Support Vector Machines

Some Questions first –

1. It is also used for classification then how it is different from logistic Regression?
2. Is it better than logistic Regression? If yes then how?

3. What are "Support Vectors" in Support Vector Machines?

4. What is the Margin of the Classification Decision Boundary?

Either we will try to answer these questions

Ans) Support-Vector Machine (SVM) is an algorithm which is used to perform regression as well as classification. We will later on see that which is regression. Right now, we will only talk about classification.

The way SVM works is totally different from Logistic Regression. Both algorithms are trying to find the equation of decision boundary,

$f(x) = \hat{\theta}_0 + \hat{\theta}^T x = 0$ or in directly, the parameters of Discriminant function, $\hat{\theta}_0$ and $\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_N \end{bmatrix}$

The cool logistic regression works is that it is trying to minimize the $KL(Q \| P)$ between two posterior distributions by actually minimizing Cross Entropy which is actually Negative log Likelihood function.

Thus, we will see later that SVM trick to find parameters $\hat{\theta}_0$ and $\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_N \end{bmatrix}$ by maximizing something

called Margin of the decision boundary. and then we have pose this problem as minimization problem

Ans 2. Yes, it is better than Logistic Regression.

It is better because it gives the most optimal decision boundary, even better than Logistic Regression.

Don't worry about the idea of most optimal decision boundary right now. We will discuss that idea later.

Moreover, we will see later that, based on training data, every time we will train SVM, it is going to give us the most optimal decision boundary and that decision boundary will be the same every time because

there are many sub optimal decision boundaries but there is one and only one most optimal decision boundary.

for a training data.

Therefore, for the same training data, every time we are going train logistic regression, we will get different sub optimal decision boundaries for one

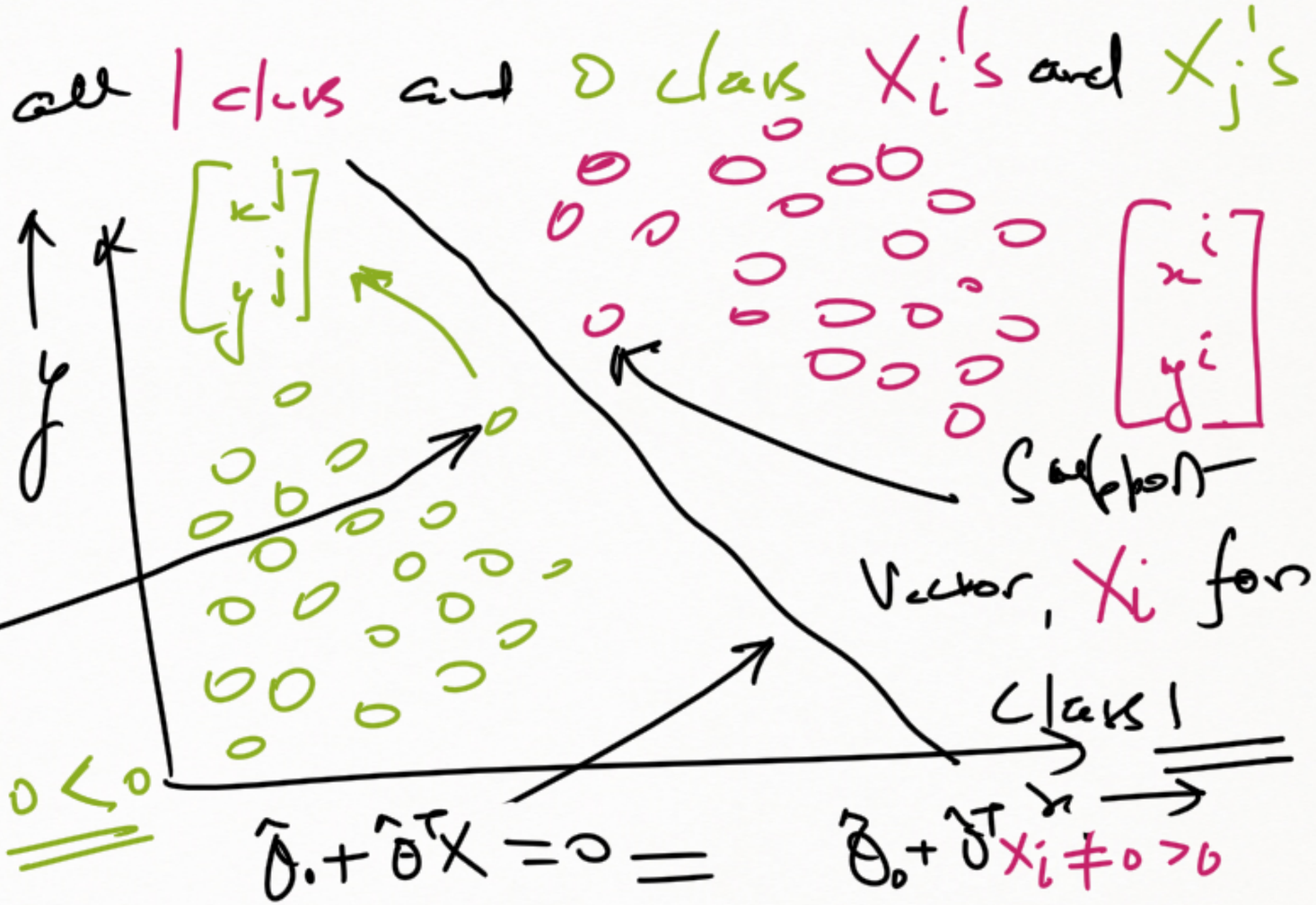
Solution is not <u>Stable</u>.

Ans 3 Support vectors are those data points of each class which have the minimum perpendicular distance from given suboptimal decision boundary, that is those points or feature vectors of each class, X_i and X_j for which

$\hat{\theta}_0 + \hat{\theta}^T X_i > 0$ and $\hat{\theta}_0 + \hat{\theta}^T X_j < 0$ but their magnitudes

will be the least among all 1 class and 0 class X_i's and X_j's

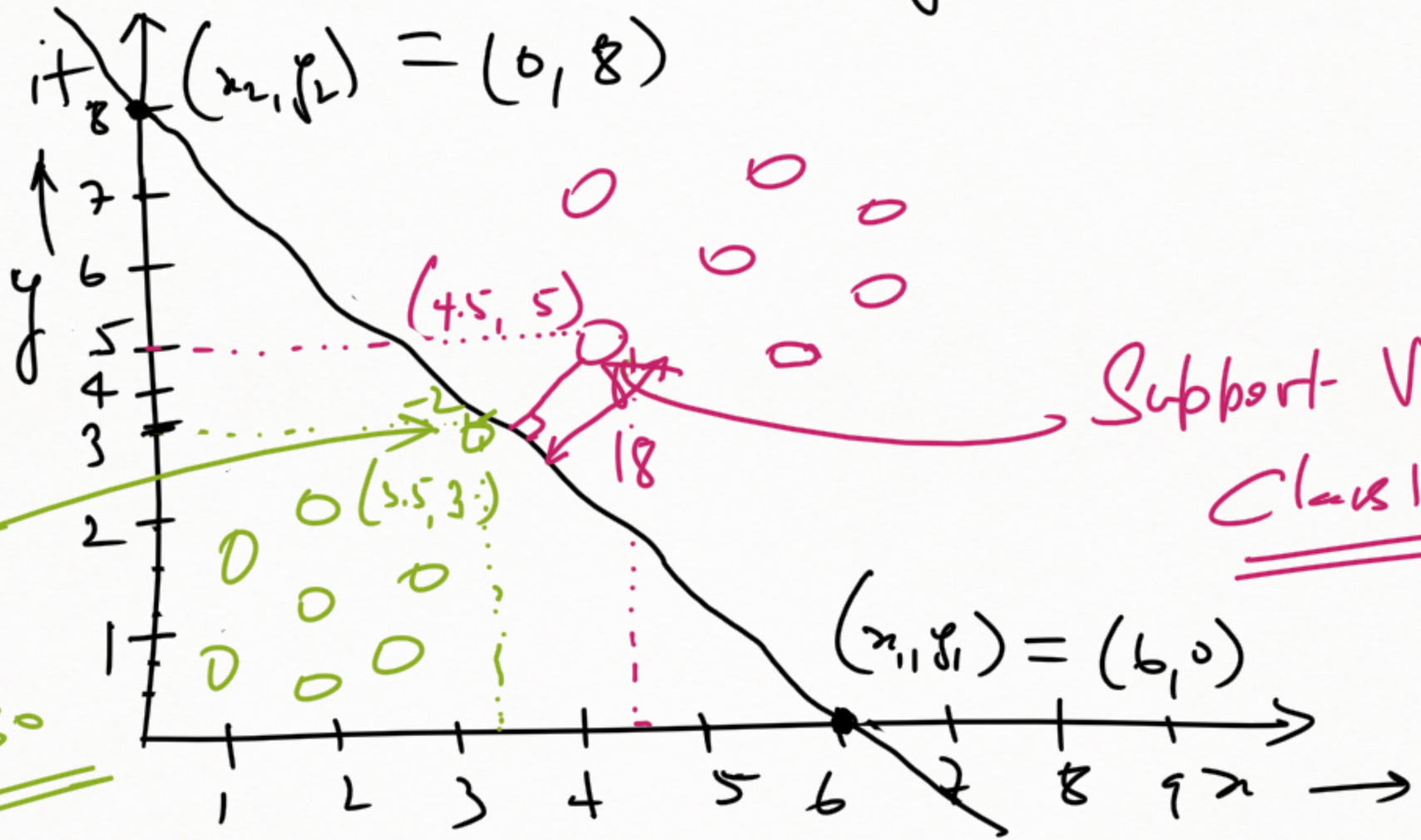
respectively. Lets see —



Support Vector X_j for Class 0 and $\hat{\theta}_0 + \hat{\theta}^T X_j \neq 0 < 0$

Ans 4 Margin of a classification decision boundary is the minimum value among the values of perpendicular distances of all the points X_i and X_j from the suboptimal decision boundary. Lets try to understand this through an **example**

Lets see that we have a very small training data and plot it



Support Vector for class 1

Support Vector for class 0

We know that the equation of decision boundary is —

$$\hat{\theta}_0 + \hat{\theta}^T X = 0$$

$$\hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 y = 0 \Rightarrow \hat{\theta}_2 y = -\hat{\theta}_1 x - \hat{\theta}_0$$

$$y = \left(-\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) x + \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right)$$

It is of the form —

$$y = mx + c$$

where $m = \left(-\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)$

$$c = \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right)$$

where $m = \left(-\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) =$ slope of the decision boundary

$c = \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right) =$ y-intercept of the decision boundary, that is where the decision boundary cuts y-axis

As the y-intercept is +8 we can say that —

$$-\frac{\hat{\theta}_0}{\hat{\theta}_2} = 8$$

We can easily calculate slope, m of the boundary as,

$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(8-0)}{(0-6)} = -\frac{8}{6} = -\frac{\hat{\theta}_1}{\hat{\theta}_2}$$

By comparison, $-\dfrac{\hat{\theta}_1}{\hat{\theta}_2} = \dfrac{-8}{6} \Rightarrow \hat{\theta}_1 = 8,\ \hat{\theta}_2 = 6$

$$C = 8 = -\frac{\hat{\theta}_0}{\hat{\theta}_2} = -\frac{(-48)}{6} \quad \text{i.e. } \hat{\theta}_0 = -48$$

In that case, equation of decision boundary is given as $\hat{\theta}_0 + \hat{\theta}^T x \Rightarrow -48 + \begin{bmatrix} 8 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$

$$-48 + 8x + 6y = 0$$

$$\Rightarrow 6y = -8x + 48$$

$$y = \frac{-8}{6}x + \frac{48}{6}$$

$$y = \frac{-8}{6}x + 8$$

Equation of decision boundary

Remember that $\hat{\theta} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$ is the score vector, W

According to the conditions followed in LDA

as dimensionality reduction before we were concerned

about the direction of W, that is its unit vector

rather than magnitude of W. Therefore we can scale

the values inside $\hat{\theta}$ up or down by multiplying them with a scalar quantity and the direction of upscaled or downscaled vector will still not change. i.e

$$\hat{\theta} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 16 \\ 12 \end{bmatrix}, \begin{bmatrix} -8 \\ -6 \end{bmatrix}, \begin{bmatrix} -4 \\ -3 \end{bmatrix} \ldots$$

all back vectors will be considered same because their direction is same or in other words, direction of the decision boundary will remain the same because for all of them $\frac{\hat{\theta}}{\|\hat{\theta}\|}$, that is the unit vector will come out to be the same and because for us direction

is more important than magnitude therefore for us

$$\hat{\theta} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 16 \\ 12 \end{bmatrix}, \begin{bmatrix} -8 \\ -6 \end{bmatrix}, \begin{bmatrix} -4 \\ -3 \end{bmatrix} \ldots$$

all are considered the same and we can choose

any of the $\hat{\theta}$

If you see the diagram then we have two support vectors, X_i and X_j where $X_i = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix} = \begin{bmatrix} 9/2 \\ 5 \end{bmatrix}$

$X_j = \begin{bmatrix} 3.5 \\ 4 \end{bmatrix} = \begin{bmatrix} 7/2 \\ 3 \end{bmatrix}$

Lets calculate the perpendicular distance of these support vectors from the decision boundary.

$$f(x_i) = \hat{\theta}_0 + \hat{\theta}^T x_i$$

$$= -48 + \begin{bmatrix} 8 & 6 \end{bmatrix} \begin{bmatrix} 9/2 \\ 5 \end{bmatrix}.$$

$$= -48 + \frac{8 \cdot 9}{2} + 6 \cdot 5 \Rightarrow -48 + 36 + 30$$

$$\Rightarrow 18 > 0 \neq 0$$

This sample-vector is lying on decision boundary

$$f(x_j) = \hat{\theta}_0 + \hat{\theta}^T x_j$$

$$= -48 + \begin{bmatrix} 8 & 6 \end{bmatrix} \begin{bmatrix} 7/2 \\ 3 \end{bmatrix}$$

$$= -48 + \frac{8 \cdot 7}{2} + 18$$

$$= -48 + 28 + 18 = -2 < 0 \neq 0$$

Why because this support vector is not lying on the decision boundary

The margin of the decision boundary is given by

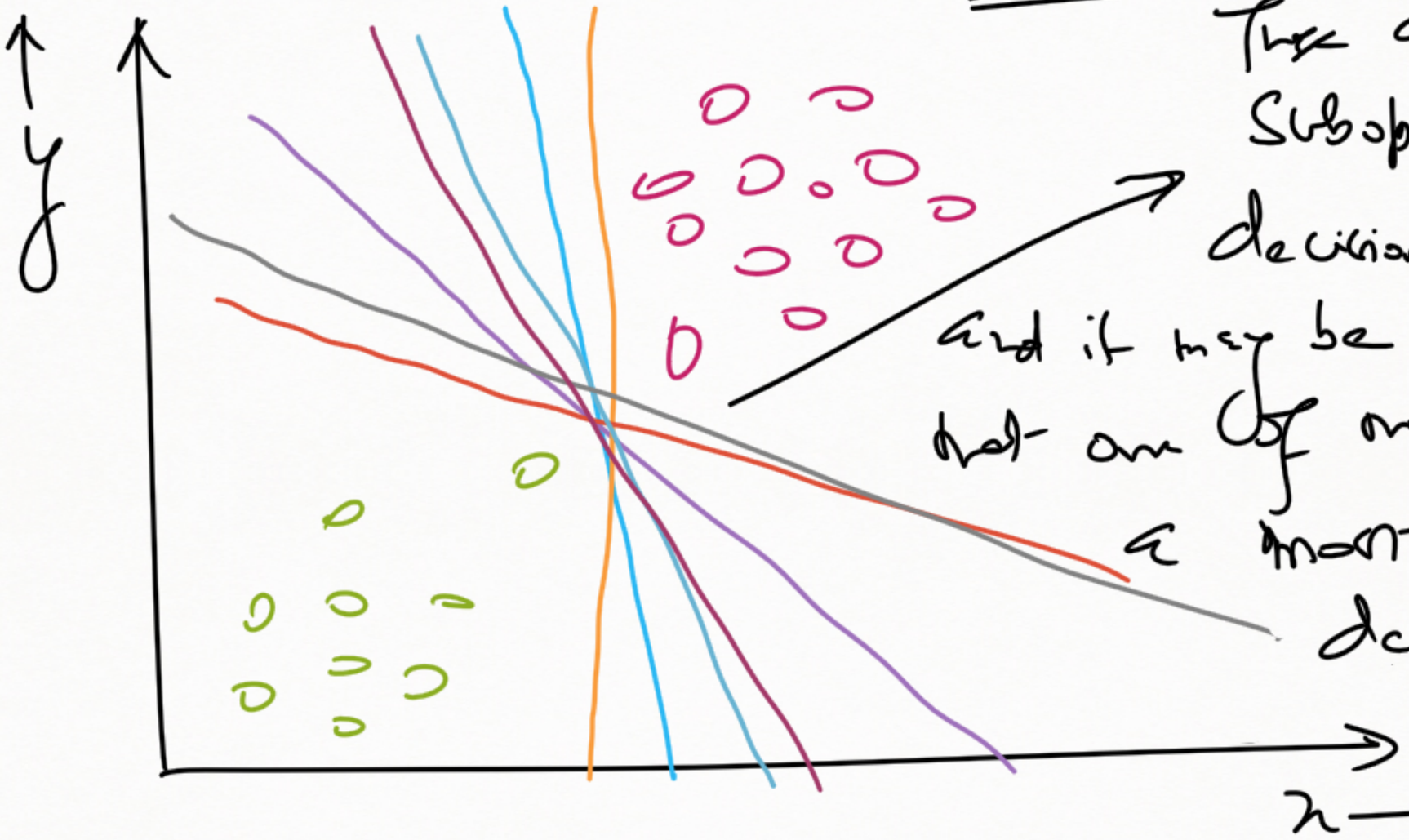
$$\min(18, -2) = \underline{\underline{-2}}$$

Similarly, the margin of any other suboptimal decision boundary can be computed.

So, what is a suboptimal decision boundary. Well, any decision boundary that can classify training data properly is called a suboptimal decision boundary. It is defined by the vector, $\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_N \end{bmatrix}$

Number of suboptimal decision boundaries can be <u>numerous</u>.

Lets consider the same data



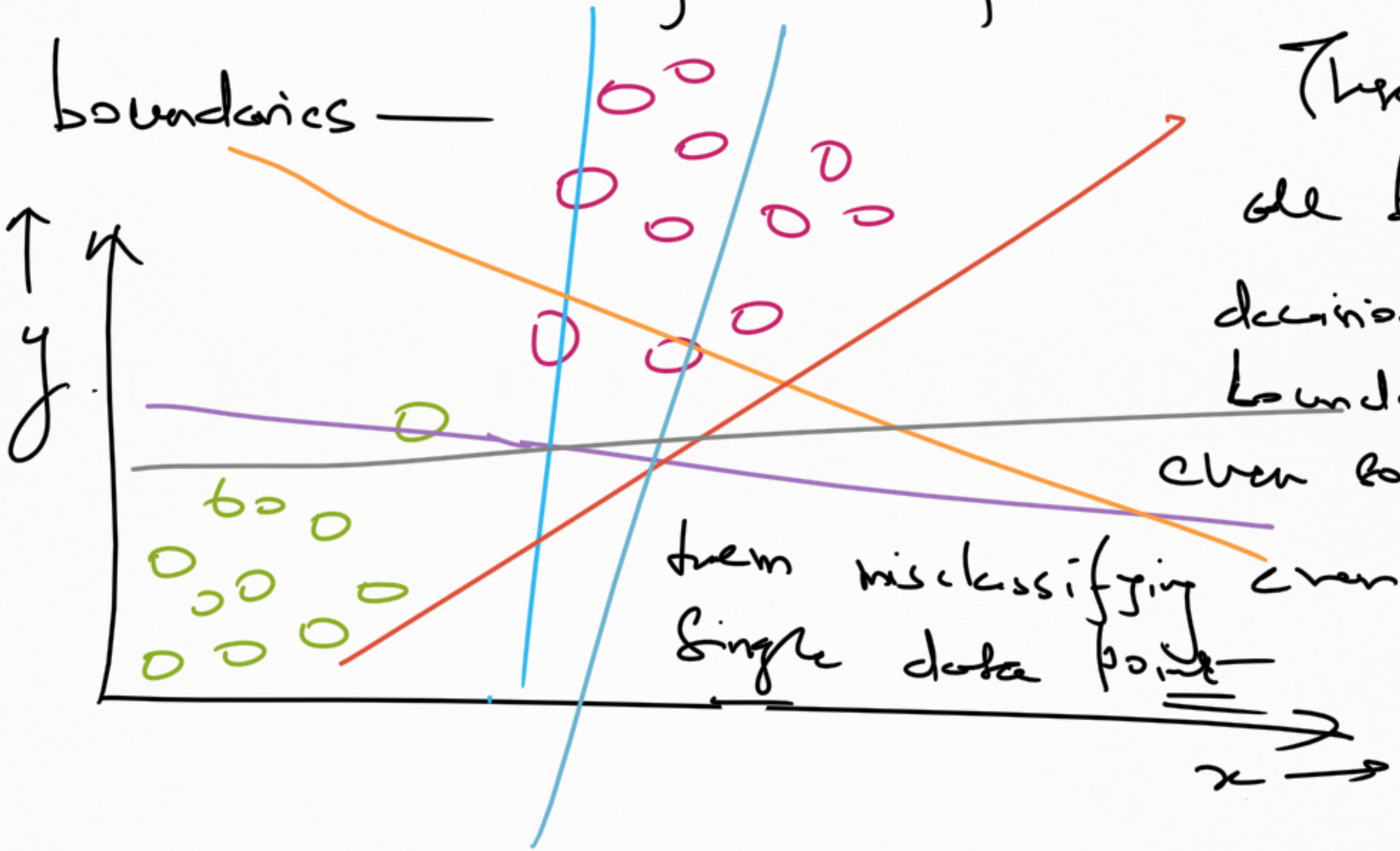
These are all suboptimal decision boundaries

and it may be the case that one of them will be a more optimal decision boundary

Logistic Regression, everytime trained on the same dataset - gives different sub-optimal decision boundaries but SVM gives the same most optimal decision boundary, when trained on the same dataset again and again. because there is only single most optimal decision boundary.

Always remember that — most optimal decision boundary will be the one from all suboptimal decision boundaries. But what is a most optimal decision boundary, that we will see <u>later</u>.

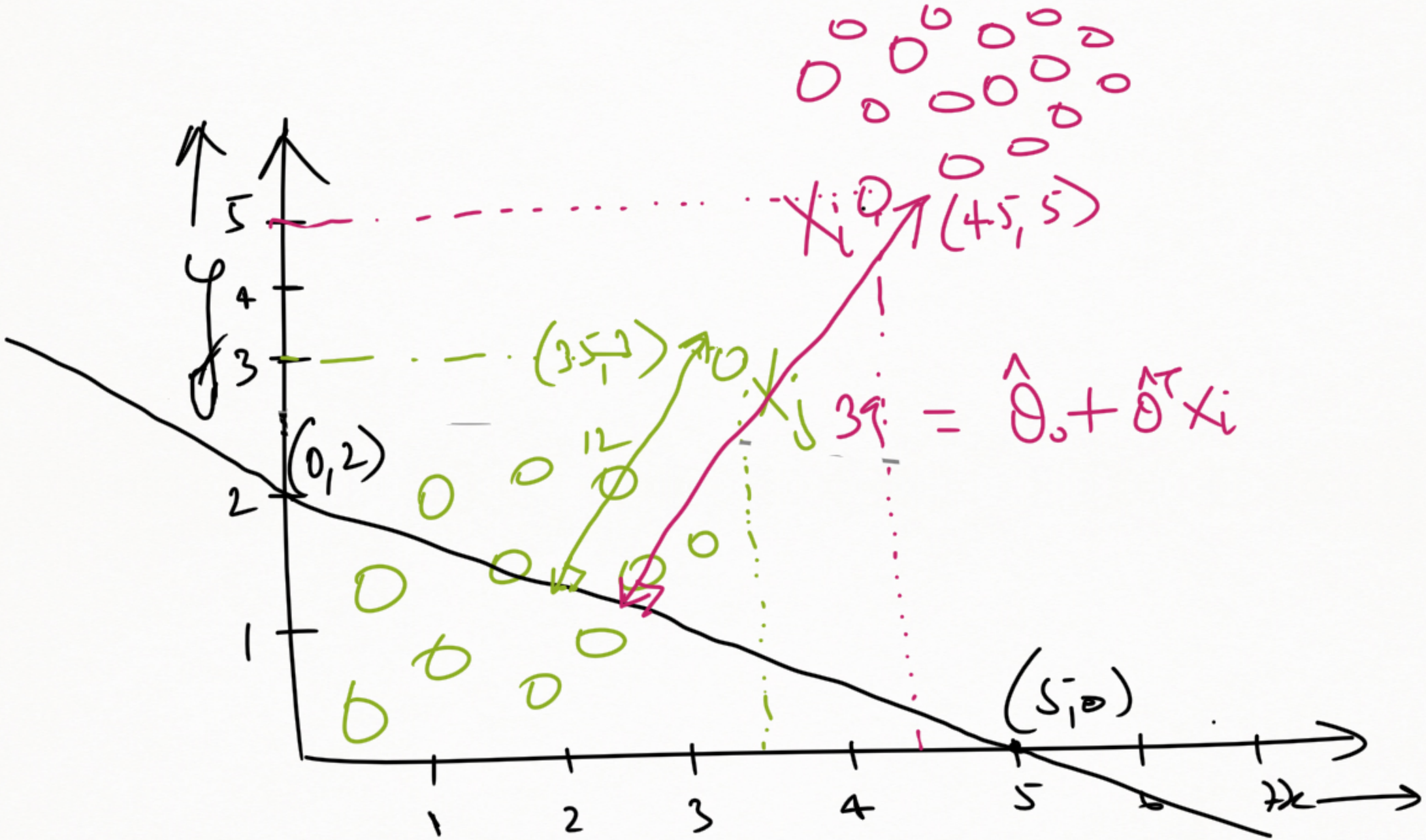
Let's see some examples of Non optimal or bad decision boundaries —



These are all bad decision boundaries even some of them misclassifying even a single data point —

So, a decision boundary is bad or non optimal even if it is misclassifying even a single training example according to the theory of <u>SVM</u>.

Lets try computing the margin of a bad decision <u>boundary</u>


y
x
(0,2)
(5,0)
(3.5,1)
(4.5,5)
x_i
y_i
$y_i = \hat{\theta}_0 + \hat{\theta}^T x_i$

The y-intercept of this line is +2 and slope, m is given by -

$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(2-0)}{(0-5)} = \frac{-2}{5} = \left(-\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)$$

$$c = +2 = \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right)$$

$$\hat{\theta}_1 = 2, \ \hat{\theta}_2 = 5$$

$$c = \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right) = -\left(\frac{-10}{5}\right) = +2, \quad \hat{\theta}_0 = \underline{\underline{10}}$$

$$f(x) = \hat{\theta}_0 + \hat{\theta}^T x$$

$$= 10 + \begin{bmatrix} 2 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Let's compute the perpendicular distance of both support vectors X_i and X_j —

$$\hat{\theta}_0 + \hat{\theta}^T X_i = 10 + \begin{bmatrix} 2 & 5 \end{bmatrix} \begin{bmatrix} 9/2 \\ 4 \end{bmatrix} = -10 + \frac{9 \cdot 2}{2} + 5 \cdot 4$$

$$= \underline{\underline{19}}$$

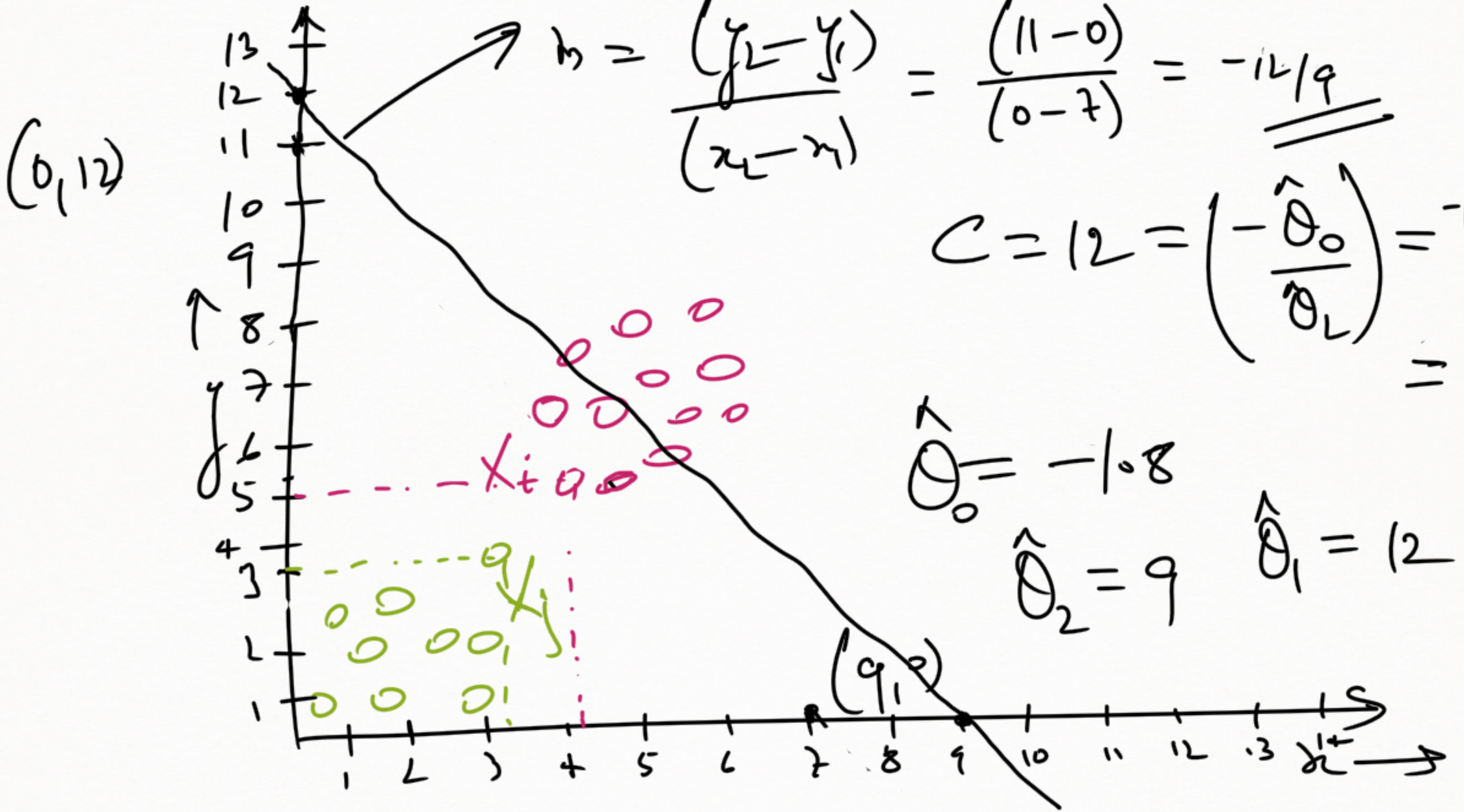
$$\hat{\theta}_0 + \hat{\theta}^T X_j = -10 + \begin{bmatrix} 2 & 5 \end{bmatrix} \begin{bmatrix} 7/2 \\ 3 \end{bmatrix} = -10 + \frac{7 \cdot 2}{2} + 5 \cdot 3$$

$$= \underline{\underline{12}}$$

The margin of decision boundary is given as -

$$\min(12, 14) = 12$$

Lets consider one last scenario



$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(11-0)}{(0-7)} = -12/9$$

$$C = 12 = \left(-\frac{\hat{\theta}_0}{\hat{\theta}_2}\right) = \frac{-(-1.8)}{9} = +12$$

$$\hat{\theta}_0 = -1.8$$

$$\hat{\theta}_2 = 9 \qquad \hat{\theta}_1 = 12$$

$$f(x) = \hat{\theta}_0 + \hat{\theta}^T x$$

$$= -108 + \begin{bmatrix} 12 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$f(x_i) = \hat{\theta}_0 + \hat{\theta}^T x_i = -108 + \begin{bmatrix} 12 & 9 \end{bmatrix} \begin{bmatrix} 9/2 \\ 5 \end{bmatrix}$$

$$= -108 + \frac{12 \cdot 9}{2} + 45 = \underline{\underline{-9}}$$

$$f(x_j) = \hat{\theta}_0 + \hat{\theta}^T x_i$$

$$= -108 + \begin{bmatrix} 12 & 9 \end{bmatrix} \begin{bmatrix} 7/2 \\ 3 \end{bmatrix}$$

$$= -108 + \frac{12 \cdot 7}{2} + 27 \Rightarrow \underline{\underline{-39}}$$

Therefore, the margin of the decision boundary is

$$\min(-9, -39)$$

$$= \underline{\underline{-39}}$$