CISC/CMPE 452 /COGS 400

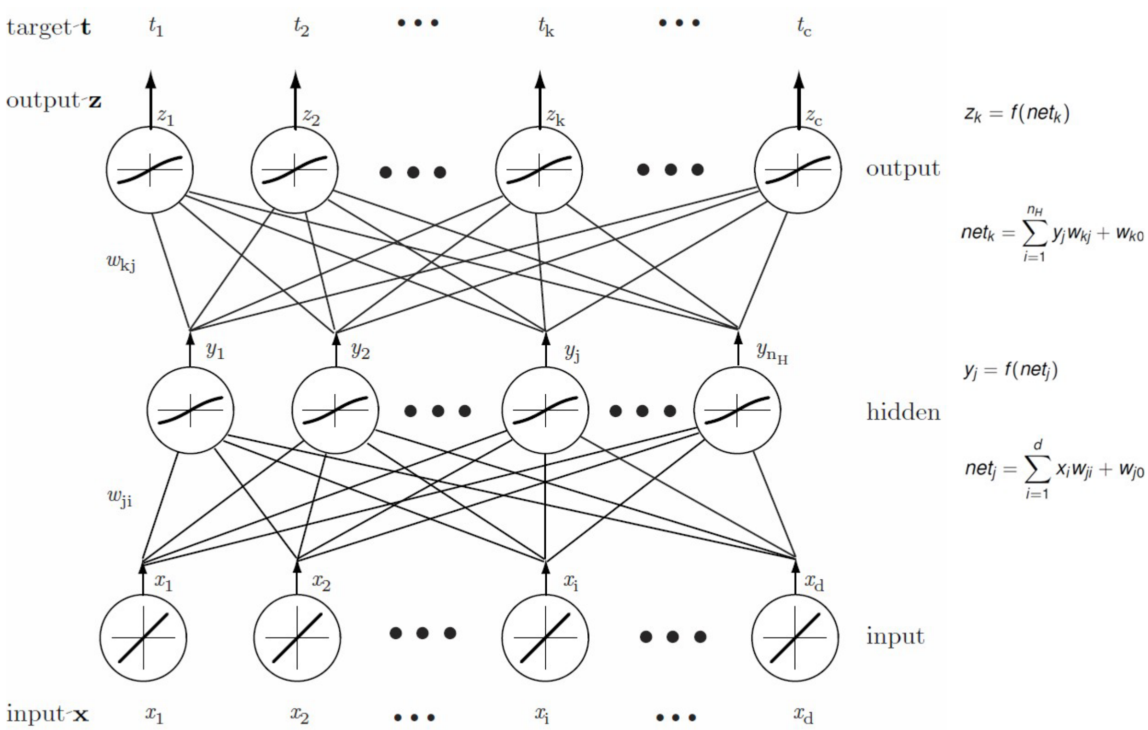
Assignment #2 – Theoretical Part

1. Given a neural network with two input neurons, x_1 and x_2, three hidden neurons, h_1, h_2 and h_3, and two output neurons, y_1 and y_2. The network acts as a classifier (2 classes). The following equations govern the network operation:

$$h_1 = \sigma(x_1 + 1) \tag{1}$$
$$h_2 = \sigma(x_2 + 1) \tag{2}$$
$$h_3 = \sigma(1 - x_1 - 2x_2) \tag{3}$$
$$y_1 = \sigma(2.5 - h_1 - h_2 - h_3) \tag{4}$$
$$y_2 = \sigma(h_1 + h_2 + h_3 - 2.5) \tag{5}$$

   where $\sigma(x) = 1$ if $x \geq 0$ and else $\sigma(x) = 0$ otherwise.
   Draw the decision region for each class and the decision boundary.

2. Consider a three layer neural network whose structure is shown in Figure 1. You are required to calculate the sensitivity $\delta_k = -\frac{\partial J}{\partial \mathrm{net}_k}$ at the output node k, where J is the objective function to be minimized and net_k is the net activation of the output node k. We consider two cases, where the objective function J and the nonlinear activation function at the output layer are chosen differently.
   a. In the first case, J is chosen as the squared error $J(W) = \frac{1}{2}\left|\left|t - z\right|\right|^2$ where z_k is the prediction at the output node k and t_k is the corresponding target value. In the classification problem, only one t_k equals to 1 (corresponding the ground truth class) and all the other t_k's are all zeros. Sigmoid $\mathrm{f(net_k)} = 1/(1 + e^{-\mathrm{net_k}})$ is chosen as the activation function at the output layer. Calculate the sensitivity δ_k in terms t_k, z_k and $\mathrm{net_k}$. Show that all the δ_k could be close to zero even if the prediction error is large and explain why this is bad.
   b. In the second case, the objective function is chosen as cross entropy, $J(W) = -\sum_{k=1}^{cC} t_k \, log(z_k)$ and the nonlinear activation function at the output layer is chosen as softmax $f(net_k) = e^{net_k} / \sum_{j=1}^{C} e^{net_j}$. Calculate the sensitivity $\delta_k$tk again. Prove that if the prediction error is large, at least one of the δ_k will be large.



target $\mathbf{t}$
t_1 t_2 ••• t_k ••• t_c
output $\mathbf{z}$
z_1 z_2 z_k z_c
$z_k = f(net_k)$
output
$net_k = \sum_{i=1}^{n_H} y_j w_{kj} + w_{k0}$
w_{kj}
y_1 y_2 y_j y_{n_H}
$y_j = f(net_j)$
hidden
$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0}$
w_{ji}
x_1 x_2 x_i x_d
input
input $\mathbf{x}$
x_1 x_2 ••• x_i ••• x_d

Figure 1

1. $h_1 = \sigma(x_1 + 1)$ $\qquad$ $y_1 = \sigma(2.5 - h_1 - h_2 - h_3)$

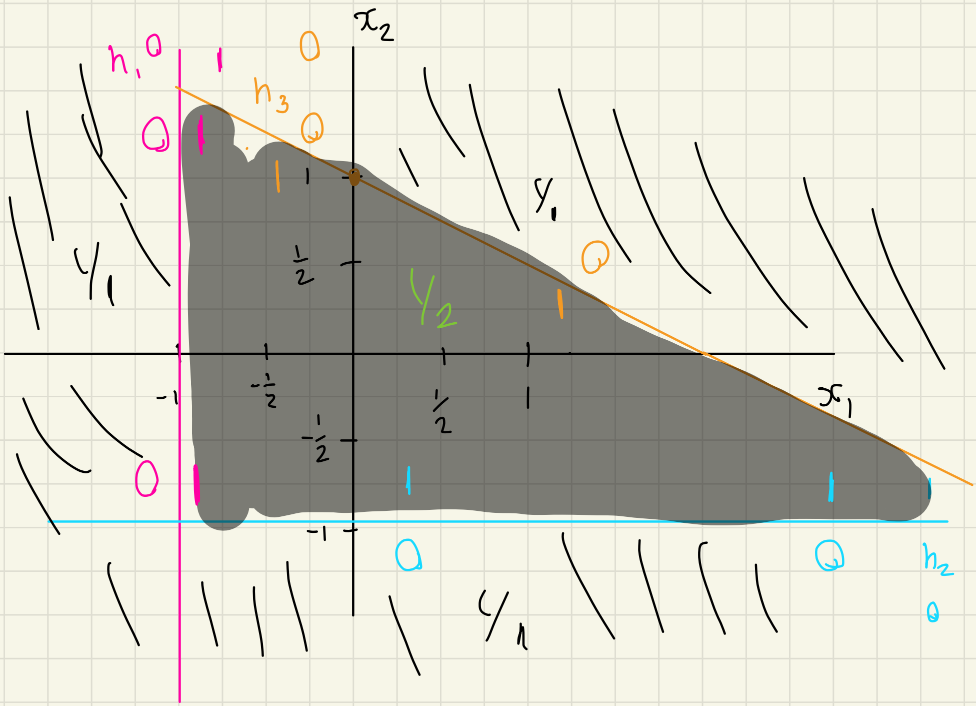
$h_2 = \sigma(x_2 + 1)$ $\qquad$ $y_2 = \sigma(h_1 + h_2 + h_3 - 2.5)$

$h_3 = \sigma(1 - x_1 - 2x_2)$

$$\Rightarrow h_1 = \begin{cases} 1 & \text{if } x_1 \geq -1 \\ 0 & \text{if } x_1 < -1 \end{cases} \qquad y_1 = \begin{cases} 1 & 2.5 \geq h_1 + h_2 + h_3 \\ 0 & 2.5 < h_1 + h_2 + h_3 \end{cases}$$

$$h_2 = \begin{cases} 1 & \text{if } x_2 \geq -1 \\ 0 & \text{if } x_2 < -1 \end{cases} \qquad y_2 = \begin{cases} 1 & h_1 + h_2 + h_3 \geq 2.5 \\ 0 & h_1 + h_2 + h_3 < 2.5 \end{cases}$$

$$h_3 = \begin{cases} 1 & 1 \geq x_1 + 2x_2 \\ 0 & 1 < x_1 + 2x_2 \end{cases} \qquad x_2 = -\frac{1}{2}x_1 + 1$$



We see that the shaded region is the region in which $h_1 + h_2 + h_3 \geq 2.5$ is of class 2 (or y_2). The \\\\ shaded region outside of the triangle is of class 1 (or y_1).

2. a) $J(W) = \frac{1}{2} \| t - z \|^2$

where z_k = prediction at output node k

t_k = desired value at node k

$$f(net_k) = \frac{1}{1+e^{-net_k}} \quad \left(\text{recall } \sigma'(x) = \sigma(x)(1-\sigma(x))\right)$$

$$\delta_k = -\frac{\partial J}{\partial net_k} = -\frac{\partial}{\partial net_k}\left(\frac{1}{2}\| t_k - z_k \|^2\right)$$

$$= -\frac{\partial}{\partial net_k}\left(\frac{1}{2}\| t_k - f(net_k) \|^2\right)$$

$$\Rightarrow -\frac{\partial J}{\partial net_k} = -\frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial net_k} = -\|t_k - z_k\| \cdot z_k(1-z_k)$$

For $t_k = 0$

$$\delta_k = -z_k^2(1-z_k) = z_k^3 - z_k^2 = \sigma(n_k)^3 - \sigma(n_k)^2$$

Now to find max sensitivity we solve for $\delta_k' = 0$

$$\frac{d}{dx}\left(\sigma(x)\sigma(x)\right) \qquad \frac{d}{dx}\left(\sigma(x)\sigma(x)\sigma(x)\right)$$

$$= 2\sigma(x)\sigma(x)(1-\sigma(x)) \qquad = 3\sigma(x)^3(1-\sigma(x))$$

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

$\Rightarrow \delta_k' = -2z_k^2(1-z_k) + 3z_k^3(1-z_k) = 0$

$\Rightarrow 2z_k^2 = 3z_k^3 \Rightarrow \boxed{\frac{2}{3} = z_k \quad \text{maximum value}}$

$\delta_k = -\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) = \frac{4}{27} = -\underline{0.148} \Rightarrow$ sensitivity of -0.148 is the greatest magnitude @ $t_k = 0$.

$\boxed{t_k = 1}$

$\delta_k = -z_k(1-z_k)^2 = z_k(1-2z_k+z_k^2) = z_k - 2z_k^2 + z_k^3$

$\delta_k' = -z_k(1-z_k) + 4z_k^2(1-z_k) - 3z_k^3(1-z_k)$

Solve for $\delta_k' = 0$

$z_k - 4z_k^2 + 3z_k^3 = 0$

$\Rightarrow 3z_k^2 - 4z_k + 1 = (1-3z_k)(1-z_k)$

$\Rightarrow z_k = 1$ or $z_k = \frac{1}{3}$

$\delta_k(1) = 0$ (hence $z_k = 1$ is a minimum) $\delta_k(\frac{1}{3}) = -\frac{2}{9} = -0.22$

So we see that the greatest magnitude of δ_k is $|0.22|$

For a large prediction error we see:

$t_k = 1,\ z_k \approx 0: \quad \delta_k = 0 \cdot (1-0)^2 \simeq 0$

$t_k = 0,\ z_k \approx 1: \quad \delta_k = 1^2(1-1) \simeq 0$

i.e. $\delta_k \rightarrow 0$ for large prediction error

Hence for large prediction error we get sensitivity, δ_x, close to zero.

This is bad because when applying backpropagation to a neural network, if our gradient vanishes then our weight updates will essentially stop which in turn causes the network to stop learning.

6) $J(W) = \sum_{k=1}^{c} t_k \log(z_k)$ $\qquad f(net_k) = \dfrac{e^{net_k}}{\sum_{j=1}^{c} e^{net_j}}$

Derivative of softmax: $(x_k = net_k)$

$g(x_k) = e^{x_k} + e^{x_1} + \cdots + e^{x_c}$ $\quad f(x_k) = e^{x_k}$

$g'(x_k) = e^{x_k}$ $\quad f'(x_k) = e^{x_k}$

$$\Rightarrow \frac{\partial z_k}{x_k} = \frac{e^{x_k} \sum_{j=1}^{c} e^{x_j} - e^{x_k} \cdot e^{x_k}}{\left(\sum_{j=1}^{c} e^{x_j}\right)^2}$$

$$= \frac{e^{x_k}}{\sum_{j=1}^{c} e^{x_j}} \cdot \frac{\sum_{j=1}^{c} e^{x_j} - e^{x_k}}{\sum_{j=1}^{c} e^{x_j}}$$

$$= z_k \cdot (1 - z_k)$$

$(i \neq k)$

$g(x_i) = e^{x_k} + e^{x_1} + \cdots + e^{x_c}$ $\quad f(x_i) = e^{x_i}$

$g'(x_i) = e^{x_k}$ $\quad f'(x_i) = 0$

$$\Rightarrow \frac{\partial z_i}{x_k} = \frac{0 - e^{x_i} e^{x_k}}{\left(\sum_{j=1}^{c} e^{x_j}\right)^2}$$

$$= -z_i z_k$$

$$= -\frac{\partial J}{\partial net_k} = -\frac{\partial J}{\partial z_i} \cdot \frac{\partial z_i}{\partial net_k} = \sum_{i=1}^{c} \frac{t_i}{z_i} \cdot \frac{\partial z_i}{\partial net_k}$$

$$= -\sum_{i \in [1,c] \setminus \{i\}} t_i \cdot z_k + t_k(1 - z_k)$$

For $z_k \approx 0,\ t_k = 1 \Rightarrow \delta_k = t_k = 1$

$z_k \approx 1,\ t_k = 0 \Rightarrow \delta_k = \sum_{i=1}^{c} t_i \approx 1$

This shows us that for large prediction error we get a large δ_k