \times h(x_1 w_{41} + x_2 w_{42})$

$$O/P_5 = h\left(W_{53} \cdot h(X_1 W_{31} + X_2 W_{32}) + W_{54} \cdot h(X_1 W_{41} + X_2 W_{42})\right)$$

b) $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

$W(1) = \begin{bmatrix} W_{3,1}, & W_{3,2} \\ W_{4,1} & W_{4,2} \end{bmatrix}$ $\quad W(2) = \begin{bmatrix} W_{5,3} & W_{5,4} \end{bmatrix}$

$W_1 \cdot x = \begin{bmatrix} W_{31} X_1 + W_{32} X_2 \\ W_{41} X_1 + W_{42} X_2 \end{bmatrix}$. O/P ~~of hidd~~

$$O/P \text{ of hidden layer } X_1 = \begin{bmatrix} h(W_{31} X_1 + W_{32} X_2) \\ h(W_{41} X_1 + W_{42} X_2) \end{bmatrix}$$

$$W_2 X_1 = \begin{bmatrix} W_{53} \cdot h(W_{31} X_1 + W_{32} X_2) + \\ W_{54} \cdot h(W_{41} X_1 + W_{42} X_2) \end{bmatrix}$$

7

O/P of $5 (y_5) = \left[h(w_{5,3} \cdot h(w_{31} X_1 + w_{32} X_2) + w_{54} \cdot h(w_{41} X_1 + w_{42} X_2)) \right]$

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

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

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

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

So, $h_t(x) = 2(h_s(2x)) - 1$
Hence, the neural N/w created using the above two activation function can generate the same function with some diff in constants.

Q1.4

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

Assuming sigmoid activation

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

$$\frac{\partial E(\vec{w})}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}} \left[\frac{1}{2} \sum_{d \in D} \sum_{k \in outputs} (t_{kd} - O_{kd})^2 + \gamma \sum_{i,j} w_{ji}^2 \right]$$

8

$$= \frac{1}{2} \times 2 \times -(t_j - o_j)(o_j)(1-o_j)\, x_{ji} - 2\gamma w_{ji}$$

$$\Delta w_{ji} = \eta\, \underbrace{(t_j - o_j)(o_j)(1-o_j)}_{\delta_j}\, x_{ji} - 2\gamma w_{ji}$$

$$\Delta w_{ji} = \eta\, \delta_j\, x_{ji} - 2\gamma w_{ji}$$

~~upate~~ update rule

$$w_{ji} = \Delta w_{ji} + w_{ji} = \eta\, \delta_j x_{ji} - 2\gamma w_{ji} + w_{ji}$$

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

$$\delta_j = (t_j - o_j)(o_j)(1-o_j) \rightarrow \text{output layer}$$

$$\delta_j = o_j(1-o_j) \sum_{k \in \text{downstream}} \delta_k w_{kj} \rightarrow \text{hidden layer}$$

Hence, it proves that update can be implemented by multiplying each weight by some constant before performing standard gradient descent.