Machine Learning - Homework 2

Savinay Shukla

Q1) Please refer to the attached .ipynb for the working python code.

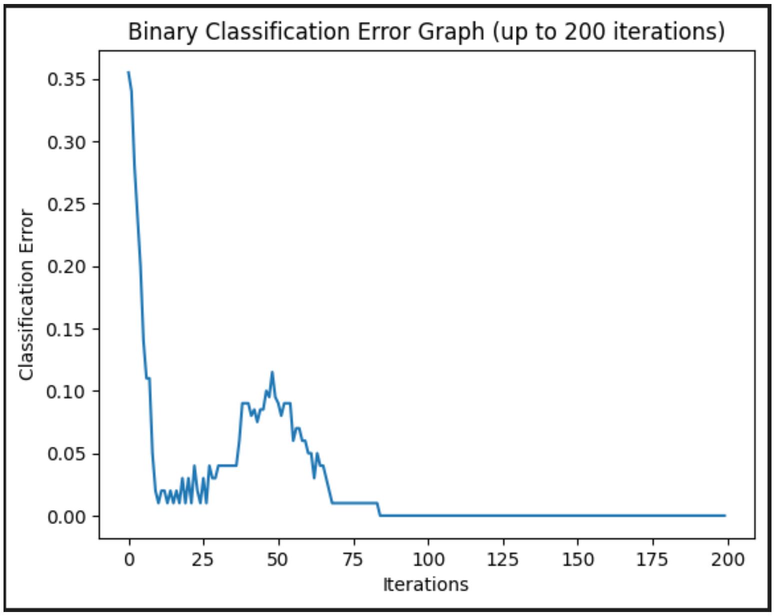


Figure 1.1 Binary Classification Error through iterations.

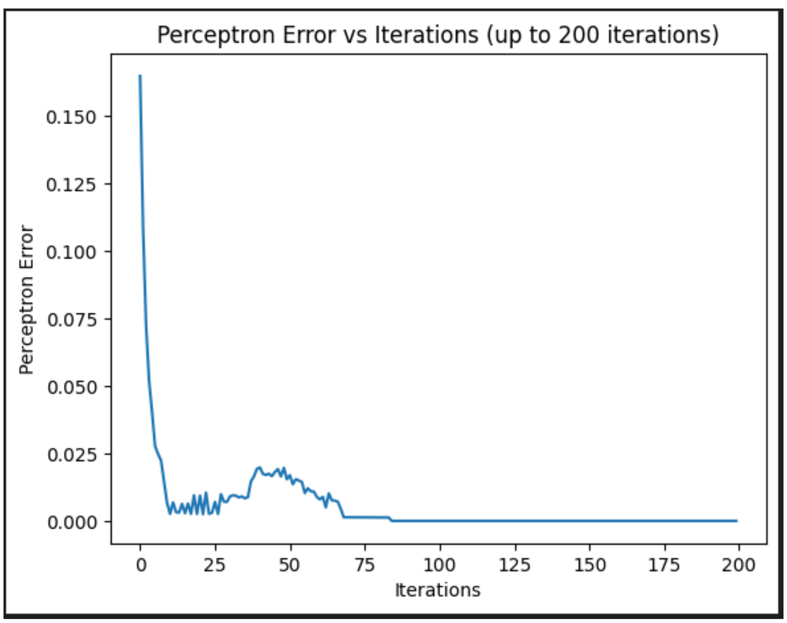


Figure 1.2 Perceptron Error through iterations.

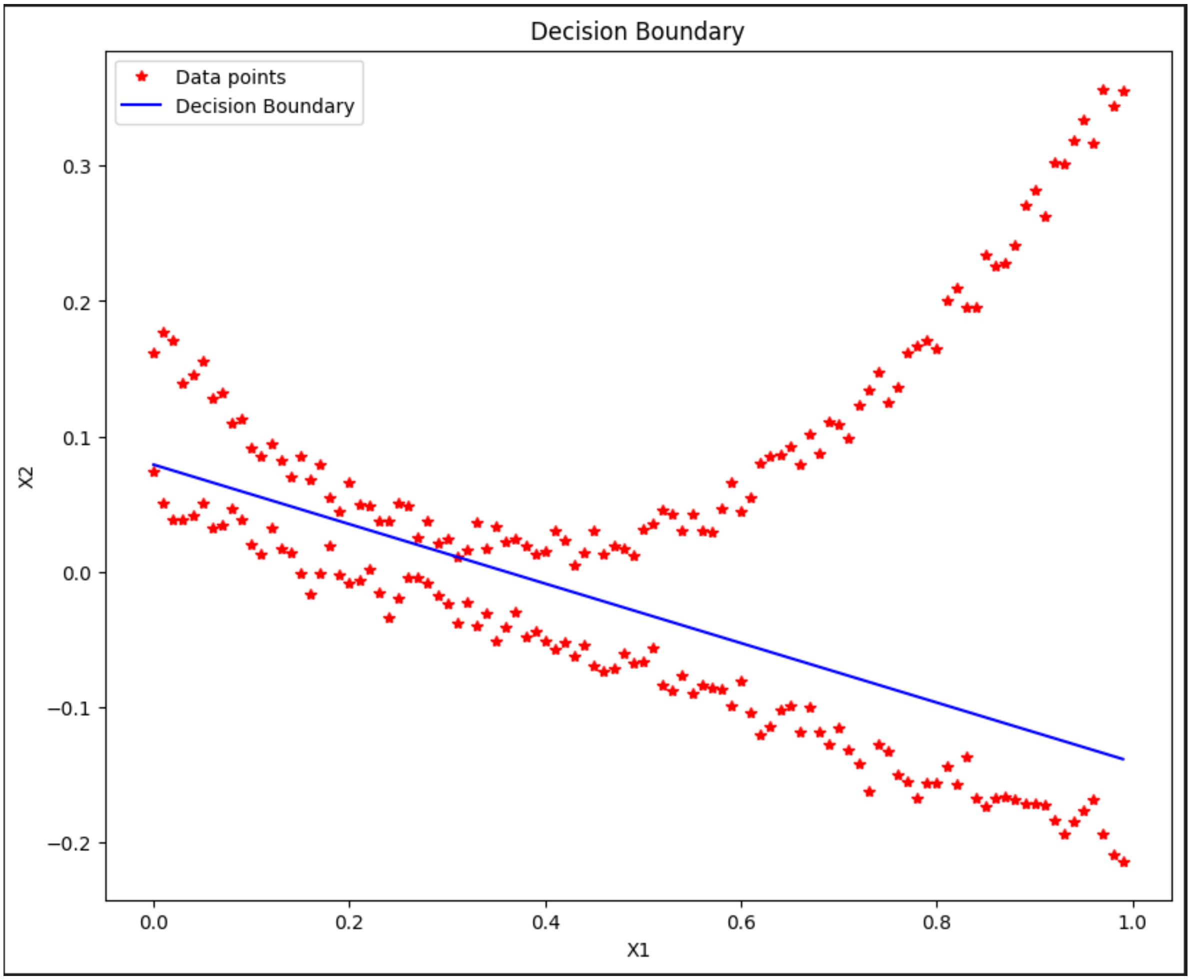


Figure 1.3 Decision Boundary for the x1 and x2 input space.

Q2 a)

Cross entropy error for a single data sample :-

$$E = -\sum_i \left(t_i \log(x_i) + (1 - t_i) \log(1 - x_i) \right)$$

Also, logistic activation function for output layer is given by :-

$$x_i = \frac{1}{1+e^{-s_i}} \quad \text{where} \quad s_i = \sum_j y_j w_{ji}$$

For hidden layer :-

$$y_j = \frac{1}{1+e^{-s_j}} \quad \text{where} \quad s_j = \sum_k z_k w_{kj}$$

NOW CALCULATING OUTER LAYER'S DERIVATIVE :-

$$\hookrightarrow \frac{\partial E}{\partial w_{ji}} = \sum_i \frac{\partial E}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$\hookrightarrow = -\sum_i \underbrace{\frac{\partial \left(t_i \log(x_i) + (1-t_i)\log(1-x_i) \right)}{\partial s_i}}_{\text{equation } 1} \cdot \underbrace{\frac{\partial s_i}{\partial w_{ji}}}_{\text{eq } 2}$$

We can further simplify equation 1 as follows :-

$$\left(\frac{t_i}{x_i} - \frac{1-t_i}{1-x_i}\right) \cdot \frac{\partial x_i}{\partial s_i}$$

Also, equation 2 can be written as :-

$$\frac{\partial s_i}{\partial w_{ji}} = y_j$$

∴ Substituting the simplified values in the derivative we get :-

$$\frac{\partial E}{\partial w_{ji}} = -\left(\frac{t_i}{x_i} - \frac{1-t_i}{1-x_i}\right) \cdot \frac{\partial x_i}{\partial s_i} \cdot y_j$$

$$= \left(\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i}\right) \cdot \frac{\partial x_i}{\partial s_i} \cdot y_j$$

$$= \frac{x_i - \cancel{x_i t_i} - t_i + \cancel{x_i t_i}}{x_i(1-x_i)} \cdot \frac{\partial x_i}{\partial s_i} \cdot y_j$$

$$= \frac{x_i - t_i}{x_i(1-x_i)} \cdot \frac{\partial x_i}{\partial s_i} \cdot y_j$$

We also know that $x_i = \dfrac{1}{1+e^{-s_i}}$

So, $\dfrac{\partial x_i}{\partial s_i} = \dfrac{1}{(1+e^{-s_i})^2} \cdot e^{-s_i}$

$$= \frac{1}{(1+e^{-s_i})} \cdot \frac{e^{-s_i}}{(1+e^{-s_i})}$$

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

Substituting the value we get :-

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

Now solving for the hidden layer :-

We can write as follows :-

$$\frac{\partial E}{\partial w_{kj}} = \sum_i \frac{\partial E}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$\hookrightarrow \sum_i \frac{\partial E}{\partial s_i} \cdot \frac{\partial s_i}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

We know :- $\dfrac{\partial s_j}{\partial w_{kj}} = z_k$ —— 3

Also since $s_i = \sum_i y_j \cdot w_{ji}$

$$\frac{\partial s_i}{\partial y_j} = w_{ji}$$

Also since $y_j = \dfrac{1}{1+e^{-s_j}}$

$$\frac{\partial y_j}{\partial s_j} = y_j \cdot (1 - y_j)$$

Using chain rule we can calculate :-

$$\frac{\partial s_i}{\partial s_j} = \frac{\partial s_i}{\partial y_j} \cdot \frac{\partial y_j}{\partial s_j}$$

$$= w_{ji} \cdot y_j \cdot (1 - y_j) \quad \text{—— 4}$$

Substituting eq 3 and 4 in $\dfrac{\partial E}{\partial w_{kj}}$ we get :-

$$\frac{\partial E}{\partial w_{kj}} = \sum_i \frac{\partial E}{\partial s_i} \cdot w_{ji} \cdot y_j \left[1 - y_j\right] \cdot z_k$$

Previously we found :-

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

$$\therefore \frac{\partial E}{\partial w_{kj}} = \sum_i (x_i - t_i) \cdot w_{ji} \cdot y_j [1 - y_j] \cdot z_k$$

Q2) B)

We have $E = -\sum_i t_i \log(x_i)$

and $x_i = \dfrac{e^{s_i}}{\sum_{c=1}^{m} e^{s_c}}$ and $s_i = \sum_j y_j w_{ji}$

For the outer layer, we can find the gradient as follows:-

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_i} \cdot \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}} \quad \text{—— } 1$$

We know $\dfrac{\partial E}{\partial x_i} = \dfrac{-t_i}{x_i}$ —— 2

Since $x_i = \dfrac{e^{s_i}}{\sum_{c=1}^{m} e^{s_c}}$

$$\frac{\partial x_i}{\partial s_i} = \begin{cases} \dfrac{e^{s_i}}{\sum_{c=1}^{m} e^{s_c}} - \left(\dfrac{e^{s_i}}{\sum_{c=1}^{m} e^{s_c}}\right)^2 & \text{for } i = k \\[2ex] -\dfrac{e^{s_i} e^{s_k}}{\left(\sum_{c=1}^{m} e^{s_c}\right)^2} & \text{for } i \neq k \end{cases}$$

$$\therefore \frac{\partial x_i}{\partial s_i} = \begin{cases} x_i(1-x_i) & \text{for } i = k \\ -x_i x_k & \text{for } i \neq k \end{cases}$$

$$\frac{\partial E}{\partial s_i} = \sum_{k}^{m} \frac{\partial E}{\partial x_k} \cdot \frac{\partial x_k}{\partial s_i}$$

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

$$= -t_i(1-x_i) + \sum_{k \neq i} t_k x_i$$

$$= -t_i + x_i \sum_{k} t_k$$

$$= x_i - t_i \quad \text{—— (3)}$$

Also, $$\frac{\partial s_i}{\partial w_{ji}} = y_i \quad \text{—— (4)}$$

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

Now for hidden layer :-

$$\frac{\partial E}{\partial w_{kj}} = \sum_j \frac{\partial E}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$= \sum_j \frac{\partial E}{\partial s_i} \cdot \frac{\partial s_i}{\partial s_j} \cdot \frac{\partial s_i}{\partial w_{kj}}$$

From previous derivation we know

$$\frac{\partial E}{\partial s_i} = (x_i - t_i)$$

$$\therefore \frac{\partial E}{\partial w_{kj}} = \sum (x_i - t_i) \cdot \frac{\partial s_i}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$\frac{\partial s_i}{\partial y_j} = w_{ji} \quad \text{and} \quad \frac{\partial y_j}{\partial s_j} = y_j(1 - y_j)$$

$$\therefore \frac{\partial s_i}{\partial s_j} = \frac{\partial s_i}{\partial y_j} \cdot \frac{\partial y_j}{\partial s_j}$$

$$= w_{ji} \cdot y_j[1 - y_j]$$

Also, $\frac{\partial s_j}{\partial w_{kj}} = z_k$

$$\therefore \frac{\partial E}{\partial w_{kj}} = (x_i - t_i) \cdot w_{ji} \cdot y_i(1 - y_i) \cdot z_k$$

Q3)

The given Entropy function :-

$$H = -\sum_{k=1}^{N} p_k \log p_k$$

Also, the given constraint :-

$$\sum_{k=1}^{N} p_k - 1 = 0$$

Using Lagrange Multipliers we get :-

$$\mathcal{L}(p_k, \lambda) = \left\{ -\sum_{k=1}^{N} p_k \log p_k - \lambda \left(\sum_{k=1}^{N} p_k - 1 \right) \right\}$$

Maximizing $\mathcal{L}(p_k, \lambda)$ with respect to $\underline{p_k}$:-

$$\frac{\partial \mathcal{L}}{\partial p_k} = -\log(p_k) - 1 - \lambda = 0$$

$$\therefore \quad p_k = e^{-(1+\lambda)} \quad \text{--- } (1)$$

Maximizing $\mathcal{L}(p_k, \lambda)$ with respect to $\underline{\lambda}$:-

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\sum_{k=1}^{N} p_k + 1 = 0$$

$$\therefore \quad \sum_{k=1}^{N} p_k = 1 \quad \text{--- } (2)$$

Plugging equation (1) into equation (2) we get:-

$$\sum_{k=1}^{N} e^{-(1+\lambda)} = 1$$

$$N \cdot e^{-(1+\lambda)} = 1$$

$$e^{-(1+\lambda)} = \frac{1}{N}$$

Also we know :→ $e^{-(1+\lambda)} = P(k)$

$$\therefore \; P_K = \frac{1}{N}$$

Hence every P_i where $i \in [1, N]$, will have the same probability $= \frac{1}{N}$.

$\therefore$ A uniform probability distribution given by $P_K = \frac{1}{N}$, will maximize the given entropy

$$H = -\sum_{k=1}^{N} P_K \log P_K$$

Max Entropy :-

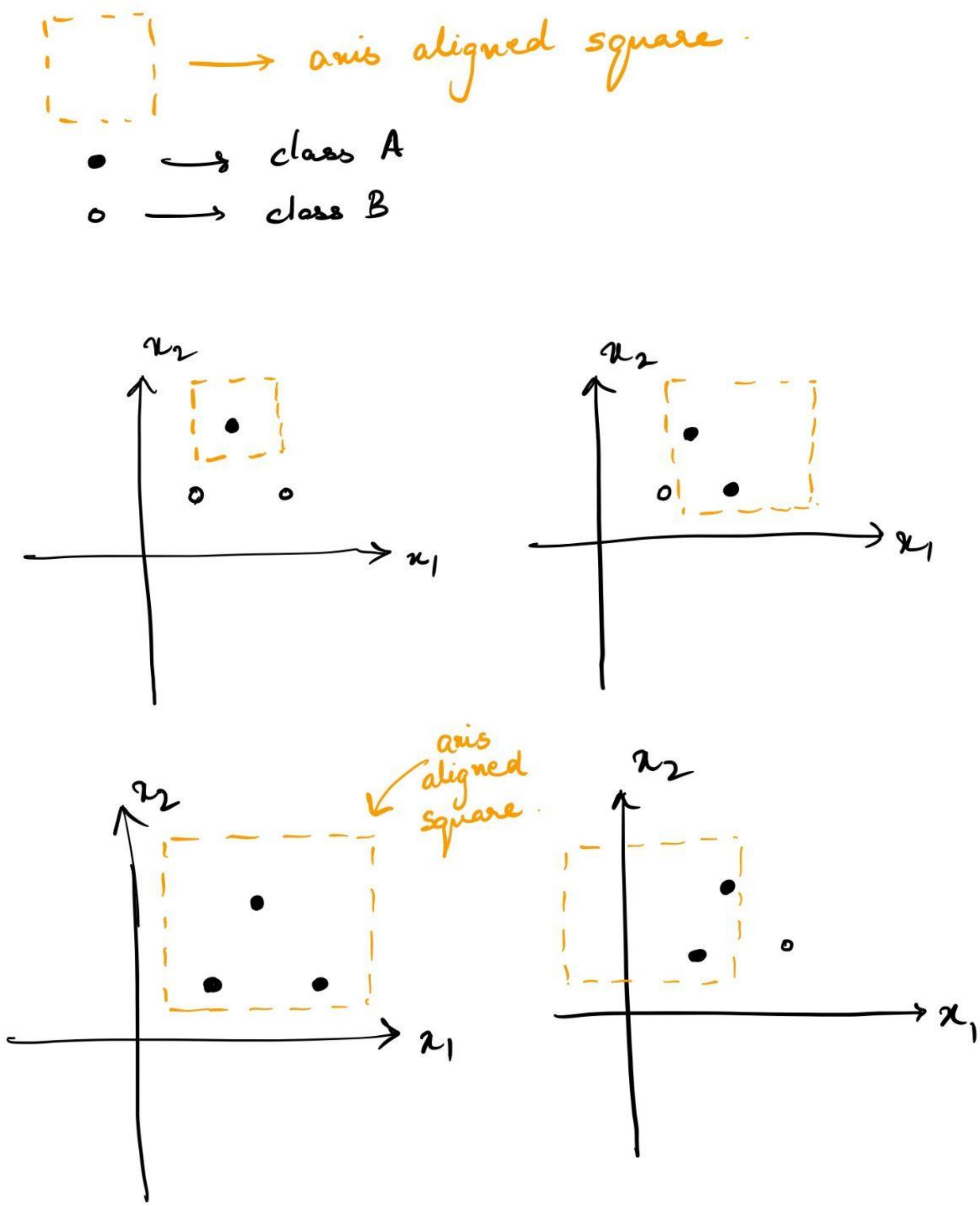
$$H_{max} = -\sum_{k=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right)$$

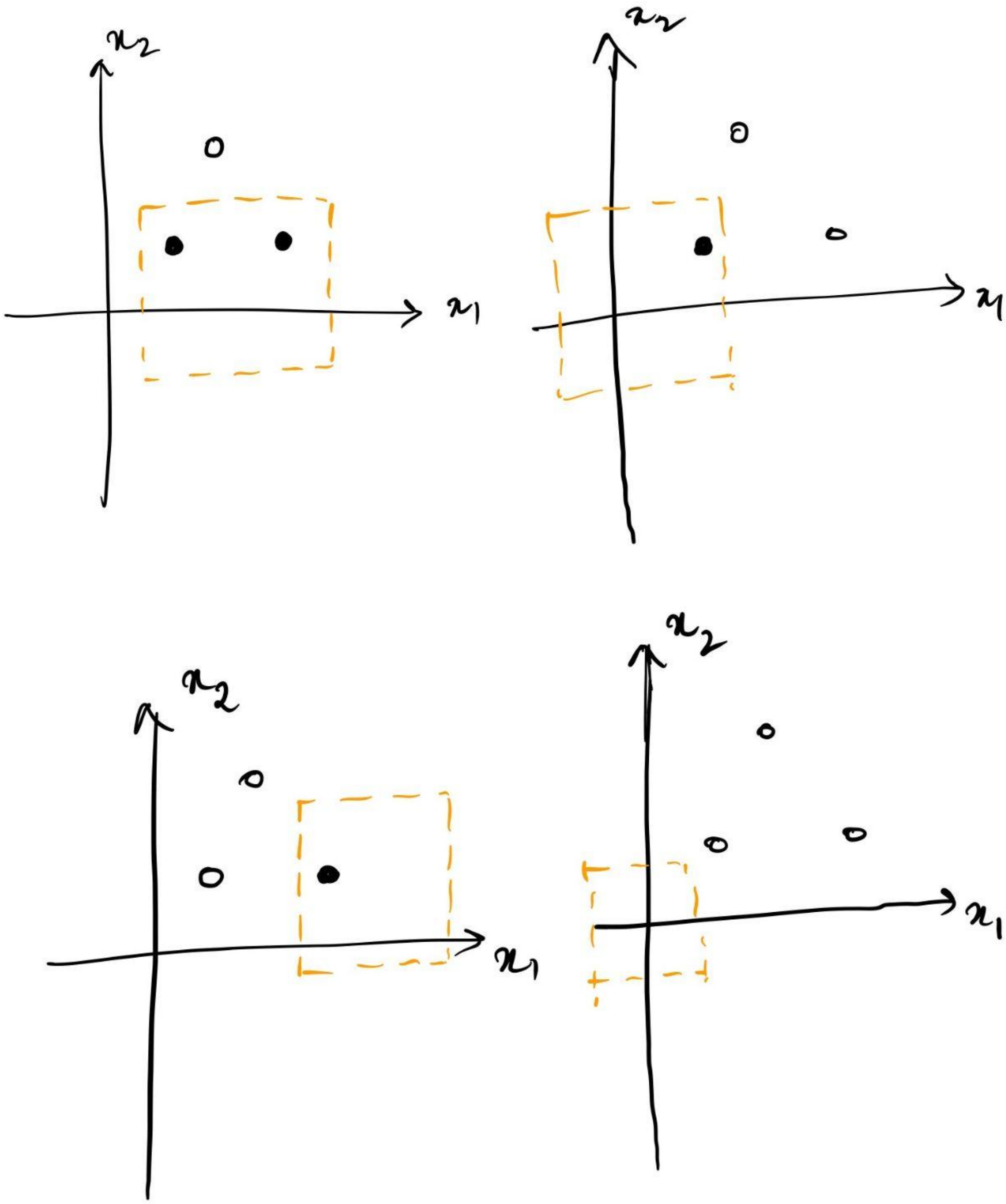
Since N does not depend on summation :-

$$_{max}H = \frac{1}{N} \times \log\left(\frac{1}{N}\right) = \log N$$

Q4)

VC Dimension for the axis-aligned squares = 3.
We can easily demonstrate that axis-aligned squares can shatter 3 data points in its 2-dimensional input space for all their possible configurations – i.e. $2^3 = 8$ data point configurations :

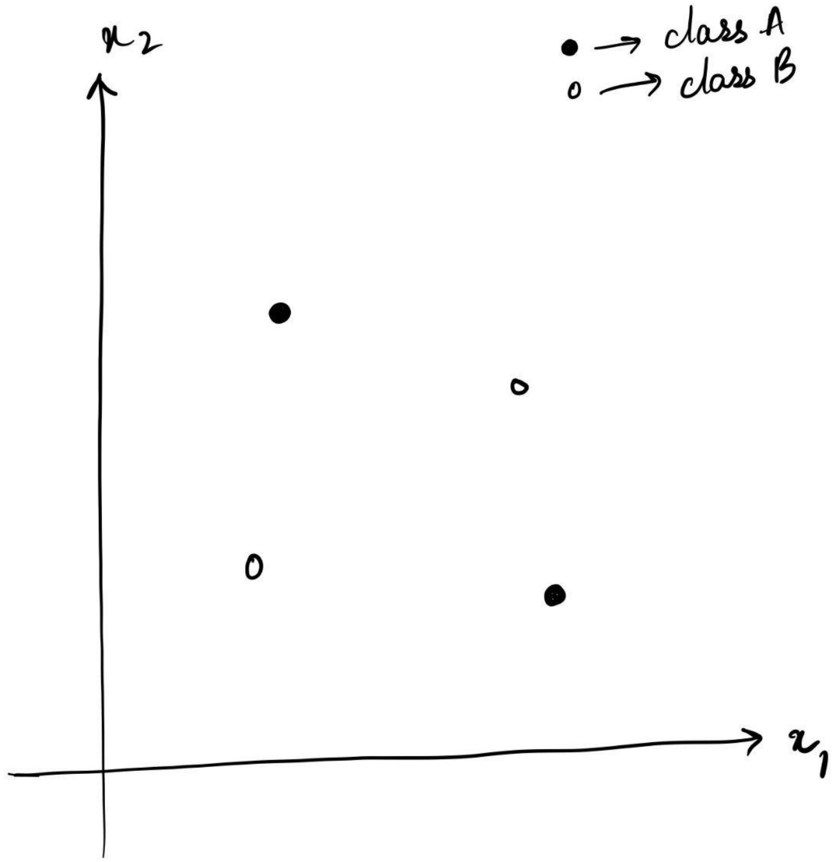




So we know for certain that VC-Dimension for the axis-aligned square **is at least 3.**

However, **when we consider 4 data points** in the 2-dimensional input space, the axis-aligned square **cannot shatter** all 2^4 data point combinations.

One particular example is the following configuration:



If we consider any two points which belong to the same class (either class A or class B), **we can't possibly create a square without including the data point belonging to the other class, which is incorrect.The minimum inclosing axis-aligned square, will contain at least one incorrectly shattered point.**

This is true for any axis-aligned orientation of the square in the 2-D input space.

$Since,\ \ VC - Dimension\ \neq\ 4$

$\Rightarrow VC - Dimension\ =\ 3$