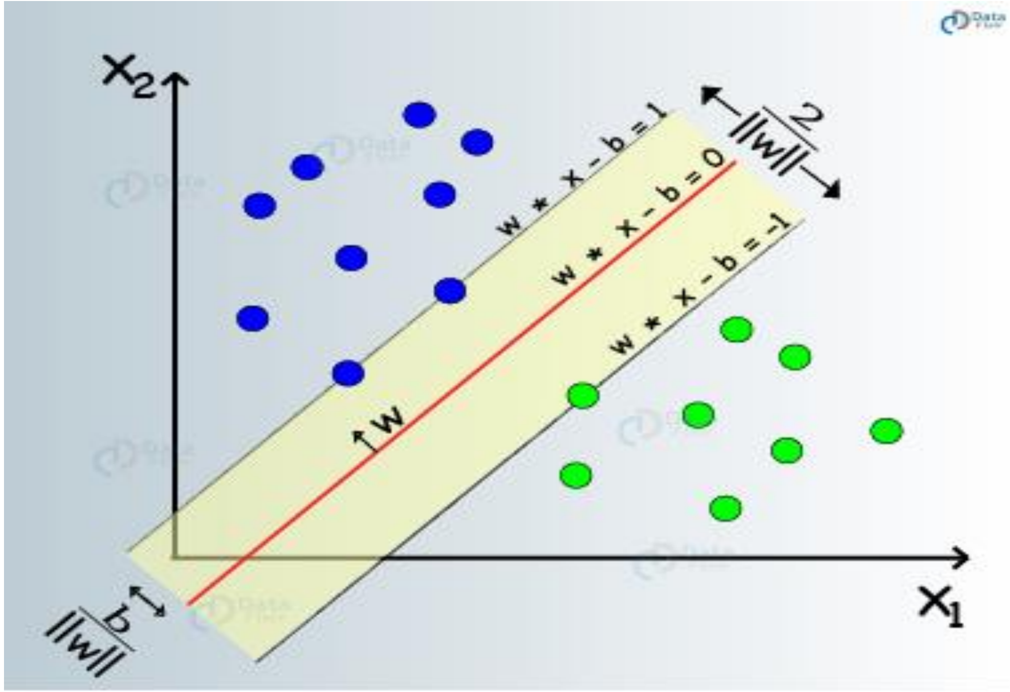
SVM Introduction

Zahoor Tanoli (PhD)
CUI Attock

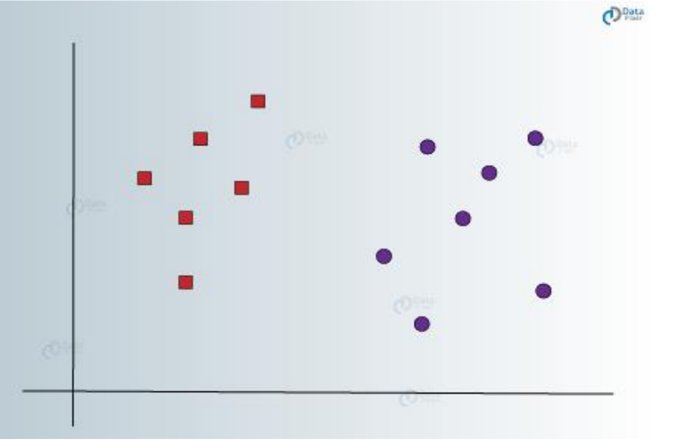
What is SVM?

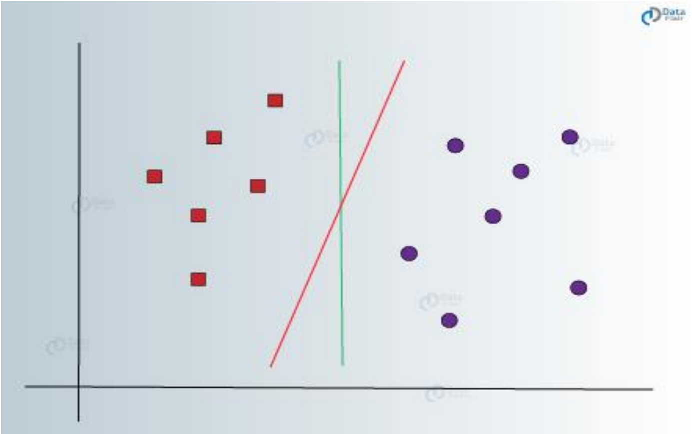
- Support Vector Machines are a type of supervised machine learning algorithm that provides analysis of data for classification and regression analysis
- Find the ideal hyperplane that differentiates between the two classes
- Datasets can be separated easily with the help of a line, called a decision boundary



How does SVM work?

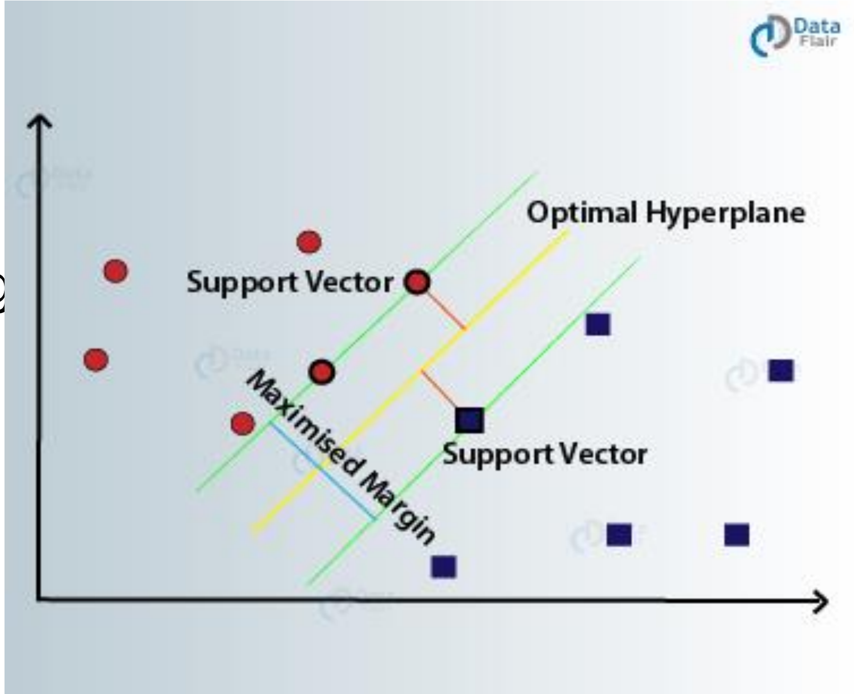
- Create a hyperplane that separates the dataset into classes
- You have to classify red triangles from blue circles
- Goal is to create a line that classifies the data into two classes, creating a distinction between red triangles and blue circles
- Visualize some of the lines that can differentiate between the two classes
- The green line cannot be the ideal line as it lies too close to the red class





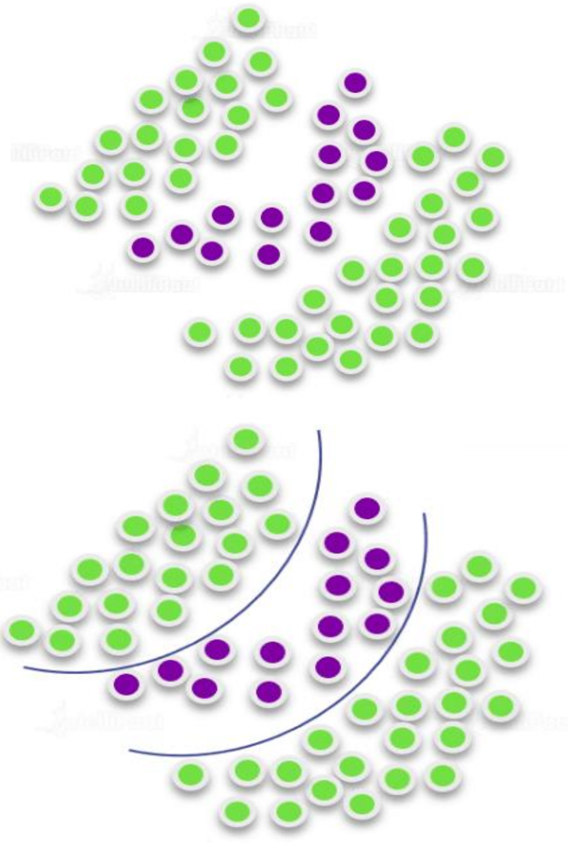
Continued...

- Find the points that lie closest to both the classes
- These points are known as support vectors
- In the next step, we find the proximity between our dividing plane and the support vectors
- The distance between the points and the dividing line is known as margin
- The aim of an SVM algorithm is to maximize this very margin
- When the margin reaches its maximum, the hyperplane becomes the optimal one



Non-linearly separable datasets

- Straight lines can't be used to classify such dataset
- Kernel SVM comes into the picture
- Kernel SVM projects the non-linearly separable datasets of lower dimensions to linearly separable data of higher dimensions



SVM – the math

$$\boxed{\max \frac{2}{||w||} \text{ such that } w^T x_i + b \begin{array}{l} \geq 1 \ \text{ if } Y_i = +1 \\ \leq -1 \text{ if } Y_i = -1 \end{array}}$$

To really understand support vector machines, you need to understand this kind of math. At the end of this lecture, I hope that you get the main ideas behind its math. Once you know this, it will be a lot easier to read more advanced stuff about SVM.

The equation of a straight line

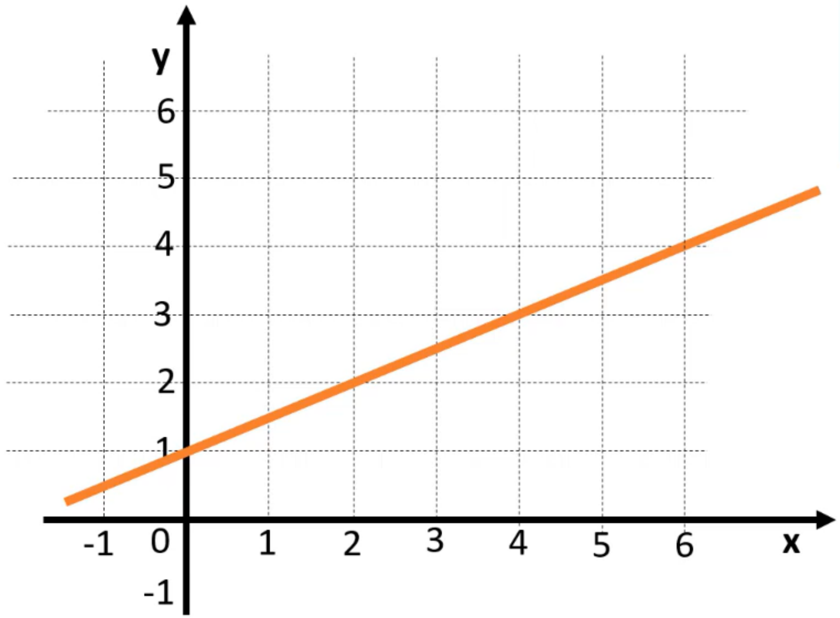
$$\boxed{y = mx + b}$$

$$\boxed{w^T x + b = 0}$$

We will therefore see how to go from the simple equation of a straight line,

to this type of equation that is used in SVM, and then see how such an equation can be used for classification.

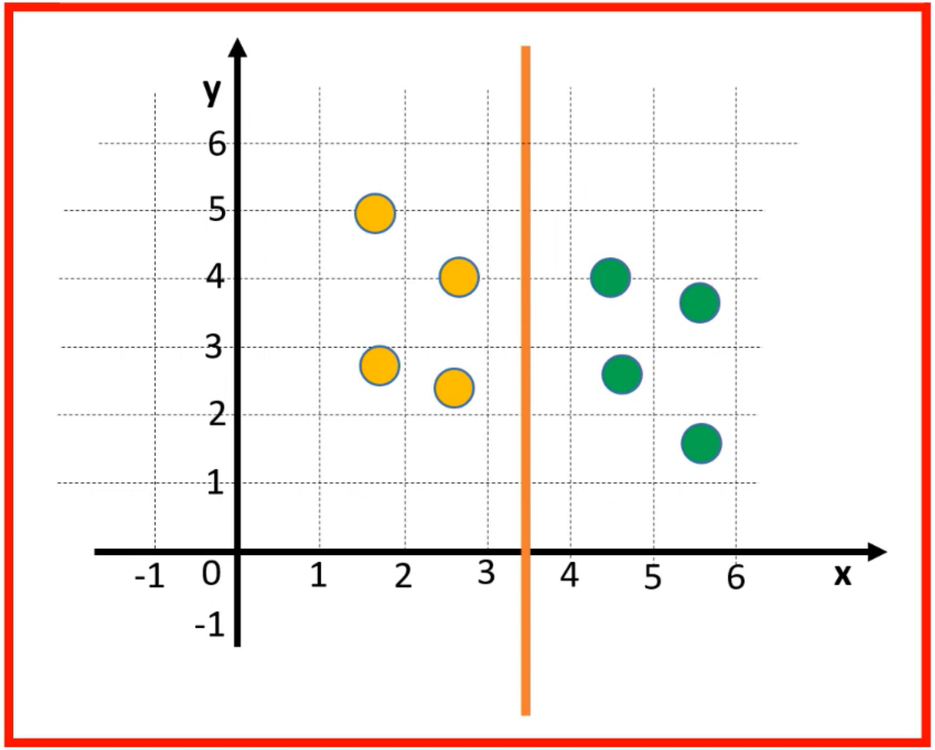
The equation of a straight line



$$y = mx + b$$

Usually, we express the equation of a straight line like this, which is called the slope-intercept form because the equation contains information about the slope and the intercept of the line.

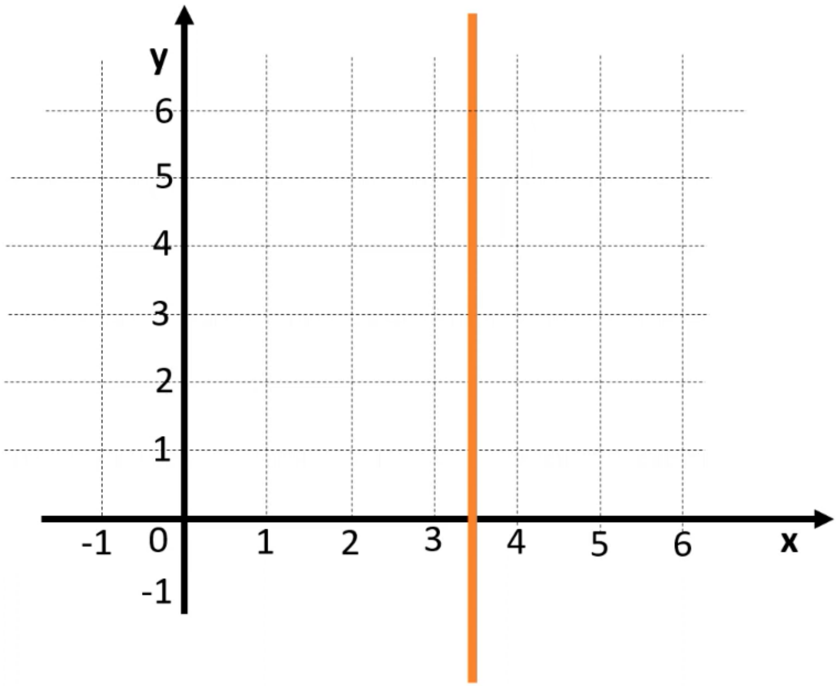
The equation of a straight line



$$y = mx + b$$

Since vertical lines can be used to separate data points of two groups, the slope-intercept form to describe the line is not appropriate to use in SVM.

The equation of a straight line

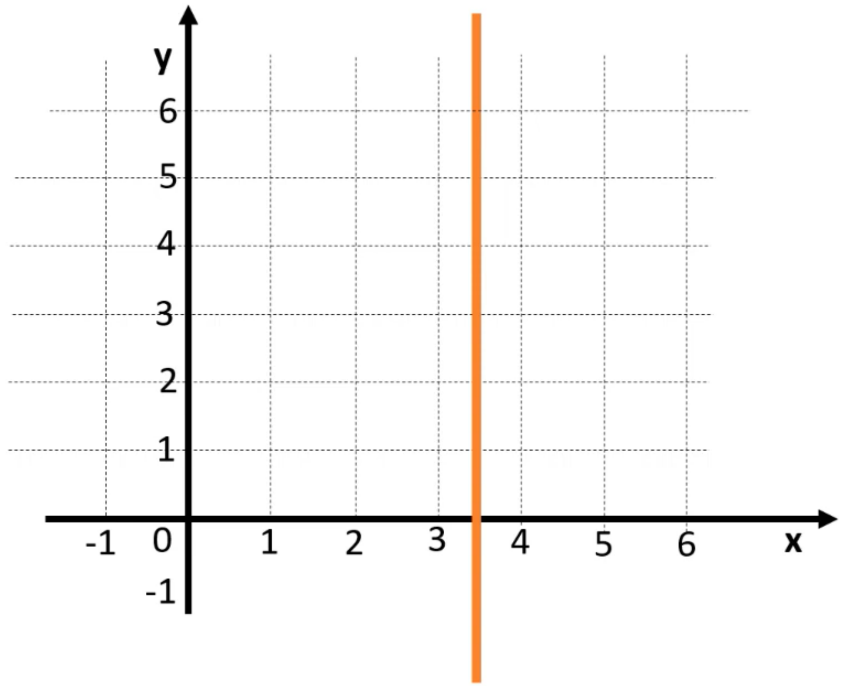


$$y = mx + b$$

$$\boxed{Ax + By + C = 0}$$

This is one reason why we need to use the so-called "general form" of the equation for a straight line.

The equation of a straight line



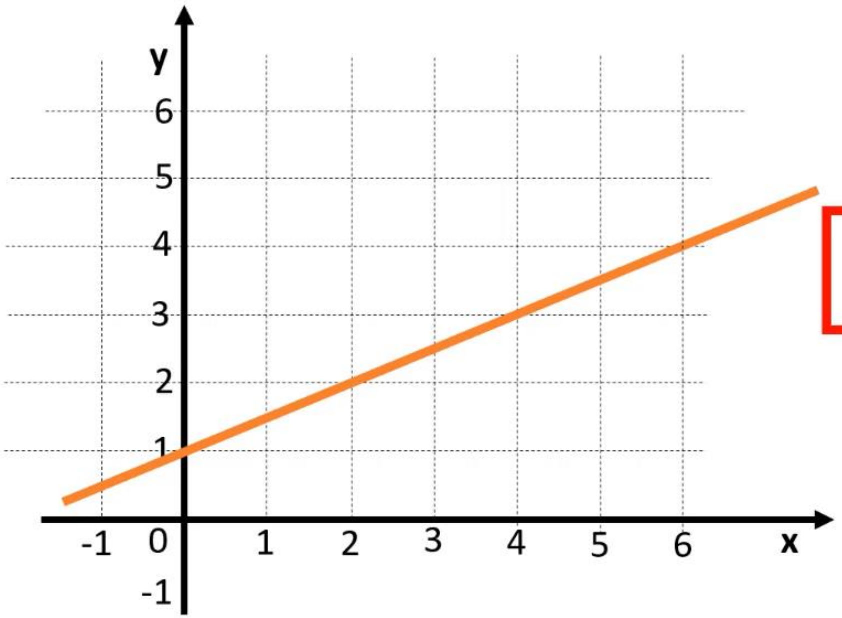
$$y = mx + b$$

$$Ax + By + C = 0$$

$$\boxed{y = -\frac{A}{B}x - \frac{C}{B}}$$

we will get the following equation,

The equation of a straight line



$y = mx + b$

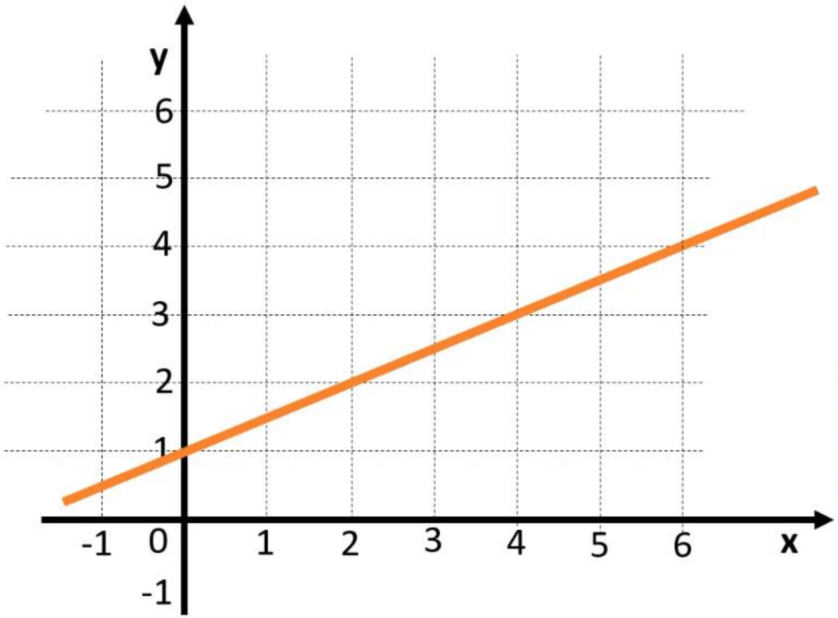
$Ax + By + C = 0$

$$y = 0.5x + 1$$

$$\boxed{-0.5x + y - 1 = 0}$$

so that we have the following equation that also describes the line.

The equation of a straight line



$$y = 0.5x + 1$$

$$-0.5x + y - 1 = 0$$

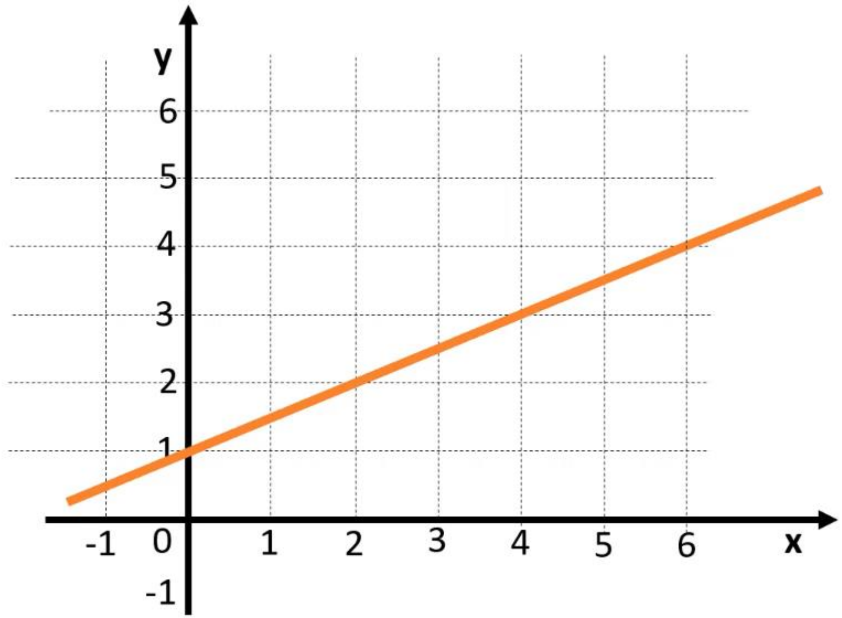
$$\boxed{-2x + 4y - 4 = 0}$$

$$y = mx + b$$

$$Ax + By + C = 0$$

we will get the following equation that also describes the same line,

The equation of a straight line



$$y = mx + b$$
$$Ax + By + C = 0$$

$$y = 0.5x + 1$$

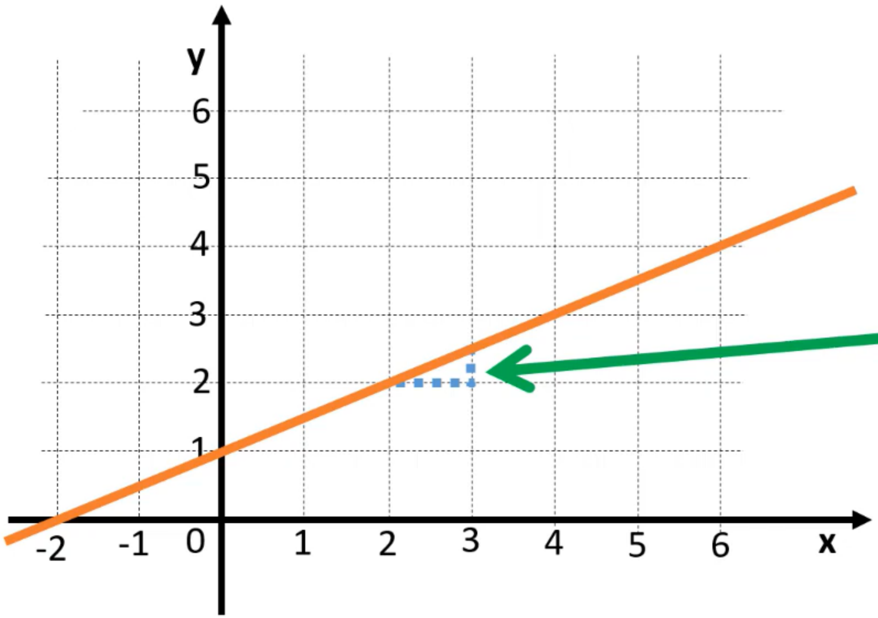
$$-0.5x + y - 1 = 0$$

$$-2x + 4y - 4 = 0 \quad \times (-1)$$

$$\boxed{2x - 4y + 4 = 0}$$

so that we get the following equation that also describes the same line.

The equation of a straight line



$$-2x + 4y - 4 = 0$$

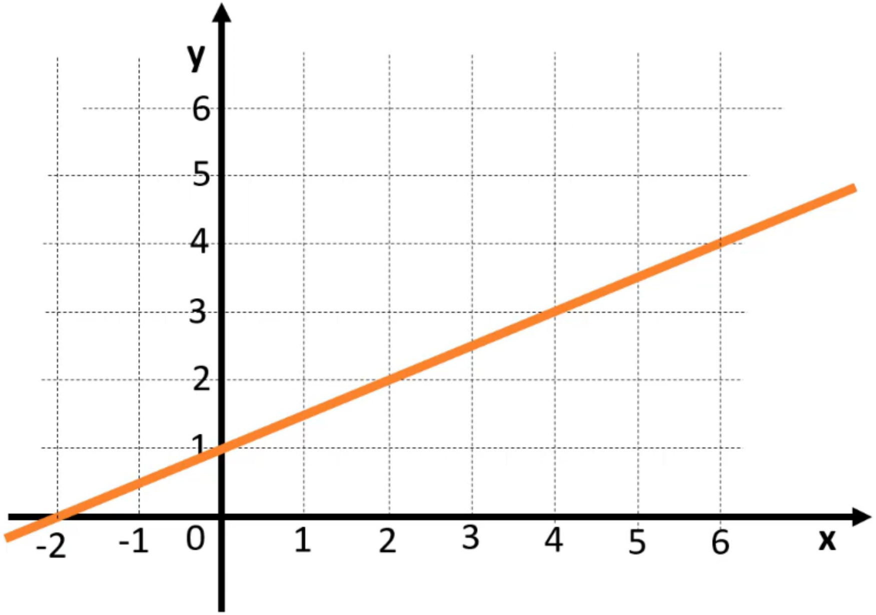
$$Ax + By + C = 0$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

$$y = \boxed{-\frac{(-2)}{4}}x - \frac{(-4)}{4}$$

We see that the slope is 0.5 because if we increase one unit in x, y is increased by 0.5,

The equation of a straight line



$$\boxed{Ax + By + C = 0}$$

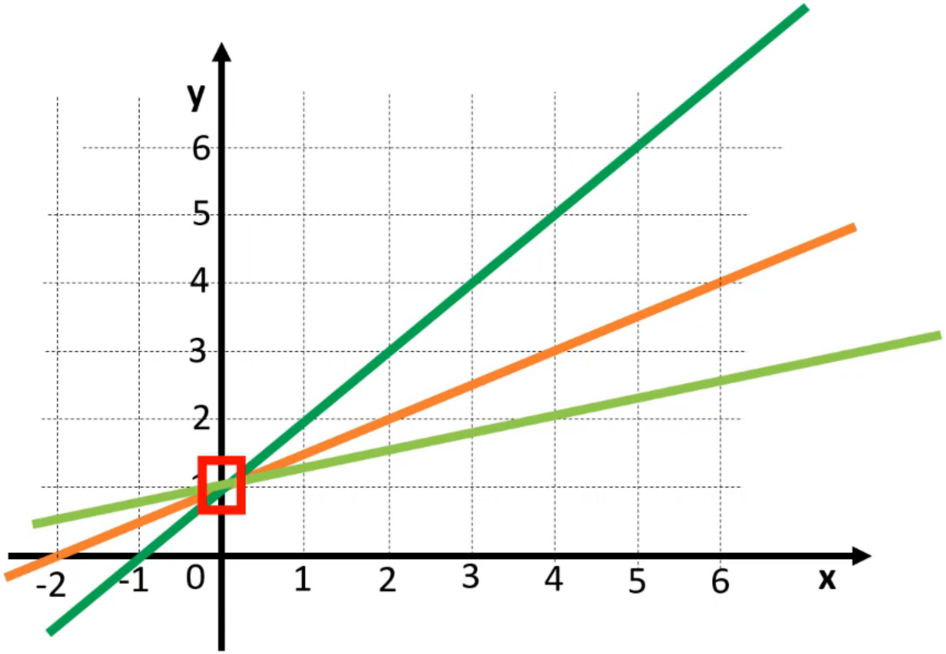
$$-2x + 4y - 4 = 0$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

$$y = -\frac{(-2)}{4}x - \frac{(-4)}{4}$$

Linear support vector machines find the optimal values of the coefficients A and B, so that the line separates the data points as good as possible. We therefore first need to understand what happens with the line if we change A, B and C.

The equation of a straight line



$$Ax + By + C = 0$$

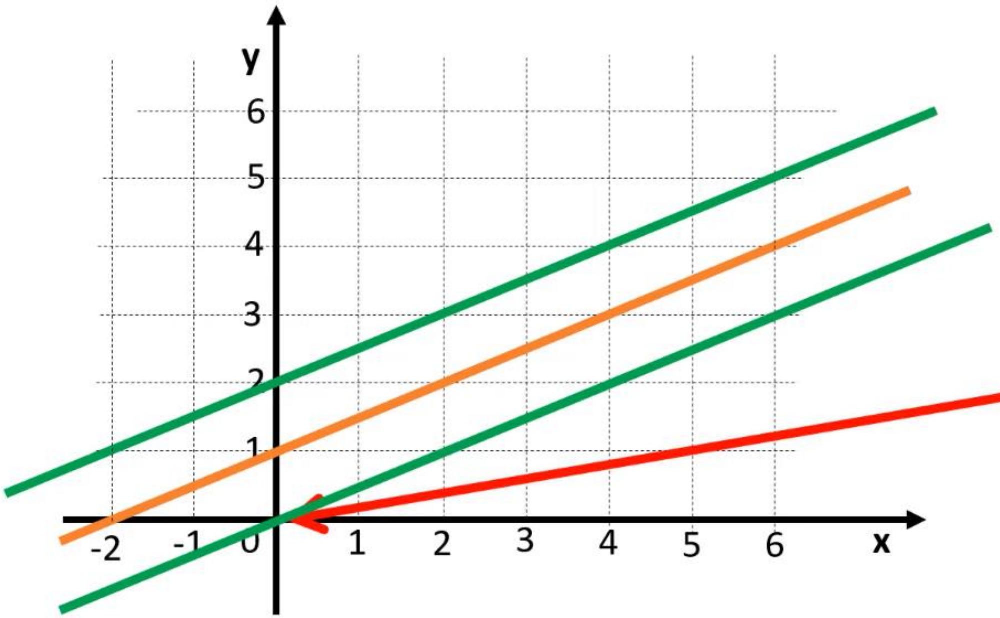
$$-2x + 4y - 4 = 0$$

$$Ax + By + C = 0$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

Note that, when we change A, the line is rotating around this point.

The equation of a straight line



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$-2x + 4y + 0 = 0$$

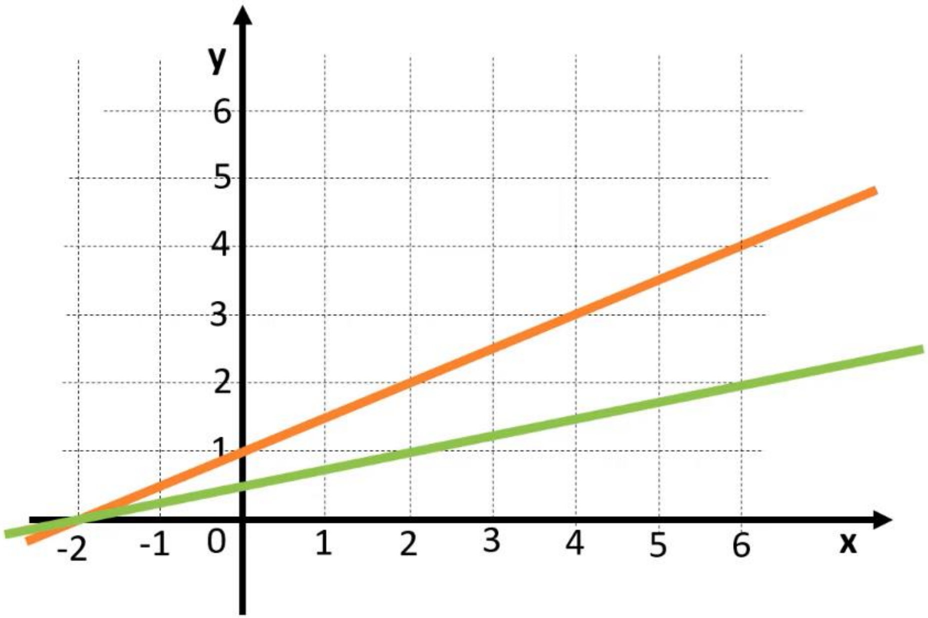
$$Ax + By + C = 0$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

$$y = -\frac{(-2)}{4}x \boxed{-\frac{0}{4}}$$

and if we set C to zero, the intercept will be reduced to zero.

The equation of a straight line



$$Ax + By + C = 0$$

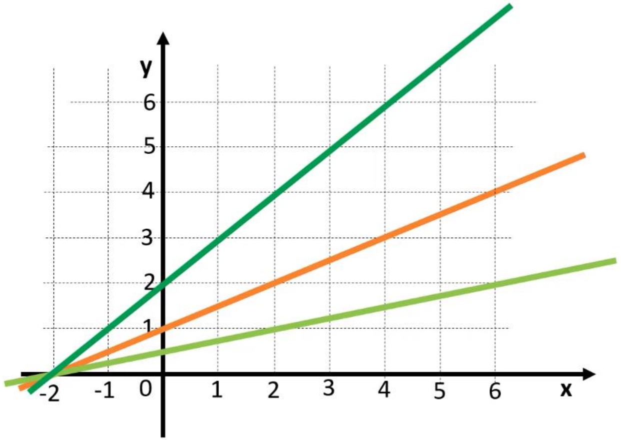
$$-2x + \boxed{4}y - 4 = 0$$

$$-2x + \boxed{2}y - 4 = 0$$

$$Ax + By + C = 0$$

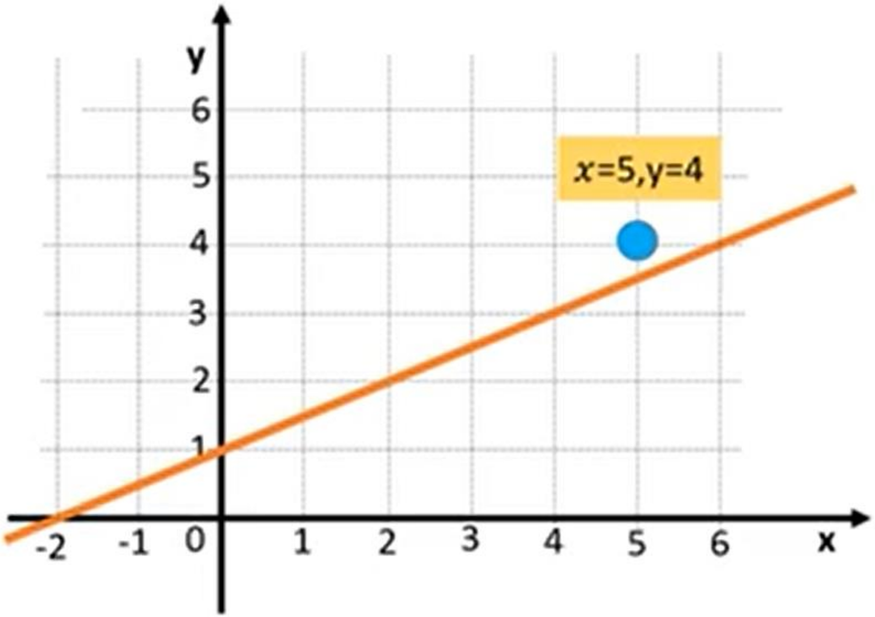
$$y = -\frac{A}{B}x - \frac{C}{B}$$

$$y = -\frac{(-2)}{\boxed{2}}x - \frac{(-4)}{\boxed{2}}$$



If we now reduce B from four to two,

The equation of a straight line

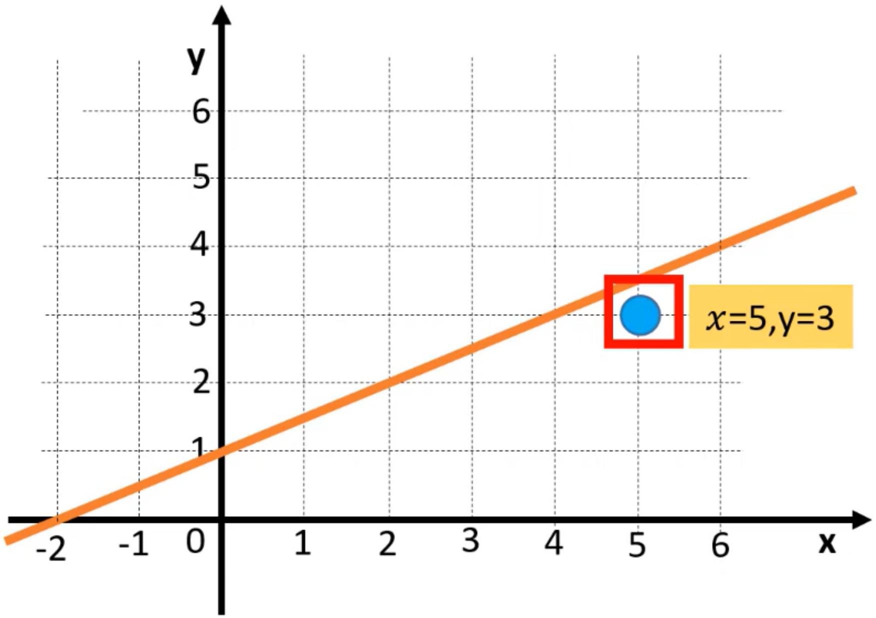


$$\boxed{\begin{array}{c} Ax + By + C = 0 \\ -2x + 4y - 4 = 0 \end{array}}$$

$$-2 \cdot 5 + 4 \cdot 4 - 4 = \boxed{2}$$

Another nice feature about this form of the equation is that it can tell us which side of the line a data point is located on.

The equation of a straight line



$$Ax + By + C = 0$$

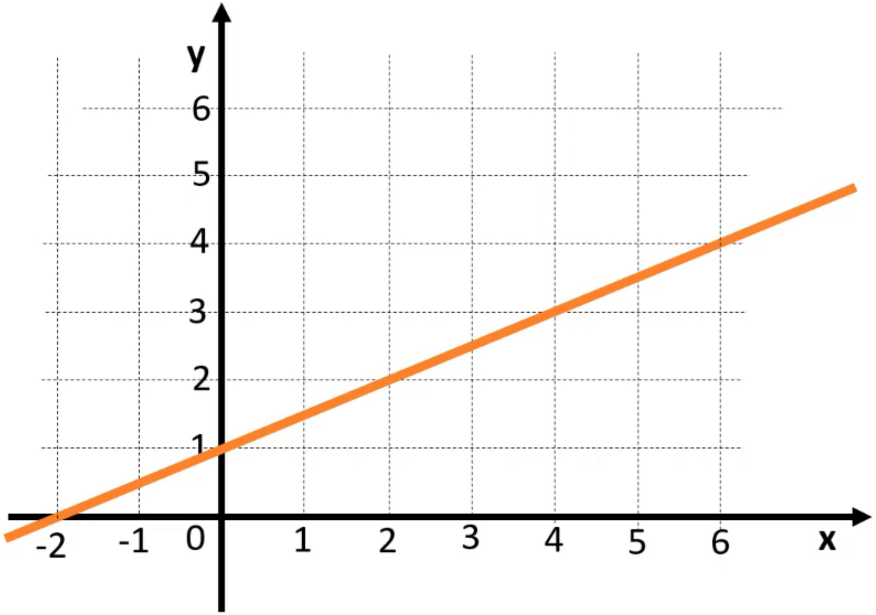
$$-2x + 4y - 4 = 0$$

$$-2 \cdot 5 + 4 \cdot 3 - 4 = \boxed{-2}$$

In comparison, this data point results in a negative value, which tells us that the data point is below the line. This is handy when we later will classify data points based on the equation of the line.

The equation of a straight line

$y = 0.5x + 1$



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$-2x + 4y - 4 = 1$$

$$4y = 2x + 5$$

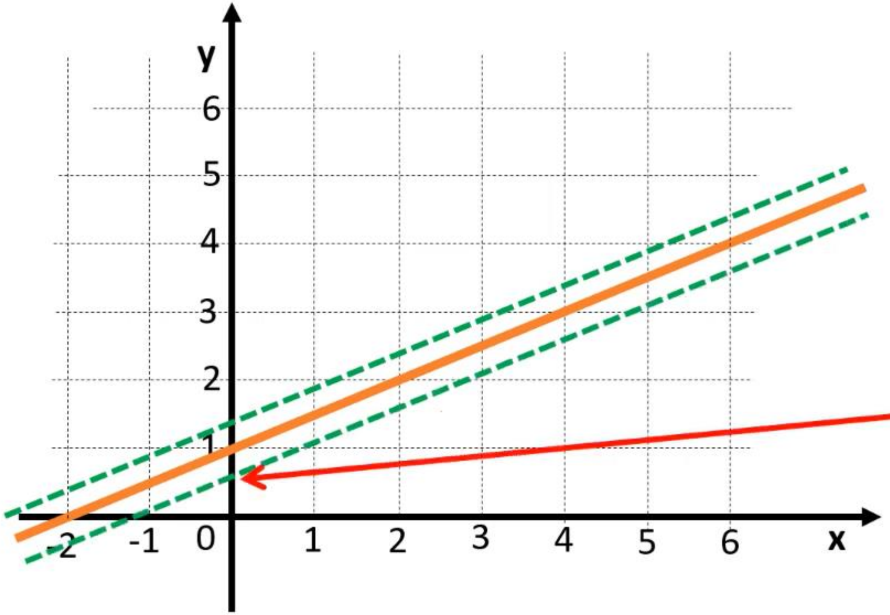
$$\boxed{y = 0.5x + 1.25}$$

If we then solve for y,

Setting the Margin

The equation of a straight line

$y = 0.5x + 1$



$Ax + By + C = 0$

$-2x + 4y - 4 = 0$

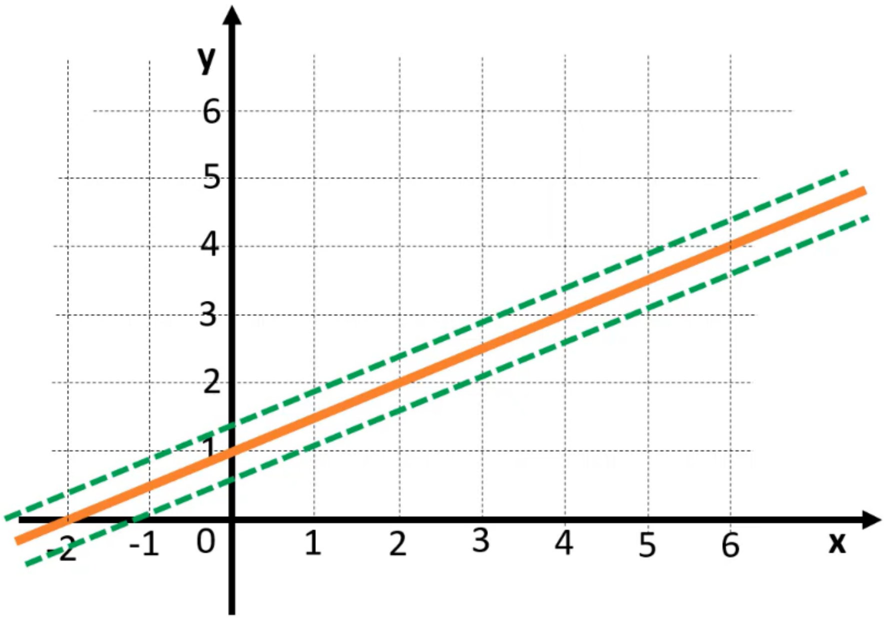
$-2x + 4y - 4 = -1$

$4y = 2x + 3$

$y = 0.5x + \boxed{0.75}$

we will get this line that intercepts the y-axis at 0.75.

The equation of a straight line

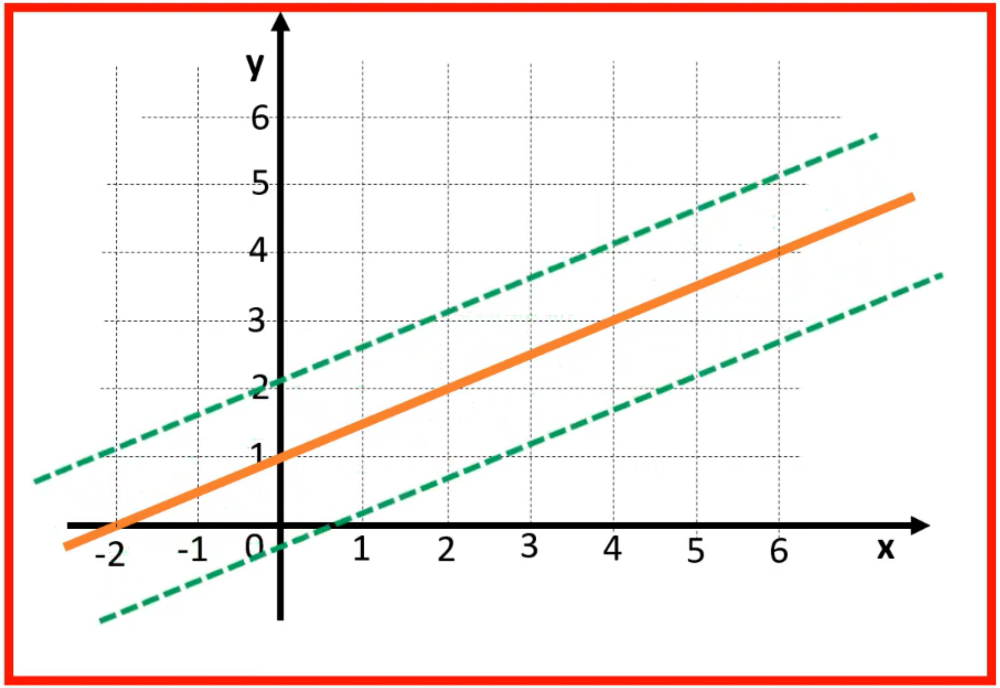


$$Ax + By + C = 0$$

$$\boxed{\begin{aligned} -2x + 4y - 4 &= 1 \\ -2x + 4y - 4 &= 0 \\ -2x + 4y - 4 &= -1 \end{aligned}}$$

If we would multiply the terms of the left-hand side of these three equations with a factor smaller than 1,

The equation of a straight line

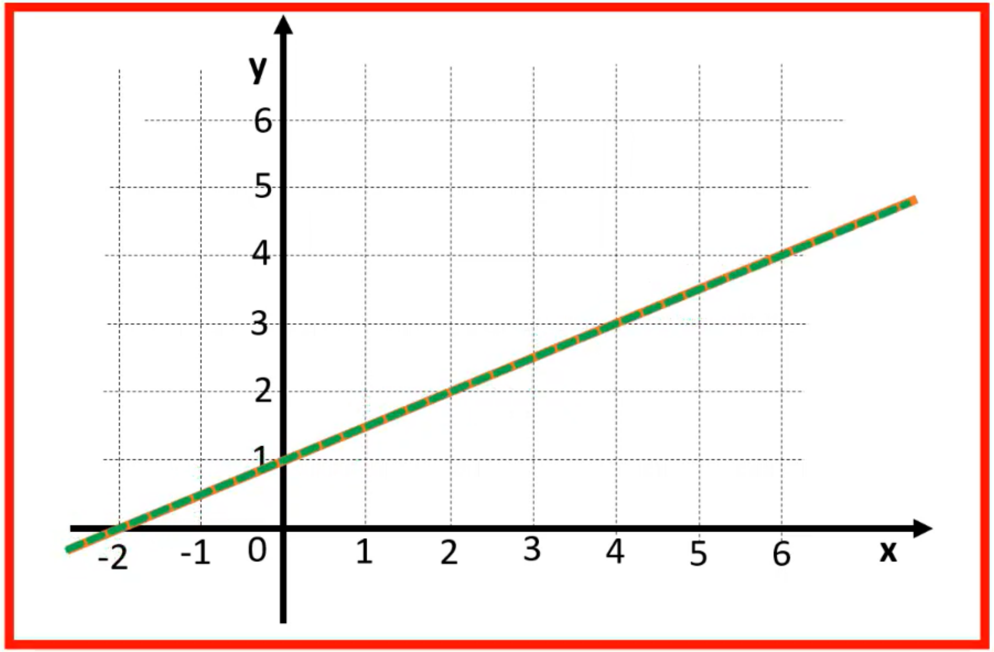


$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 1$$
$$-2x + 4y - 4 = 0$$
$$-2x + 4y - 4 = -1$$

the green lines would move away from the original line, which stays in the same position.

The equation of a straight line



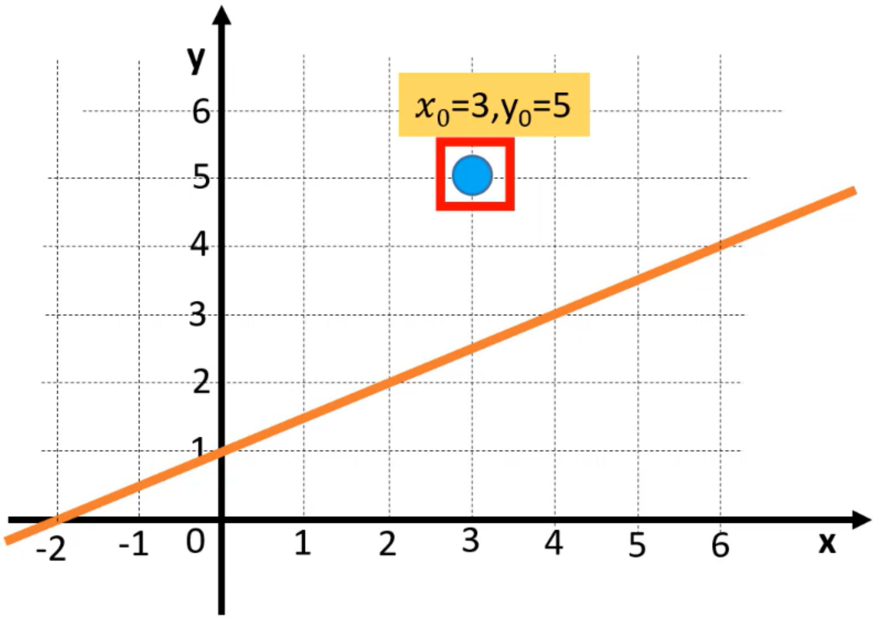
$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 1$$
$$-2x + 4y - 4 = 0$$
$$-2x + 4y - 4 = -1$$

If we multiply with the factor greater than 1

the green lines would move closer to the original line. This method is used on SVM to increase or decrease the so-called margin that we will discuss later on.

Distance between a point and a line

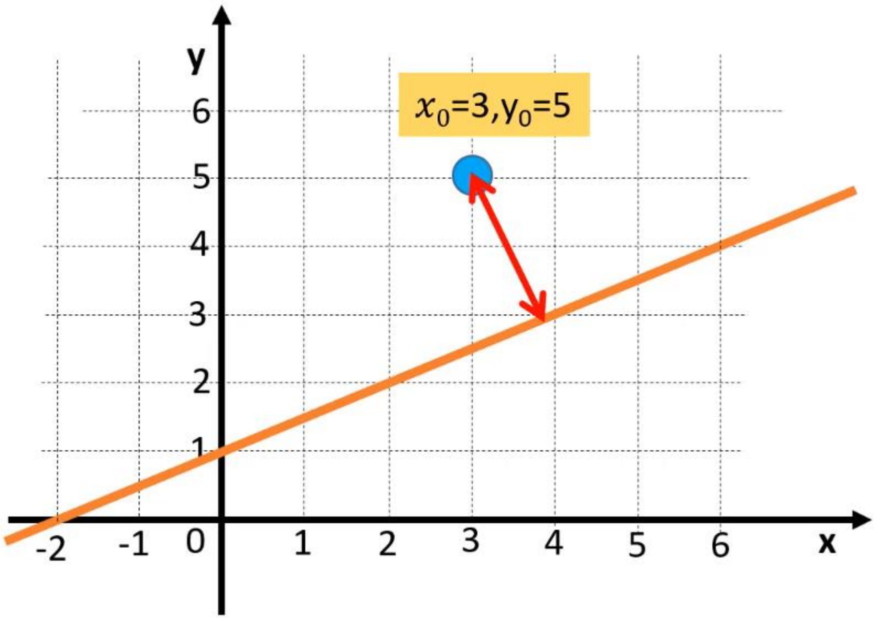


$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

Another thing we need to know before we look into the details about SVM is how to calculate the distance between a data point and a line.

Distance between a point and a line



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

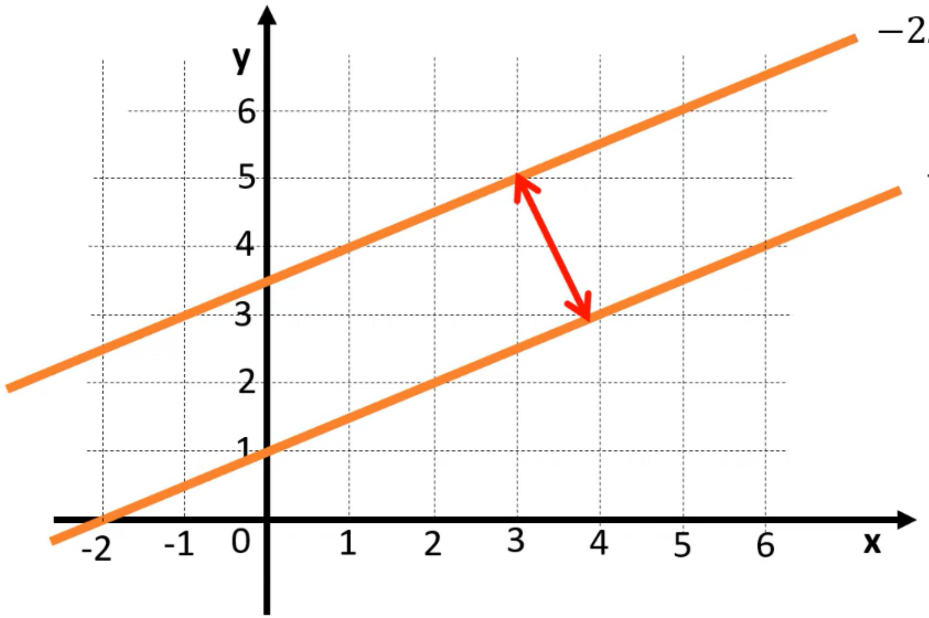
$$d = \frac{|Ax_0 + By_o + C|}{\sqrt{A^2 + B^2}}$$

$$d = \frac{|(-2)\cdot 3 + 4 \cdot 5 + (-4)|}{\sqrt{(-2)^2 + 4^2}} = \frac{10}{\sqrt{20}} = 2.236$$

If we plug in the numbers from the equation of the line,

Distance between two parallel lines

$$Ax + By + C = 0$$



$$-2x + 4y - 14 = 0$$

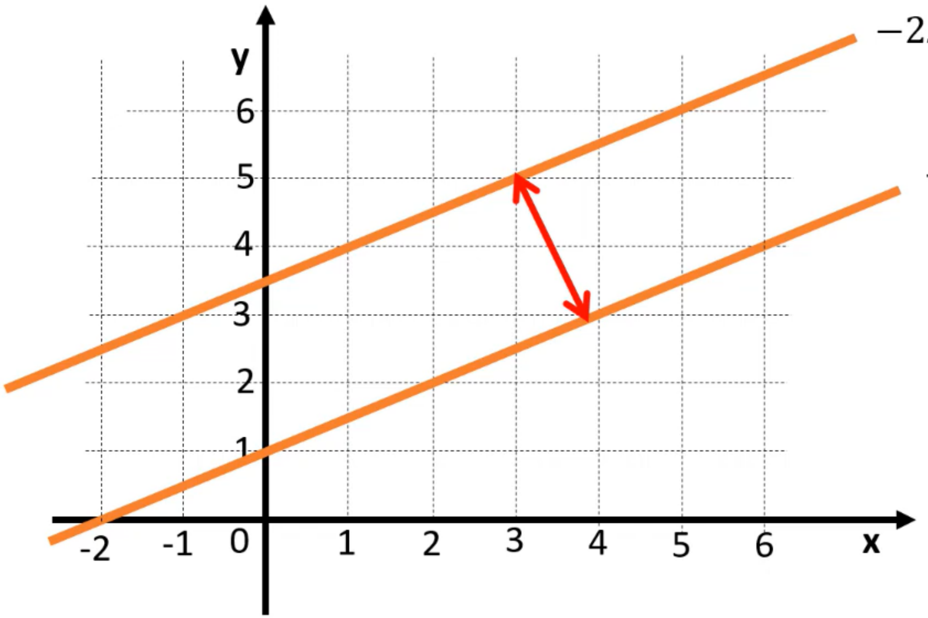
$$-2x + 4y - 4 = 0$$

$$\boxed{d = \frac{|C_2 - C_1|}{\sqrt{A^2 + B^2}}}$$

Similarly, the distance between two parallel lines can be calculated by the following formula.

Distance between two parallel lines

$$Ax + By + C = 0$$



$$-2x + 4y - 14 = 0$$

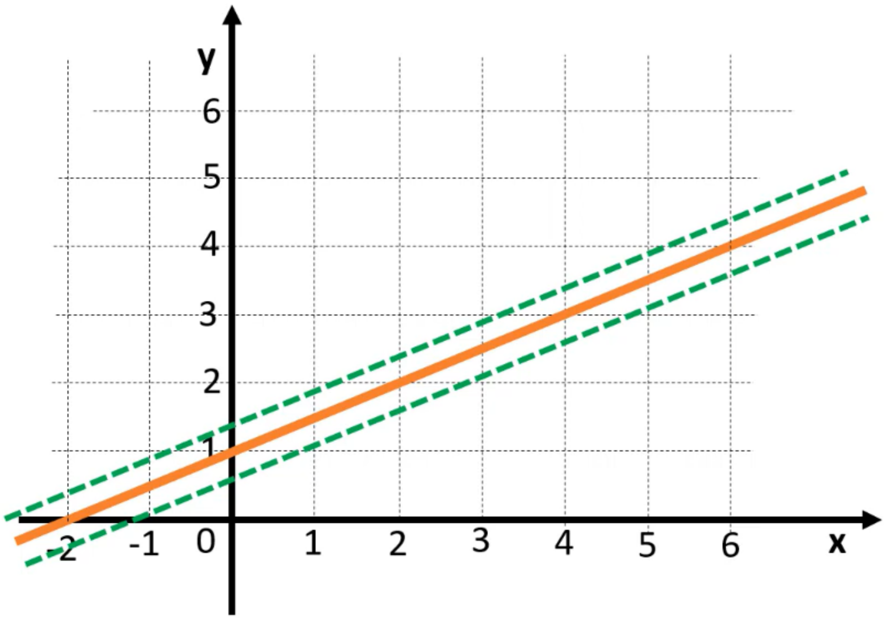
$$-2x + 4y - 4 = 0$$

$$d = \frac{|C_2 - C_1|}{\sqrt{A^2 + B^2}}$$

$$d = \frac{|(-4) - (-14)|}{\boxed{\sqrt{(-2)^2 + 4^2}}} = \frac{10}{\sqrt{20}} = 2.236$$

If we plug in the coefficients of the equations of the two parallel lines,

Distance between two parallel lines

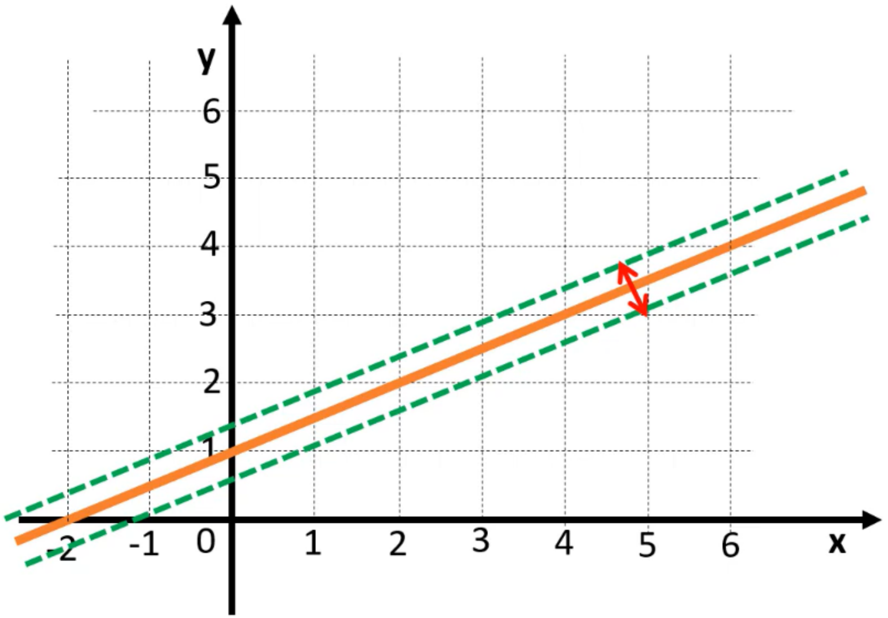


$$Ax + By + C = 0$$

$$-2x + 4y - 4 \boxed{= 1}$$
$$-2x + 4y - 4 \boxed{= 0}$$
$$-2x + 4y - 4 \boxed{= -1}$$

Remember that we previously added a one and a negative on the right-hand side to obtain the two green lines.

Distance between two parallel lines



$$Ax + By + C = 0$$

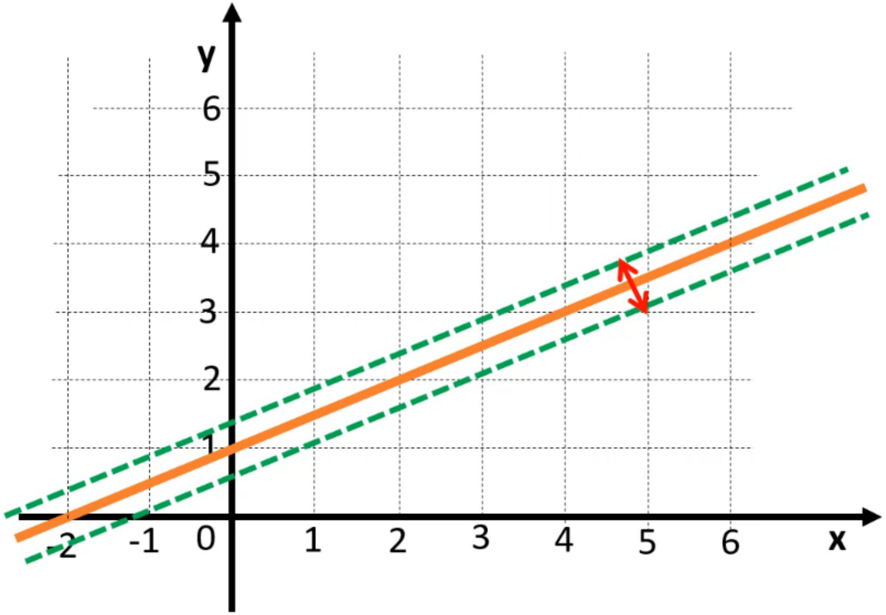
$$-2x + 4y - 4 = 1$$
$$-2x + 4y - 4 = 0$$
$$-2x + 4y - 4 = -1$$

$$d = \frac{|C_2 - C_1|}{\sqrt{A^2 + B^2}}$$

$$\boxed{d = \frac{|(-3) - (-5)|}{\sqrt{(-2)^2 + 4^2}} = \frac{2}{\sqrt{20}} = 0.447}$$

The distance between these two green lines is calculated like this.

Distance between two parallel lines



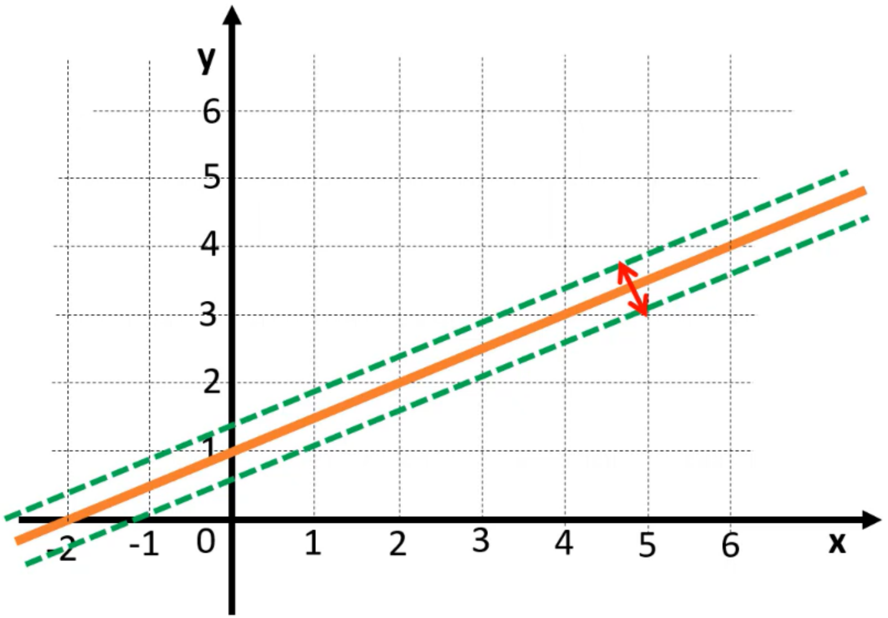
$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 1$$
$$-2x + 4y - 4 = 0$$
$$-2x + 4y - 4 = -1$$

$$\boxed{d = \frac{2}{\sqrt{A^2 + B^2}}}$$

For such lines, we can simplify the equation to this,

Distance between two parallel lines



$$Ax + By + C = 0$$

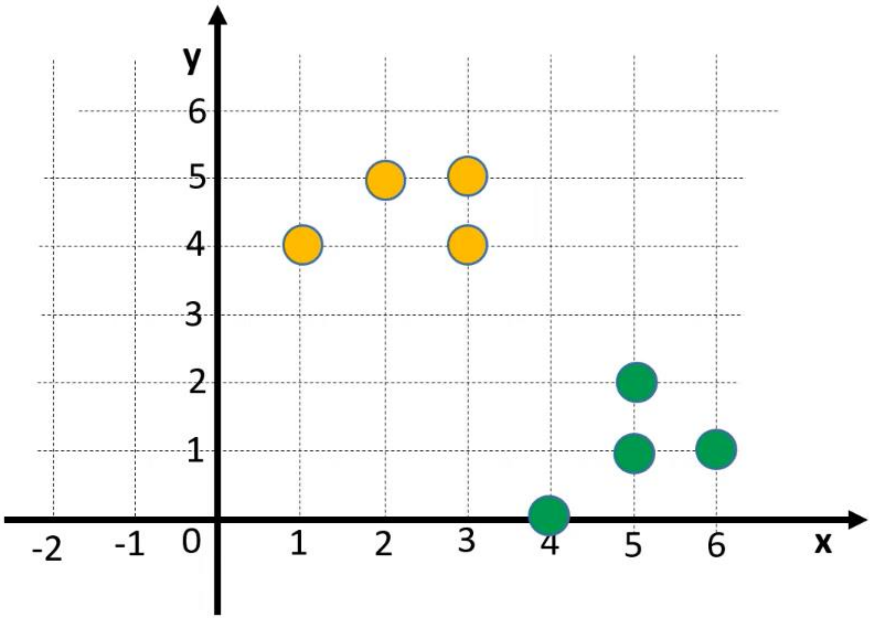
$$-2x + 4y - 4 = 1$$
$$-2x + 4y - 4 = 0$$
$$-2x + 4y - 4 = -1$$

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \boxed{\frac{2}{\|w\|}}$$

or like this, which you will see later on.

How to Predict

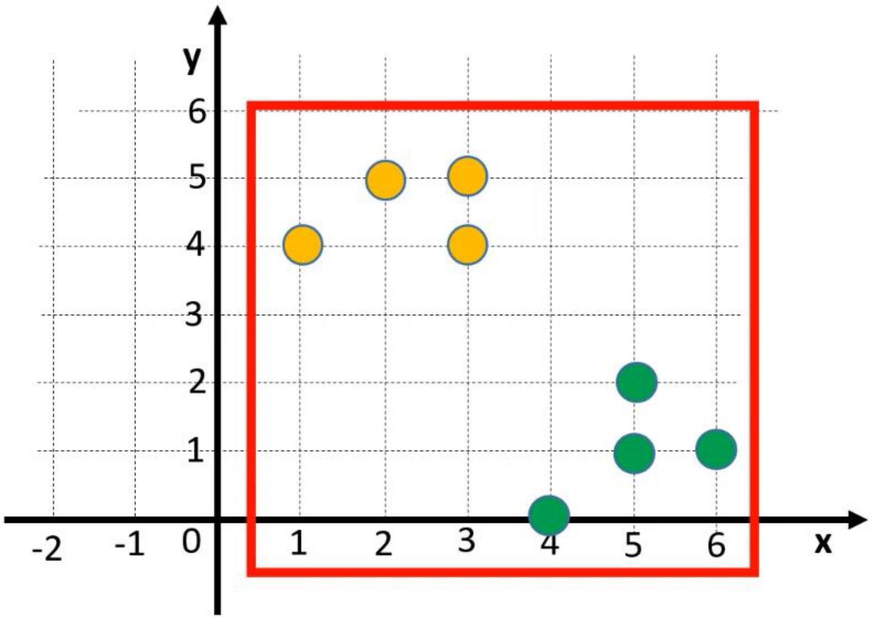
SVM



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

We will now have a look at a real data set, where we will use a SVM to generate a line that separates the two groups as good as possible.

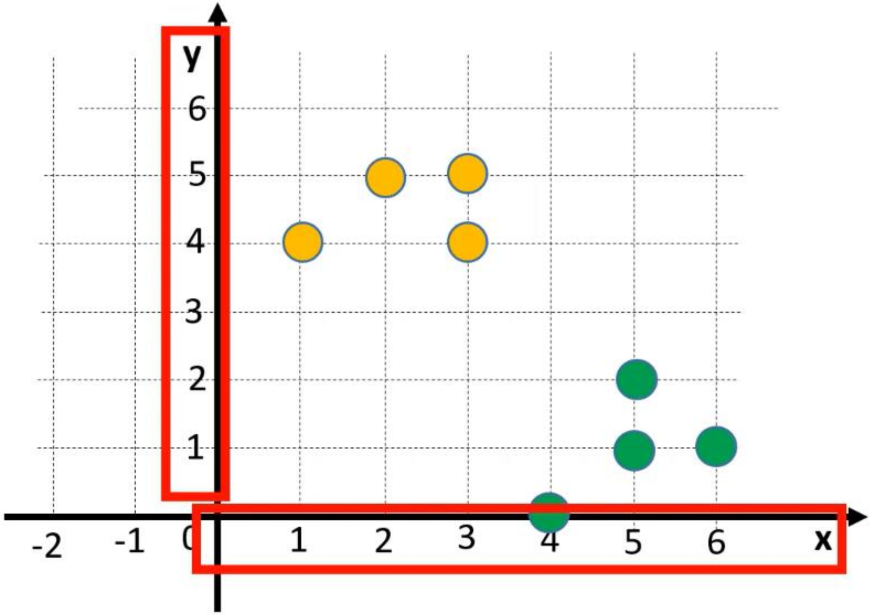
SVM



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Since we know which group the data points belong to, we call this our training data because we will let the SVM train on this data. The training means that the SVM finds a line that separate the two groups in an optimal way.

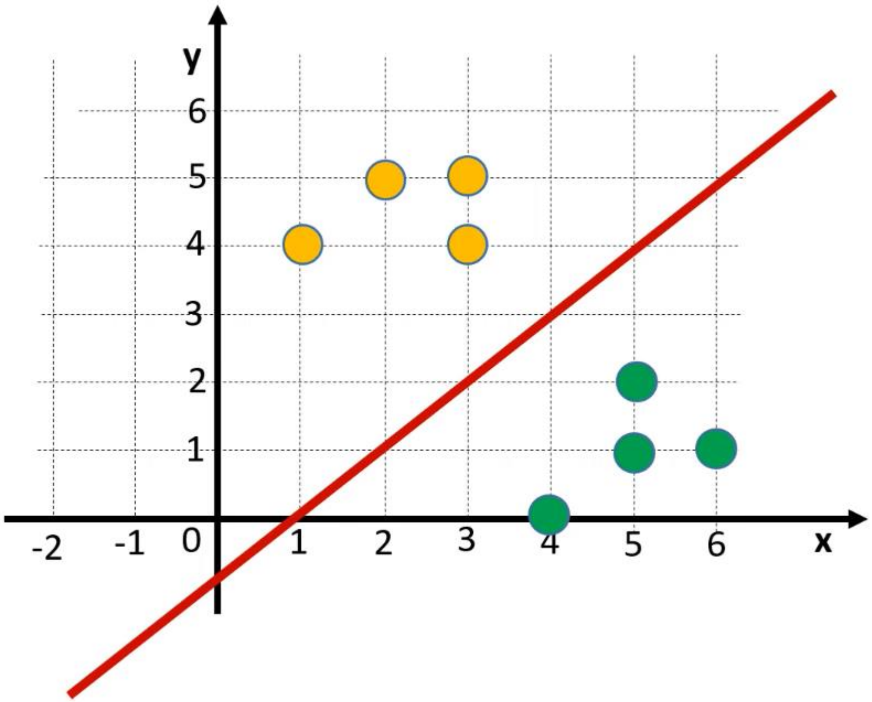
SVM



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

x and y are two variables, that could represent two measurements on the eight individuals, for example, blood pressure and cholesterol level.

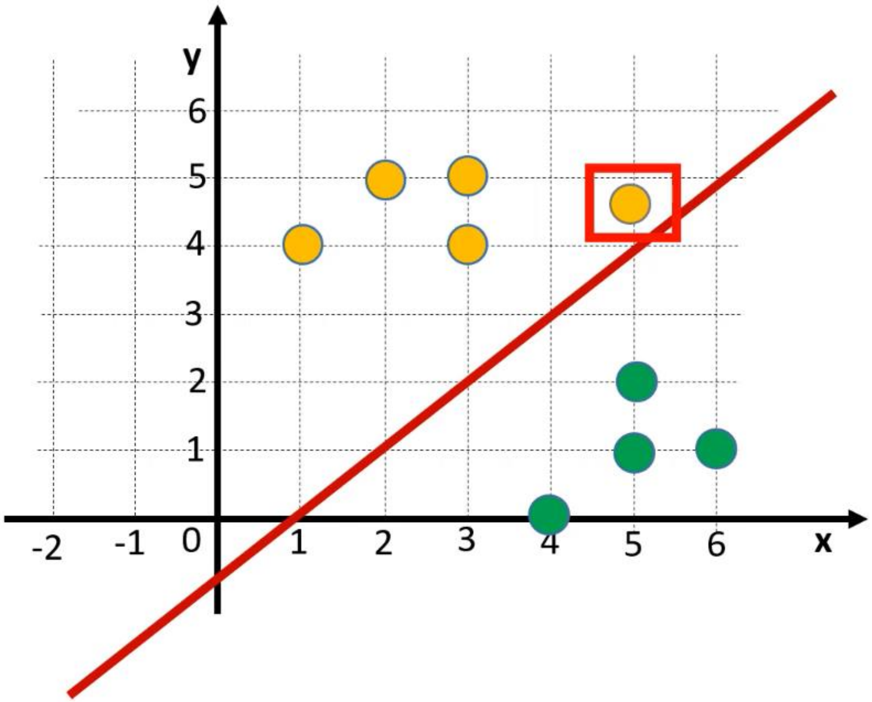
SVM



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

The following line is the best line to separate the two groups based on the training data.

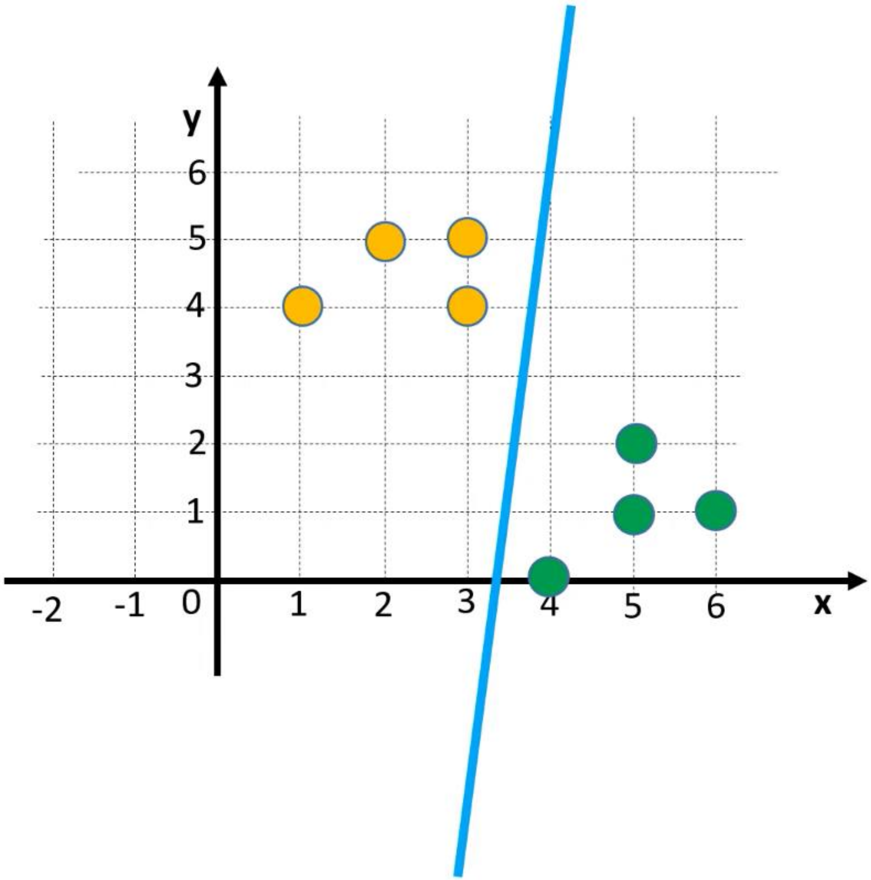
Predict



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Based on the measured values of x and y, the blood pressure and cholesterol level, the SVM can help us to predict if someone has the disease or not.

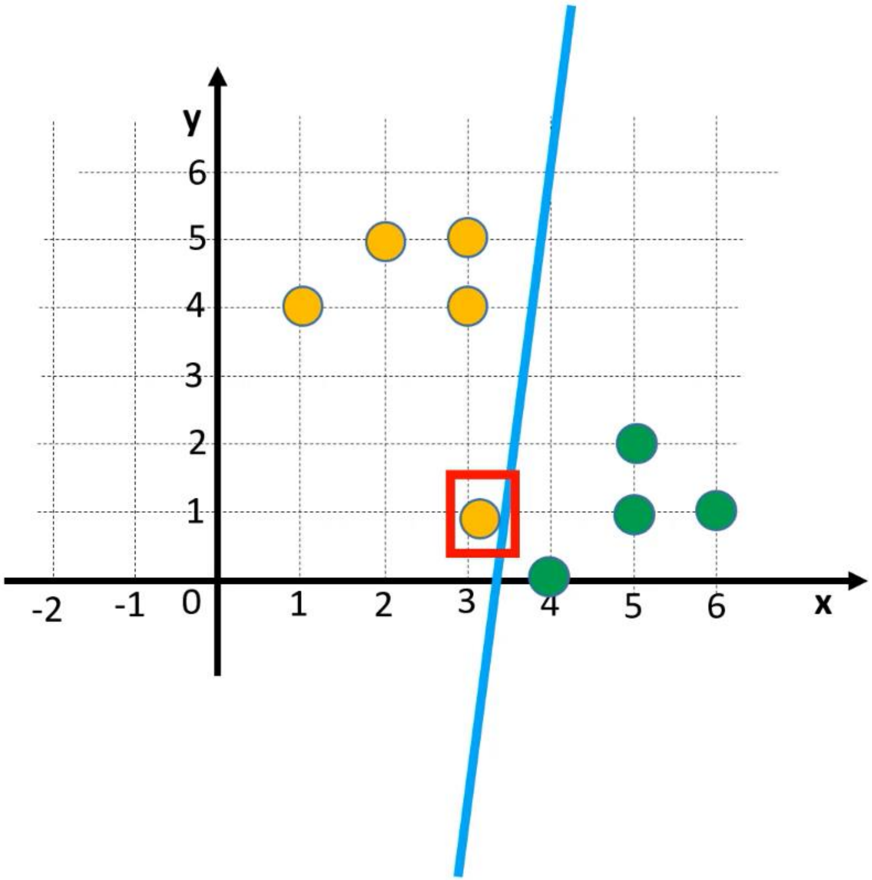
Predict



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Although this blue line can also separate the two groups completely, it will not be good for predicting new data,

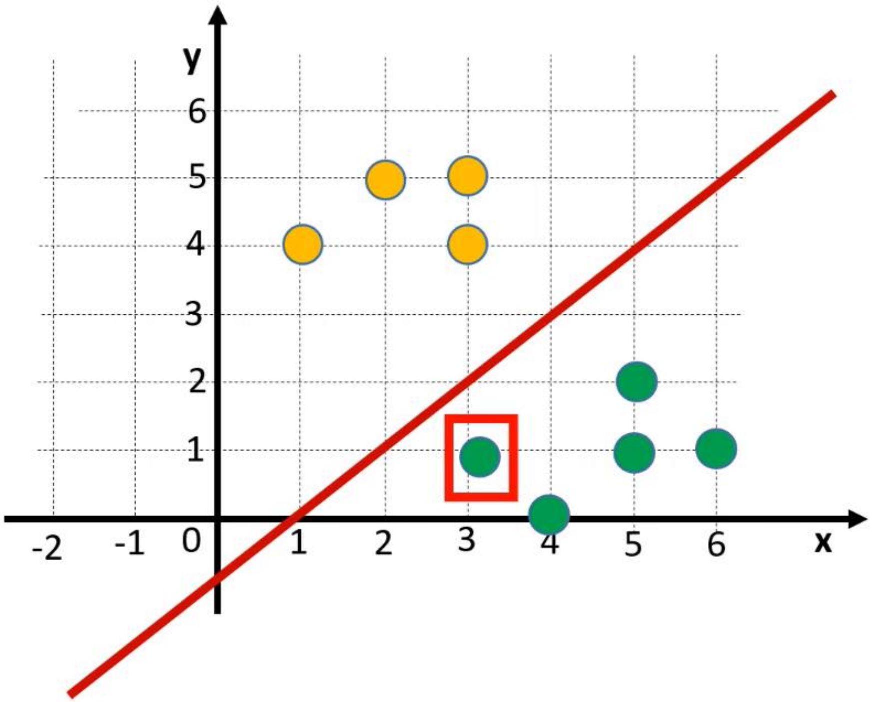
Predict



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

we would predict that it belongs to group A, which means that the person is predicted to have the disease. This does not make sense because the data point is much closer to the green points, which represent the healthy individuals.

Predict

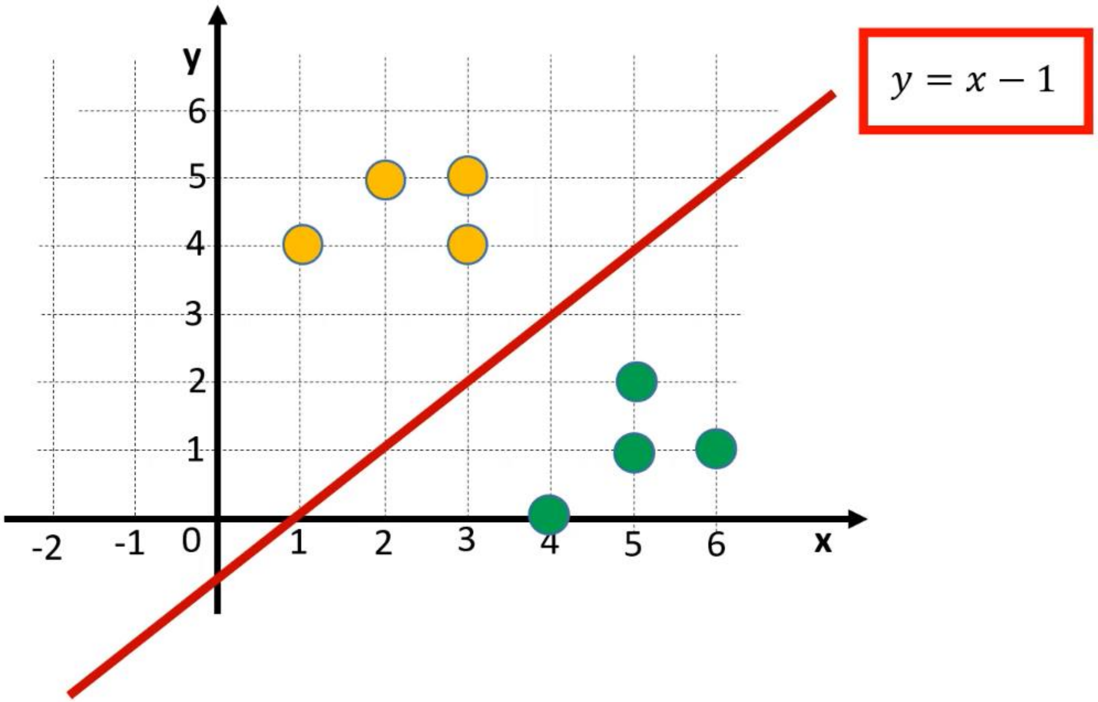


Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

we would classify the data point as belonging to group B, the healthy group, which seems like a better prediction because the point is closer to the green data points.

SVM Math

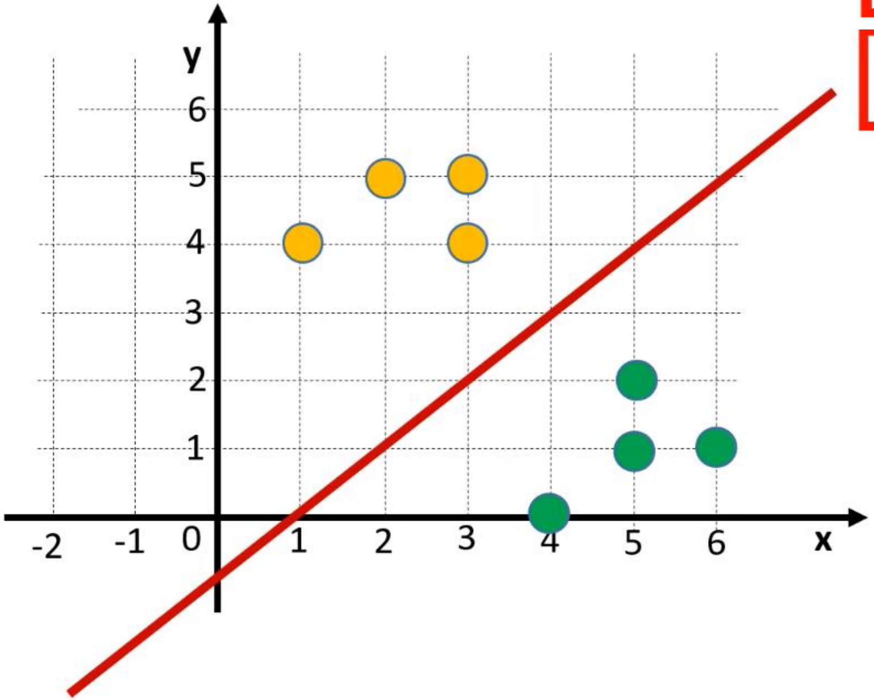
SVM – the math



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

This is the equation of the red line, which has an intercept of negative one and a slope of one.

SVM – the math



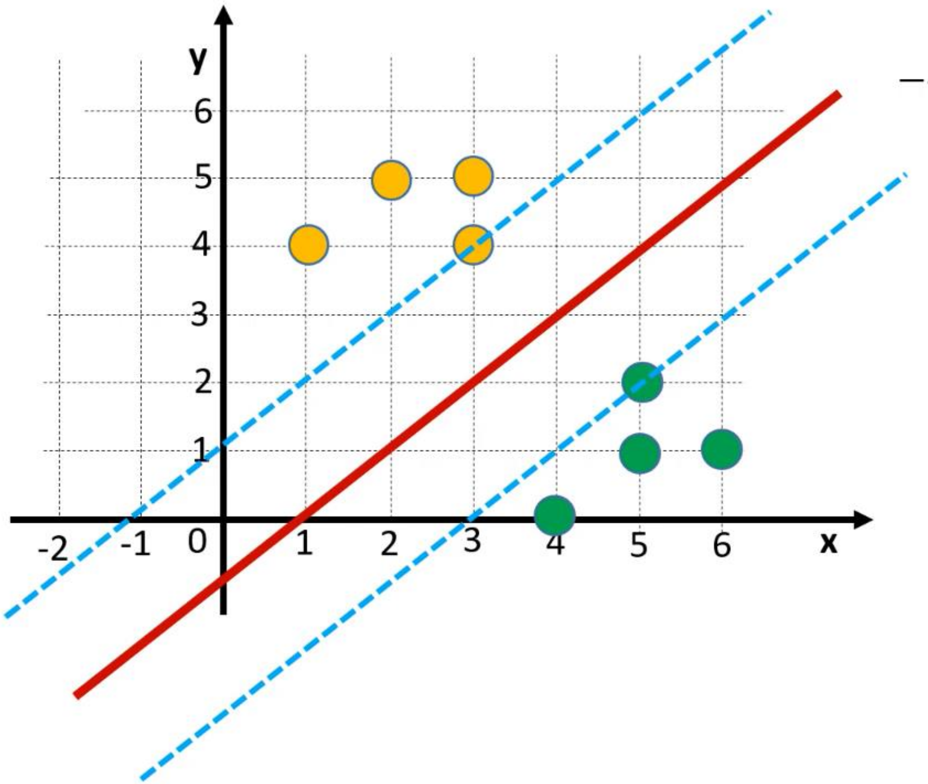
$$-x + y + 1 = 0$$

$$-4x + 4y + 4 = 0$$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

We can, for example, multiply the terms by 4 because that does not change the position of the line.

SVM – the math

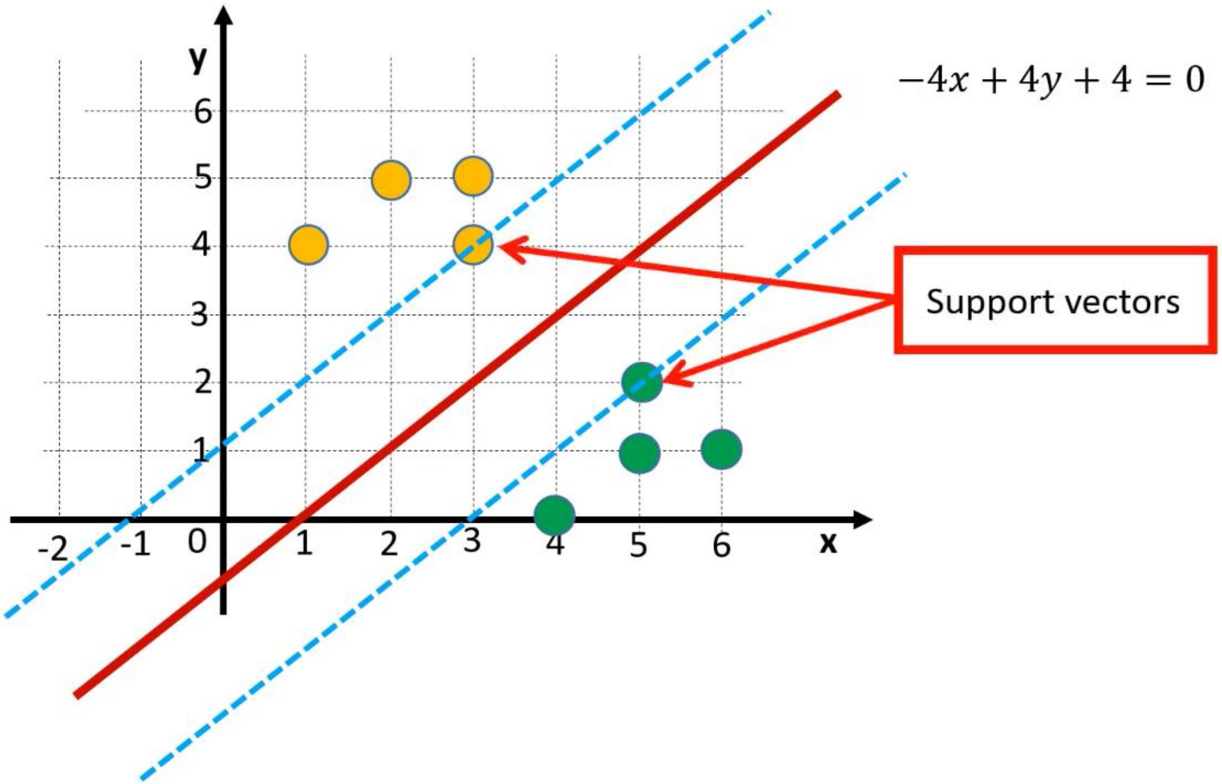


$$-4x + 4y + 4 = 0$$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

We will now draw two parallel lines that intercept the two closest data points to the hyperplane.

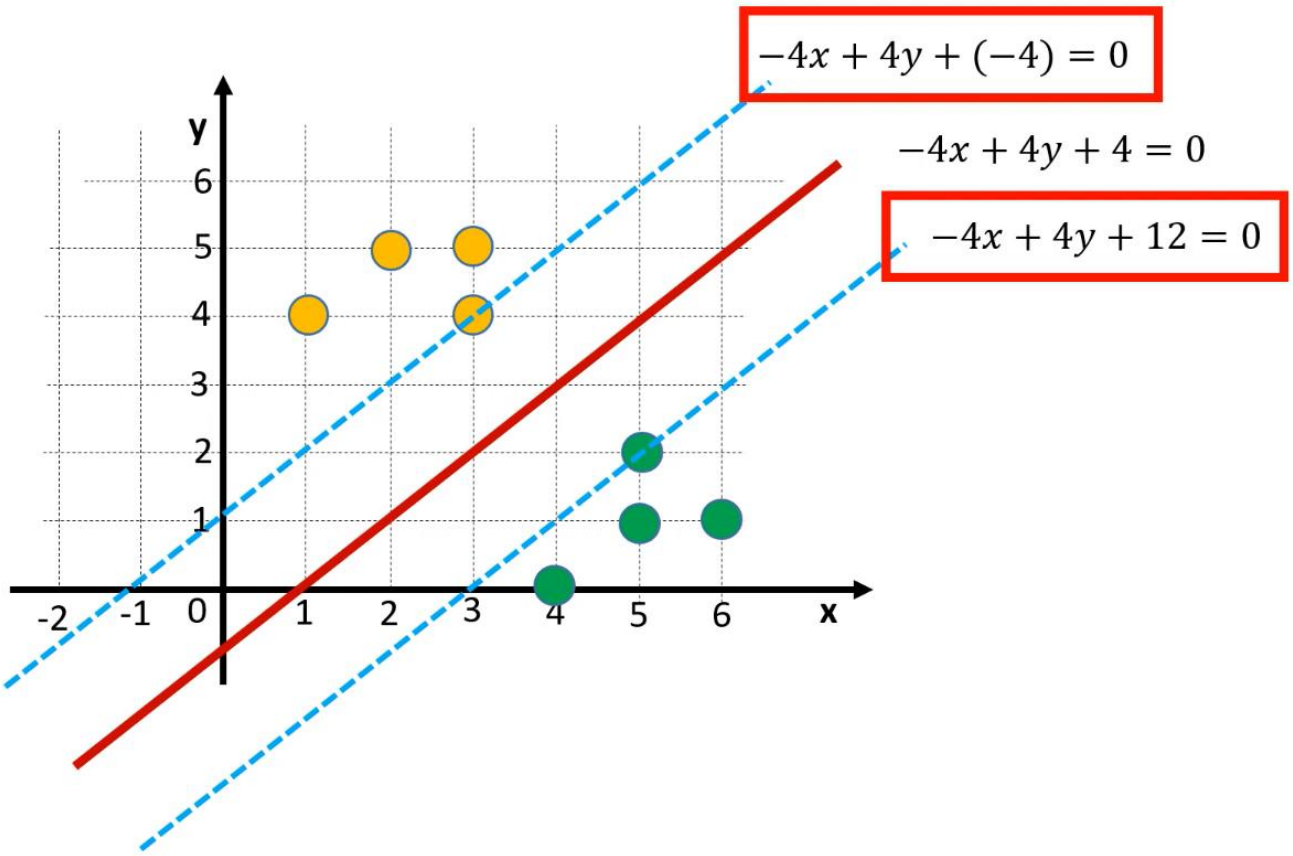
SVM – the math



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Such data points are called support vectors.

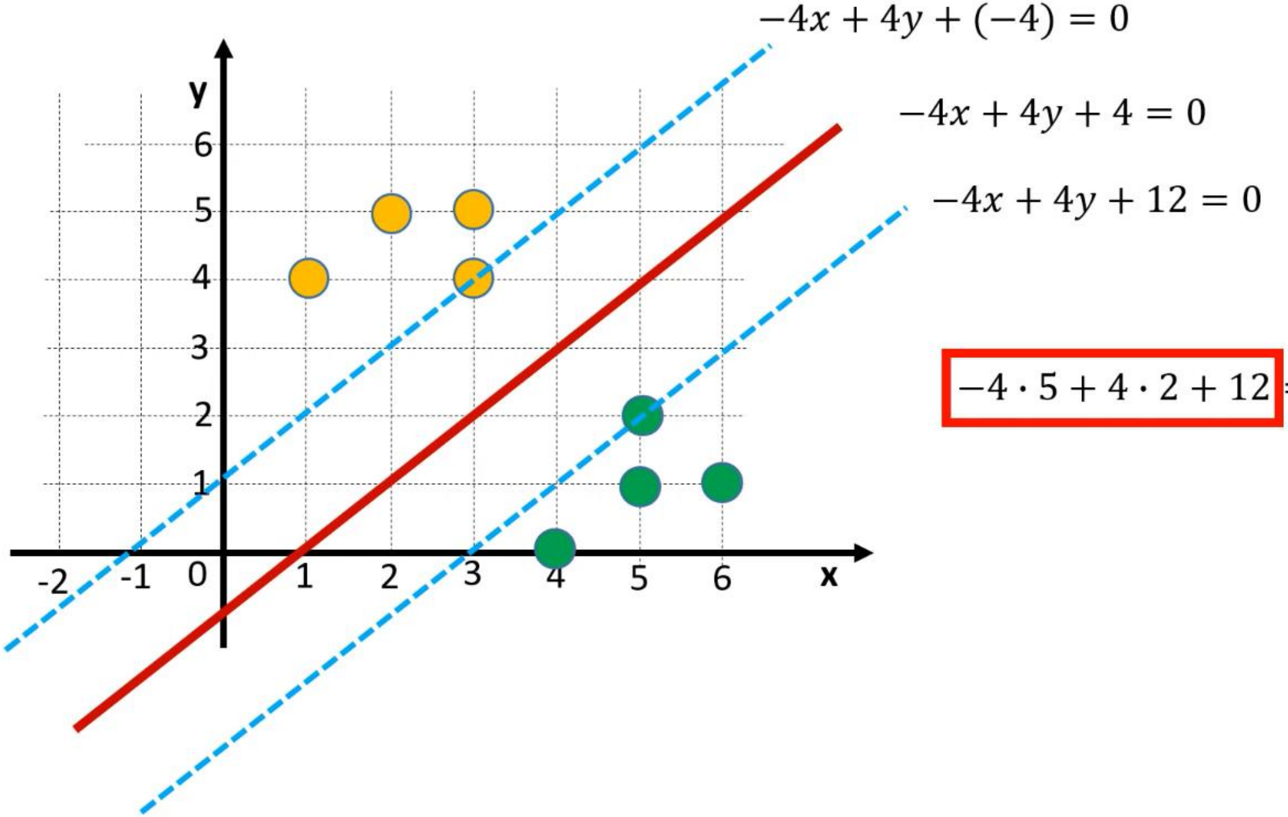
SVM – the math



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

The equations of these two blue lines look like this.

SVM – the math

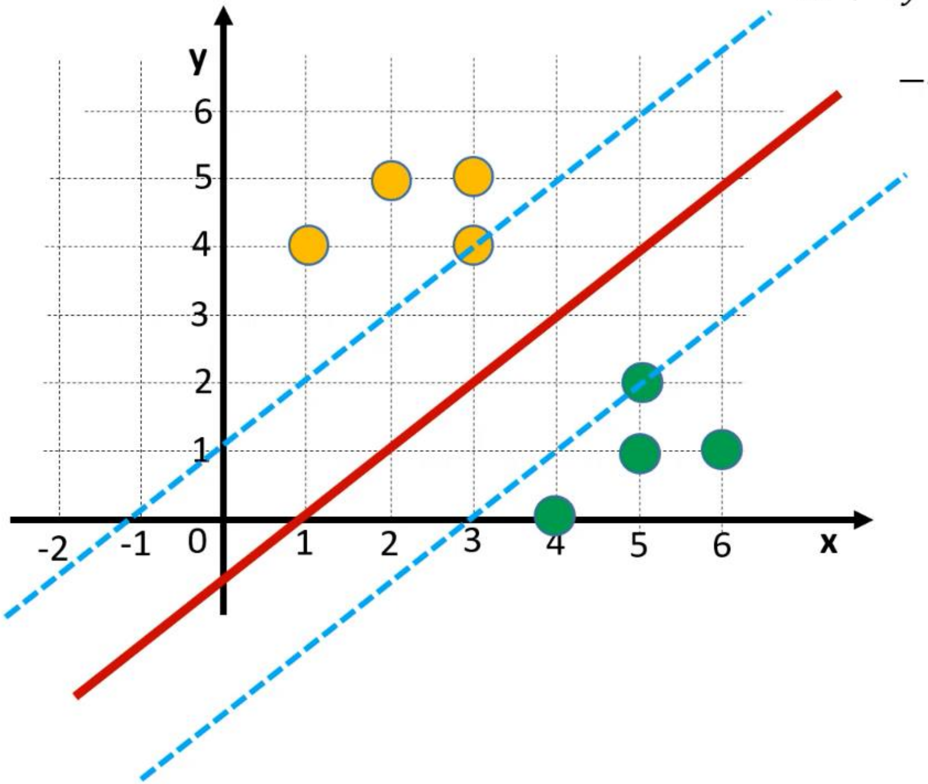


Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$\boxed{-4 \cdot 5 + 4 \cdot 2 + 12} = 0$$

the left-hand side should be equal to zero, which is true in this case.

SVM – the math



$$-4x + 4y + \boxed{(-4)} = 0$$

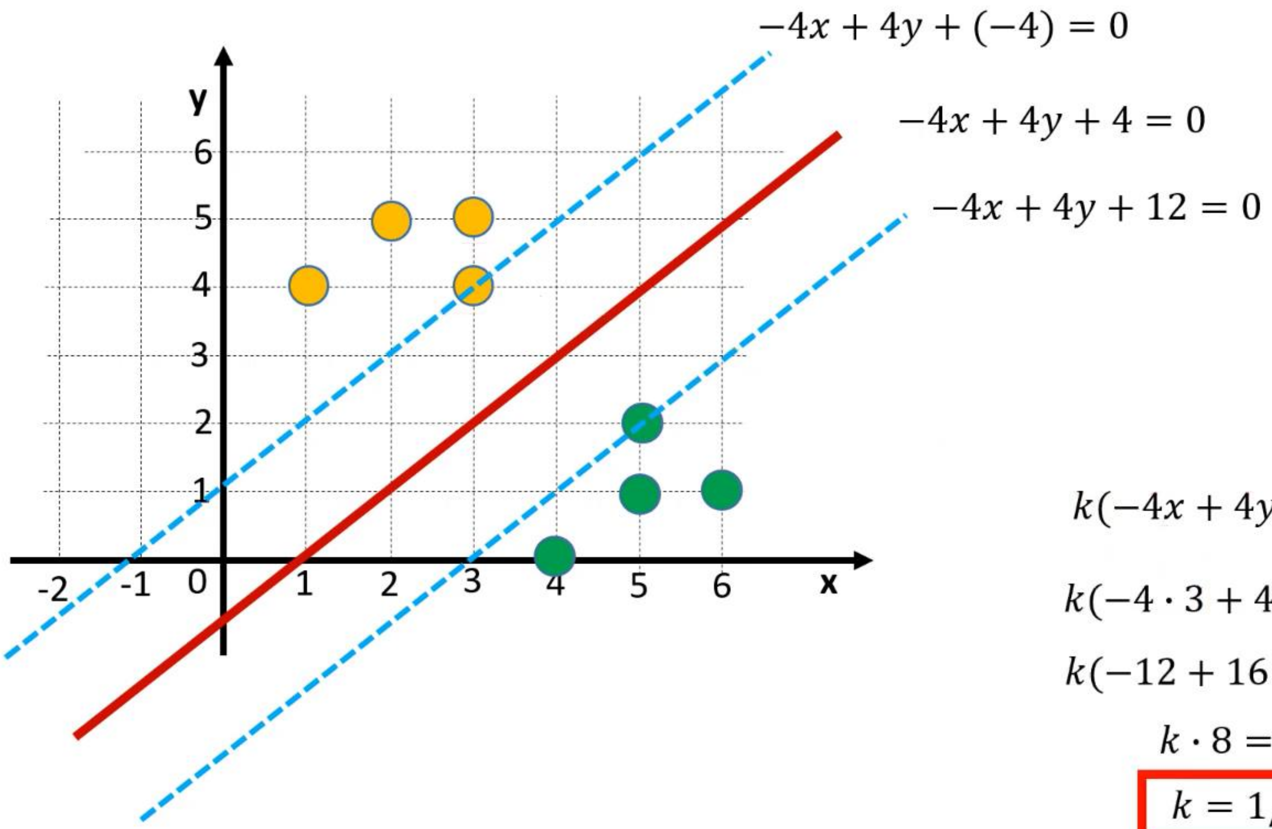
$$-4x + 4y \boxed{+ 4} = 0$$

$$-4x + 4y \boxed{+ 12} = 0$$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

given that all three equations have the same constant term on the left-hand side as the equation of the hyperplane?

SVM – the math



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$k(-4x + 4y + 4) = 1$$

$$k(-4 \cdot 3 + 4 \cdot 4 + 4) = 1$$

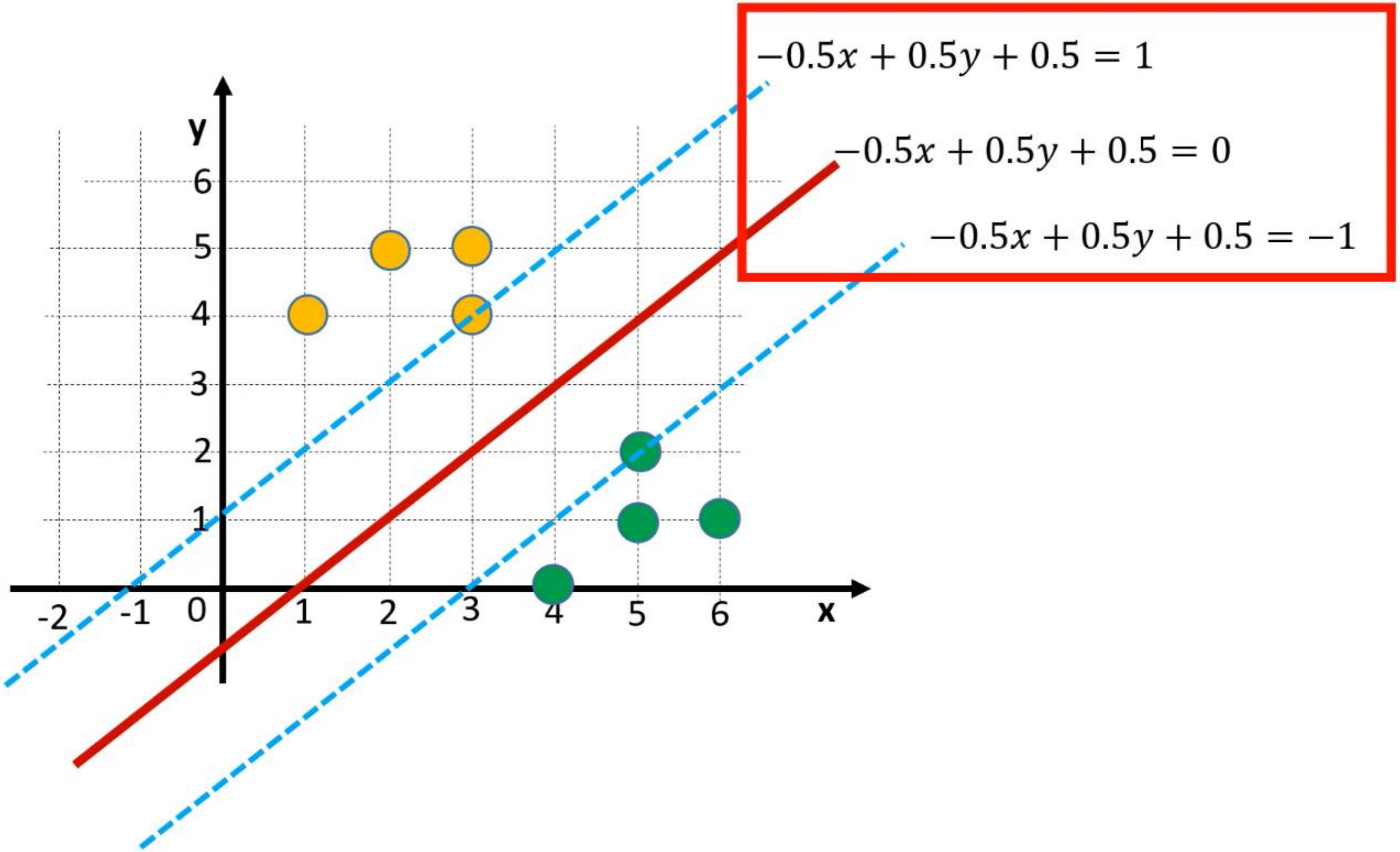
$$k(-12 + 16 + 4) = 1$$

$$k \cdot 8 = 1$$

$$\boxed{k = 1/8}$$

If we solve this equation, we see that we should multiply the terms on the left-hand side by one over eight, or simply divide the terms by eight.

SVM – the math



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

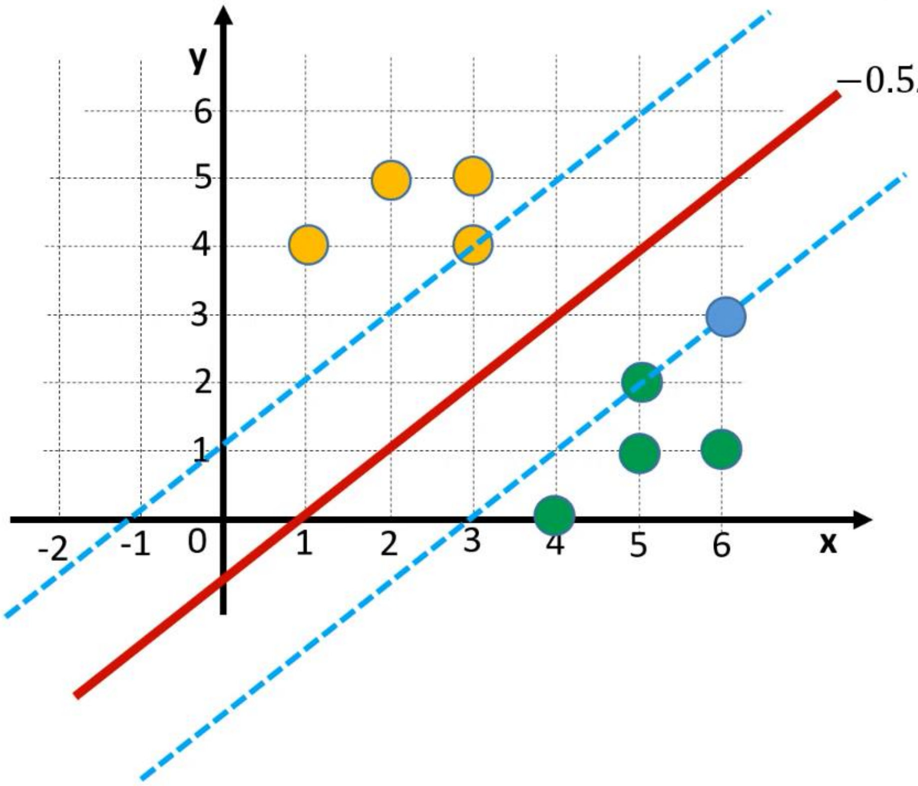
After dividing the terms by eight, the equations look like this, which is the standard form of the equations in SVM,

SVM – the math

$$-0.5x + 0.5y + 0.5 = 1$$

$$-0.5x + 0.5y + 0.5 = 0$$

$$-0.5x + 0.5y + 0.5 = -1$$



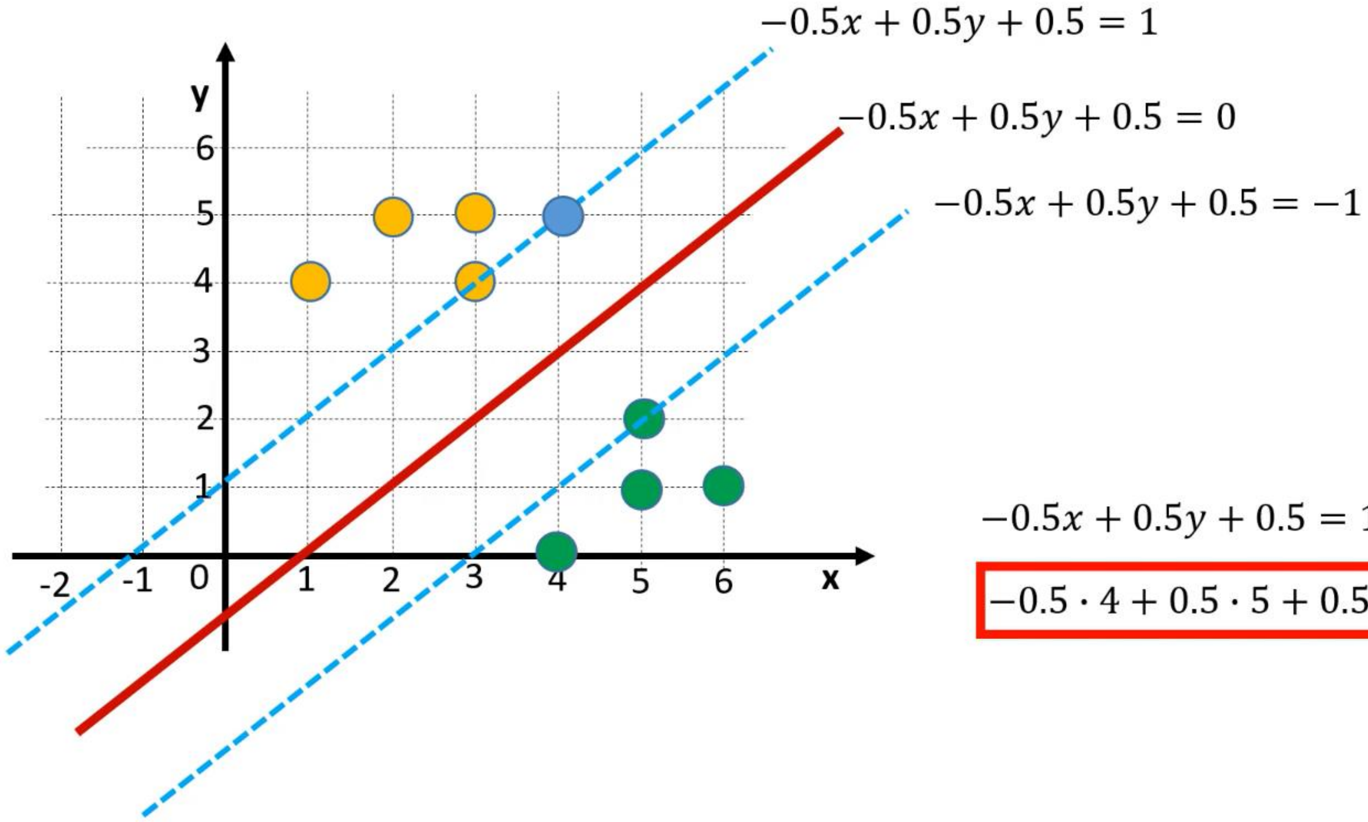
Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$-0.5x + 0.5y + 0.5 = -1$$

$$\boxed{-0.5 \cdot 6 + 0.5 \cdot 3 + 0.5} = -1$$

we see that the left-hand side is now equal to negative one.

SVM – the math



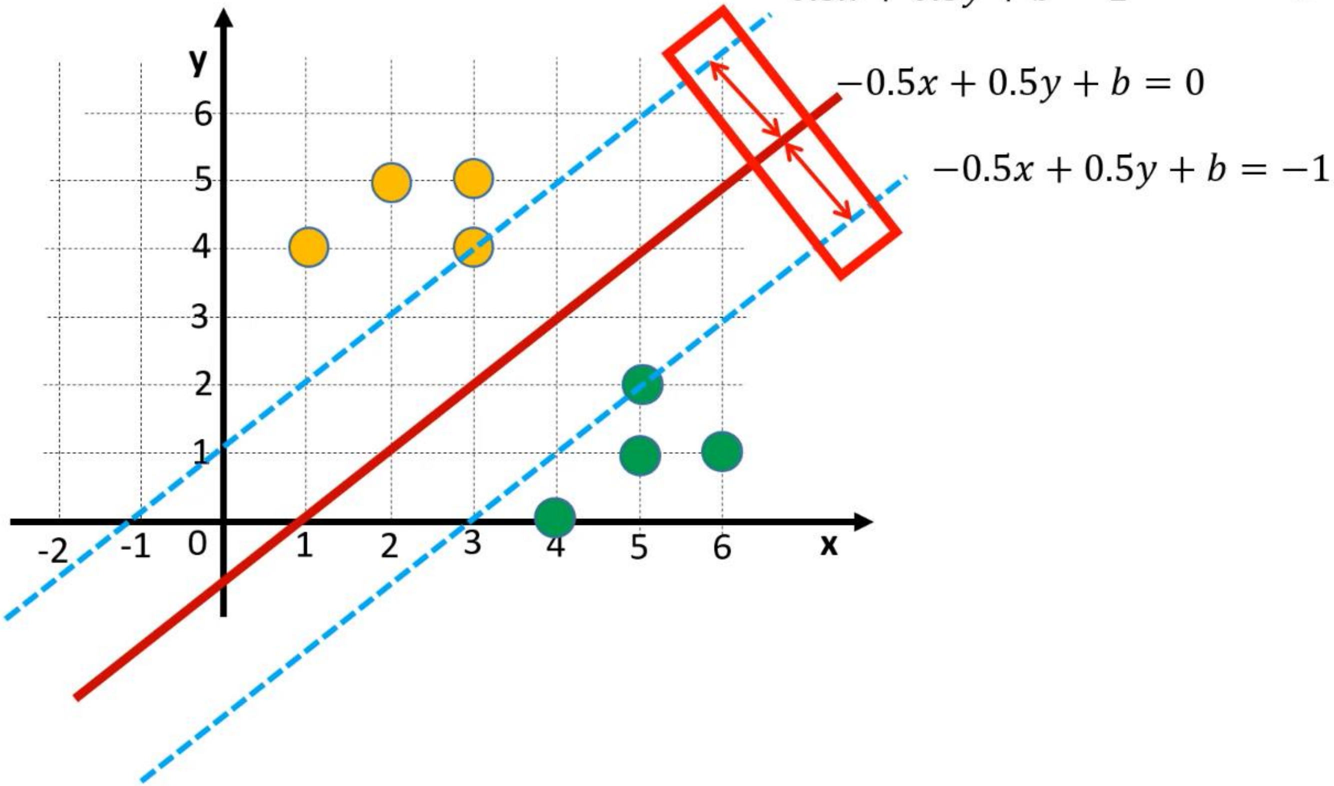
Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$-0.5x + 0.5y + 0.5 = 1$$

$$\boxed{-0.5 \cdot 4 + 0.5 \cdot 5 + 0.5} = 1$$

we see that the left-hand side is now equal to positive one.

SVM – the math

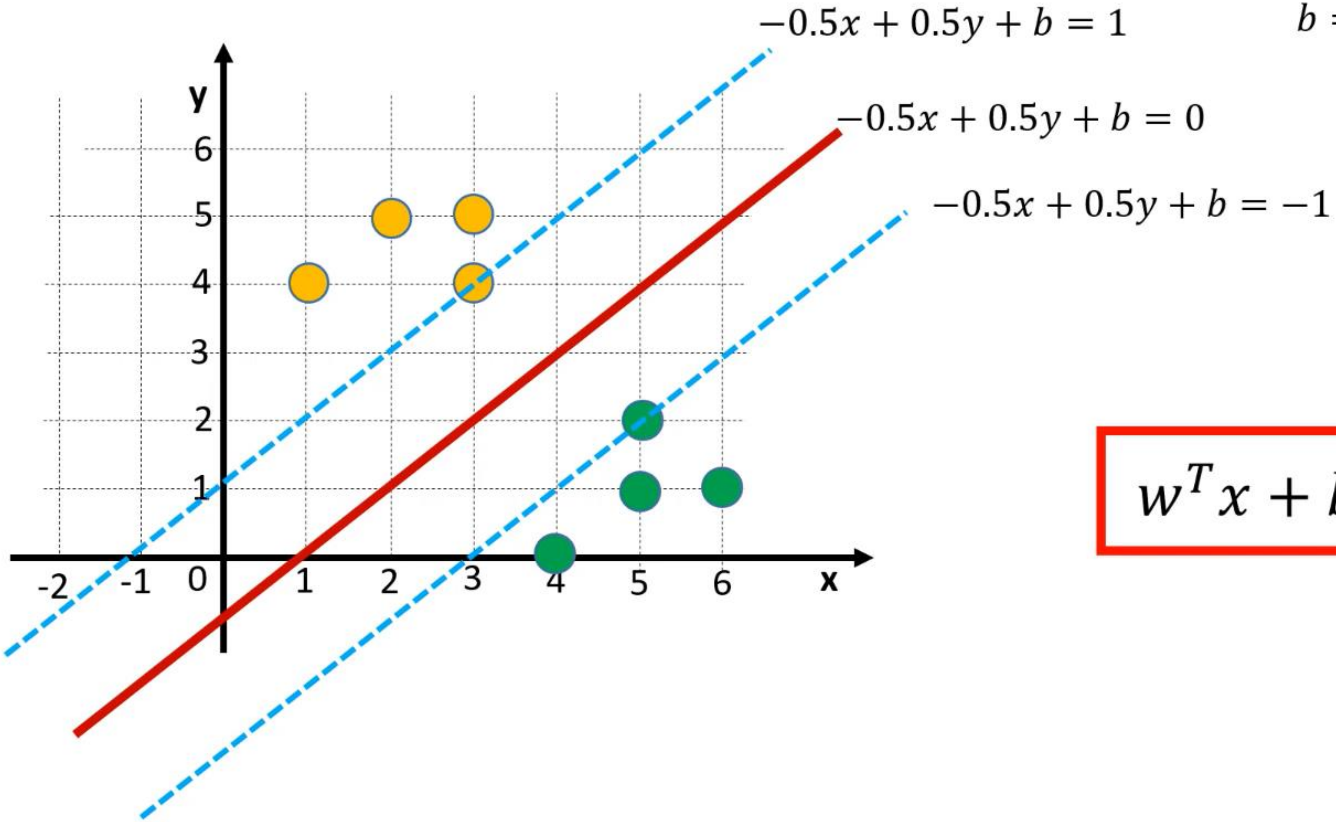


$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Note that the two distances between the red line, our hyperplane, and the blue lines should always be equal so that the red line is in the center.

SVM – the math



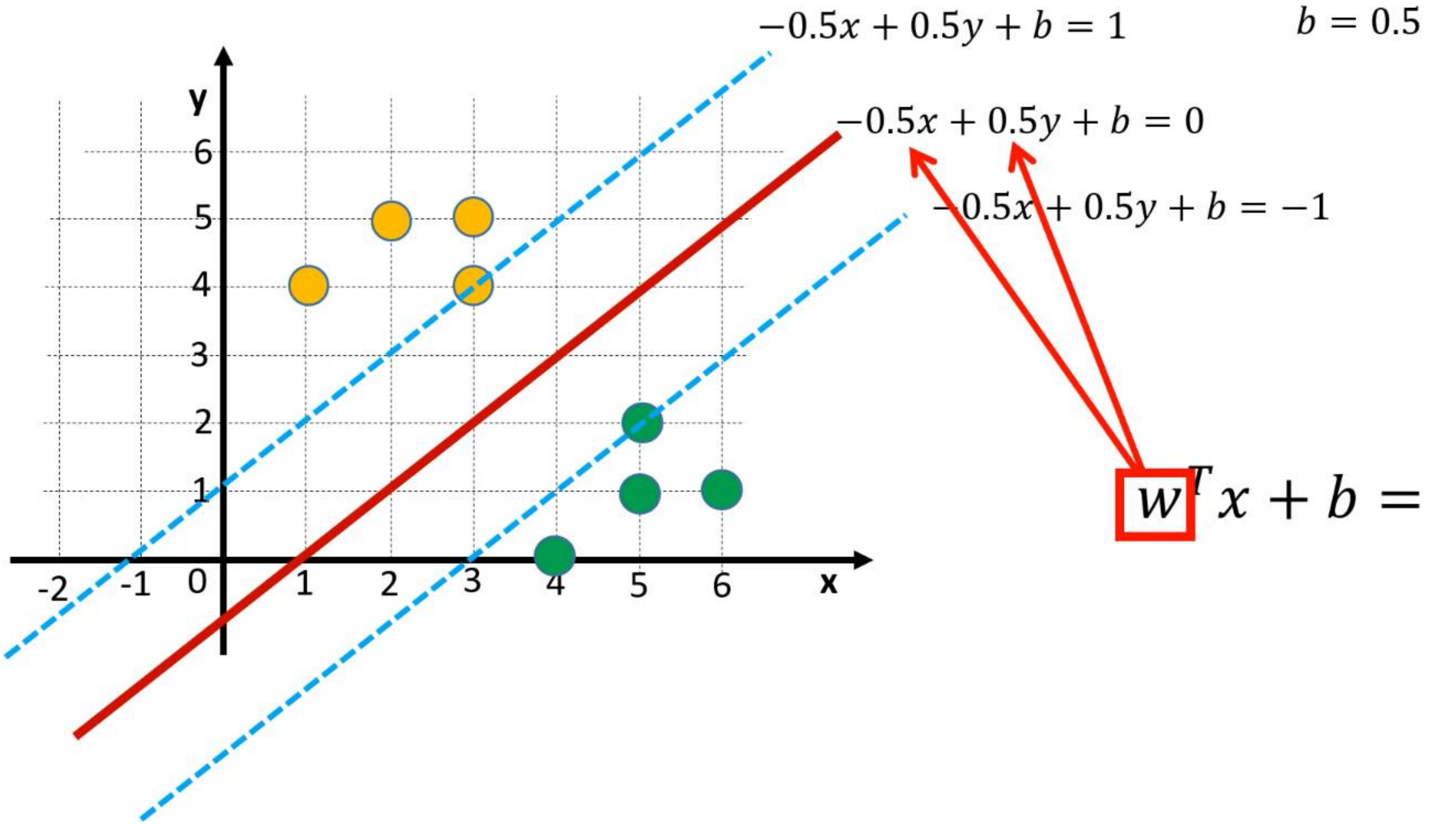
$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$\boxed{w^T x + b = 0}$$

The equation of the hyperplane is usually expressed like this in SVM.

SVM – the math

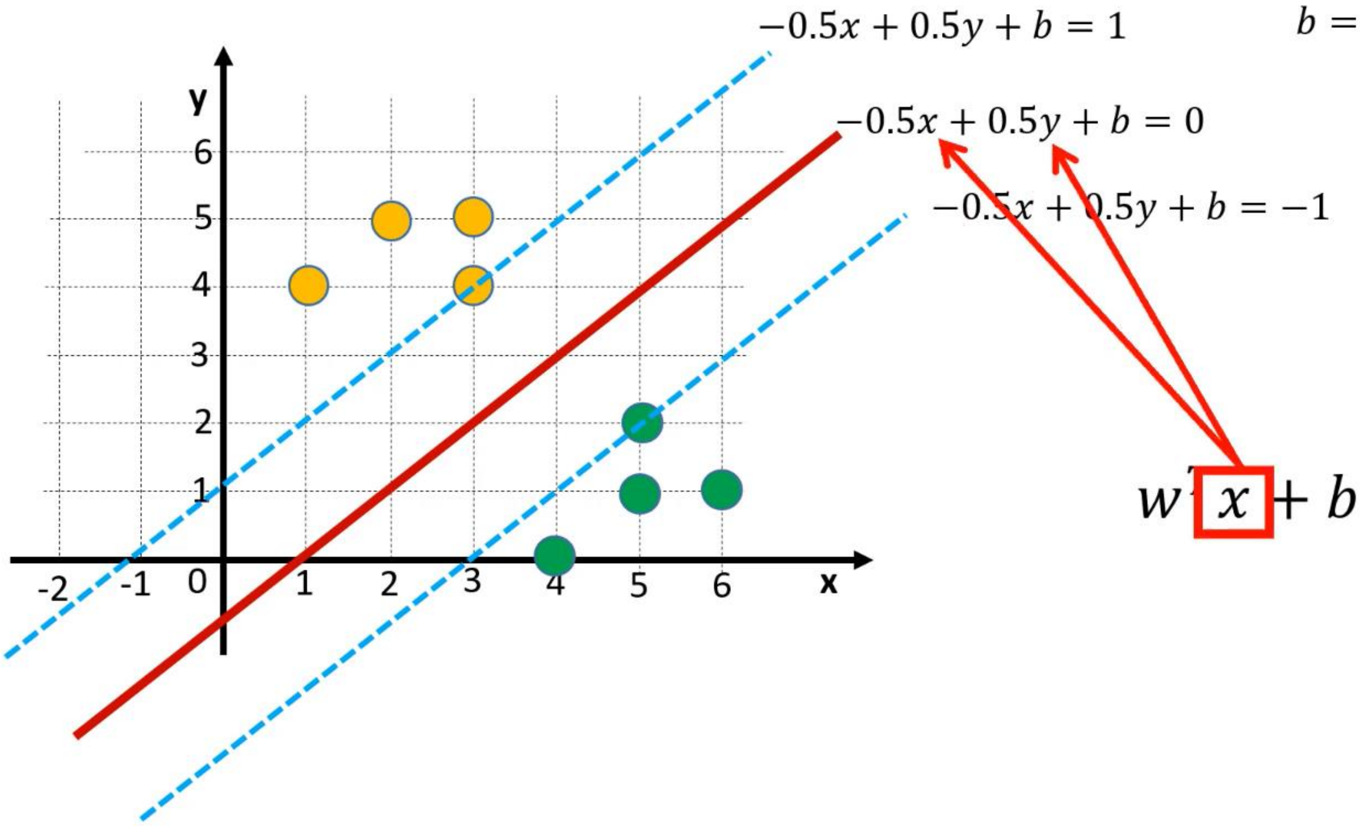


Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$\boxed{w}^T x + b = 0$$

This represents a vector that holds our coefficients,

SVM – the math



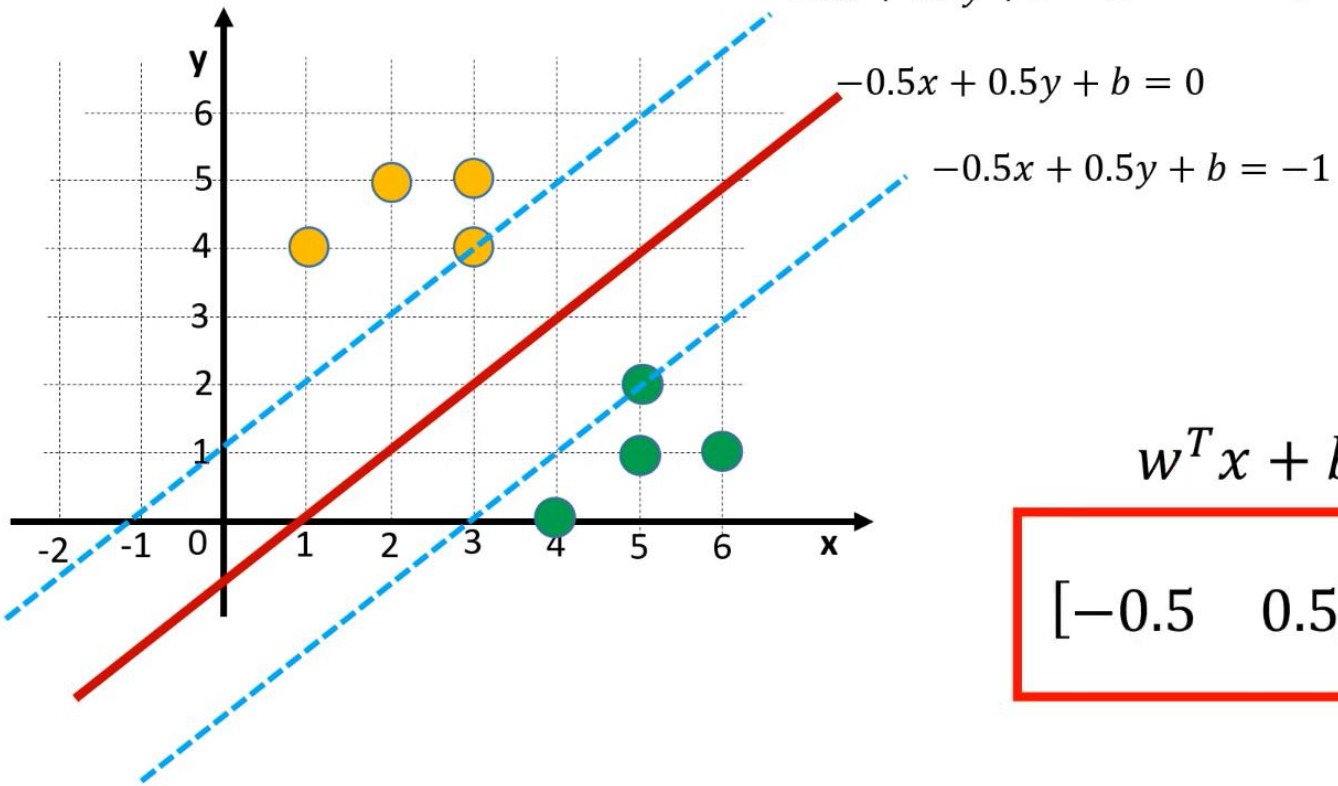
$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$w^T \boxed{x} + b = 0$$

and this is a vector that includes the variables x and y in our example.

SVM – the math



$b = 0.5$

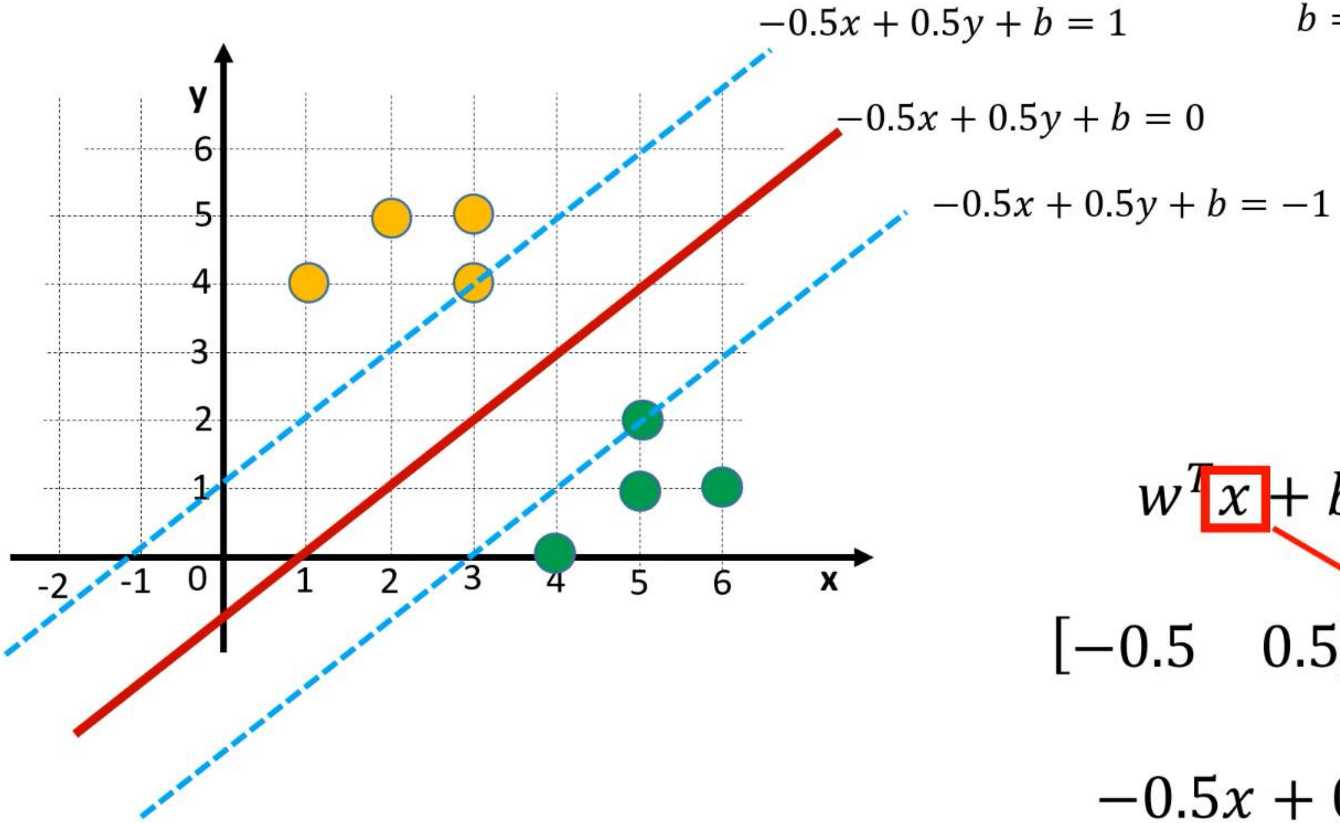
Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$w^T x + b = 0$$

$$\boxed{\begin{bmatrix} -0.5 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + b = 0}$$

If we plug in the coefficients and x and y in this equation,

SVM – the math



$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$w^T \boxed{x} + b = 0$$

$$\begin{bmatrix} -0.5 & 0.5 \end{bmatrix} \cdot \boxed{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} + b = 0$$

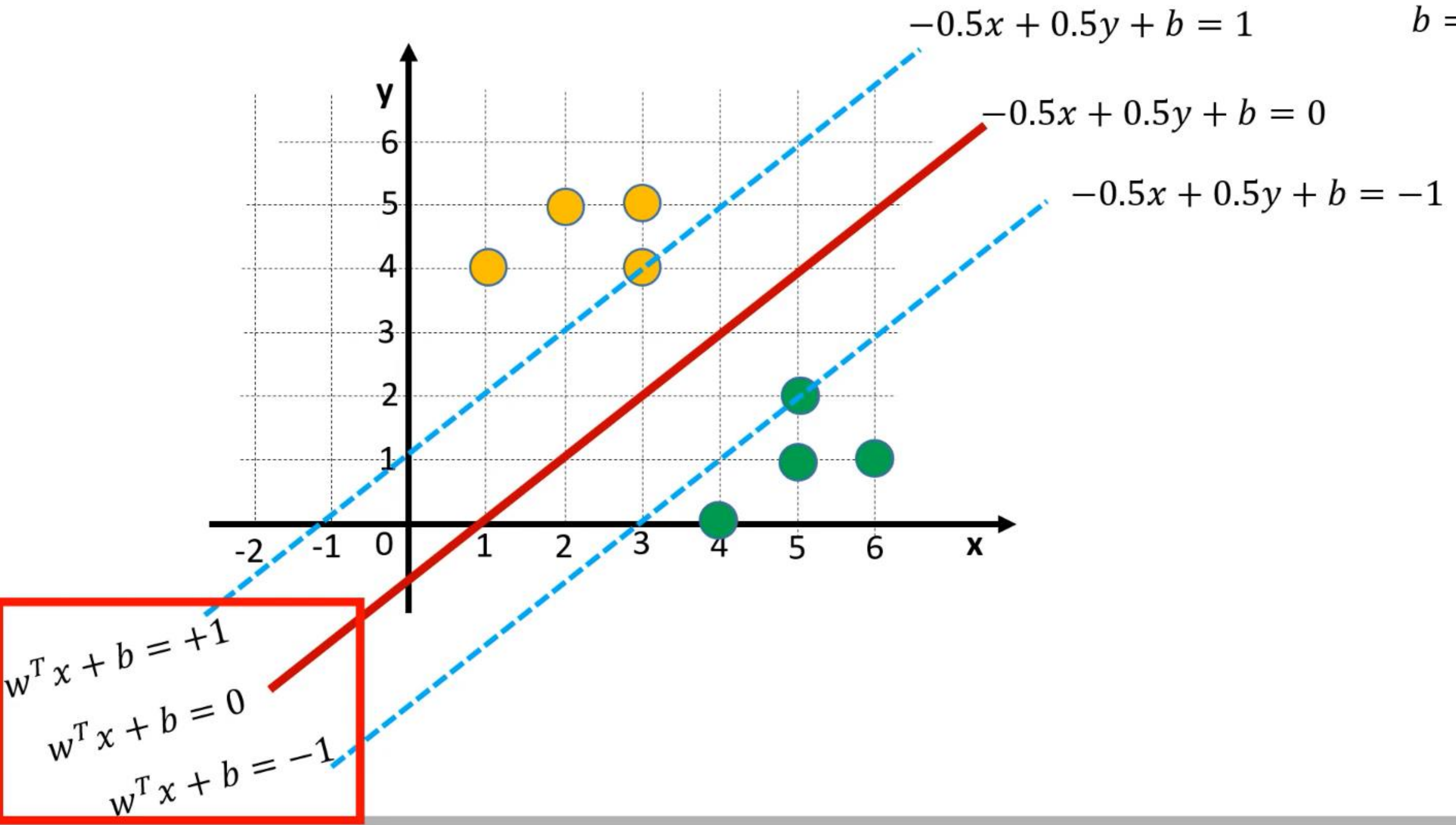
$$-0.5x + 0.5y + b = 0$$

Usually, one defines the two dimensions as x1 and x2 in SVM instead of x and y but to keep things simple, I have here named the axes as x and y.

SVM – the math

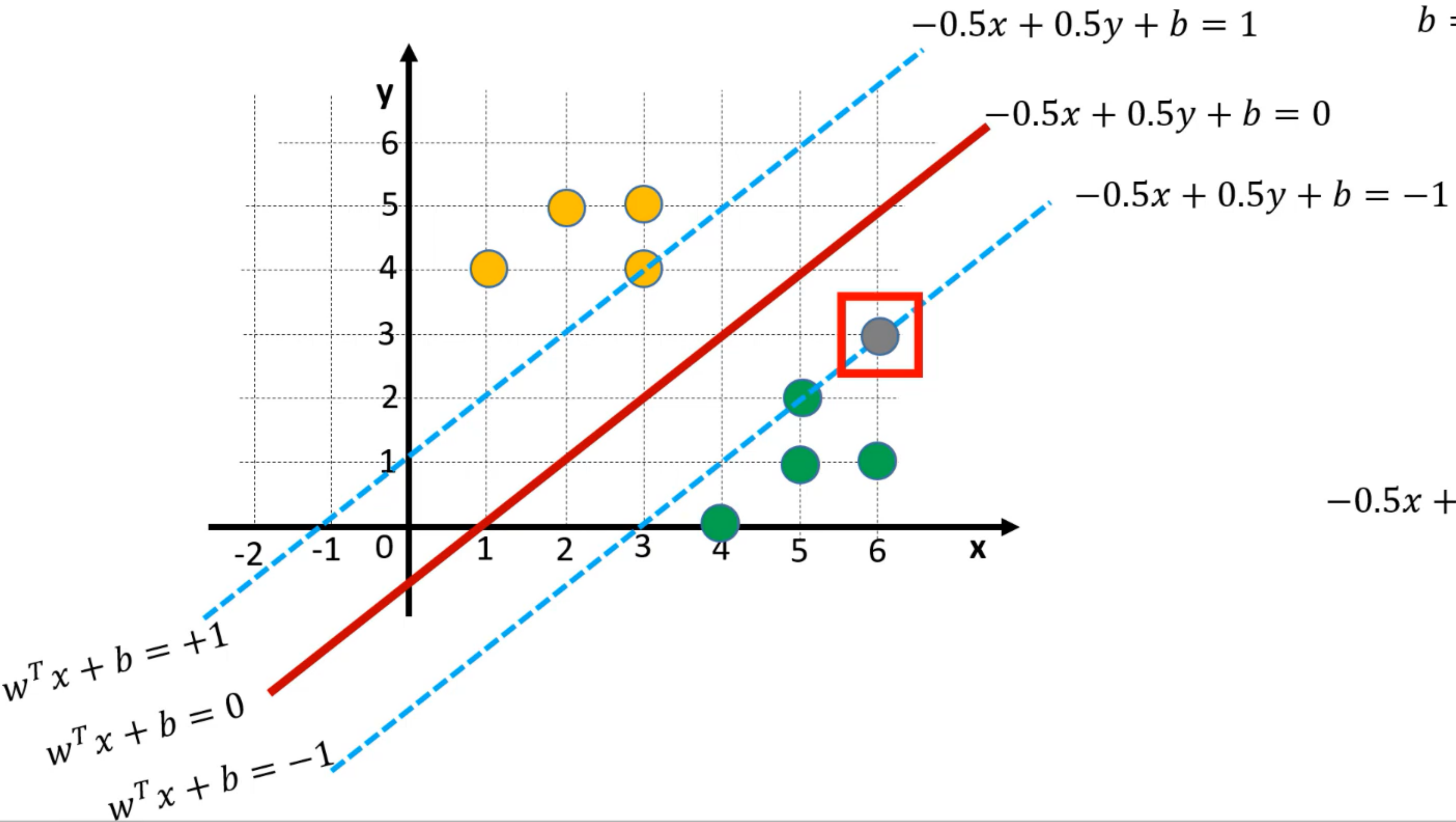
$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1



In the standard form, the equations of the three lines are defined like this, which is good to know because you will see these equations a lot when you read more about SVM.

SVM – the math



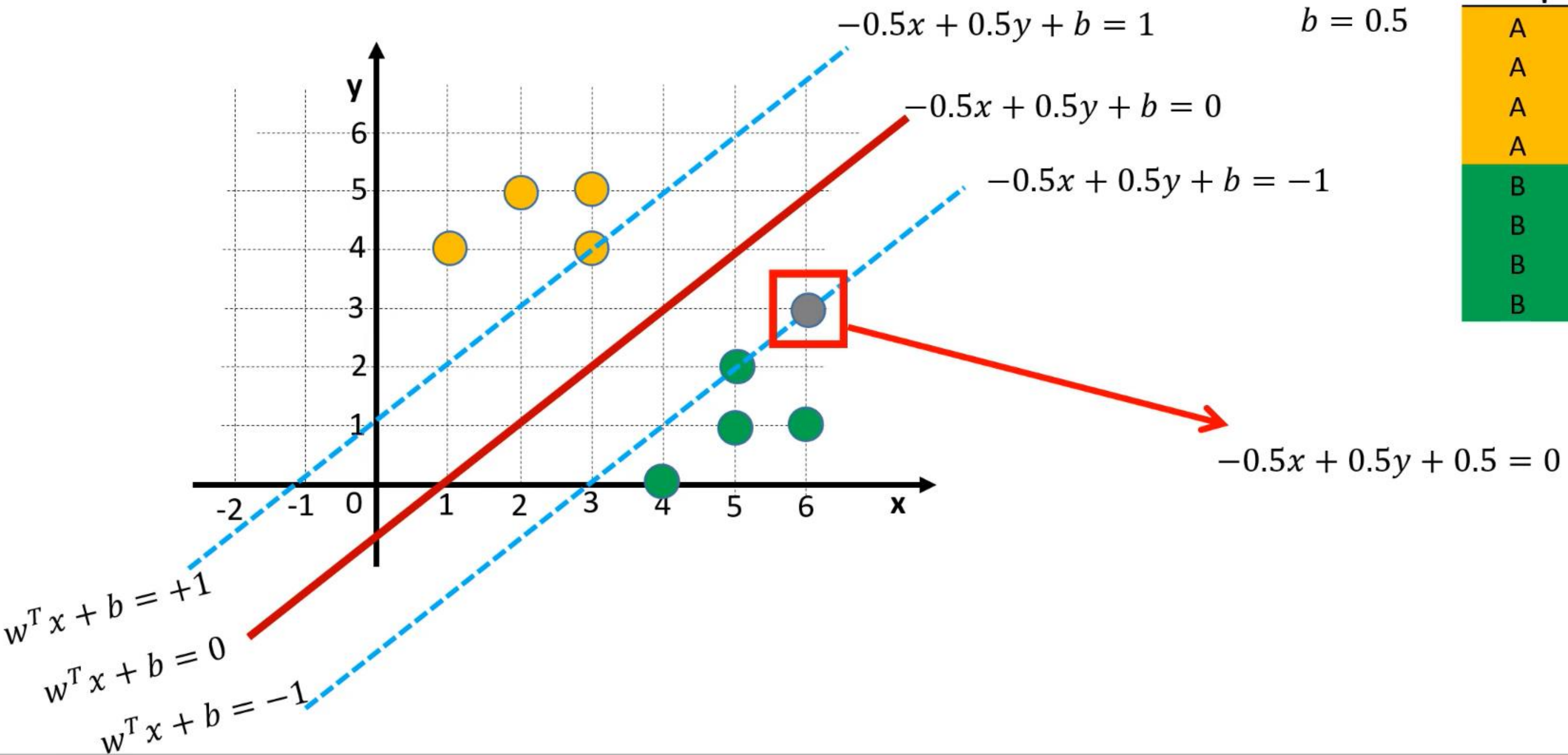
$b = 0.5$

Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

$$-0.5x + 0.5y + 0.5 = 0$$

Now, let's use our hyperplane to classify if a data point belongs to group A, the yellow points, or group B, the green points.

SVM – the math

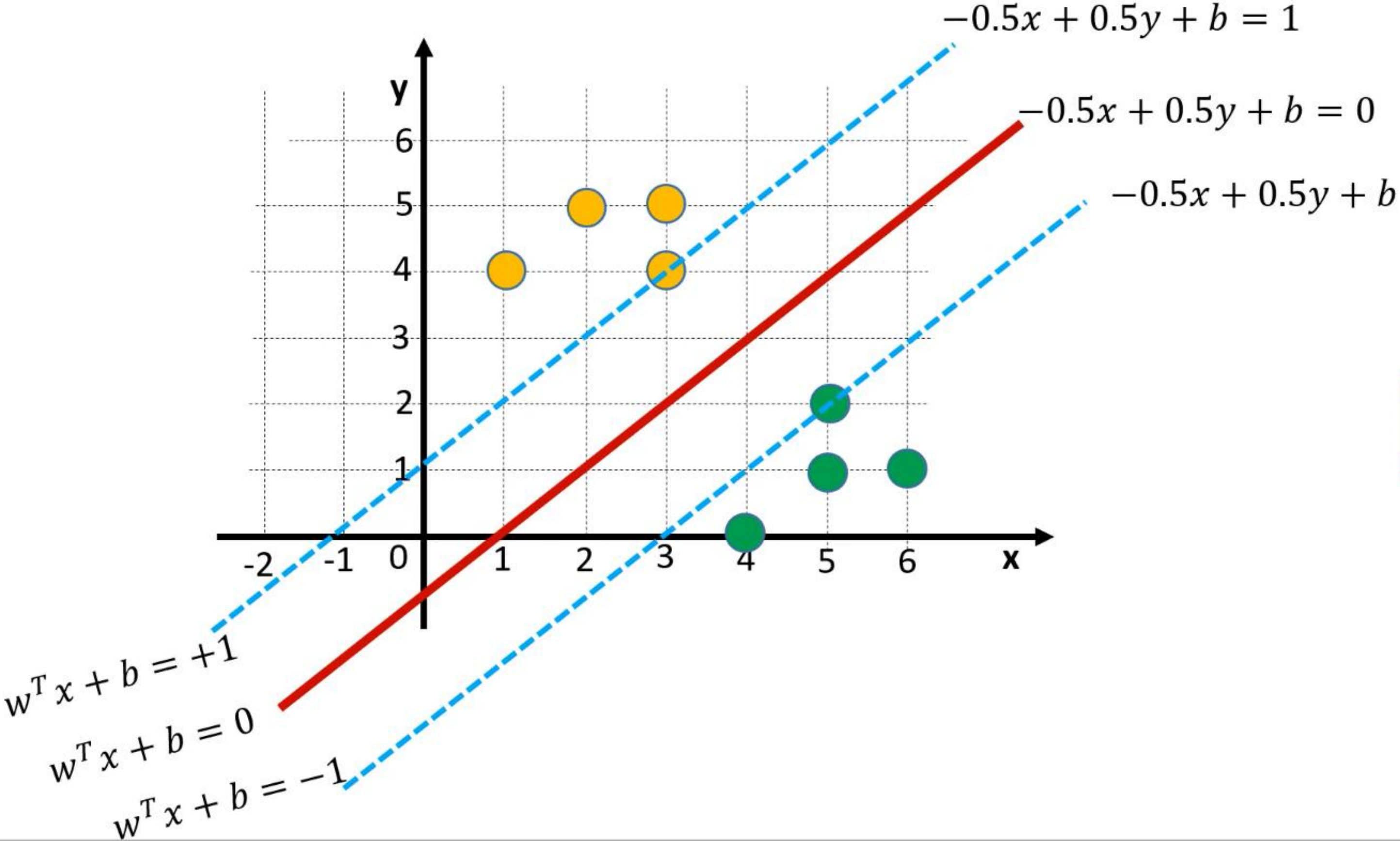


$b = 0.5$

Group	x	y
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

We plug in the x and y-coordinates of this data point of an unknown class in the equation of the hyperplane.

SVM – the math



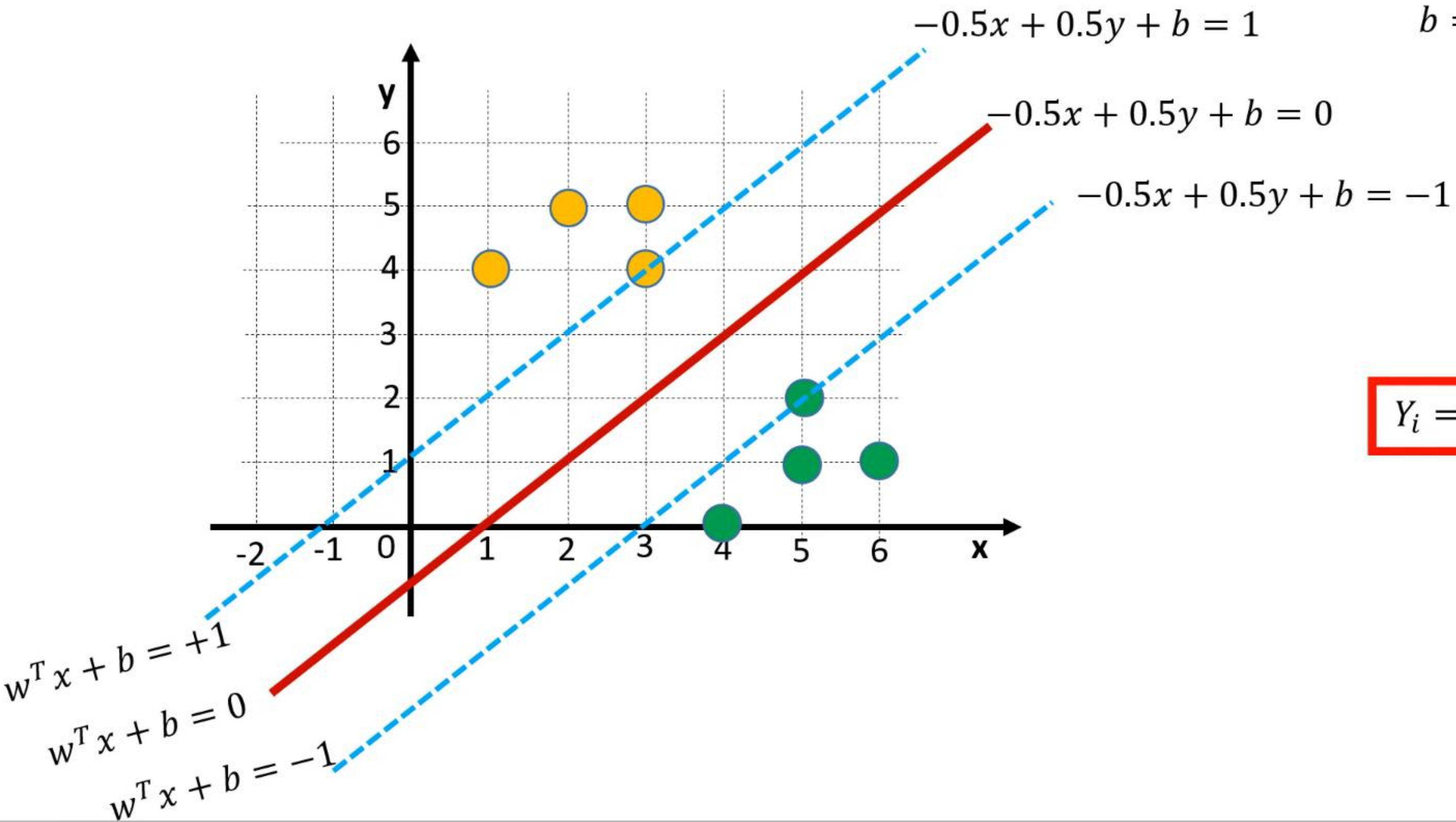
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{Y_i = [\,1, 1, 1, 1, -1, -1, -1, -1]}$$

In support vector machines, Y-sub-i usually defines the class of the training data as one and negative one.

SVM – the math



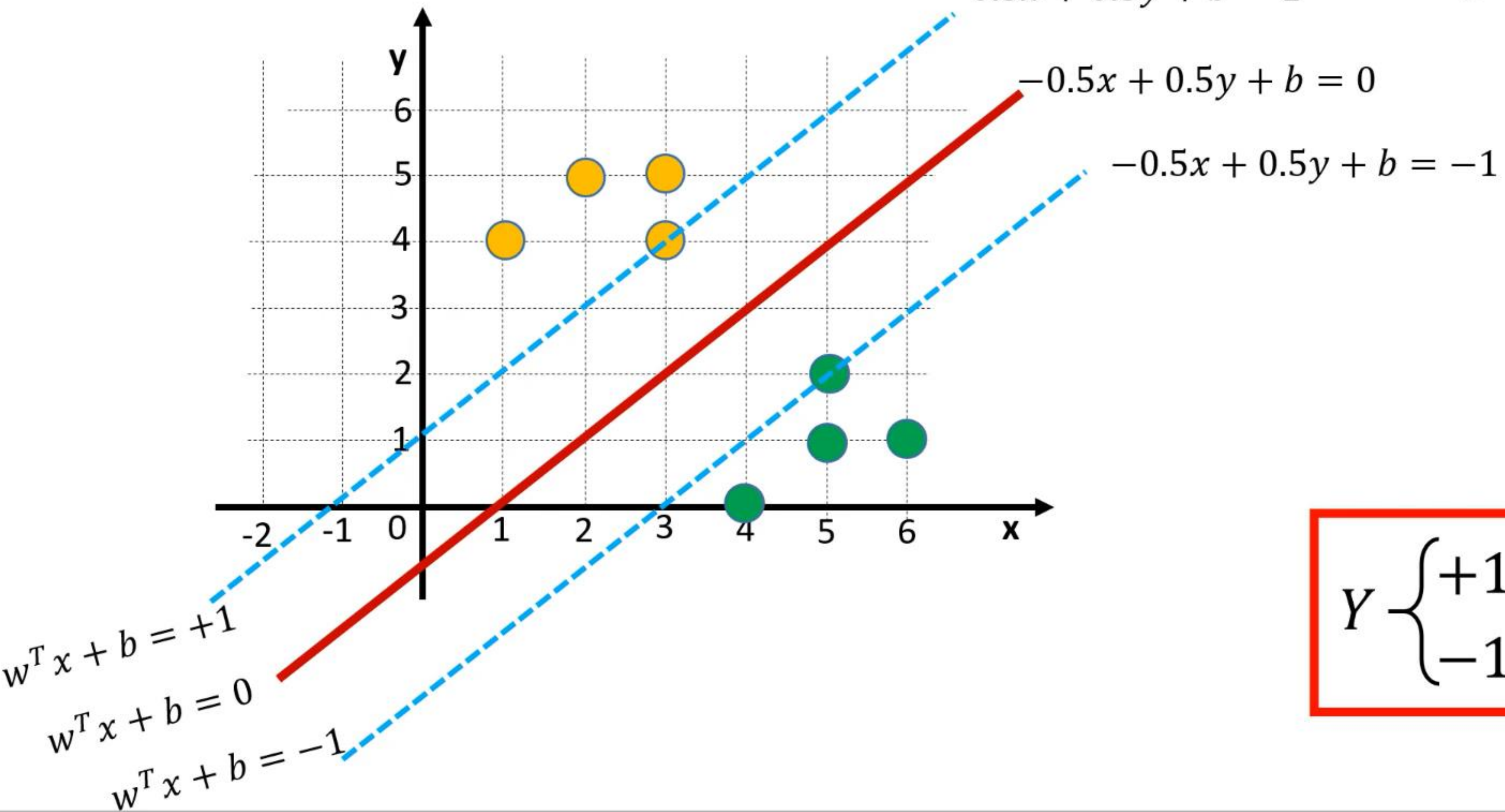
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{Y_i = [\,1, 1, 1, 1, -1, -1, -1, -1]}$$

This is just another way to tell if a data point in our training data belongs to group A or B.

SVM – the math



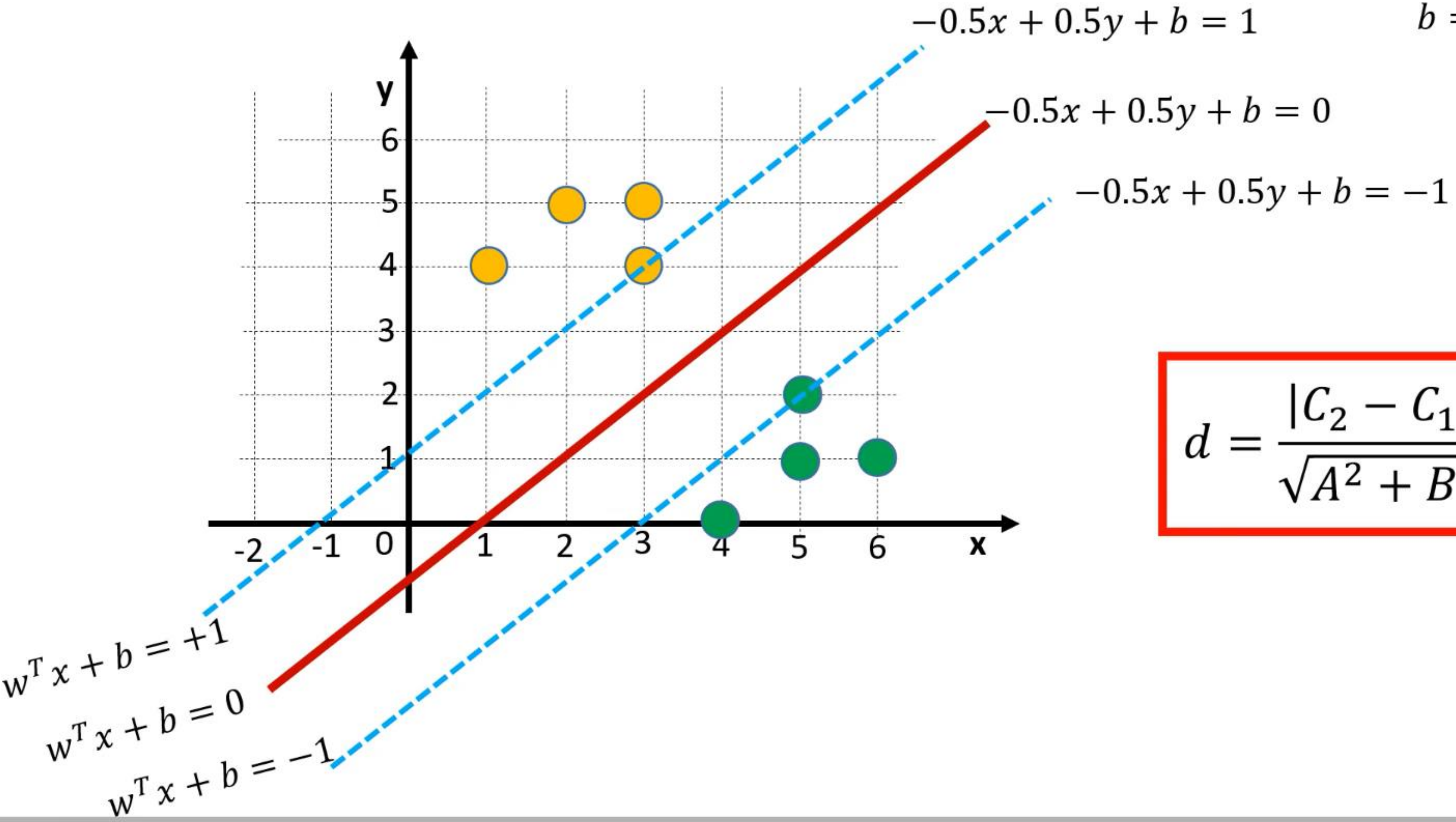
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{Y \begin{cases} +1 & \text{if } w^T x + b \geq 0 \\ -1 & \text{if } w^T x + b < 0 \end{cases}}$$

The classification in SVM is usually defined like this.

SVM – the math



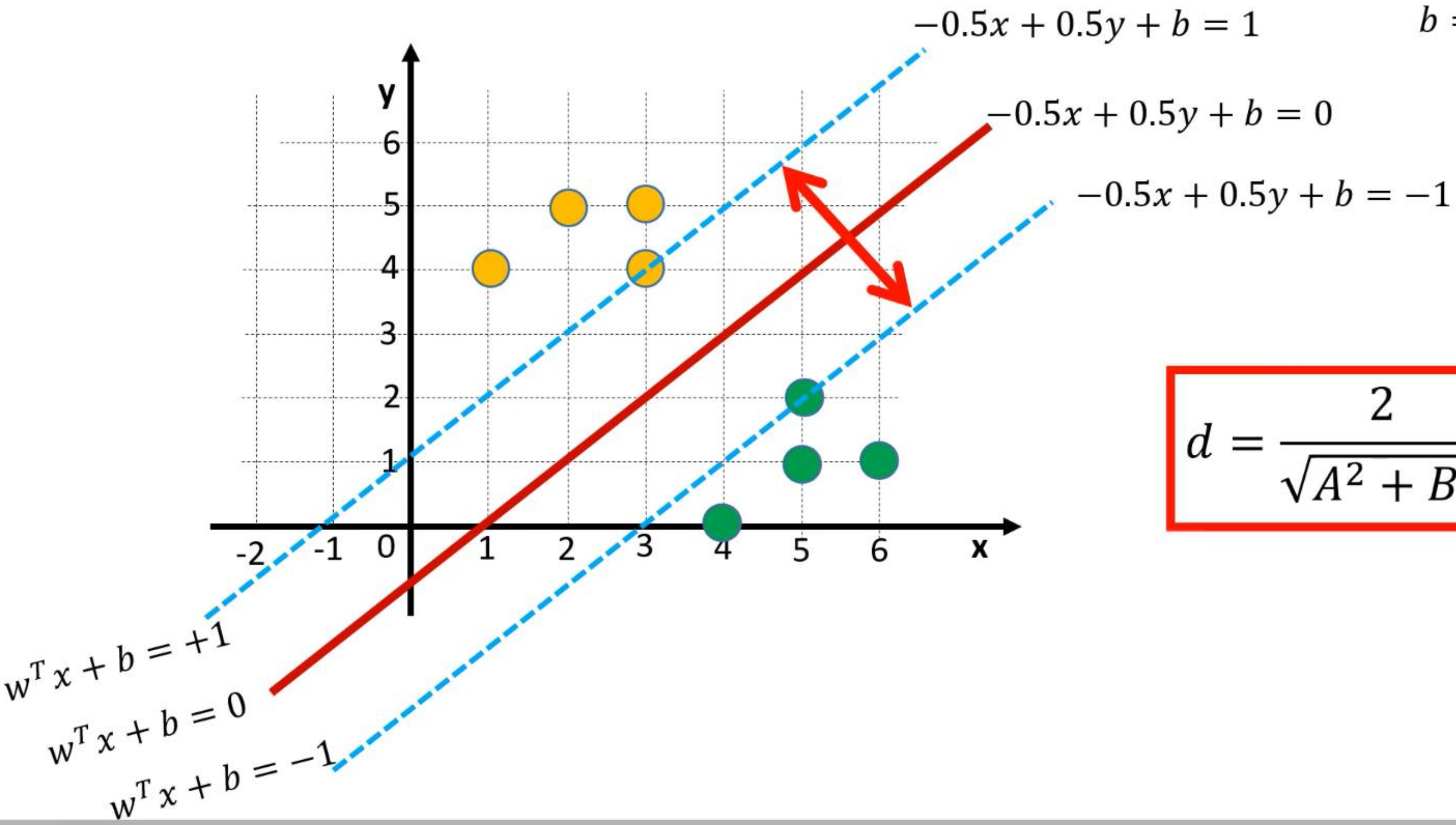
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{d = \frac{|C_2 - C_1|}{\sqrt{A^2 + B^2}}}$$

Remember that we can calculate the distance between two parallel lines with this equation.

SVM – the math



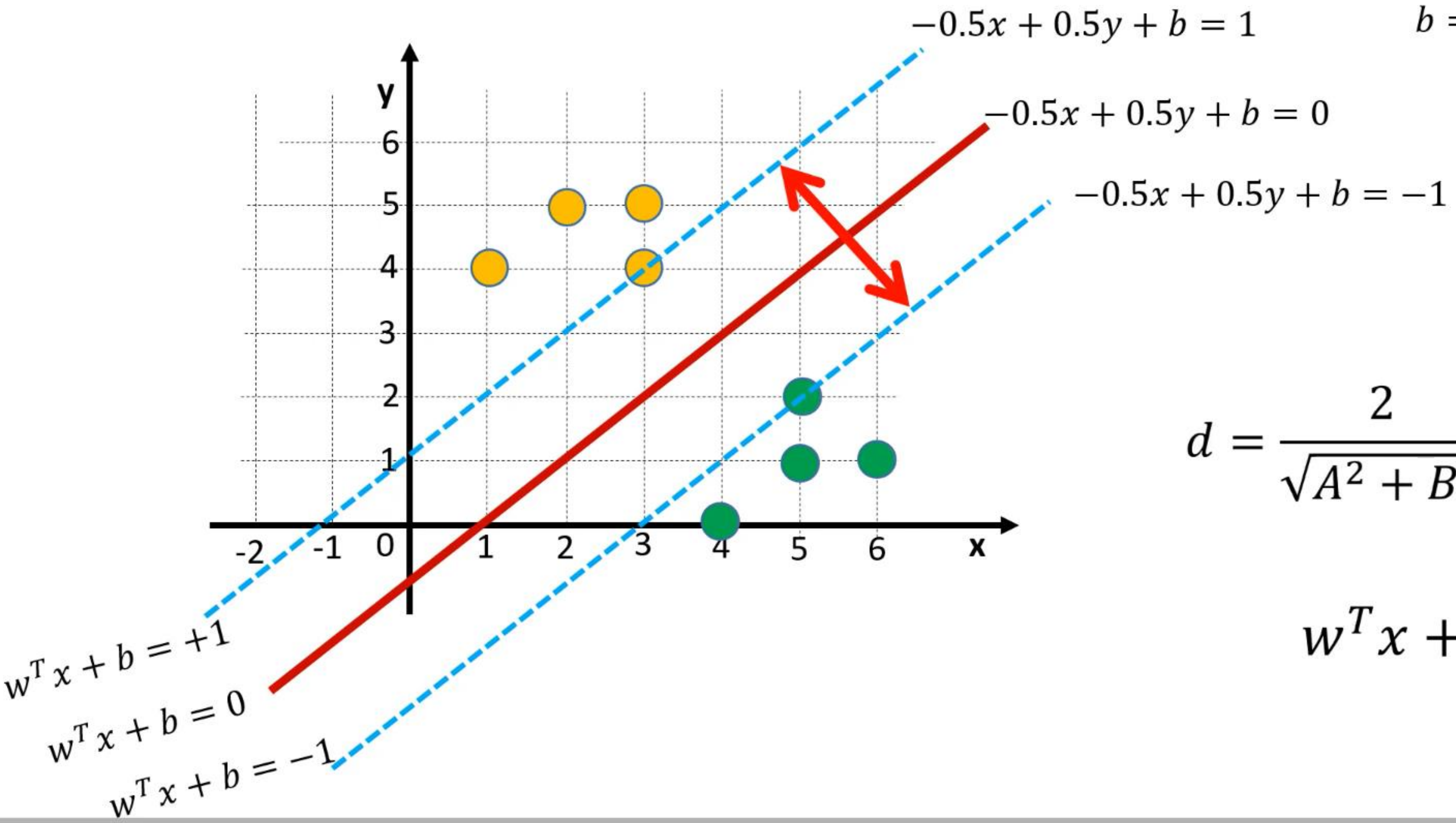
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{d = \frac{2}{\sqrt{A^2 + B^2}}}$$

we can simplify the equation to this,

SVM – the math



$b = 0.5$

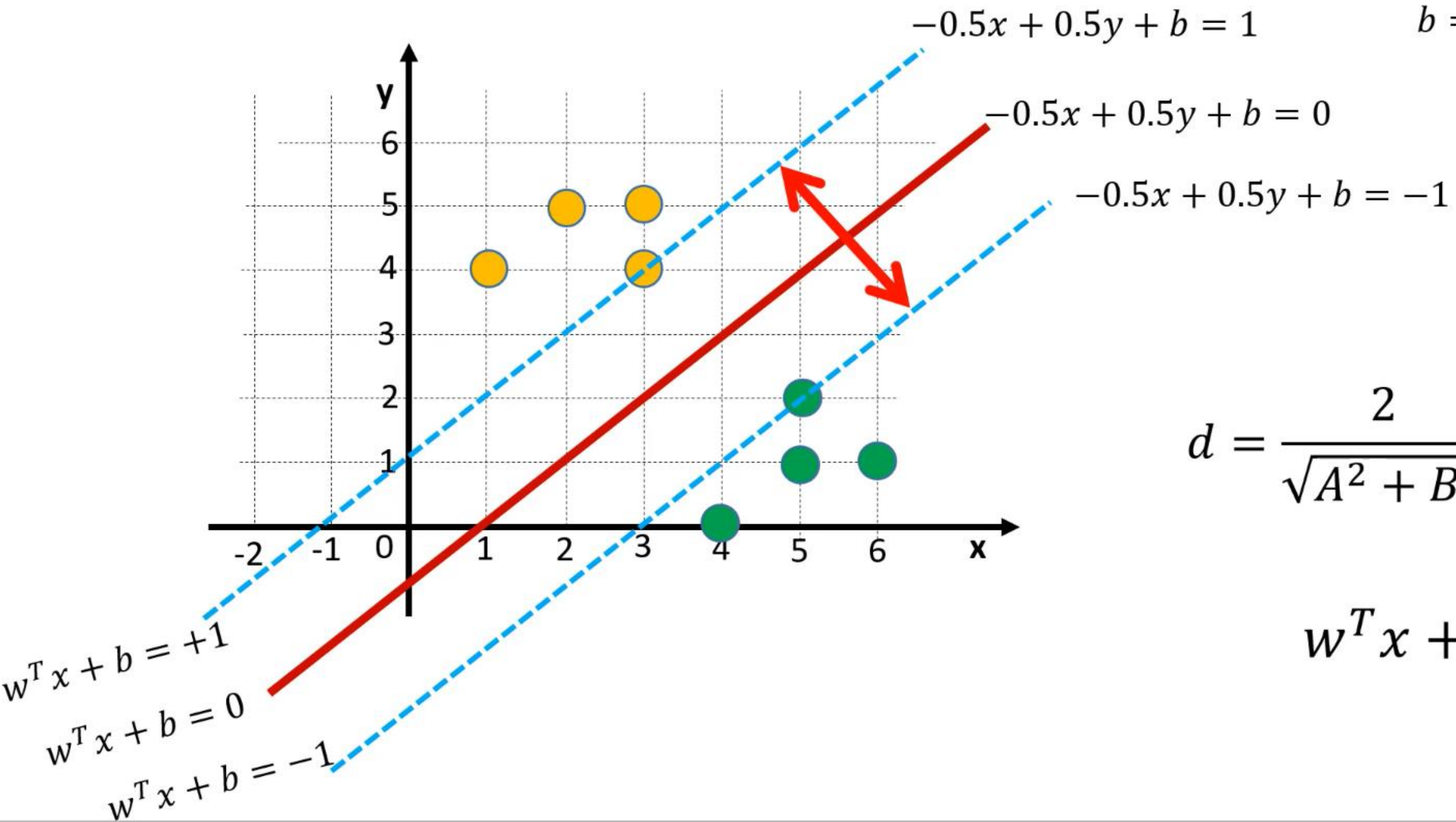
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \boxed{\frac{2}{\|w\|}}$$

$$w^T x + b = 0$$

which explains why the distance between the two blue lines is usually expressed like this.

SVM – the math



$b = 0.5$

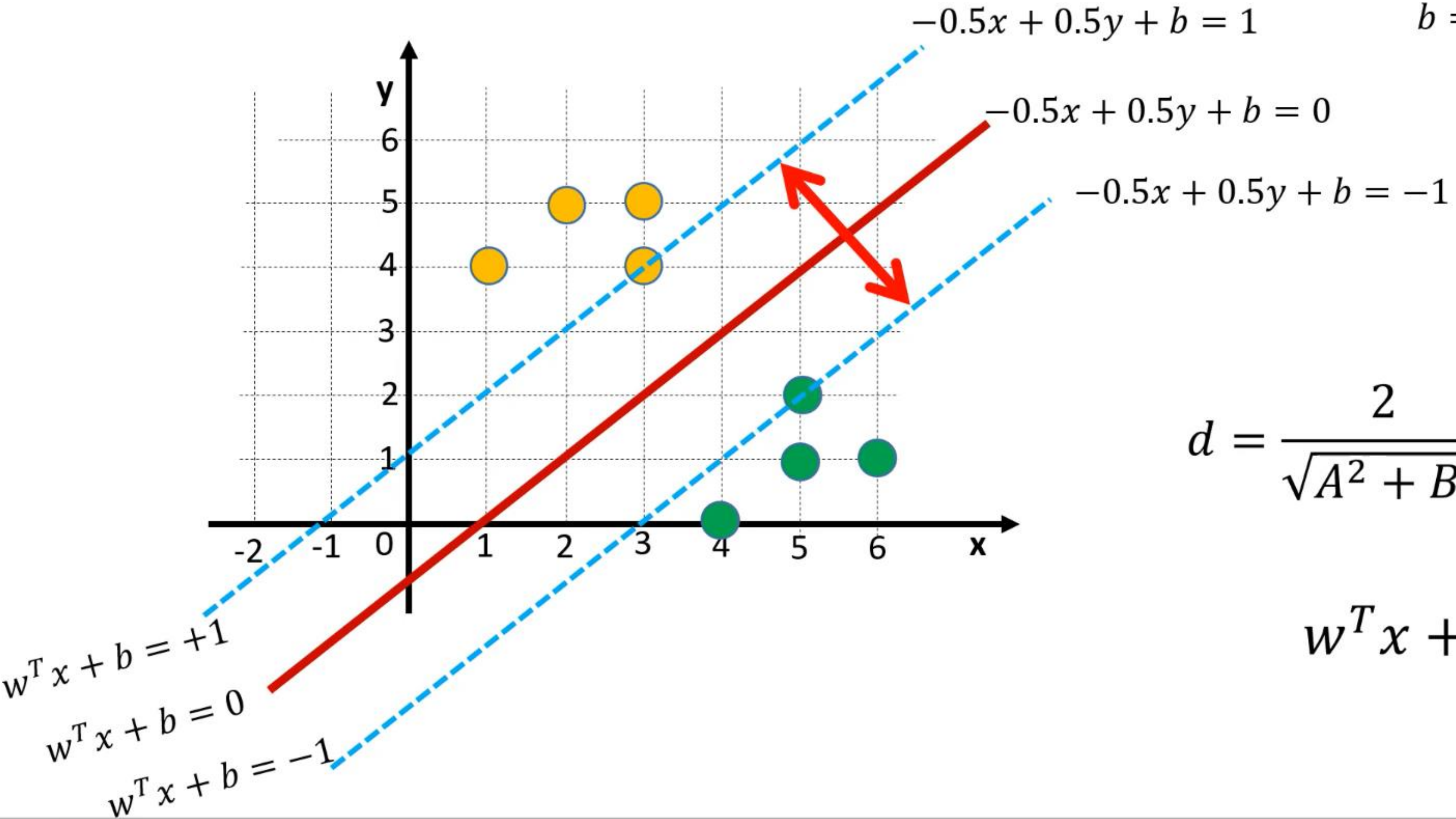
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \boxed{\frac{2}{\|w\|}}$$

$$w^T x + b = 0$$

If there is a line that can separate the two groups completely, SVM tries to find the optimal values of A and B so that the distance between the two blue lines is maximized.

SVM – the math



$b = 0.5$

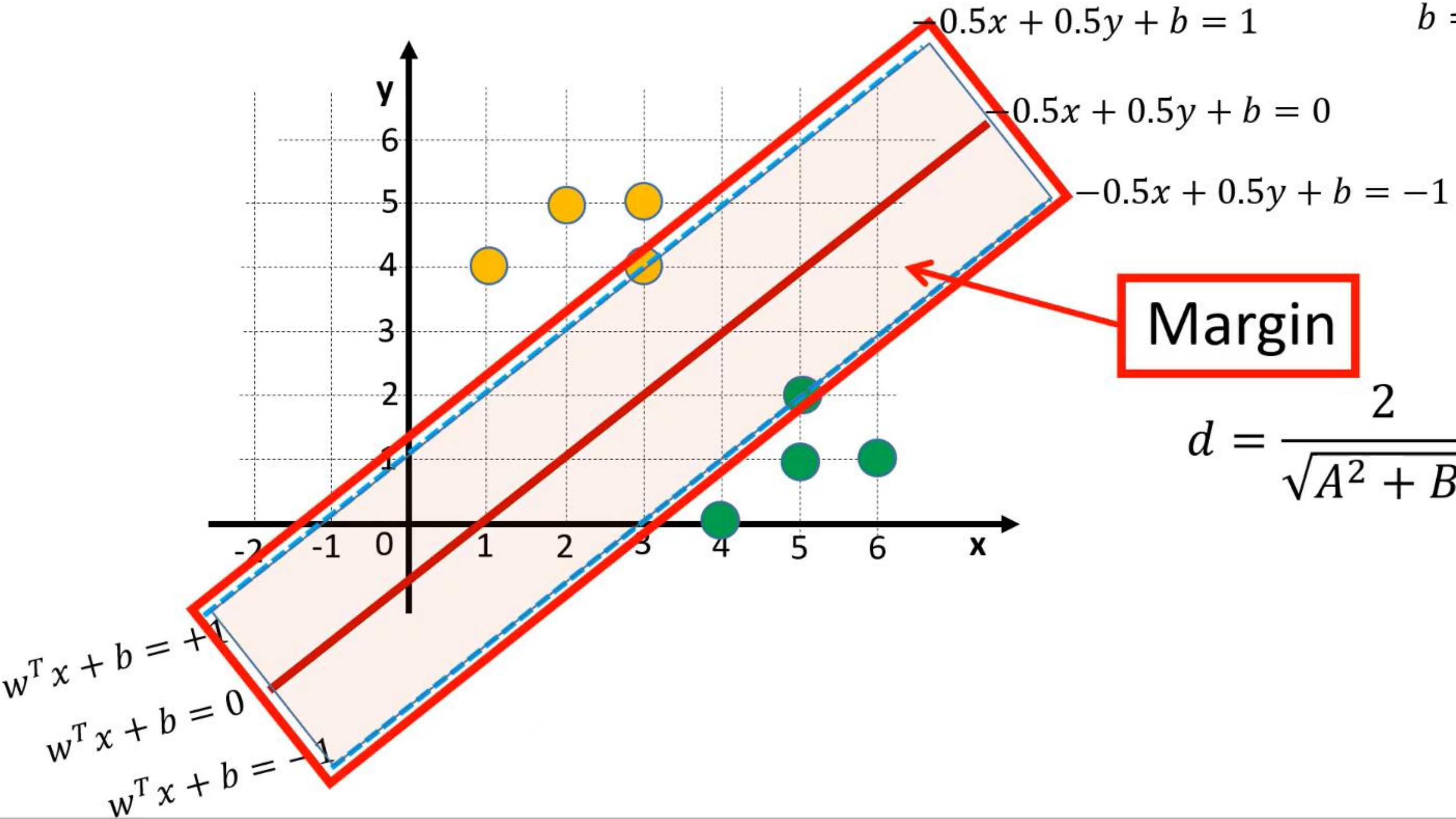
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \boxed{\frac{2}{\|w\|}}$$

$$w^T x + b = 0$$

Or, we can also minimize the denominator because that will also maximize the distance between the two blue lines.

SVM – the math



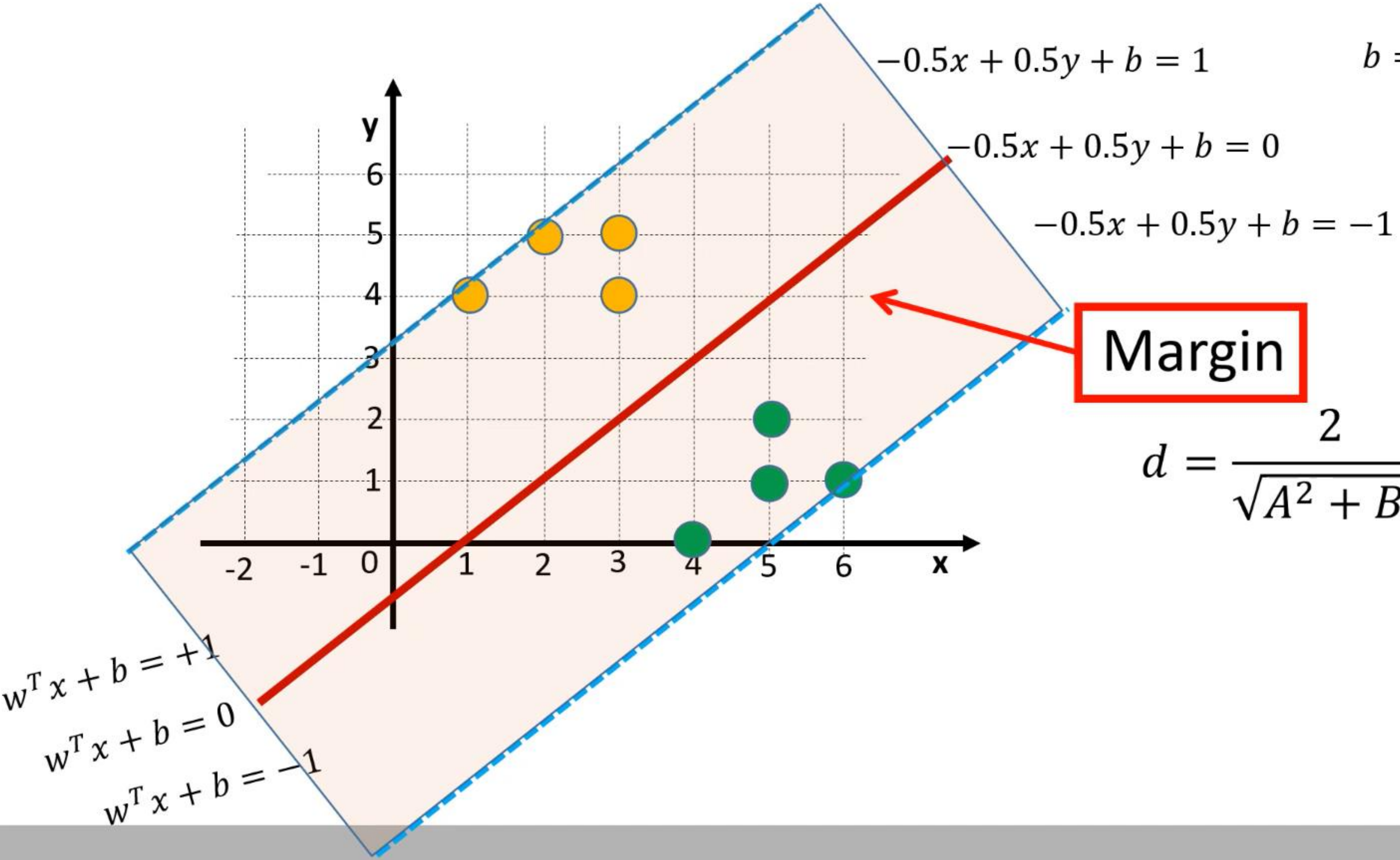
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \frac{2}{\|w\|}$$

The area between the two blue lines is called the margin. The SVM therefore tries to find a hyperplane that maximizes the width of this margin.

SVM – the math



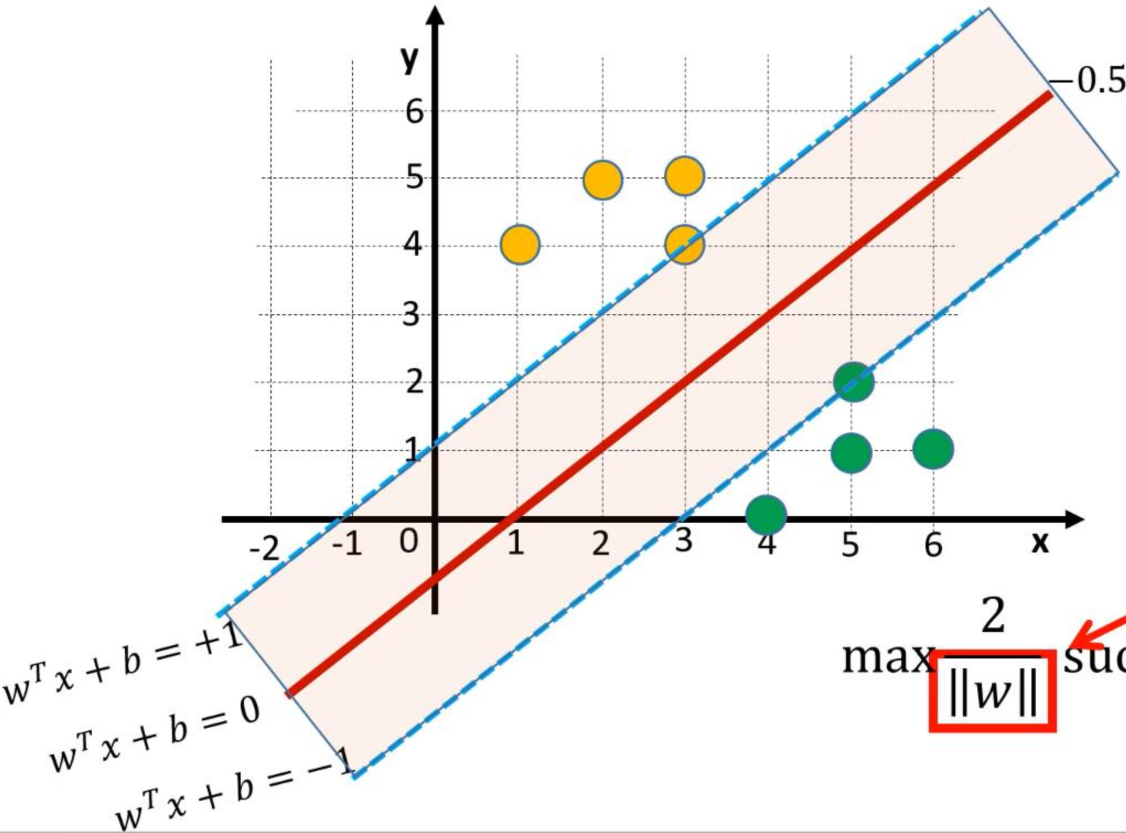
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \frac{2}{\|w\|}$$

However, we need some constraints because the margin can be infinitely large.

SVM – the math



$b = 0.5$

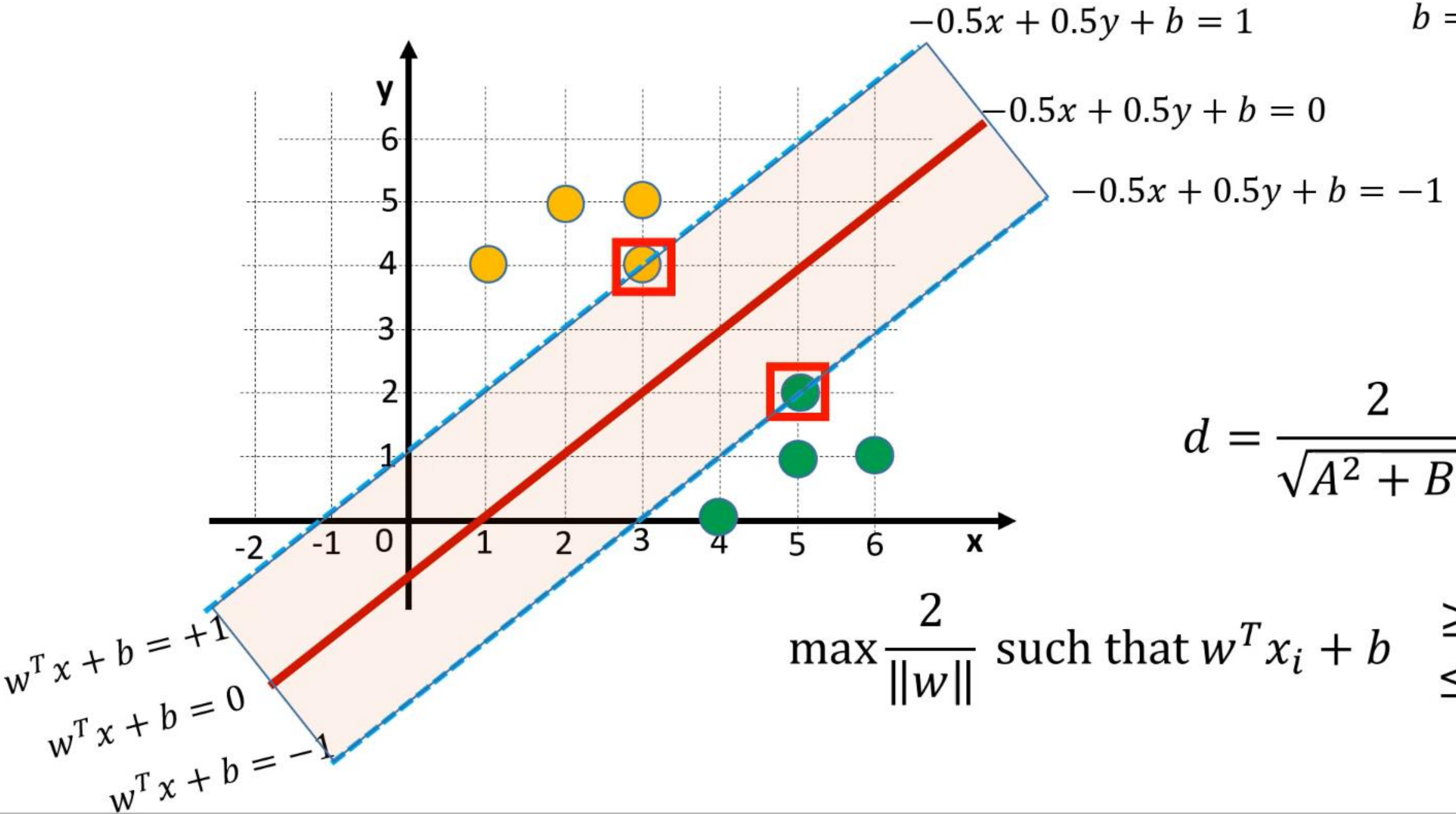
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\boxed{\sqrt{A^2 + B^2}}} = \frac{2}{\|w\|}$$

$$\max \frac{2}{\boxed{\|w\|}} \text{ such that } w^T x_i + b \begin{array}{l} \geq 1 \ \text{ if } Y_i = +1 \\ \leq -1 \text{ if } Y_i = -1 \end{array}$$

Training the SVM can be defined as we try to find the optimal values of A and B of the hyperplane,

SVM – the math



$b = 0.5$

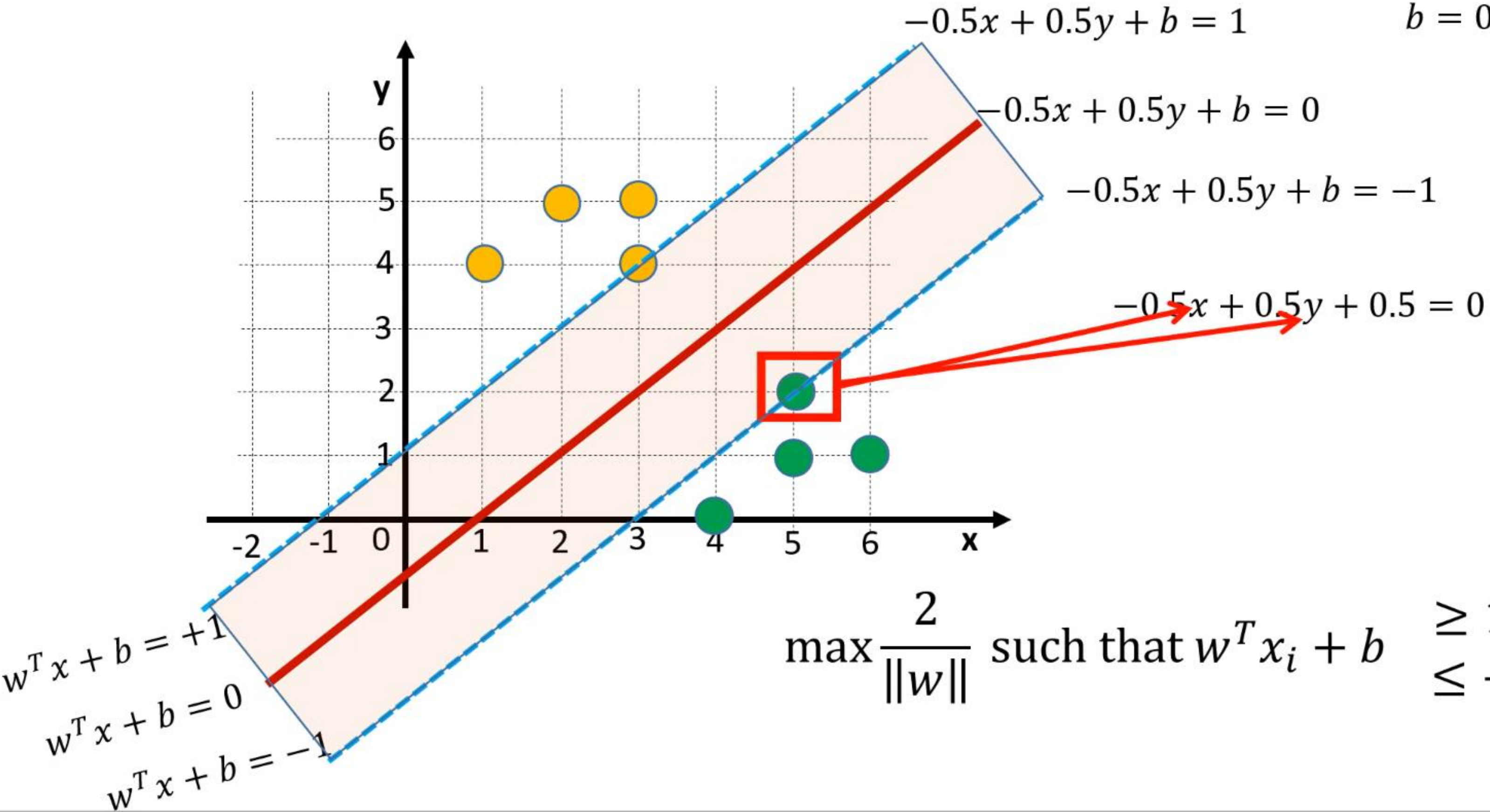
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$d = \frac{2}{\sqrt{A^2 + B^2}} = \frac{2}{\|w\|}$$

$$\max \frac{2}{\|w\|} \text{ such that } w^T x_i + b \begin{array}{l} \geq 1 \;\; \text{if } Y_i = +1 \\ \leq -1 \text{ if } Y_i = -1 \end{array}$$

This means that the margin should not span beyond the data points closest to the hyperplane. Remember that these points are called support vectors.

SVM – the math



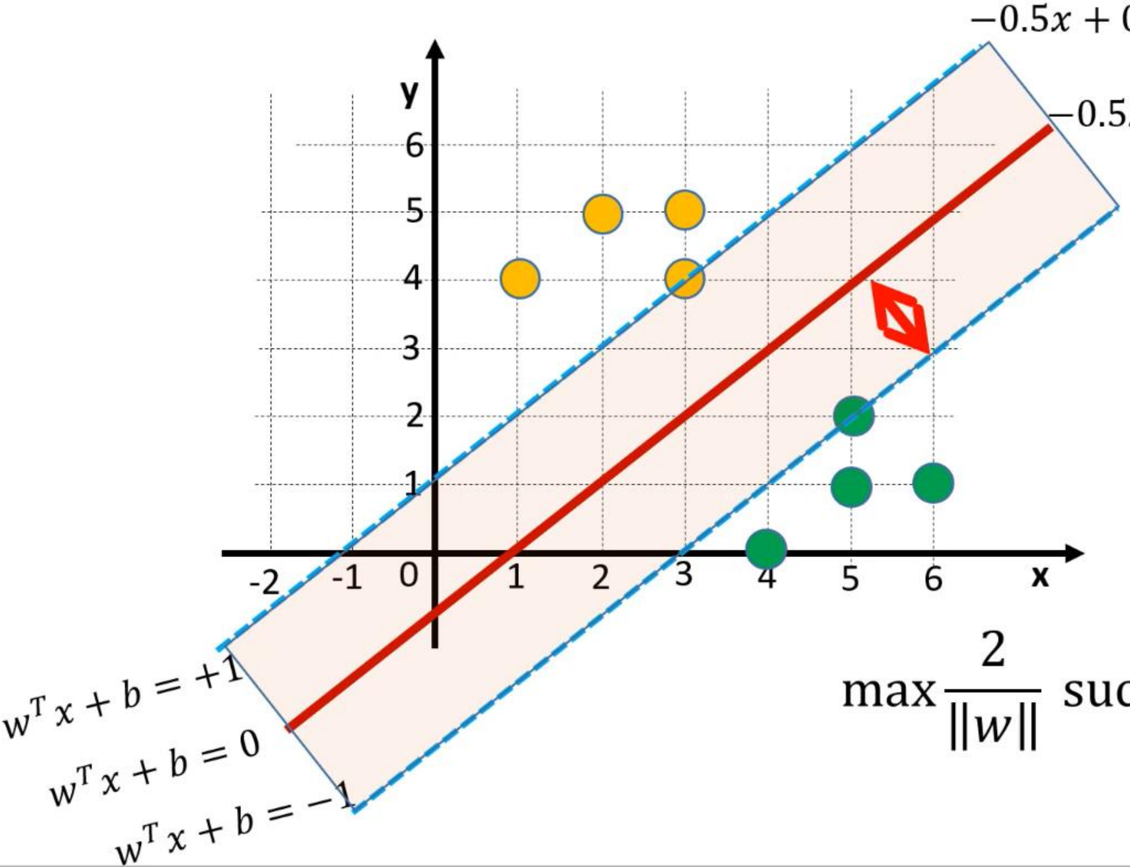
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\max \frac{2}{\|w\|} \text{ such that } w^T x_i + b \begin{cases} \geq 1 & \text{if } Y_i = +1 \\ \leq -1 & \text{if } Y_i = -1 \end{cases}$$

Let's see what value we get if we plug in the x and y coordinates of this data point in the equation of our hyperplane.

SVM – the math



$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

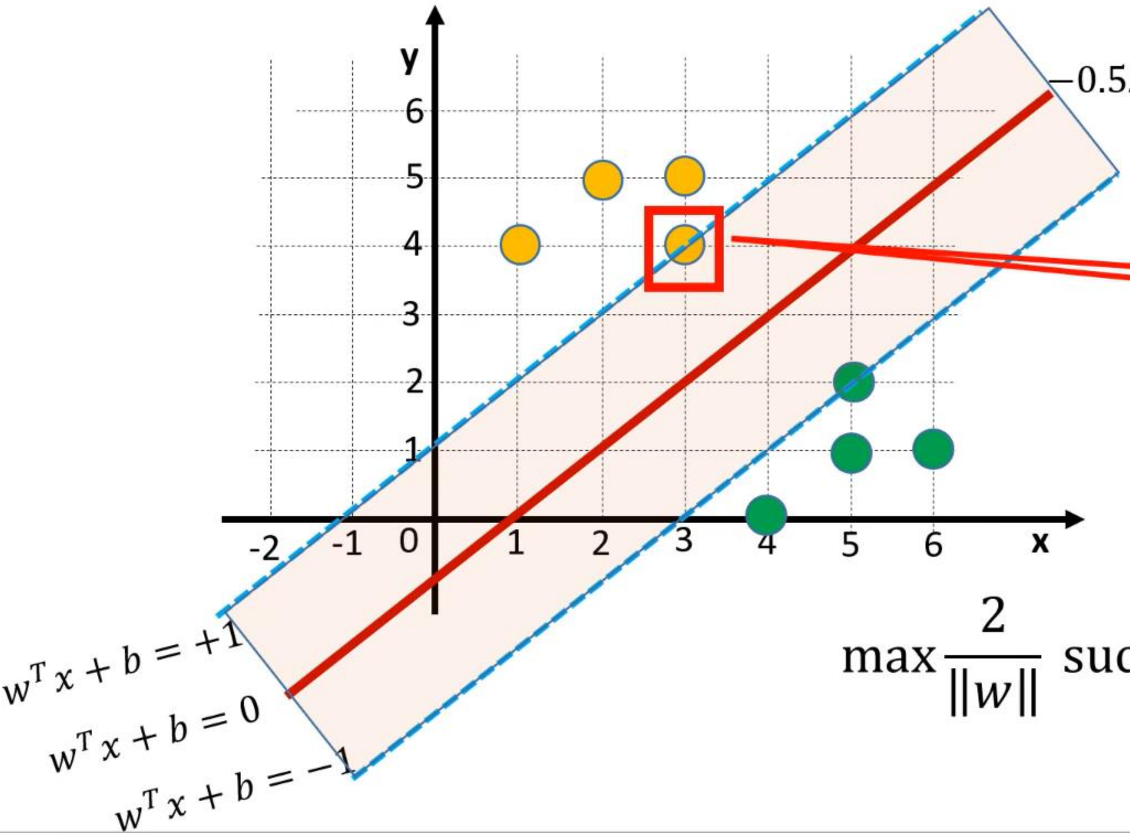
$$-0.5x + 0.5y + 0.5 = 0$$

$$-0.5 \cdot 5 + 0.5 \cdot 2 + 0.5 = -1$$

$$\max \frac{2}{\|w\|} \text{ such that } w^T x_i + b \begin{cases} \geq 1 & \text{if } Y_i = +1 \\ \leq -1 & \text{if } Y_i = -1 \end{cases}$$

The width of the margin is therefore constrained so that it cannot span beyond the support vector if the hyperplane can separate the two groups completely.

SVM – the math



$-0.5x + 0.5y + b = 1$

$b = 0.5$

$-0.5x + 0.5y + b = 0$

$-0.5x + 0.5y + b = -1$

$-0.5x + 0.5y + 0.5 = 0$

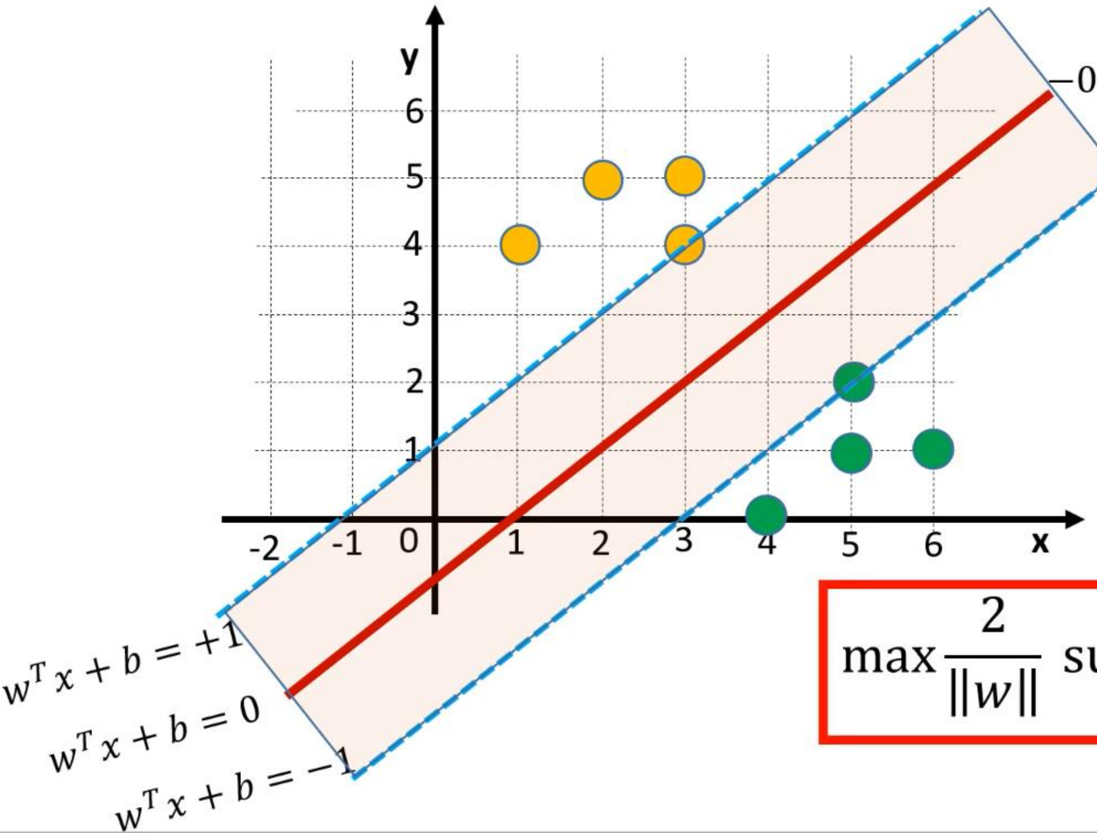
$\boxed{-0.5 \cdot 3 + 0.5 \cdot 4 + 0.5 = 1}$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\max \frac{2}{\|w\|} \text{ such that } w^T x_i + b \begin{cases} \geq 1 & \text{if } Y_i = +1 \\ \leq -1 & \text{if } Y_i = -1 \end{cases}$$

Similarly, the line that defines the boundary above our hyperplane cannot be further away because if we plug in the x and y coordinates of this support vector in the equation of the hyperplane, that results in a value that is equal to positive one.

SVM – the math



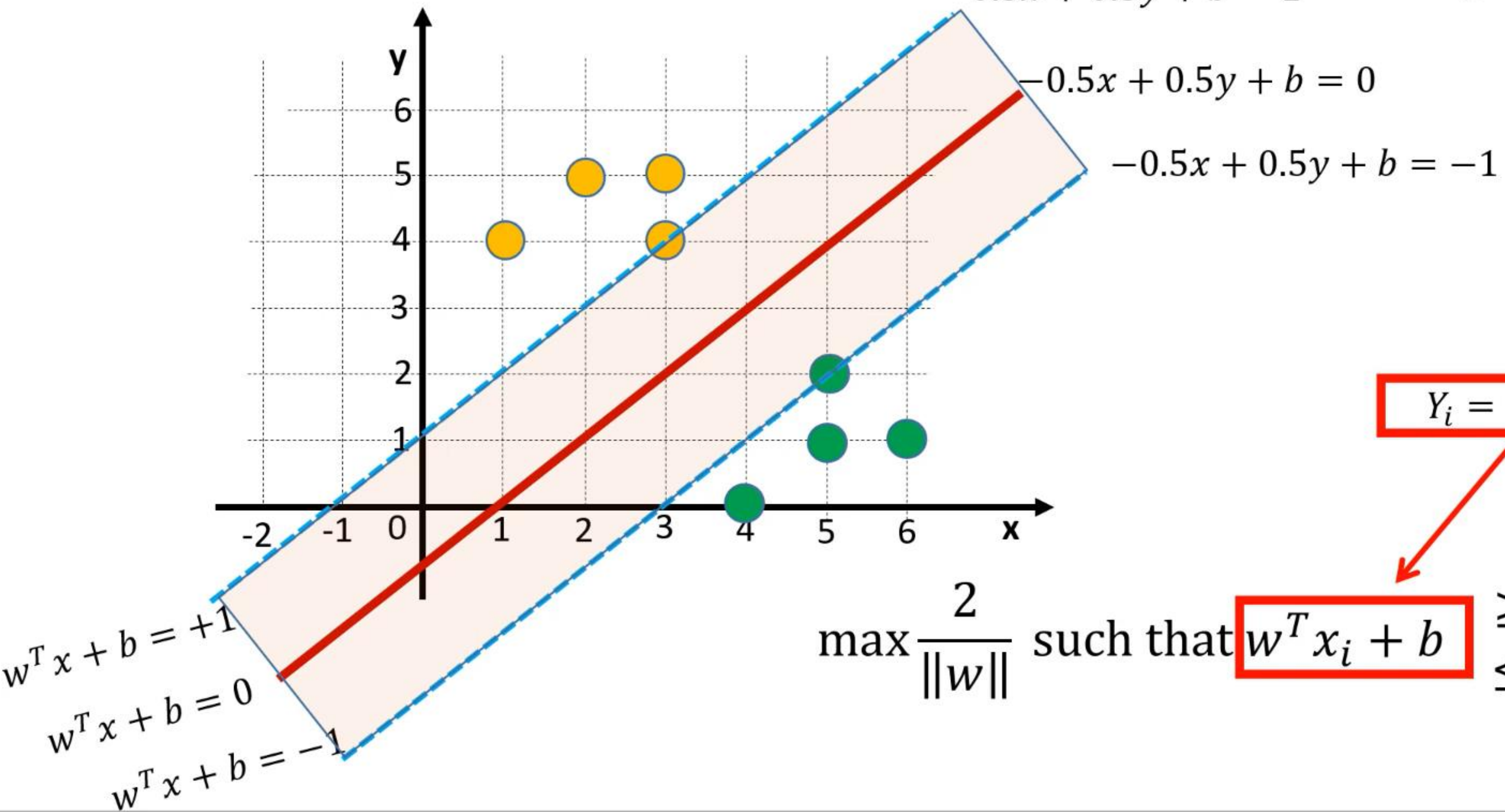
$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\max \frac{2}{\|w\|} \text{ such that } w^T x_i + b \begin{array}{l} \geq 1 \text{ if } Y_i = +1 \\ \leq -1 \text{ if } Y_i = -1 \end{array}$$

We can simplify this equation a bit,

SVM – the math



$b = 0.5$

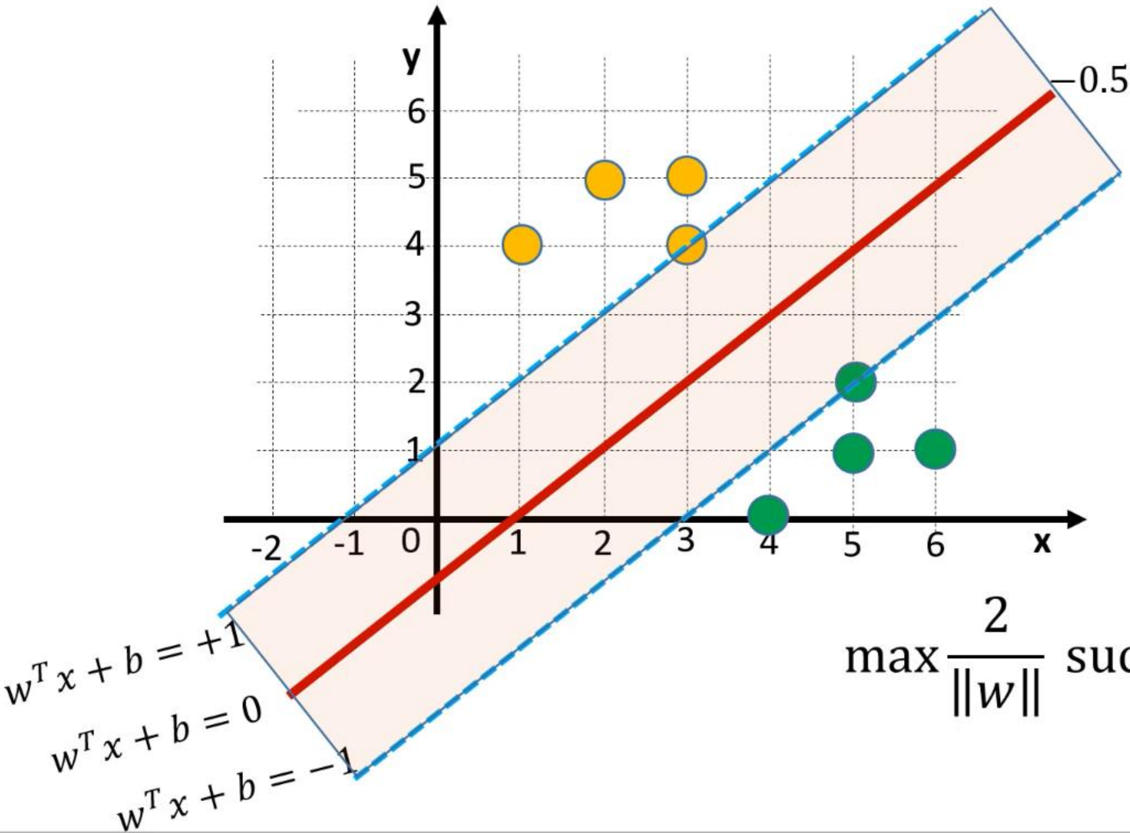
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$$\boxed{Y_i = [\,1, 1, 1, 1, -1, -1, -1, -1]}$$

$$\max \frac{2}{\|w\|} \text{ such that } \boxed{w^T x_i + b} \begin{array}{l} \geq 1 \;\text{ if } Y_i = +1 \\ \leq -1 \text{ if } Y_i = -1 \end{array}$$

if we multiply the equation by Y-sub-i,

SVM – the math



$b = 0.5$

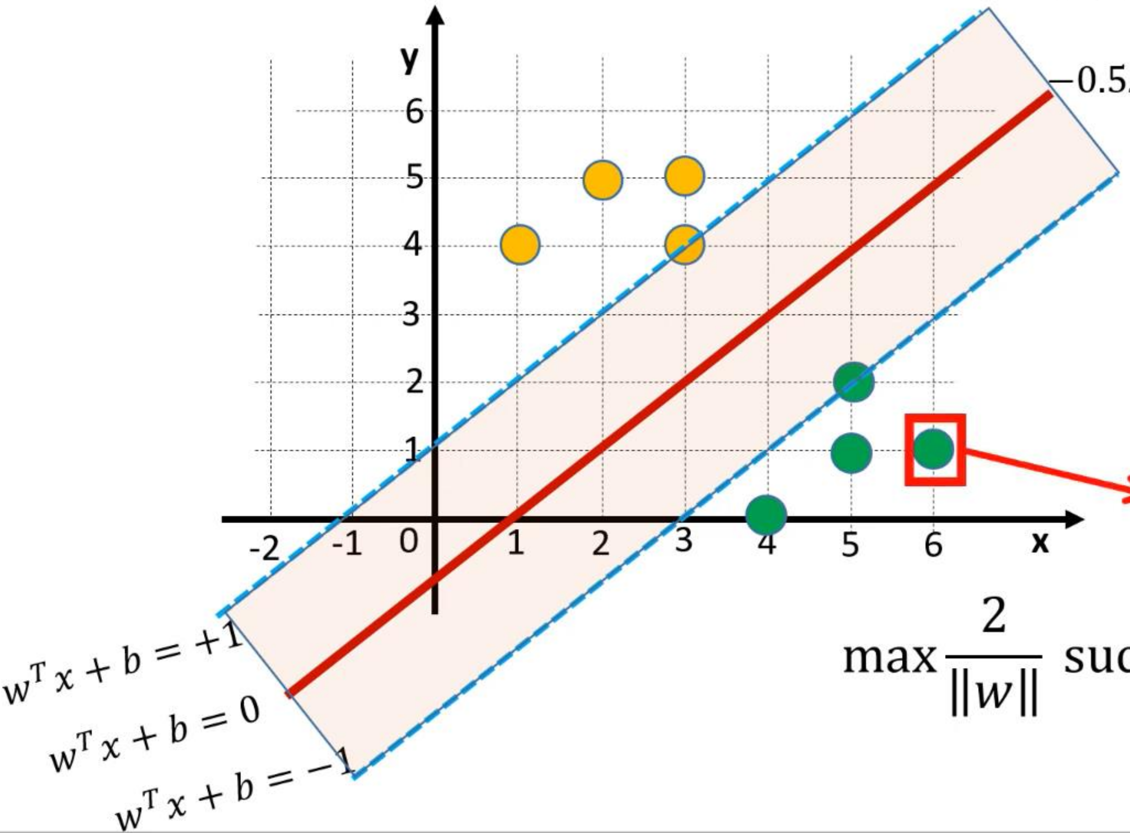
Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

$Y_i = [\,1, 1, 1, 1, -1, -1, -1, -1]$

$$\max \frac{2}{\|w\|} \text{ such that } \boxed{Y_i}(w^T x_i + b) \geq 1$$

because Y-sub-i is equal to one if the data point belongs to the positive group, and equal to negative one if the data point belongs to the negative group.

SVM – the math



$b = 0.5$

Group (Y_i)	*x*	*y*
A (+1)	1	4
A (+1)	2	5
A (+1)	3	5
A (+1)	3	4
B (-1)	6	1
B (-1)	4	0
B (-1)	5	2
B (-1)	5	1

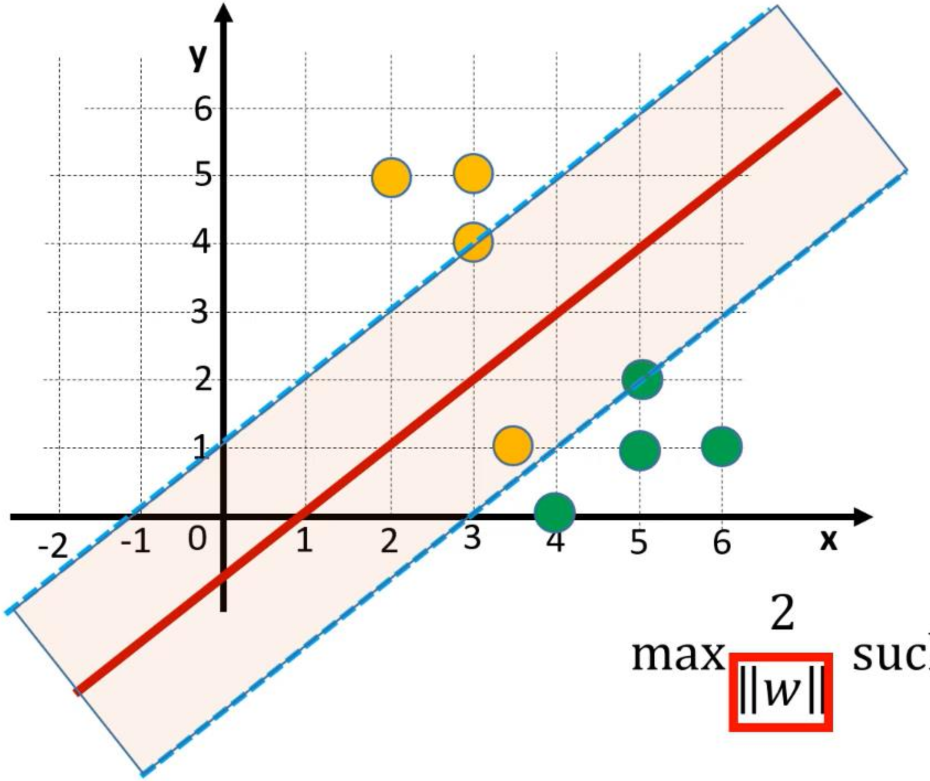
$Y_i = [\,1, 1, 1, 1, -1, -1, -1, -1]$

$$\boxed{(-1) \cdot (-0.5 \cdot 6 + 0.5 \cdot 1 + 0.5) = 2}$$

$$\max \frac{2}{\|w\|} \text{ such that } Y_i(w^T x_i + b) \geq 1$$

For example, this data point will now result in a positive value,

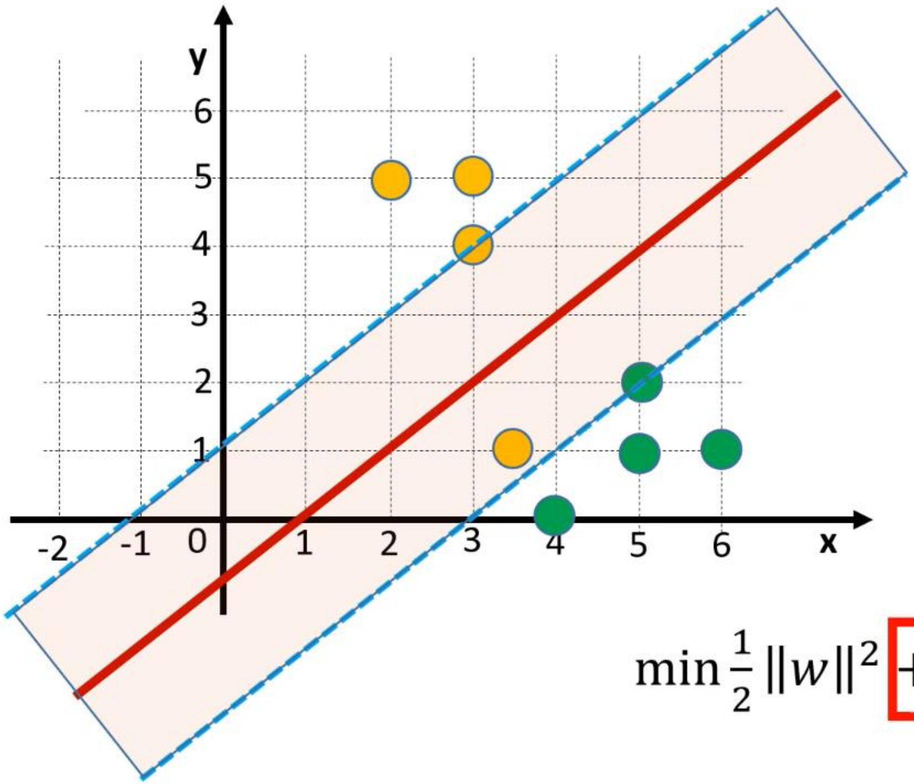
SVM – the math



$$\max \frac{2}{\boxed{||w||}} \text{ such that } Y_i(w^T x_i + b) \geq 1$$

Remember that, to maximize the margin, we need to minimize what is in the denominator.

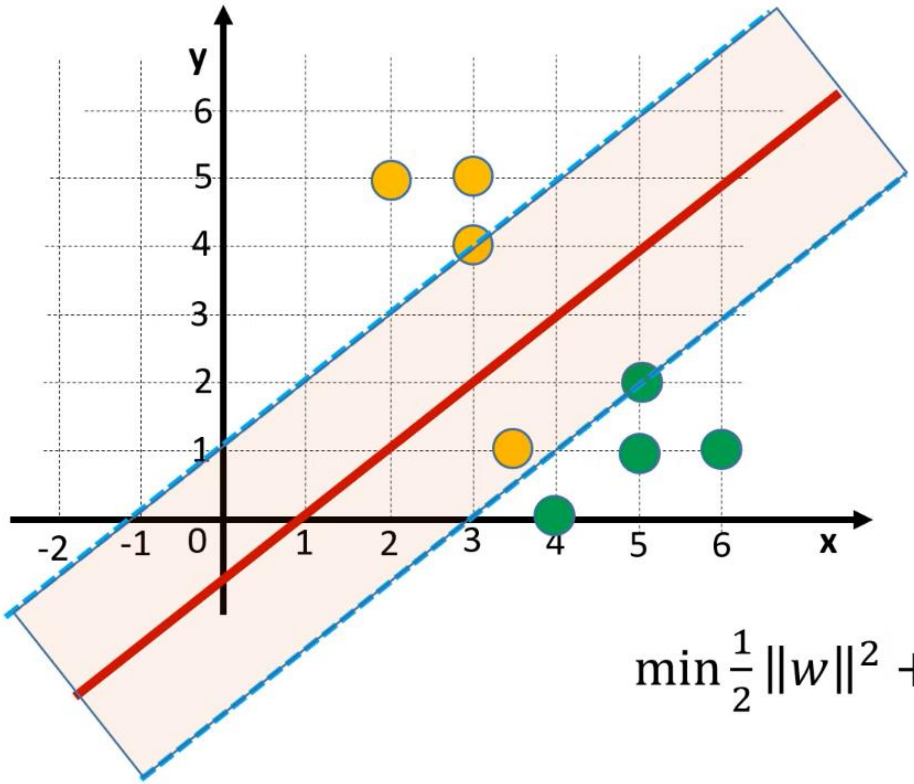
SVM – the math



$$\min \frac{1}{2}\|w\|^2 \boxed{+ C\sum_i^N \varepsilon_i}$$

And to allow for misclassification, we can add this term,

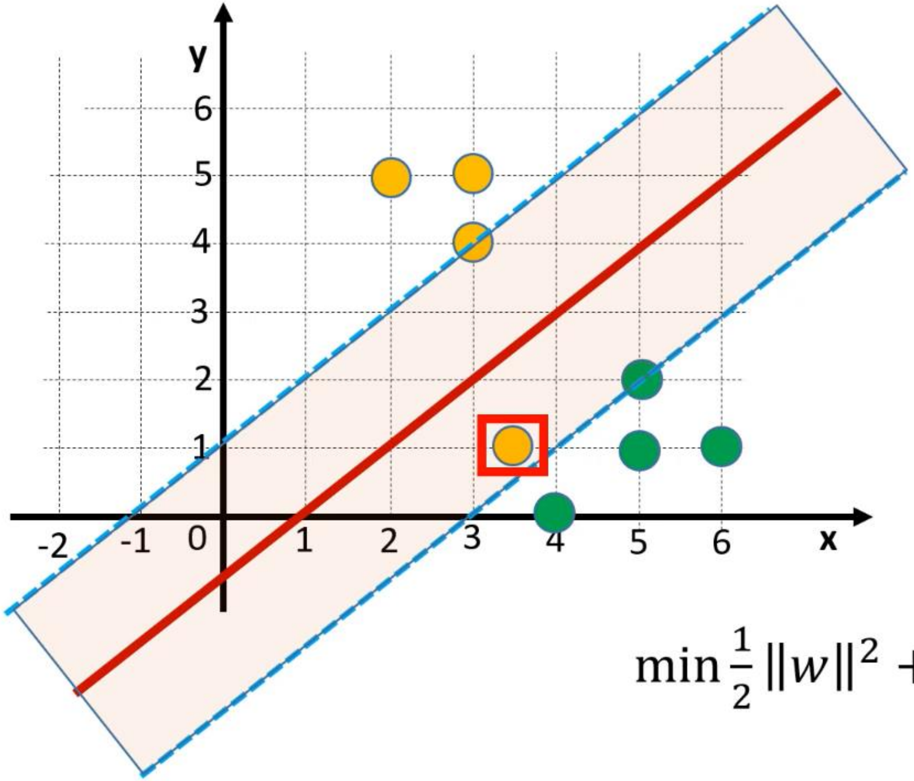
SVM – the math



$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \boxed{\varepsilon_i}$$

where epsilon is a distance measure of the data points from their corresponding blue line. This is called a slack variable in SVM.

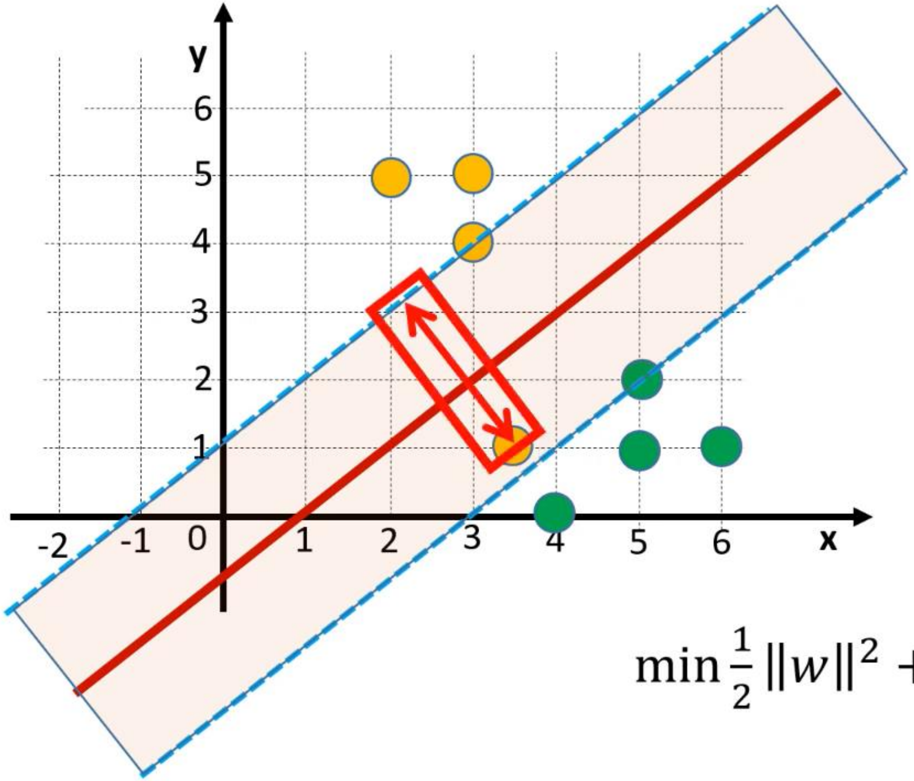
SVM – the math



$$\min \frac{1}{2}\|w\|^2 + C\sum_i^N \varepsilon_i$$

In this example, there is only one data point that is incorrectly classified because it is on the wrong side of the hyperplane.

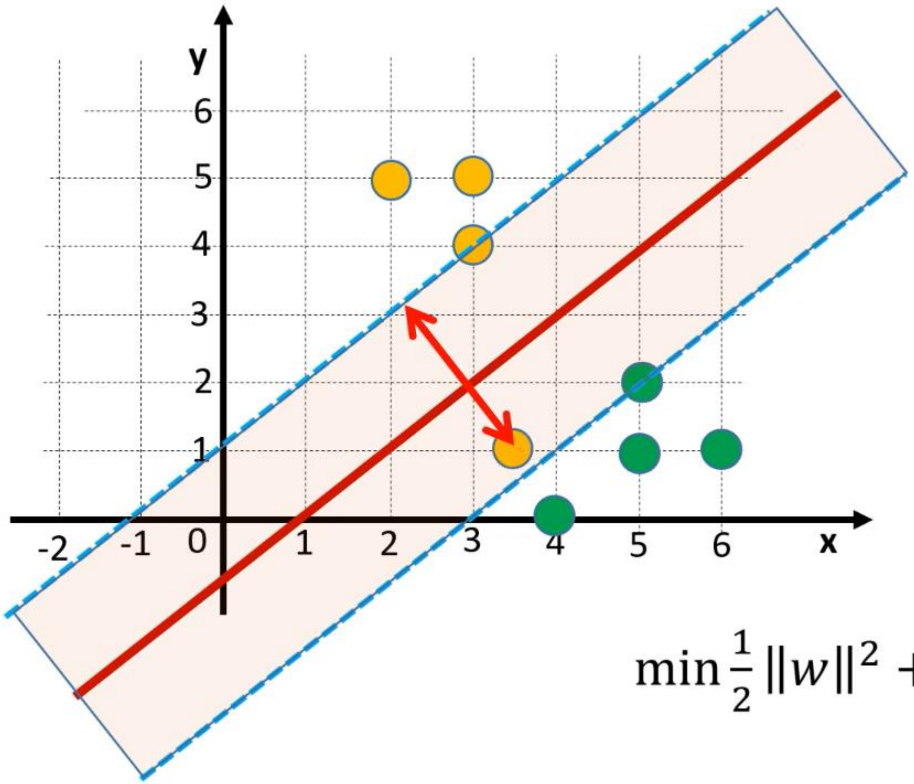
SVM – the math



$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \varepsilon_i$$

The distance between this data point and its corresponding blue line,

SVM – the math

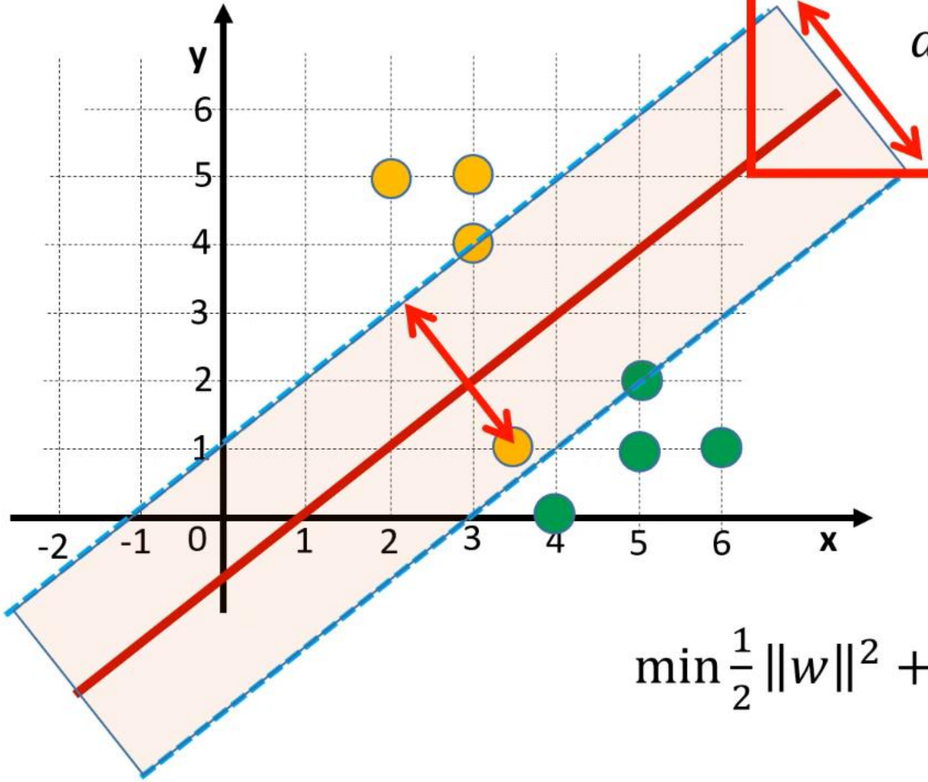


$$\boxed{d = \frac{\varepsilon}{\|w\|}}$$

$$\min \frac{1}{2}\|w\|^2 + C\sum_i^N \varepsilon_i$$

can be described by the following equation.

SVM – the math



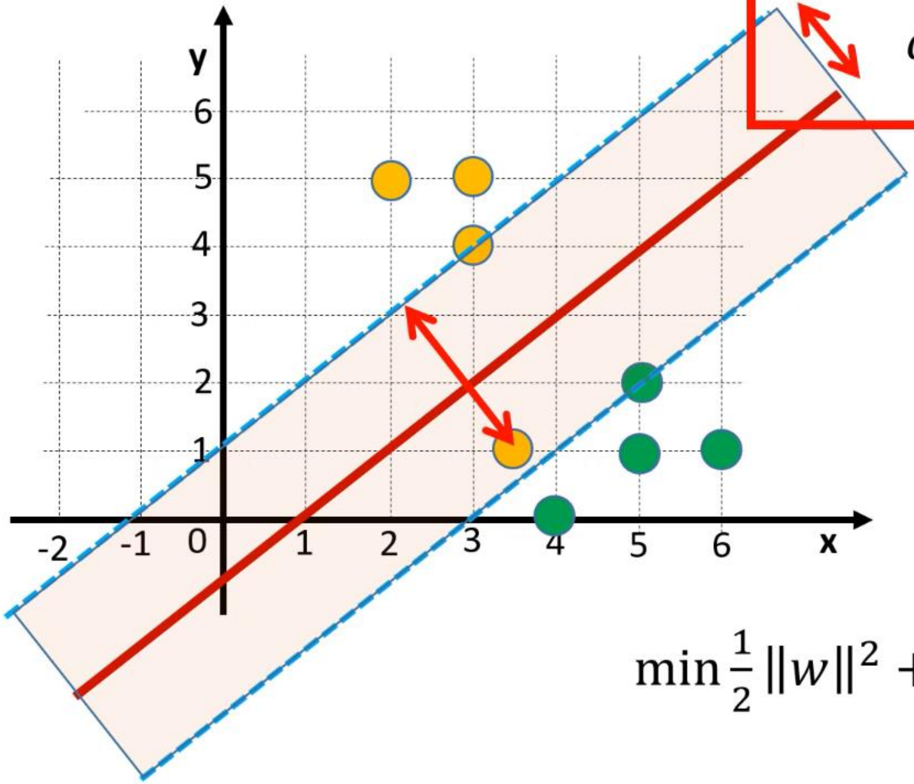
$$d = \frac{2}{\sqrt{A^2 + B^2}} = \frac{2}{\|w\|}$$

$$d = \frac{\varepsilon}{\|w\|}$$

$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \varepsilon_i$$

Remember that the width of the margin can be calculated by the following formula,

SVM – the math



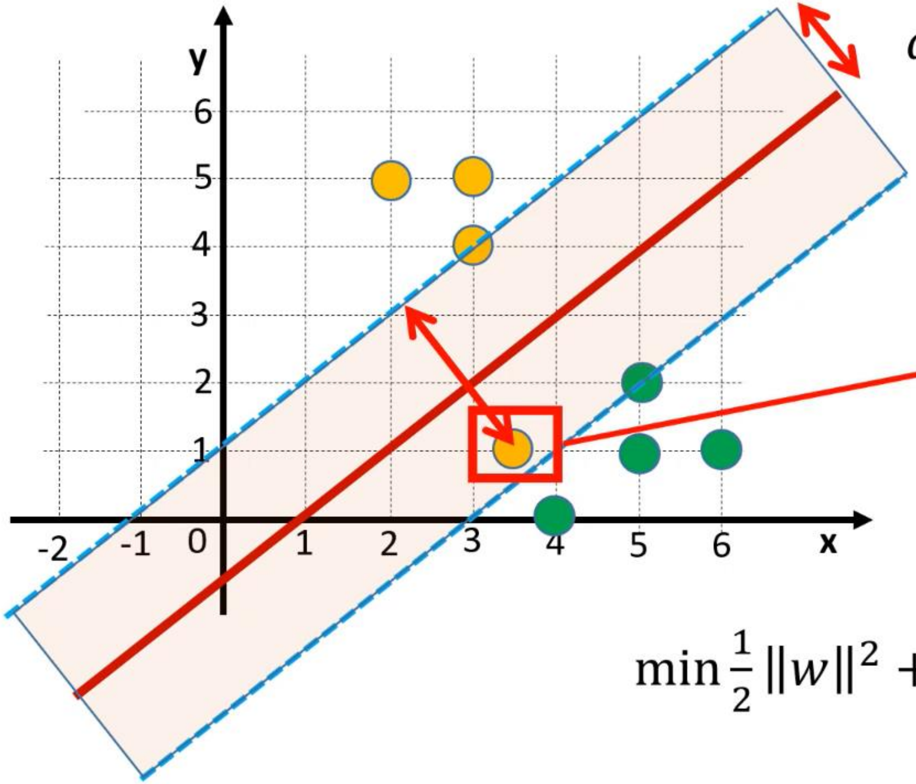
$$d = \frac{1}{\sqrt{A^2 + B^2}} = \frac{1}{\|w\|}$$

$$d = \frac{\varepsilon}{\|w\|}$$

$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \varepsilon_i$$

whereas half the width of the margin, which is the distance between one of the blue lines and the hyperplane, is calculated like this.

SVM – the math



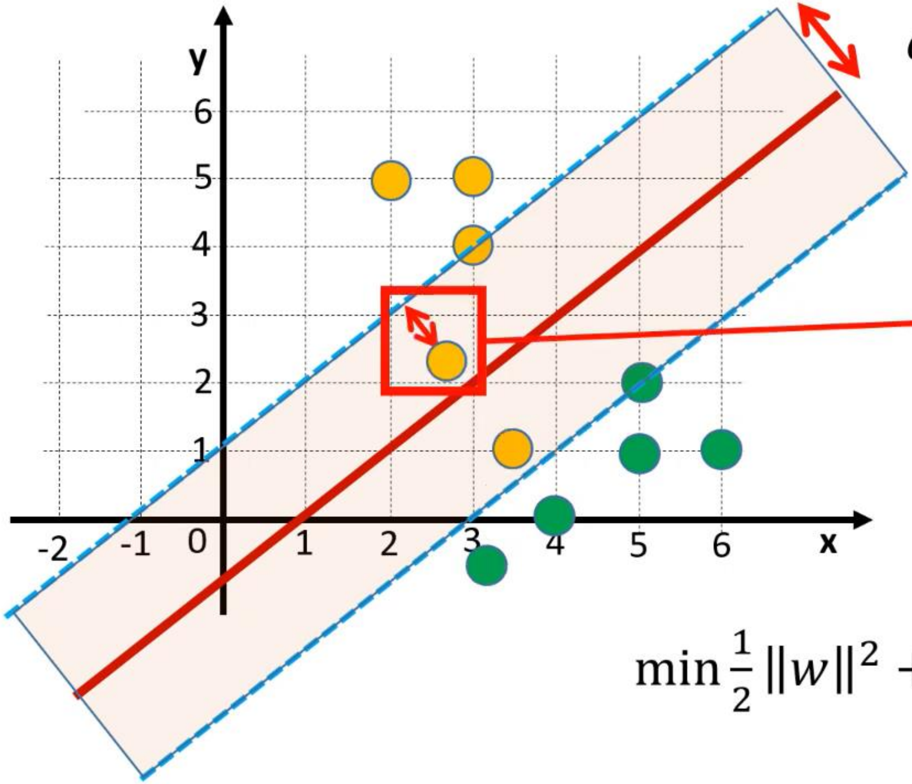
$$d = \frac{1}{\sqrt{A^2 + B^2}} = \frac{1}{\|w\|}$$

$$d = \frac{\varepsilon}{\|w\|}$$

$$\min \frac{1}{2}\|w\|^2 + C\sum_i^N \varepsilon_i$$

This means that epsilon will be greater than one for data points that are on the wrong side of the hyperplane,

SVM – the math



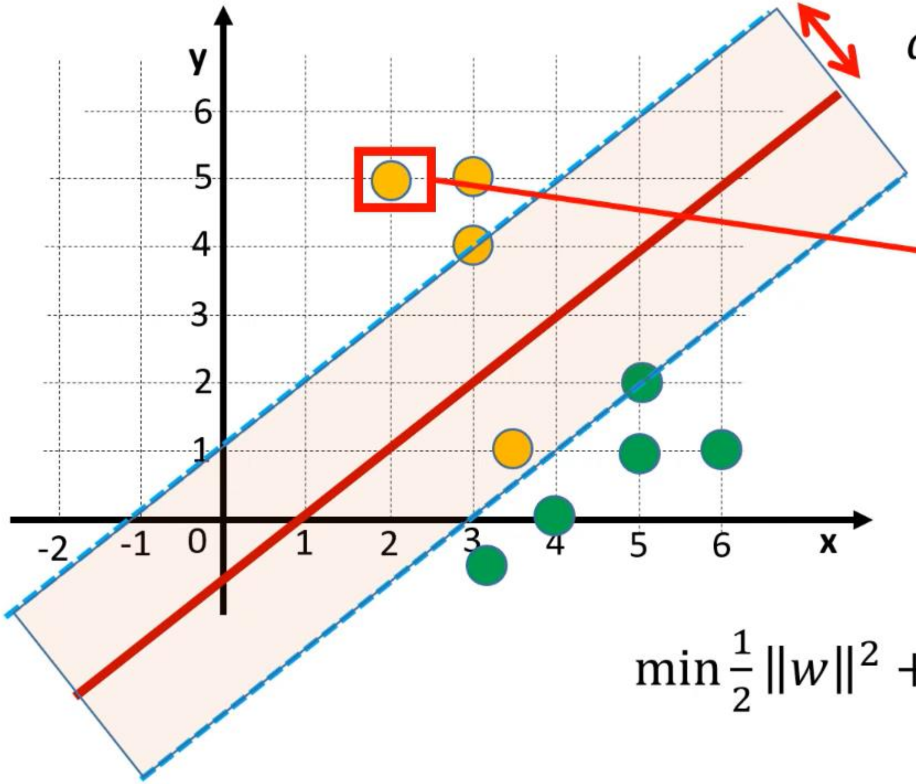
$$d = \frac{1}{\sqrt{A^2 + B^2}} = \frac{1}{\|w\|}$$

$$d = \frac{\varepsilon}{\|w\|}$$

$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \varepsilon_i$$

whereas a data point within the margin, but on the correct side of the hyperplane, will have an epsilon value that is between zero and one.

SVM – the math



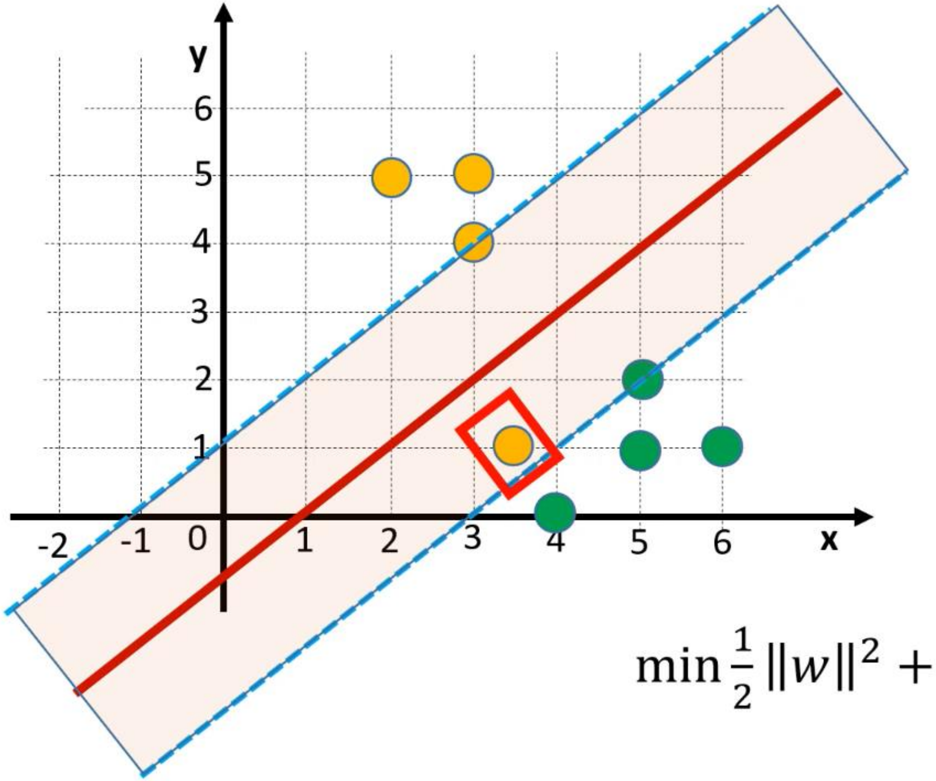
$$d = \frac{1}{\sqrt{A^2 + B^2}} = \frac{1}{\|w\|}$$

$$d = \frac{\varepsilon}{\|w\|}$$

$$\min \frac{1}{2}\|w\|^2 + C\sum_i^N \varepsilon_i$$

Epsilon is set to zero for all data points that are correctly classified and that are outside the margin.

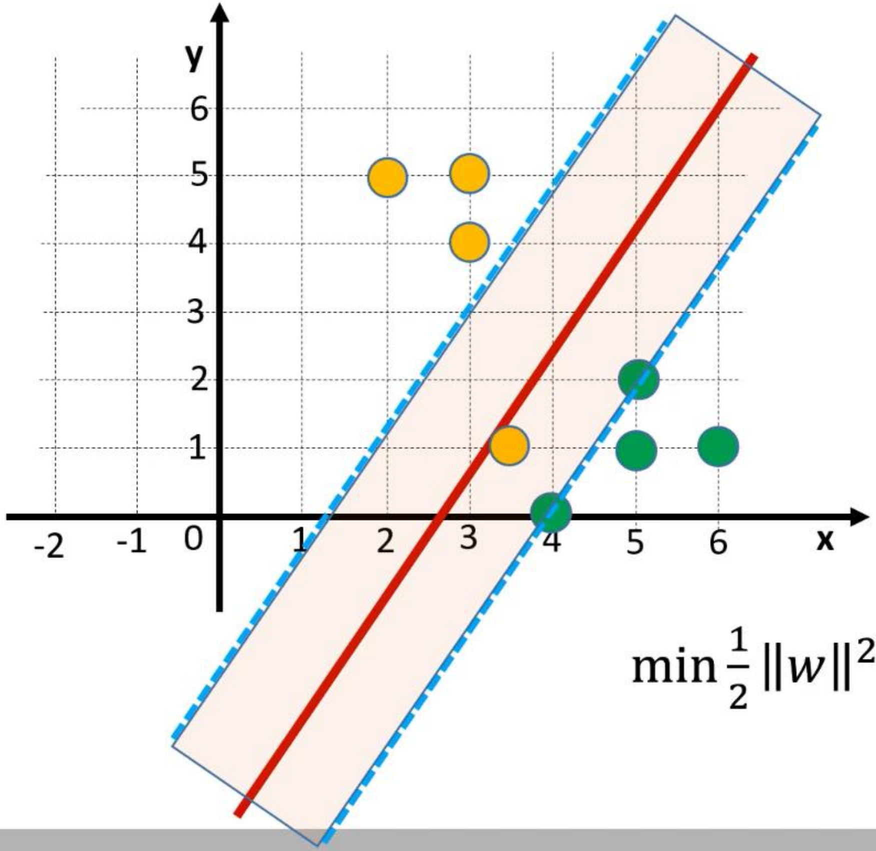
SVM – the math



$$\min \frac{1}{2} \|w\|^2 + C \sum_i^N \varepsilon_i$$

Note that this point is now also considered as a support vector because it will influence the orientation of the hyperplane,

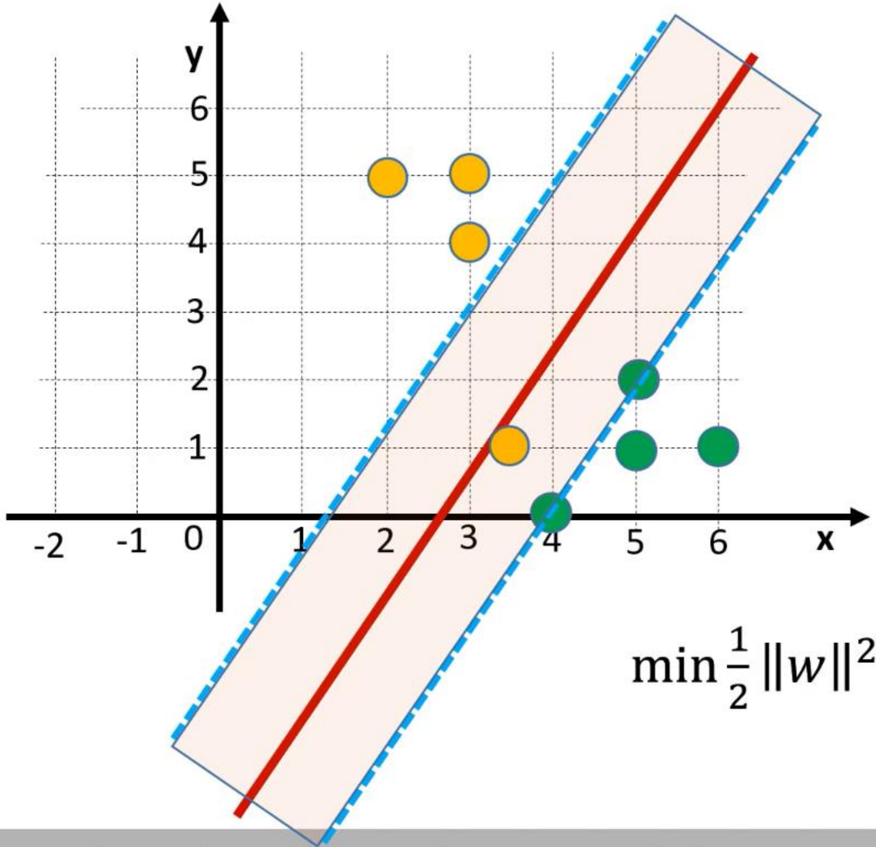
SVM – the math



$$\min \frac{1}{2}\|w\|^2 + \boxed{C} \sum_i^N \varepsilon_i$$

C is a tuning parameter that controls how much weight we should put on the data points that are incorrectly classified.

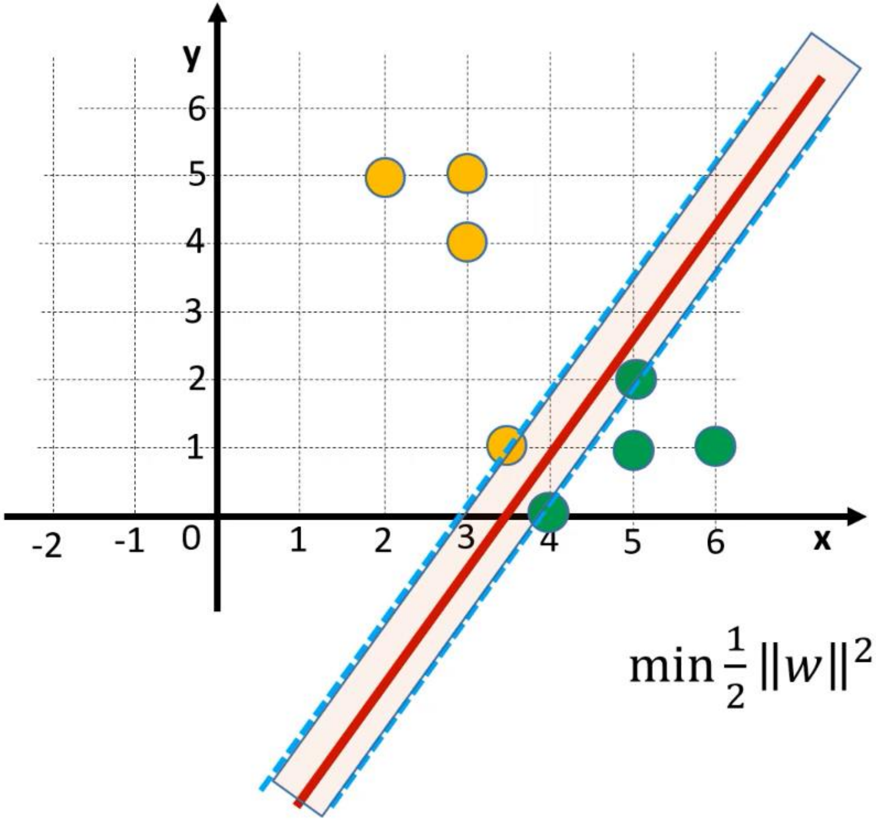
SVM – the math



$$\min \frac{1}{2}\|w\|^2 + \boxed{C} \sum_i^N \varepsilon_i$$

If C is small, we will get a large margin where we put little weight on data points that are misclassified. This will result in a so-called soft margin SVM.

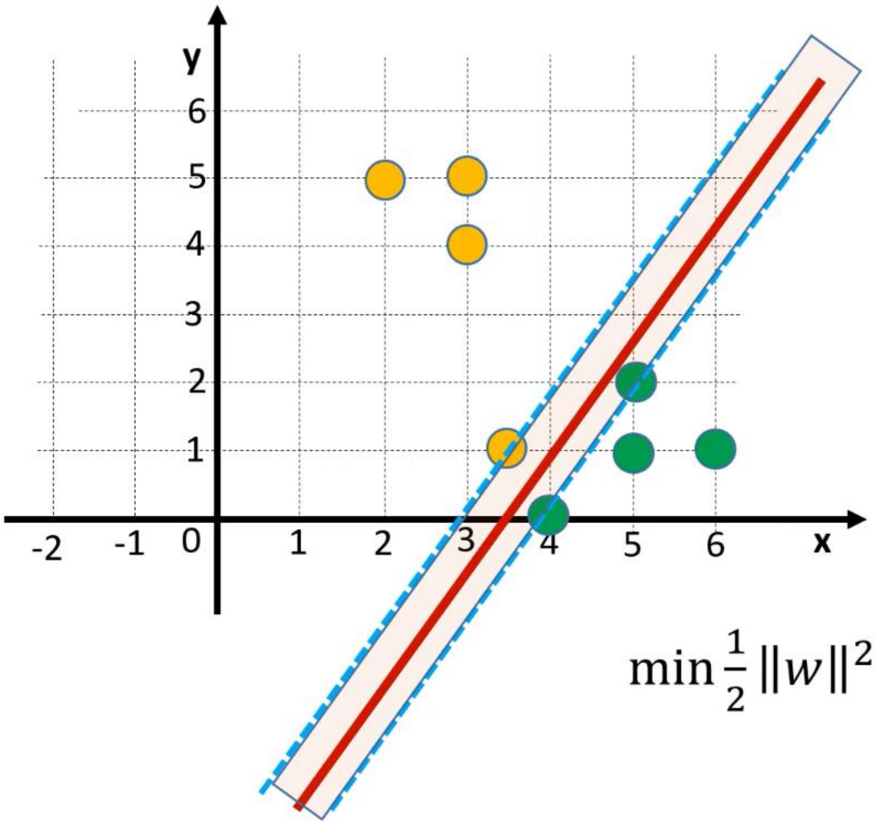
SVM – the math



$$\min \frac{1}{2} \|w\|^2 + \boxed{C} \sum_i^N \varepsilon_i$$

In contrast, if C is large, which puts more weight to avoid a misclassification, will result in a smaller margin. A large value of C will result in a SVM that is less tolerant to outliers that might cause overfitting.

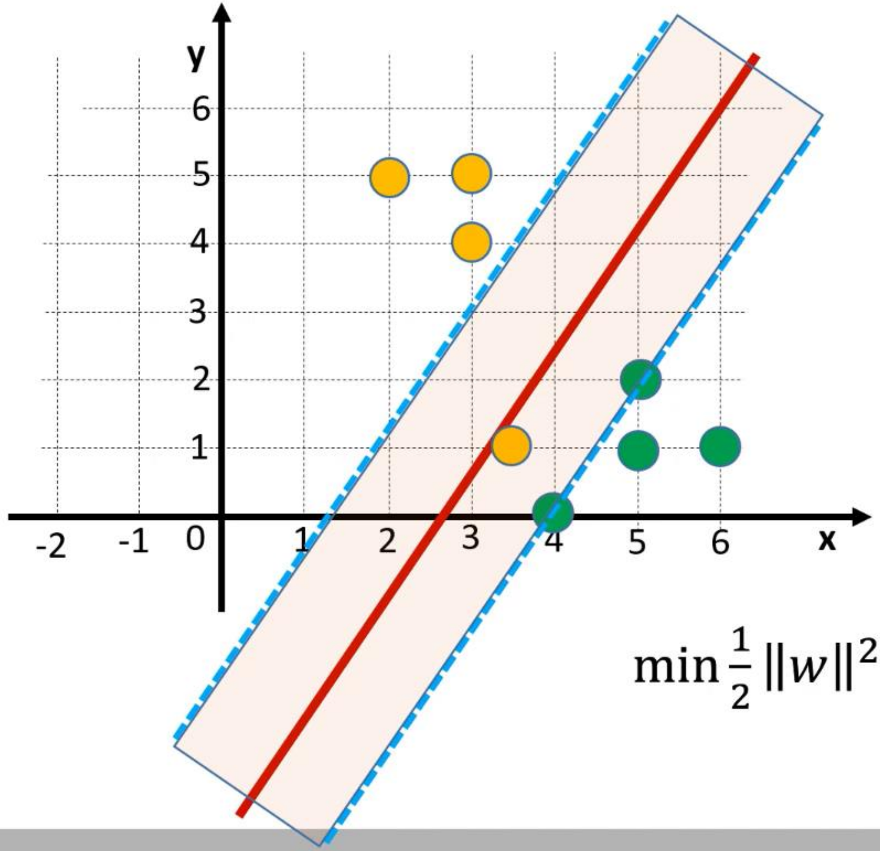
SVM – the math



$$\min \frac{1}{2} \|w\|^2 + \boxed{C} \sum_i^N \varepsilon_i$$

The value of C can be optimized by using, for example, cross-validation where we pick the value of C that results in the best performance according to the cross-validation.

SVM – the math



$$\min \frac{1}{2}\|w\|^2 + C\sum_i^N \varepsilon_i \quad \boxed{\text{such that } Y_i(w^T x_i + b) \geq 1 - \varepsilon_i}$$

Similar to before, optimizing the SVM is constrained to the following condition,