[Machine Learning Assignment 2]

1.1 a) We want the error of the network w.r.t a weight w_{ji}

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}} \quad \text{(Chain rule)}$$

$$= \frac{\partial E_d}{\partial net_j} \cdot x_{ji}$$

<u>Case 1</u>: j is an output neuron

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \cdot \frac{\partial o_j}{\partial net_j} \quad \text{(Chain rule)}$$

$$= \frac{\partial}{\partial o_j}\left[\frac{1}{2}(t_j - o_j)^2\right] \cdot \frac{\partial o_j}{\partial net_j}$$

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

$$= -(t_j - o_j) \cdot (1 - o_j^2)$$

<u>Case 2</u>: j is a hidden layer neuron, k is one of the neurons downstream from j.

$$\frac{\partial E_d}{\partial net_j} = \sum_k \frac{\partial E_d}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_j}$$

$$= \sum_k \frac{\partial E_d}{\partial net_k} \cdot \left(\frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j}\right)$$

$$= \sum_k \frac{\partial E_d}{\partial net_k} \cdot w_{kj} \cdot (1-o_j)^2$$

$$= (1-o_j)^2 \sum_k \delta_k w_{kj}$$

for all k downstream of j

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

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

Case 1: j output $\delta_j = (1-o_j)^2 (t_j - o_j)$

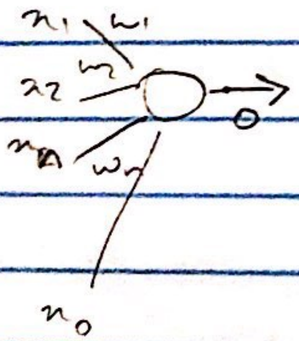
Case 2: j hidden $\delta_j = (1-o_j)^2 \sum_k \delta_k w_{kj}$

b) The derivative of the ReLU is 1 for $x > 0$ and 0 elsewhere. Similar to previous part, but

Case 1: j output $\delta_j = (t_j - o_j)$

Case 2: j hidden $\delta_j = \sum_k \delta_k w_{kj}$

1.2



Suppose t is the value we wanted. The error $E = \frac{1}{2}(t - o)^2$. So the training rule is

$$w_{i\text{-}new} = w_{i\text{-}old} + \Delta w_i$$

[Homework cont'd]

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

$$= -\eta \frac{\partial}{\partial w_i}\left[\frac{1}{2}(t-o)^2\right]$$

$$= -\eta \, (t-o) \cdot \frac{\partial}{\partial w_i}(t-o)$$

$$= -\eta \, (t-o) \cdot \frac{\partial}{\partial w_i}\left(t - (w_0 + w_1(x_1+x_1^2) + \dots w_n(x_n + x_n^2))\right)$$

$$= -\eta (t-o)(-x_i - x_i^2)$$

summed over all the data points.

1.3 a) $(x_1 \cdot w_{41} + x_2 \cdot w_{42}) \leftarrow$ input to 3

$(x_2 \cdot w_{52} + x_1 \cdot w_{51}) \leftarrow$ input to 4

$$h\left(w_{53} \cdot h(x_1 \cdot w_{51} + x_2 w_{52}) + h(x_1 \cdot w_{41} + x_2 \cdot w_{42}) \cdot w_{54}\right)$$

b) $h\left(\begin{bmatrix} w_{53} & w_{54}\end{bmatrix}\begin{bmatrix} x_3 \\ x_4 \end{bmatrix}\right) = h\left(\begin{bmatrix} w_{53} & w_{54}\end{bmatrix}\begin{pmatrix} w_{31} & w_{32} \\ w_{41} & w_{42}\end{pmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$

c) We know $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

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

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

This seems to be $1-2\left(\frac{1}{e^{2x}+1}\right)$

$= 1-2\sigma(-2x)$.

The outputs are

$\sigma(\sigma(x \cdot W^{(1)}) \cdot W^{(2)})$ and

$1-2\sigma(1-2\sigma(4x \cdot W^{(1)} \cdot W^{(2)}))$

These are just linear transformations of each other

1.4 $E(\vec{w}) = \frac{1}{2}\sum_{d\in D}\sum_{k\in outputs}(t_{kd}-O_{kd})^2 + \gamma\sum_{i,j} w_{ji}^2$

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}}\left(\text{''} \quad \text{''} \right)$$

$$= \sum_{d\in D}\sum_{k}(t_{kd}-O_{kd}) + 2\sum w_{ji}$$

1. For an output unit k, its δ_k is

$$\delta_k \leftarrow O_k(1-O_k)(t_k-O_k) + 2\sum w_{ji}$$

2. For a hidden unit,

$$\delta_k \leftarrow O_k(1-O_k)\sum_{j\in outputs} w_{kj}\delta_k + 2\sum w_{ji}$$

Then do $w_{ji}^{new} \leftarrow w_{ji}^{old} - \eta\frac{\partial E}{\partial w_{ji}}$