Yinhong Qin
yq2021@nyu.edu

Homework 2

ECE 6143A-Machine Learning
October 6, 2021

Problem 1

1. The training results and classification boundary with different learning rates and gradient descent methods are listed below:

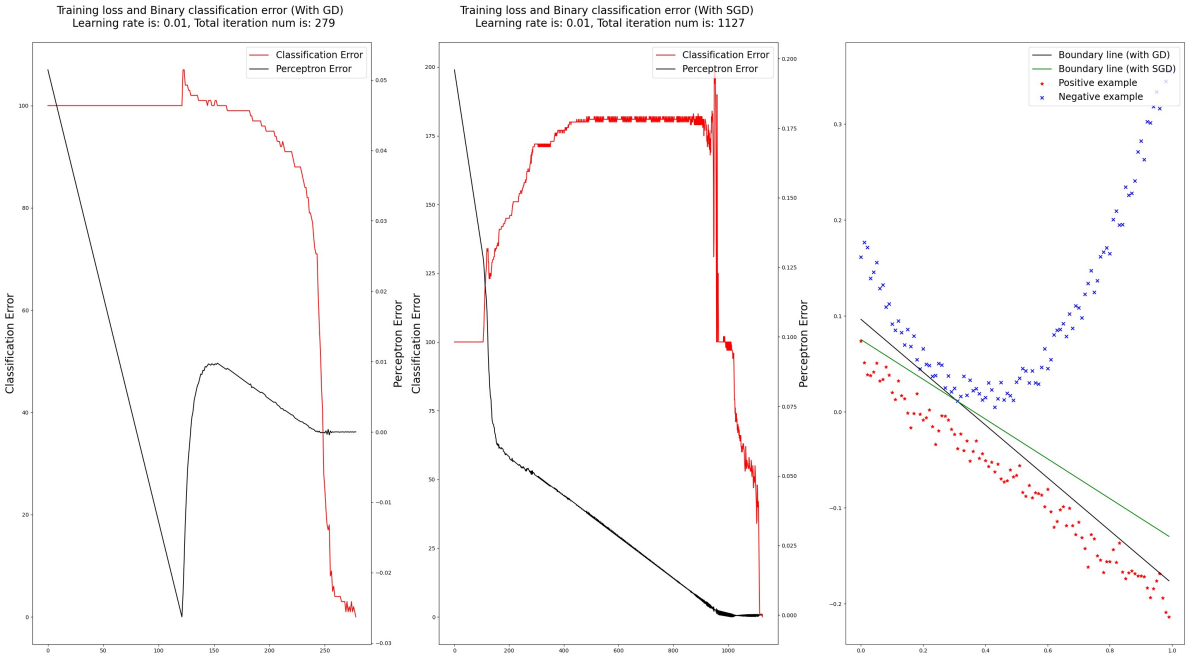

Figure 1.1 Learning rate = 0.01 when using Gradient Descent

When using gradient descent, the smaller the learning rate is, the iteration will become larger because the condition provided by problem 1 is linear separable. The perceptron loss is oscillating when the learning rate is larger.

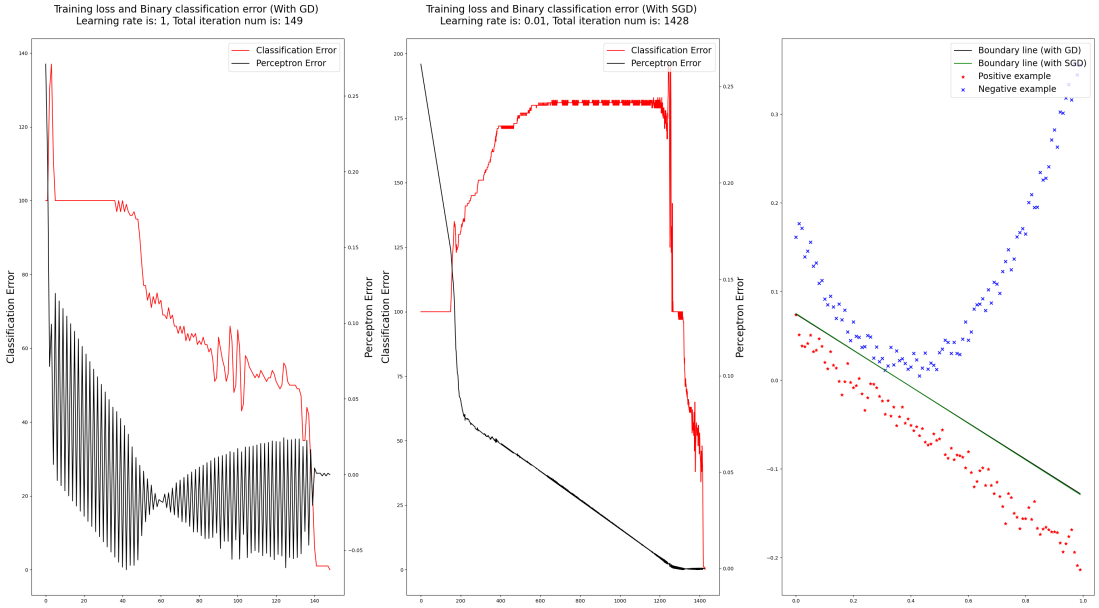

Figure 1.2 Learning rate = 1 when using Gradient Descent

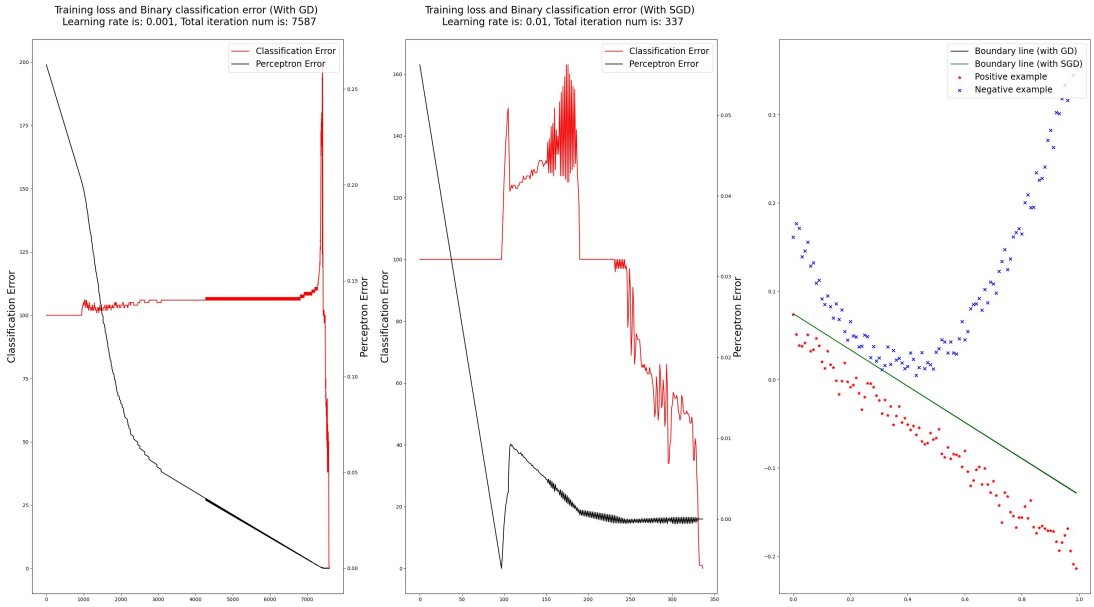

Figure 1.3 Learning rate = 0.001 when using Gradient Descent

Problem 2

Assume the weight between input layer and hidden layer is m_{kj}.

1. For subtask a)

   The cross entropy loss is:

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

   (a) Computing the partial derivative of cross entropy loss E with respect to w_{ji}, which is equal to $\frac{\partial E}{\partial w_{ji}}$.

      i. The partial derivative of cross entropy loss with respect to x_i is:

$$\frac{\partial E}{\partial x_i} = -(\frac{t_i}{x_i} + \frac{t_i - 1}{1 - x_i})$$

      ii. The partial derivative of x_i with respect to s_i is:

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

      iii. The partial derivative of s_i with respect to w_{ji} is:

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

   Therefore, according to the chain rule, the partial derivative of cross entropy loss E with respect to w_{ji} is:

$$\begin{aligned}\frac{\partial E}{\partial w_{ji}} &= \frac{\partial E}{\partial x_i} \cdot \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}} \\ &= -(\frac{t_i}{x_i} + \frac{t_i - 1}{1 - x_i}) \cdot x_i(1 - x_i) \cdot y_j\end{aligned}$$

   (b) Computing the partial derivative of cross entropy loss E with respect to m_{kj}, which is equal to $\frac{\partial E}{\partial m_{kj}}$.
   Assume the input value of hidden layer is p_j, which satisfies

$$p_j = \sum_k z_k m_{kj}$$

$$y_j = \frac{1}{1 + \exp(-p_j)}$$

Therefore,

$$\begin{aligned}\frac{\partial p_j}{\partial m_{kj}} &= z_k \\ \frac{\partial y_j}{\partial p_j} &= y_j(1-y_j) \\ \frac{\partial y_j}{\partial m_{kj}} &= \frac{\partial y_j}{\partial p_j}\cdot\frac{\partial p_j}{\partial m_{kj}} \\ &= y_j(1-y_j)\cdot z_k\end{aligned}$$

From case 1, we can know that:

$$\frac{\partial E}{\partial x_i} = -(\frac{t_i}{x_i} + \frac{t_i - 1}{1-x_i})$$

and

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

and

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

Finally, the partial derivative of E with respect to m_{kj} is:

$$\begin{aligned}\frac{\partial E}{\partial m_{kj}} &= -\sum_i (\frac{\partial E}{\partial x_i}\cdot\frac{\partial x_i}{\partial s_i}\cdot\frac{\partial s_i}{\partial y_j}\cdot\frac{\partial y_j}{\partial p_j}\cdot\frac{\partial p_j}{\partial m_{kj}}) \\ &= -\sum_i (-(\frac{t_i}{x_i} + \frac{t_i-1}{1-x_i})\cdot x_i(1-x_i)\cdot w_{ji}\cdot y_j(1-y_j)\cdot z_k \\ &= -\sum_i ((x_i - t_i)\cdot w_{ji}\cdot y_j(1-y_j)\cdot z_k\end{aligned}$$

2. For subtask b)
   The modified cross entropy loss is:

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

and

$$x_i = \frac{\exp(s_i)}{\sum_{c=1}^{m}\exp(s_c)}$$

   (a) The partial derivative of E with respect to x_i is:

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

(b) The partial derivative of x_i with respect to s_j is:

i. if $i \neq j$

$$\begin{aligned}
\frac{\partial x_i}{\partial s_j} &= \frac{\frac{\partial \exp(s_i)}{\partial s_j} \cdot \sum_{c=1}^{m} \exp(s_c) - \exp(s_i) \cdot \frac{\partial \sum_{c=1}^{m} \exp(s_c)}{\partial s_j}}{(\sum_{c=1}^{m} \exp(s_c))^2} \\
&= \frac{0 \cdot \sum_{c=1}^{m} \exp(s_c) - \exp(s_i) \cdot \exp(s_j)}{(\sum_{c=1}^{m} \exp(s_c))^2} \\
&= -\frac{\exp(s_i)}{\sum_{c=1}^{m} \exp(s_c)} \cdot \frac{\exp(s_j)}{\sum_{c=1}^{m} \exp(s_c)} \\
&= -x_i \cdot x_j
\end{aligned}$$

ii. if $i = j$

$$\begin{aligned}
\frac{\partial x_i}{\partial s_j} &= \frac{\partial x_i}{\partial s_i} \\
&= \frac{\frac{\partial \exp(s_i)}{\partial s_i} \cdot \sum_{c=1}^{m} \exp(s_c) - \exp(s_i) \cdot \frac{\partial \sum_{c=1}^{m} \exp(s_c)}{\partial s_i}}{(\sum_{c=1}^{m} \exp(s_c))^2} \\
&= \frac{\exp(s_i) \cdot \sum_{c=1}^{m} \exp(s_c) - \exp(s_i) \cdot \exp(s_i)}{(\sum_{c=1}^{m} \exp(s_c))^2} \\
&= \frac{\exp(s_i)}{\sum_{c=1}^{m} \exp(s_c)} - \frac{\exp(s_i)}{\sum_{c=1}^{m} \exp(s_c)} \cdot \frac{\exp(s_i)}{\sum_{c=1}^{m} \exp(s_c)} \\
&= x_i - x_i \cdot x_i \\
&= x_i(1 - x_i)
\end{aligned}$$

(c) The partial derivative of s_i with respect to w_{ji} is:

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

Finally, the partial derivative of E with respect to w_{ji} is:

$$\begin{aligned}
\frac{\partial E}{\partial w_{ji}} &= (\sum_{c \neq i} -\frac{t_c}{x_c} \cdot \frac{\partial x_c}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}) + (-\frac{t_i}{x_i} \cdot \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}) \\
&= (\sum_{c \neq i} -\frac{t_c}{x_c} \cdot (-x_c x_i) \cdot y_j) + (-\frac{t_i}{x_i} \cdot x_i(1 - x_i) \cdot y_j) \\
&= (\sum_{c \neq i} t_c x_i - t_i(1 - x_i)) y_j
\end{aligned}$$

From subtask a), we can know that,

$$\begin{aligned} p_j &= \sum_k z_k m_{kj} \\ y_j &= \frac{1}{1+\exp(-p_j)} \\ \frac{\partial p_j}{\partial m_{kj}} &= z_k \\ \frac{\partial y_j}{\partial p_j} &= y_j(1-y_j) \\ \frac{\partial y_j}{\partial m_{kj}} &= \frac{\partial y_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial m_{kj}} \\ &= y_j(1-y_j) \cdot z_k \end{aligned}$$

The partial derivative of E with respect to m_{kj}, which is

$$\begin{aligned} \frac{\partial E}{\partial m_{kj}} &= \sum_i [\frac{\partial E}{\partial x_i} \cdot (\sum_c \frac{\partial x_i}{\partial s_c} \cdot \frac{\partial s_c}{\partial y_j} \cdot \frac{\partial y_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial m_{kj}})] \\ &= \sum_i [-\frac{t_i}{x_i} \cdot (\sum_{c \neq i} (\frac{\partial x_i}{\partial s_c} \cdot \frac{\partial s_c}{\partial y_j} \cdot \frac{\partial y_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial m_{kj}}) + \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial y_j} \cdot \frac{\partial y_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial m_{kj}})] \\ &= \sum_i [-\frac{t_i}{x_i} \cdot (\sum_{c \neq i} (-x_i x_c \cdot w_{jc} \cdot y_j(1-y_j) \cdot z_k) + x_i(1-x_i) \cdot w_{ji} \cdot y_j(1-y_j) \cdot z_k)] \\ &= \sum_i [-\frac{t_i}{x_i} \cdot (\sum_{c \neq i} (-x_i x_c \cdot w_{jc}) + x_i(1-x_i) \cdot w_{ji}) \cdot y_j(1-y_j) \cdot z_k] \\ &= \sum_i [t_i \cdot (\sum_{c \neq i} (-x_c \cdot w_{jc}) - (1-x_i) \cdot w_{ji}) \cdot y_j(1-y_j) \cdot z_k] \end{aligned}$$

Problem 3

We want to find the maximum value of:

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

with constraint condition that p_k is a discrete distribution such that

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

Let the λ be the Lagrange multiplier, therefore the Lagrangian function is:

$$\begin{aligned} L(\boldsymbol{p_k}, \lambda) &= \boldsymbol{H} + \lambda \left(\sum_{k=1}^{N} p_k - 1 \right) \\ &= -\sum_{k=1}^{N} p_k \log p_k + \lambda \left(\sum_{k=1}^{N} p_k - 1 \right) \end{aligned}$$

The partial derivative of L with respect to an arbitrary p_k is:

$$\begin{aligned} \frac{\partial L(\boldsymbol{p_k}, \lambda)}{\partial p_k} &= \frac{\partial}{\partial p_k}(-p_k \log p_k + \lambda \cdot p_k) \\ &= -\log p_k + -p_k \cdot \frac{1}{p_k} + \lambda \\ &= \lambda - \log p_k - 1 \end{aligned}$$

Therefore we can have these $N + 1$ equations that:

$$\begin{cases} \frac{\partial L(\boldsymbol{p_k}, \lambda)}{\partial p_1} & = \lambda - \log p_1 - 1 & = 0 \\ & \vdots & \\ \frac{\partial L(\boldsymbol{p_k}, \lambda)}{\partial p_N} & = \lambda - \log p_N - 1 & = 0 \\ & \sum_{k=1}^{N} p_k = 1 & \end{cases}$$

From the first N equations, we can have:

$$\begin{aligned} \lambda - \log p_k - 1 &= 0 \\ \log p_k &= \lambda - 1 \\ p_k &= \exp(\lambda - 1) \end{aligned}$$

It should be noticed that this relation holds for any $k \in [1, N]$, thus

$$p_1 = p_2 = \cdots = p_N$$

Since

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

Therefore, to reach the maximum of:

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

The discrete distribution p_k is:

$$\{p_k = \frac{1}{N}, k \in [1, N]\}$$

Problem 4

The VC dimension of an axis-aligned squares is 3.

Proof

1. Any sets that containing 3 points can be shattered.

   For sets containing 1 positive point or 3 positive points, it is trivial to prove they can be shattered.

   For sets containing 2 positive points, the corresponding squares are shown below:

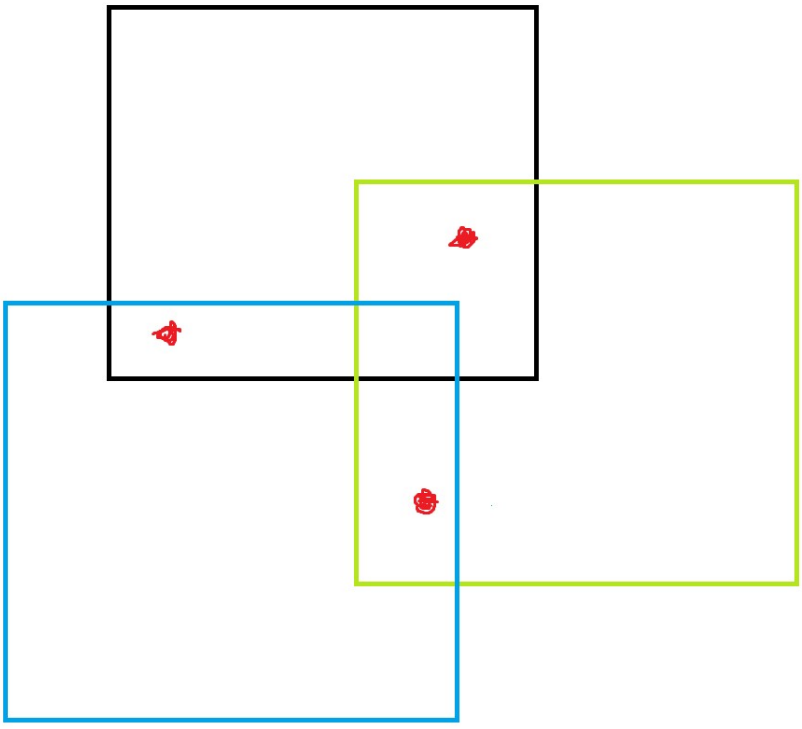

Figure 4.1 Square containing 2 positive points.

2. There do not exist any sets of 4 points or more than 4 points that can be shattered by a square.

   For any 4 points that do not line on the same line, there exists a minimum enclosing rectangle like below:

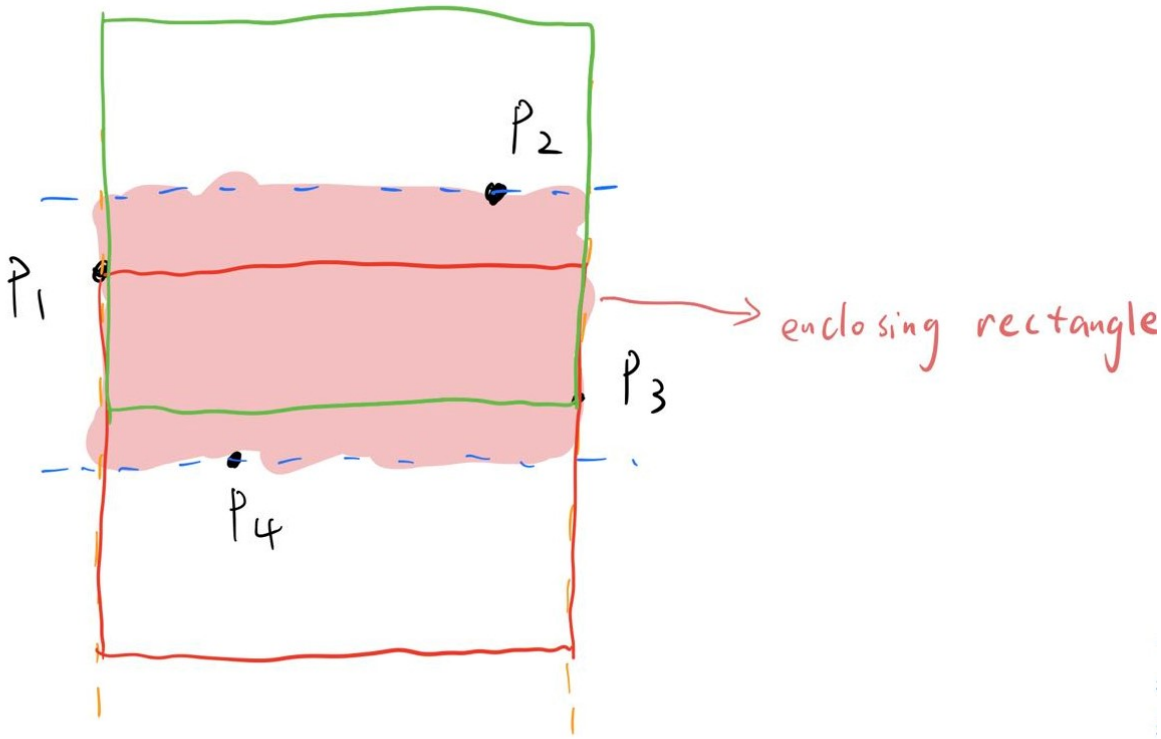

Figure 4.2 4 points situation.

To find the enclosing rectangle of these 4 points, draw the vertical line through the most left and most right points (In this case is P_1 and P_3) and draw horizontal line through the most top and most bottom points (In this case is P_2 and P_4). The intersection area of these 4 lines is the minimum enclosing rectangle.

Suppose P_1 and P_3 is the positive points, the green square and the red square are the minimum enclosing square of these 2 points. However, both squares also contain a negative point P_2 or P_4. Therefore, an axis-aligned square cannot shatter this set of 4 points.

If 4 points lie in the same line like below:

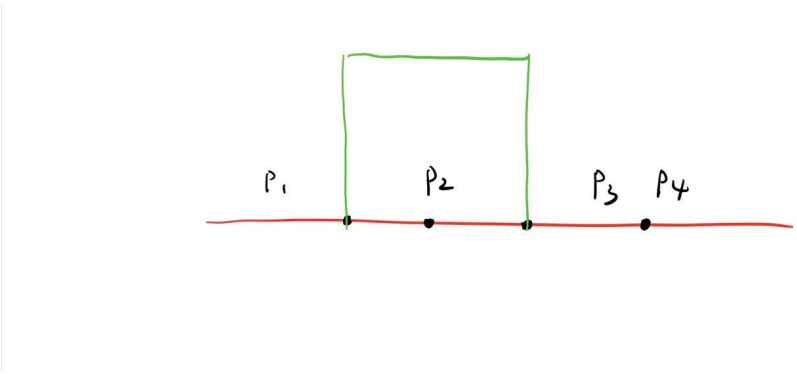

Figure 4.2 4 points situation.

Suppose P_1 and P_3 is the positive points, the green square is the minimum enclosing square of these 2 points. However, it will enclose the negative point P_2 at the same time. Therefore, an axis-aligned square also cannot shatter this set of 4 points.

3. Thus, the VC dimension of an axis-aligned squares is 3.