NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

# Homework 2

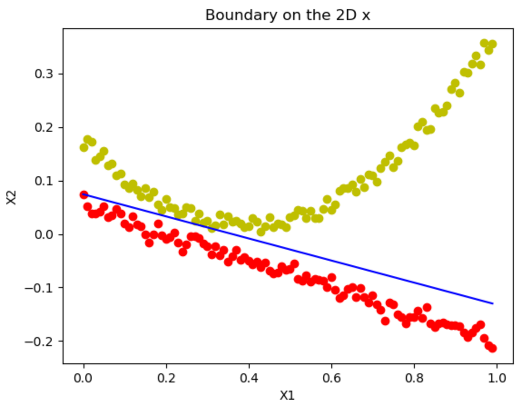
Student name: *Zihan Zeng*

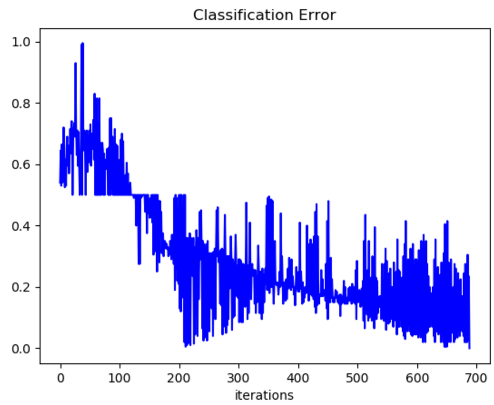
Course: *ECE 6143 Machine Learning* – Professor: *Yury Dvorkin*
Due date: *Oct 7th, 2021*

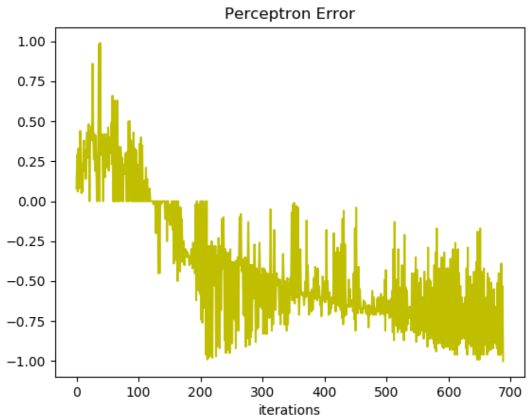
## Problem 1

**Answer.**

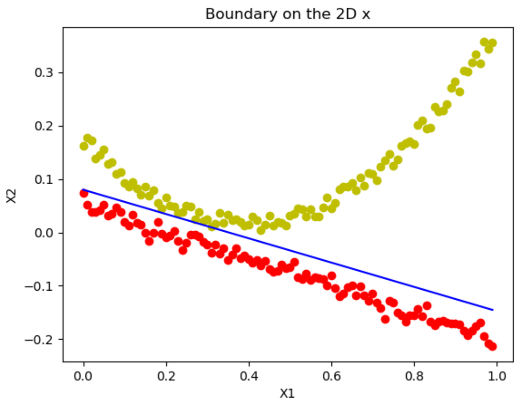
(*a*) When learning rate is 0.1:

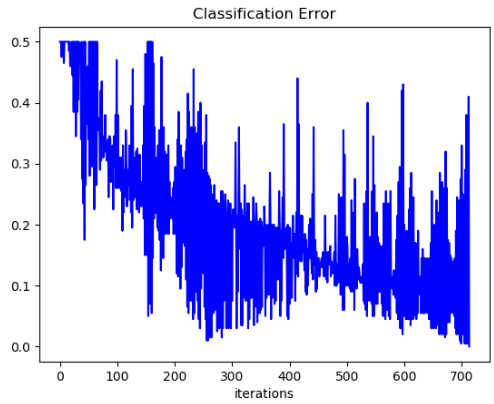


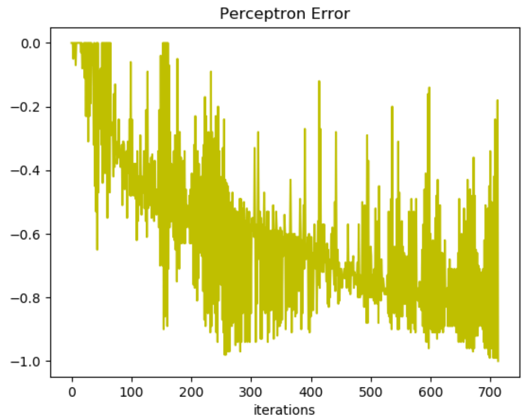




(*b*) When learning rate is 0.5:




## Problem 2

**Answer.**

$(a)$ For the hidden layer:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}}$$

According to the express in the problem, we can know that:

$$\frac{\partial E}{\partial x_i} = -\frac{t_i}{x_i} + \frac{1-t_i}{1-x_i}, \quad \frac{\partial x_i}{\partial s_i} = \frac{e^{-s_i}+1-1}{(1+e^{-s_i})^2} = x_i\,(1-x_i)\,, \quad \frac{\partial s_i}{\partial w_{ji}} = y_j$$

So that:

$$\frac{\partial E}{\partial w_{ji}} = (x_i - t_i)\,y_j$$

Then about the weight update, we can get($\eta$ is Learning Rate):

$$w_{ji}^{t+1} = w_{ji}^{t} - \eta\frac{\partial E}{\partial w_{ji}}$$

For the Input Layer:

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial s_j}\frac{\partial s_j}{\partial w_{kj}}$$

$$\frac{\partial E}{\partial y_j} = \sum_i \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial s_i}\frac{\partial s_i}{\partial y_j}, \quad \frac{\partial s_i}{\partial y_j} = w_{ji}, \quad \frac{\partial y_j}{\partial s_j} = \frac{e^{-s_j}+1-1}{\left(1+e^{-s_j}\right)^2} = y_j\left(1-y_j\right), \quad \frac{\partial s_j}{\partial w_{kj}} = z_k$$

So that we can get:

$$\frac{\partial E}{\partial w_{kj}} = \sum_i \left(x_i - t_i\right) w_{ji} y_j \left(1-y_j\right) z_k$$

Then about the weight update, we can get($\eta$ is Learning Rate):

$$w_{kj}^{t+1} = w_{kj}^{t} - \eta \frac{\partial E}{\partial w_{kj}}$$

($b$) For the hidden layer:

$$\frac{\partial E}{\partial x_i} = -\frac{t_i}{x_i}$$

Here we assume that the kth neuron is the correct label, so that we have two different cases:

$$\frac{\partial x_i}{\partial s_k} = \begin{cases} \frac{e^{s_i}}{\sum_c^n e^{s_c}} - \left(\frac{e^{s_i}}{\sum_c^n e^{s_c}}\right)^2 \\ -\frac{e^{s_i}e^{s_k}}{\left(\sum_n^c e^{s_c}\right)^2} \end{cases} = \begin{cases} x_i - x_i^2, i = k \\ -x_i x_k, i \neq k \end{cases}$$

$$\frac{\partial E}{\partial s_i} = \sum_k \frac{\partial E}{\partial x_k}\frac{\partial x_k}{\partial s_i} = \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial s_i} - \sum_{k\neq i}\frac{\partial E}{\partial x_k}\frac{\partial x_k}{\partial s_i}$$

$$\frac{\partial x_i}{\partial s_k} = t_i\left(1-x_i\right) + x_i \sum_{k\neq i} t_k = -t_i + x_i \sum_k t_k = x_i - t_i$$

$$\frac{\partial E}{\partial w_{ji}} = \sum_i \frac{\partial E}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}} = \left(x_i - t_i\right) x_j$$

For the Input layer:

$$\frac{\partial E}{\partial s_j} = \frac{\partial E}{\partial s_i}\frac{\partial s_i}{\partial x_j}\frac{\partial x_j}{\partial s_j} = \left(x_i - t_i\right) w_{ji}\left(x_j - x_j^2\right)$$

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial s_j}\frac{\partial s_j}{\partial w_{kj}} = \left(x_i - t_i\right) w_{ji}\left(x_j - x_j^2\right) x_k$$

## Problem 3

**Answer.** According to the description of the problem, the entropy of the discrete distribution is expressed as:

$$H(x_1, x_2, \ldots, x_n) = \sum_{K=1}^{n} p_k log p_k$$
$$g(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{N} p_k = 1$$

According to the Lagrange multiplier method, suppose:

$$F(p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n) + \lambda[g(p_1, p_2, \ldots, p_n) - 1]$$

Taking partial derivative to all$p_k$, and then let the formula equal to 0, we get:

$$\frac{\partial}{\partial p_k}\left(-\sum_{k=1}^{n} p_k \log p_k + \lambda\left(\sum_{k=1}^{n} p_k - 1\right)\right) = 0$$

Then, we can know:

$$\begin{aligned}
&\because (p_k \log p_k)' = p_k' \log p_k + p_k (\log p_k)' = \log p_k + \tfrac{1}{\ln 2} \\
&\therefore -\left(\tfrac{1}{\ln 2} + \log p_k\right) + \lambda = 0 \\
&\because \textstyle\sum_{k=1}^{n} p_k = 1 \\
&\therefore p_k = \tfrac{1}{n}
\end{aligned}$$

This shows that all $p_k$ are equal, so that $p_k = \frac{1}{n}$. This also shows that when the system is uniformly distributed, the system is in the most chaotic state, and the entropy is the largest.

## Problem 4

**Answer.** Similar to the straight line model, we do not consider the situation that all the points are in the side of the square. I drew two pictures to illustrate my thoughts:

We can know from the Figure 1 that axis-aligned square can shatter 3 points. For example, if point A and point B are in the same set, we can use the red square to distinguish them.

From Figure 2, if point B and point C are in the same set, and we want to distinguish them, we can not use a square. As the figure illustrate, the square becomes a rectangle.

As a result, the VC dimension of the axis-aligned square is 3. In addition, We can also infer that the VC dimension of the rectangle is 4, and the VC dimension of the rotatable rectangle is 7.

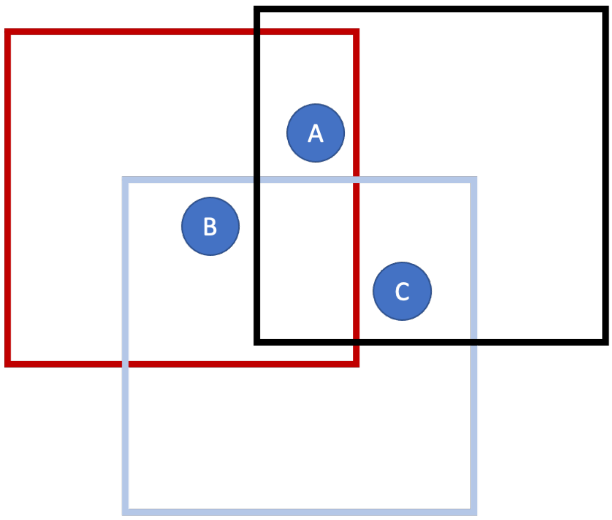



Figure 1: An axis-aligned square shatter 3 points

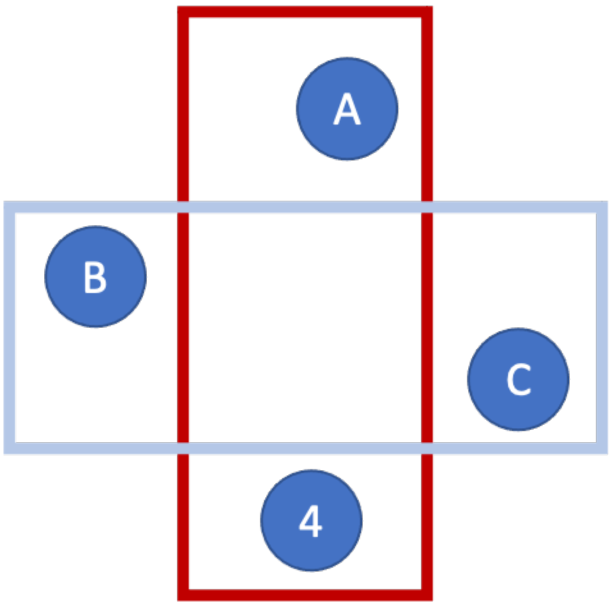



Figure 2: Using an axis-aligned square try to shatter 4 points