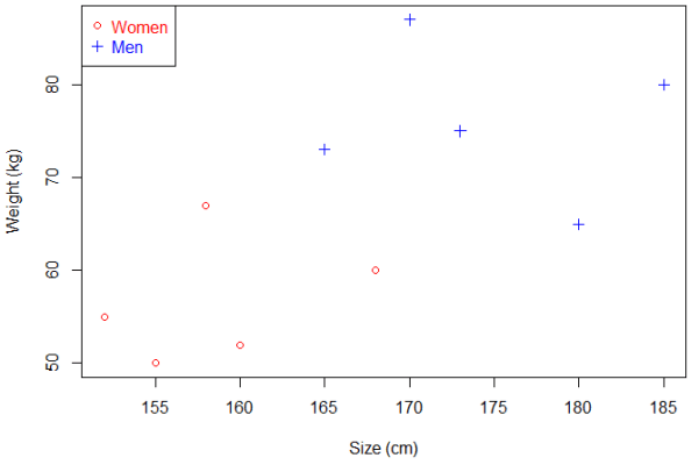
SVM - Understanding the math - Part 1 - The margin

What is the goal of the Support Vector Machine (SVM)? Alexandre KOWALCZYK

The goal of a support vector machine is to find the optimal separating hyperplane which maximizes the margin of the training data.

The first thing we can see from this definition, is that a SVM needs training data. Which means it is a supervised learning algorithm. It is also important to know that SVM is a classification algorithm. Which means we will use it to predict if something belongs to a particular class. For instance, we can have the training data below:



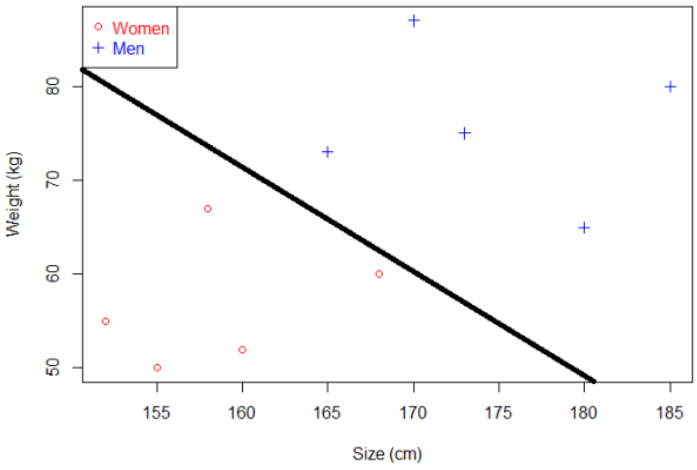
We have plotted the size and weight of several people, and there is also a way to distinguish between men and women. With such data, using a SVM will allow us to answer the following question:

Given a particular data point (weight and size), is the person a man or a woman ?

For instance: if someone measures 175 cm and weights 80 kg, is it a man of a woman?

What is a separating hyperplane?

Just by looking at the plot, we can see that it is possible to separate the data. For instance, we could trace a line and then all the data points representing men will be above the line, and all the data points representing women will be below the line. Such a line is called a **separating hyperplane** and is depicted below:

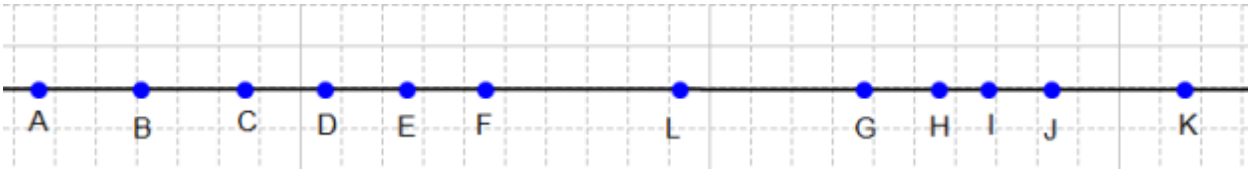


If it is just a line, why do we call it an hyperplane ?

Even though we use a very simple example with data points laying in R^2 R2 the support vector machine can work with any number of dimensions !

An hyperplane is a generalization of a plane.

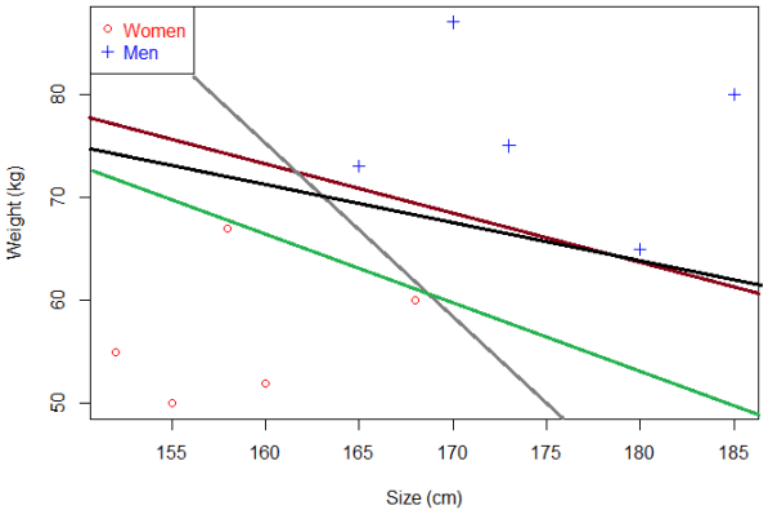
- in one dimension, an hyperplane is called a point
- in two dimensions, it is a line
- in three dimensions, it is a plane
- in more dimensions you can call it an hyperplane



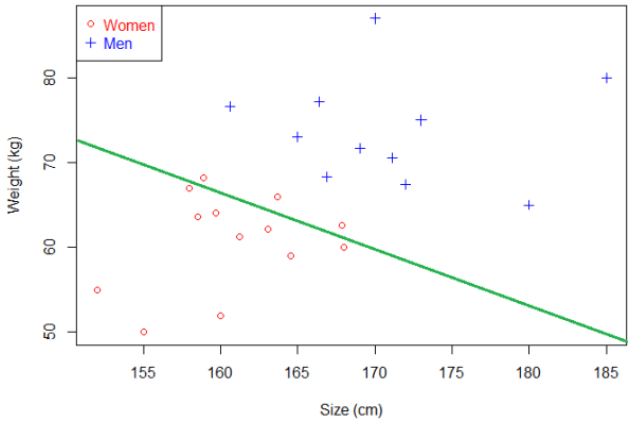
The point L is a separating hyperplane in one dimension

What is the *optimal* separating hyperplane?

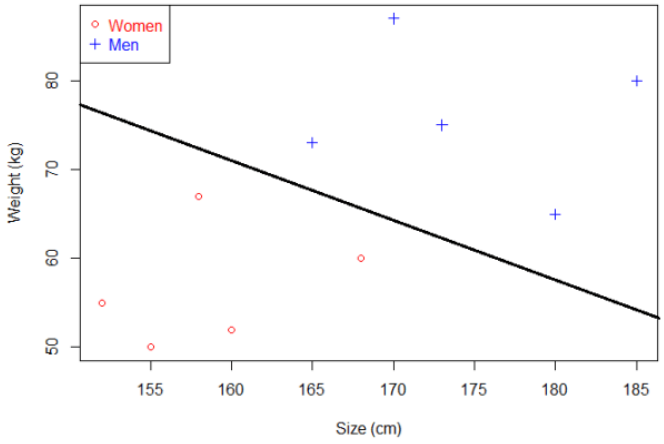
The fact that you can find a **separating hyperplane**, does not mean it is the best one ! In the example below there is several separating hyperplanes. Each of them is valid as it successfully separates our data set with men on one side and women on the other side.



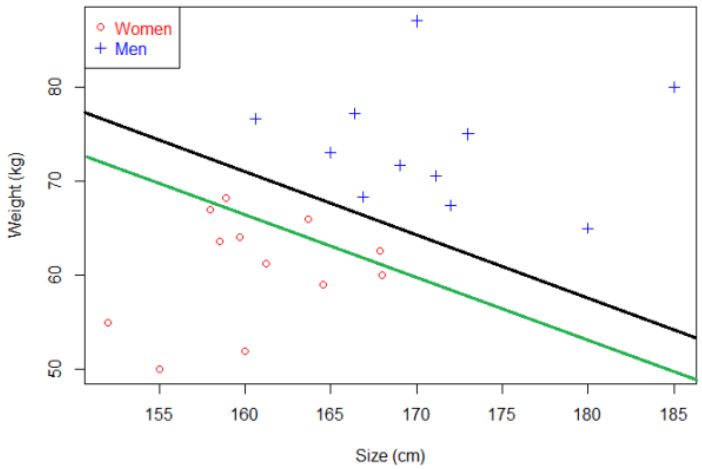
There can be a lot of separating hyperplanes. Suppose we select the green hyperplane and use it to classify on real life data.



This hyperplane does not generalize well. This time, it makes some mistakes as it wrongly classify three women. Intuitively, we can see that *if we select an hyperplane which is close to the data points of one class, then it might not generalize well.* So we will try to select an hyperplane **as far as possible from data points from each category:**



This one looks better. When we use it with real life data, we can see it still make perfect classification.

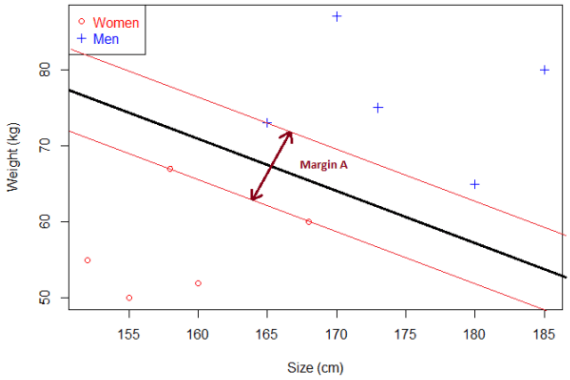


The black hyperplane classifies more accurately than the green one

That's why the objective of a SVM is to **find the optimal separating hyperplane**:

- because it correctly classifies the training data
- and because it is the one which will generalize better with unseen data

What is the margin and how does it help choosing the optimal hyperplane?

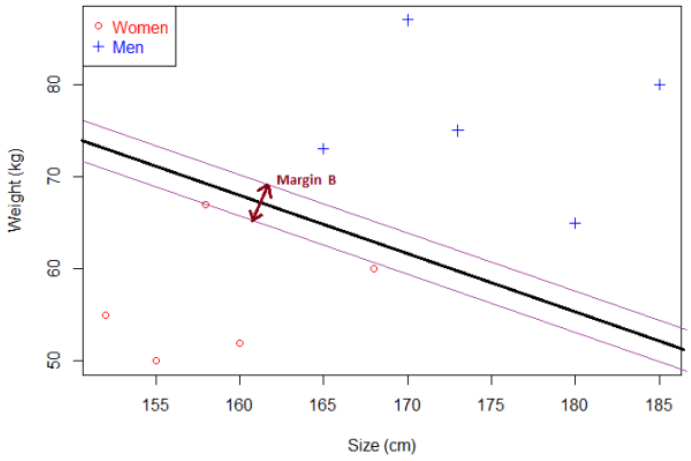


The margin of our optimal hyperplane

Given a particular hyperplane, we can compute the distance between the hyperplane and the closest data point. Once we have this value, if we double it we will get what is called the **margin**.

Basically the margin is a no man's land. There will never be any data point inside the margin. (Note: this can cause some problems when data is noisy, and this is why soft margin classifier will be introduced later)

For another hyperplane, the margin will look like this :



As you can see, Margin B is smaller than Margin A.

We can make the following observations:

- If an hyperplane is very close to a data point, its margin will be small.
- The further an hyperplane is from a data point, the larger its margin will be.

This means that **the optimal hyperplane will be the one with the biggest margin.**

That is why the objective of the SVM **is to find the optimal separating hyperplane which maximizes the margin of the training data.**

This concludes this introductory post about the math behind SVM. There was not a lot of formula, but in the next article we will put on some numbers and try to get the mathematical view of this using geometry and vectors.

But how do we calculate this margin?

SVM = Support VECTOR Machine

In Support Vector Machine, there is the word **vector.**
That means it is important to understand vector well and how to use them.

Here a short sum-up of what we will see today:

- What is a vector?
  - its norm
  - its direction
- How to add and subtract vectors ?
- What is the dot product ?
- How to project a vector onto another ?

Once we have all these tools in our toolbox, we will then see:

- What is the equation of the hyperplane?
- How to compute the margin?

What is a vector?

If we define a point $A(3,4)$ A(3,4) in $\mathbb{R}^2$ R2 we can plot it like this.

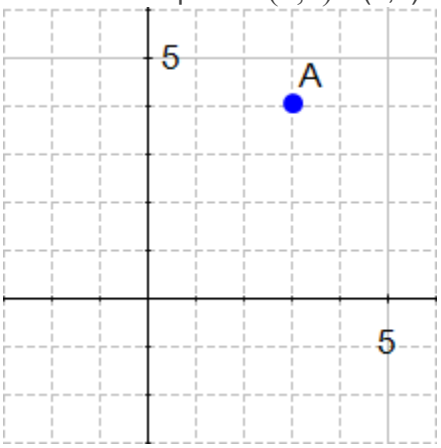


Figure 1: a point

Definition: Any point $x=(x_1,x_2),x\neq 0$ x=(x1,x2),x≠0 *, in* $\mathbb{R}^2$ R2 *specifies a vector in the plane, namely the vector starting at the origin and ending at x.*

This definition means that there exists a vector between the origin and A.

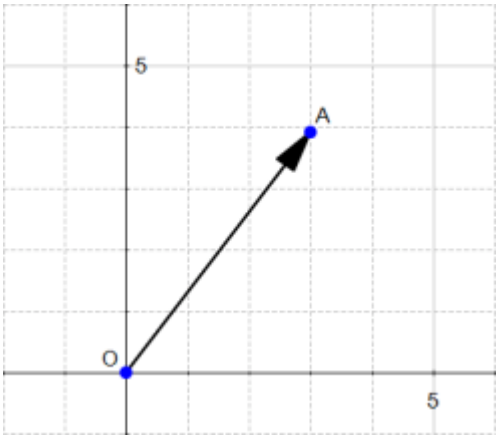


Figure 2 - a vector

If we say that the point at the origin is the point $O(0,0)$ O(0,0) then the vector above is the vector $OA \rightarrow$ OA→ . We could also give it an arbitrary name such as $\mathbf{u}$ u .

Note: You can notice that we write vector either with an arrow on top of them, or in bold, in the rest of this text I will use the arrow when there is two letters like $OA \rightarrow$ OA→ and the bold notation otherwise.

Ok so now we know that there is a vector, but we still don't know what **IS** a vector.

Definition: A vector is an object that has both a magnitude and a direction.

We will now look at these two concepts.

1) The magnitude of a vector

The magnitude or length of a vector x x **is written** $\|x\|$ ‖x‖ **and is called its norm.**

For our vector $OA \rightarrow$ OA→ , $\|OA\|$ ‖OA‖ is the length of the segment OA OA

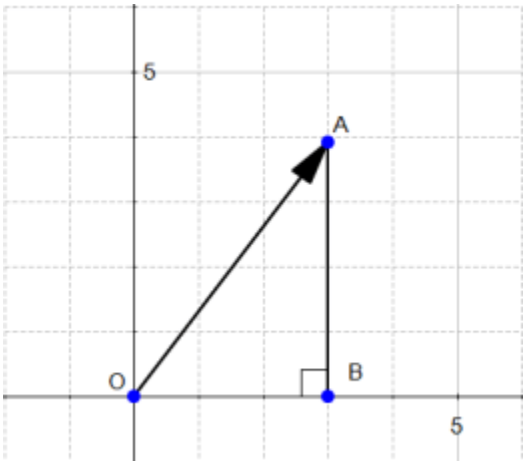


Figure 3

From Figure 3 we can easily calculate the distance OA using Pythagoras' theorem:

$$OA^2 = OB^2 + AB^2 \quad \text{OA2=OB2+AB2}$$

$$OA^2 = 3^2 + 4^2 \quad \text{OA2=32+42}$$

$$OA^2 = 25 \quad \text{OA2=25}$$

$$OA=25 -- \sqrt{} \quad \text{OA=25}$$

$$\|OA\|=OA=5 \quad \|\text{OA}\|=\text{OA}=5$$

2) The direction of a vector

The direction is the second component of a vector.

*Definition : The **direction** of a vector* $\mathbf{u}(u_1,u_2)$ u(u1,u2) *is the vector* $\mathbf{w}(u_1\|u\|,u_2\|u\|)$ w(u1‖u‖,u2‖u‖)

Where does the coordinates of $\mathbf{w}$ w come from ?

Understanding the definition

To find the direction of a vector, we need to use its angles.

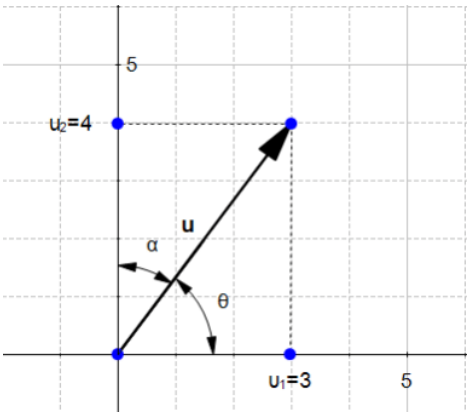


Figure 4 - direction of a vector

Figure 4 displays the vector $\mathbf{u}(u_1,u_2)$ u(u1,u2) with $u_1=3$ u1=3 and $u_2=4$ u2=4
We could say that :

Naive definition 1: The direction of the vector $\mathbf{u}$ u *is defined by the angle* θ θ *with respect to the horizontal axis, and with the angle* α α *with respect to the vertical axis.*
This is tedious. Instead of that we will use the cosine of the angles.

In a right triangle, the cosine of an angle β β is defined by :

$$cos(\beta)=adjacenthypotenuse \quad \cos(\beta)=\text{adjacenthypotenuse}$$

In Figure 4 we can see that we can form two right triangles, and in both case the adjacent side will be on one of the axis. Which means that the definition of the cosine implicitly contains the axis related to an angle. We can rephrase our naïve definition to :

Naive definition 2: The direction of the vector $\mathbf{u}$ u *is defined by the cosine of the angle* θ θ *and the cosine of the angle* α α *.*
Now if we look at their values :

$$cos(\theta)=u_1\|u\| \quad \cos(\theta)=u1\|u\|$$

$$cos(\alpha)=u_2 \|u\| \quad \cos(\alpha)=u2\|u\|$$

Hence the original definition of the vector $\mathbf{w}$ w . That's why its coordinates are also called *direction cosine*.

Computing the direction vector

We will now compute the direction of the vector $\mathbf{u}$ u from Figure 4.:

$$cos(\theta)=u_1 \|u\| =35 =0.6 \quad \cos(\theta)=u1\|u\|=35=0.6$$

and

$$cos(\alpha)=u_2 \|u\| =45 =0.8 \quad \cos(\alpha)=u2\|u\|=45=0.8$$

The direction of $\mathbf{u}(3,4)$ u(3,4) is the vector $\mathbf{w}(0.6,0.8)$ w(0.6,0.8)
If we draw this vector we get Figure 5:

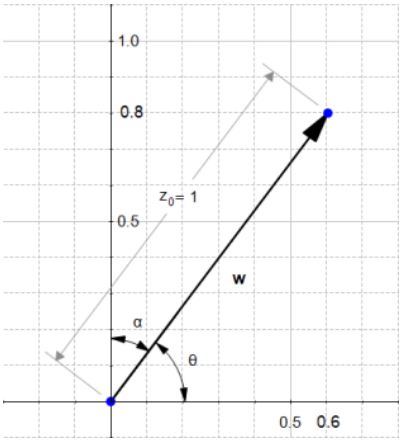


Figure 5: the direction of u

We can see that $\mathbf{w}$ w as indeed the same look as $\mathbf{u}$ u except it is smaller. Something interesting about direction vectors like $\mathbf{w}$ w is that their norm is equal to 1. That's why we often call them **unit vectors**.

The sum of two vectors

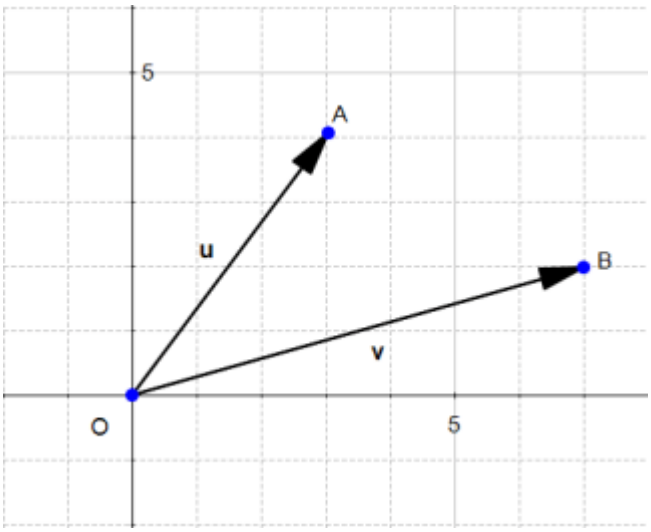


Figure 6: two vectors u and v

Given two vectors $\mathbf{u}(u_1, u_2)$ u(u1,u2) and $\mathbf{v}(v_1, v_2)$ v(v1,v2) then :

$$\mathbf{u}+\mathbf{v}=(u_1+v_1, u_2+v_2)$$ u+v=(u1+v1,u2+v2)

Which means that adding two vectors gives us **a third vector** whose coordinate are the sum of the coordinates of the original vectors.

You can convince yourself with the example below:

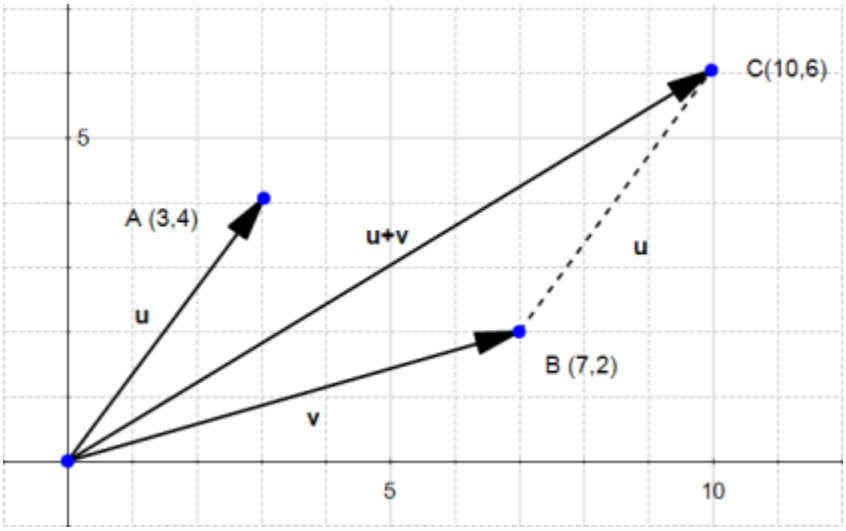


Figure 7: the sum of two vectors

The difference between two vectors

The difference works the same way :

$$\mathbf{u}-\mathbf{v}=(u_1-v_1, u_2-v_2)$$ u−v=(u1−v1,u2−v2)

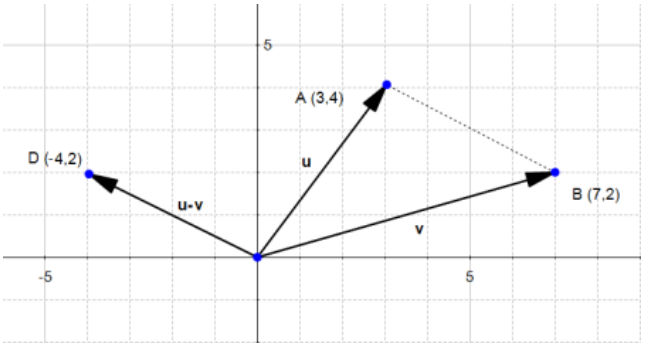


Figure 8: the difference of two vectors

Since the subtraction is not commutative, we can also consider the other case:

$$\mathbf{v}-\mathbf{u}=(v_1-u_1,v_2-u_2)$$ v−u=(v1−u1,v2−u2)

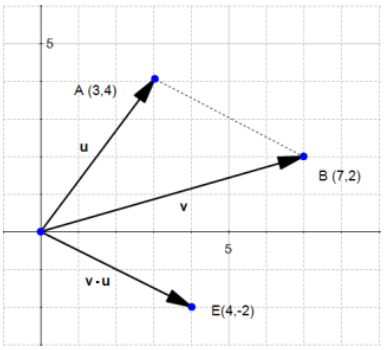


Figure 9: the difference v-u

The last two pictures describe the "*true*" vectors generated by the difference of $\mathbf{u}$ u and $\mathbf{v}$ v .
However, since a vector has a magnitude and a direction, we often consider that parallel translate of a given vector (vectors with the same magnitude and direction but with a different origin) are the same vector, just drawn in a different place in space.

So don't be surprised if you meet the following :

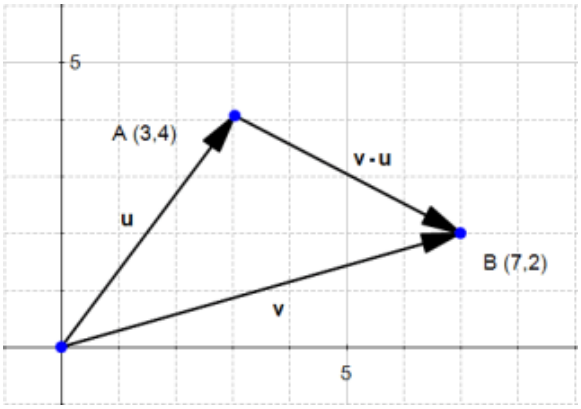


Figure 10: another way to view the difference v-u

and

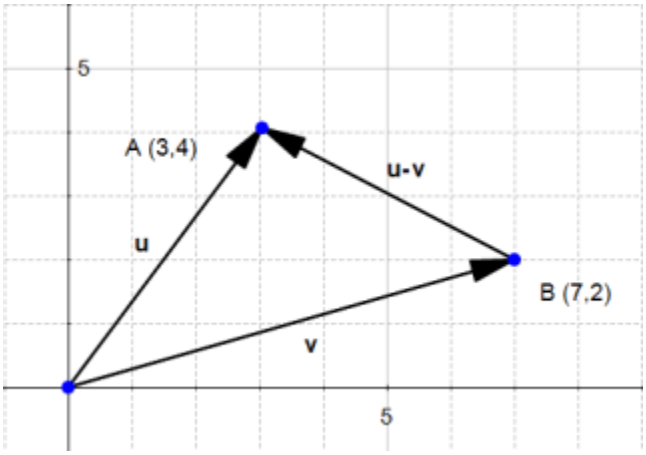


Figure 11: another way to view the difference u-v

If you do the math, it looks wrong, because the end of the vector $\mathbf{u}-\mathbf{v}$ u−v is not in the right point, but it is a convenient way of thinking about vectors which you'll encounter often.

The dot product

One **very** important notion to understand SVM is the dot product.

Definition: Geometrically, it is the product of the Euclidian magnitudes of the two vectors and the cosine of the angle between them

Which means if we have two vectors $\mathbf{x}$ x and $\mathbf{y}$ y and there is an angle θ θ (theta) between them, their dot product is :

$$\mathbf{x}\cdot\mathbf{y}=\|x\|\|y\|cos(\theta)$$ x·y=‖x‖‖y‖cos(θ)

Why ?

To understand let's look at the problem geometrically.

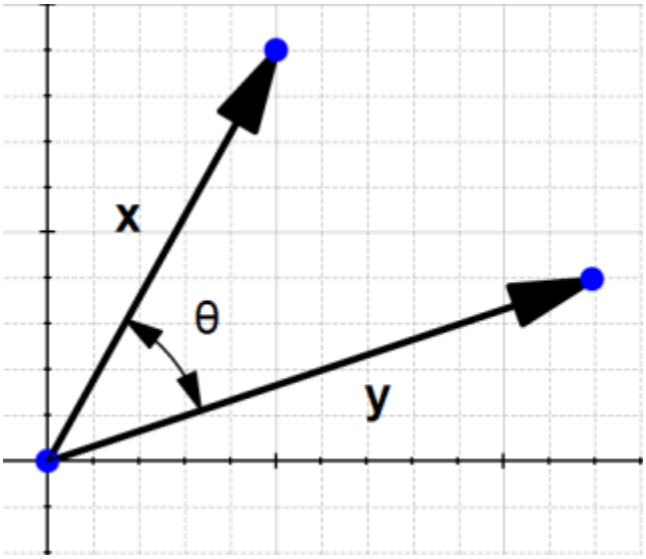


Figure 12

In the definition, they talk about $cos(\theta)$ cos(θ) , let's see what it is.
By definition we know that in a right-angled triangle:

$$cos(\theta)=adjacenthypotenuse$$ cos(θ)=adjacenthypotenuse

In our example, we don't have a right-angled triangle.

However if we take a different look Figure 12 we can find two right-angled triangles formed by each vector with the horizontal axis.

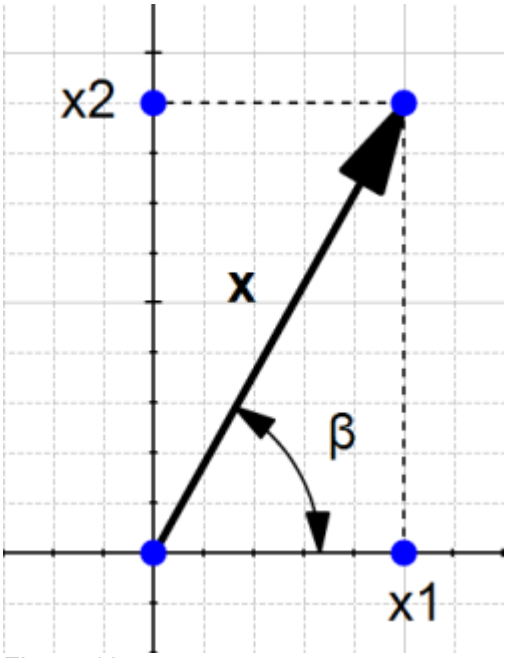


Figure 13

and

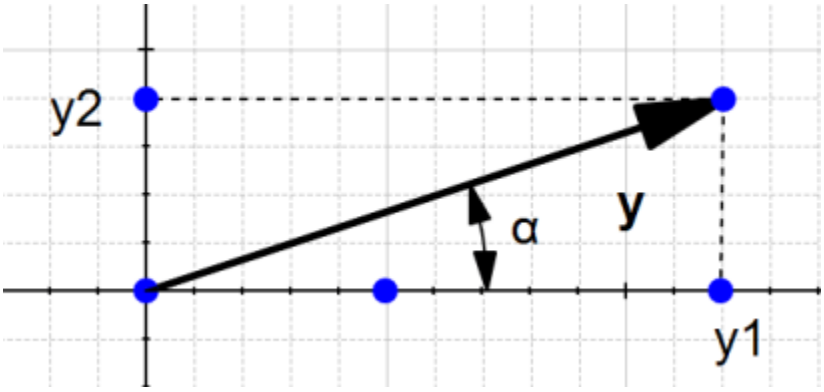


Figure 14

So now we can view our original schema like this:

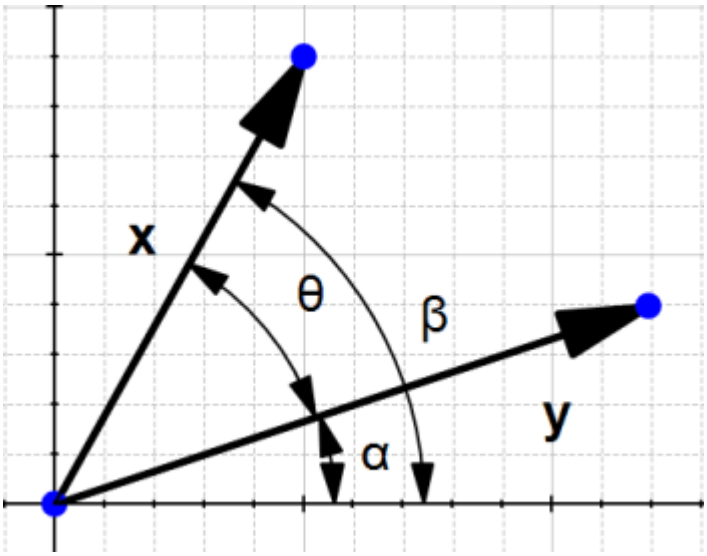


Figure 15

We can see that

$$\theta=\beta-\alpha \quad \theta=\beta-\alpha$$

So computing $cos(\theta)$ cos(θ) is like computing $cos(\beta-\alpha)$ cos(β−α)
There is a special formula called the *difference identity for cosine* which says that:

$$cos(\beta-\alpha)=cos(\beta)cos(\alpha)+sin(\beta)sin(\alpha) \quad \cos(\beta-\alpha)=\cos(\beta)\cos(\alpha)+\sin(\beta)\sin(\alpha)$$

(if you want you can read the demonstration here)

Let's use this formula!

$$cos(\beta)=adjacenthypotenuse=x_1\|x\| \quad \cos(\beta)=\text{adjacenthypotenuse}=x1\|x\|$$

$$sin(\beta)=oppositehypotenuse=x_2\|x\| \quad \sin(\beta)=\text{oppositehypotenuse}=x2\|x\|$$

$$cos(\alpha)=adjacenthypotenuse=y_1\|y\| \quad \cos(\alpha)=\text{adjacenthypotenuse}=y1\|y\|$$

$$sin(\alpha)=oppositehypotenuse=y_2\|y\| \quad \sin(\alpha)=\text{oppositehypotenuse}=y2\|y\|$$

So if we replace each term

$$cos(\theta)=cos(\beta-\alpha)=cos(\beta)cos(\alpha)+sin(\beta)sin(\alpha) \quad \cos(\theta)=\cos(\beta-\alpha)=\cos(\beta)\cos(\alpha)+\sin(\beta)\sin(\alpha)$$

$$cos(\theta)=x_1\|x\|y_1\|y\|+x_2\|x\|y_2\|y\| \quad \cos(\theta)=x1\|x\|y1\|y\|+x2\|x\|y2\|y\|$$

$$cos(\theta)=x_1y_1+x_2y_2\|x\|\|y\| \quad \cos(\theta)=x1y1+x2y2\|x\|\|y\|$$

If we multiply both sides by $\|x\|\|y\|$ ‖x‖‖y‖ we get:

$$\|x\|\|y\|cos(\theta)=x_1y_1+x_2y_2 \quad \|x\|\|y\|\cos(\theta)=x1y1+x2y2$$

Which is the same as :

$$\|x\|\|y\|cos(\theta)=\mathbf{x}\cdot\mathbf{y}$$ ‖x‖‖y‖cos(θ)=x·y

We just found the geometric definition of the dot product !

Eventually from the two last equations we can see that :

$$\mathbf{x}\cdot\mathbf{y}=x_1y_1+x_2y_2=\sum_{i=1}^{2}(x_iy_i)$$ x·y=x1y1+x2y2=∑i=12(xiyi)

This is the algebraic definition of the dot product !

A few words on notation

The dot product is called like that because we write a dot between the two vectors.
Talking about the dot product $\mathbf{x}\cdot\mathbf{y}$ x·y is the same as talking about

- the **inner product** $\langle x,y\rangle$ ⟨x,y⟩ (in linear algebra)
- **scalar product** because we take the product of two vectors and it returns a scalar (a real number)

The orthogonal projection of a vector

Given two vectors $\mathbf{x}$ x and $\mathbf{y}$ y , we would like to find the orthogonal projection of $\mathbf{x}$ x onto $\mathbf{y}$ y .

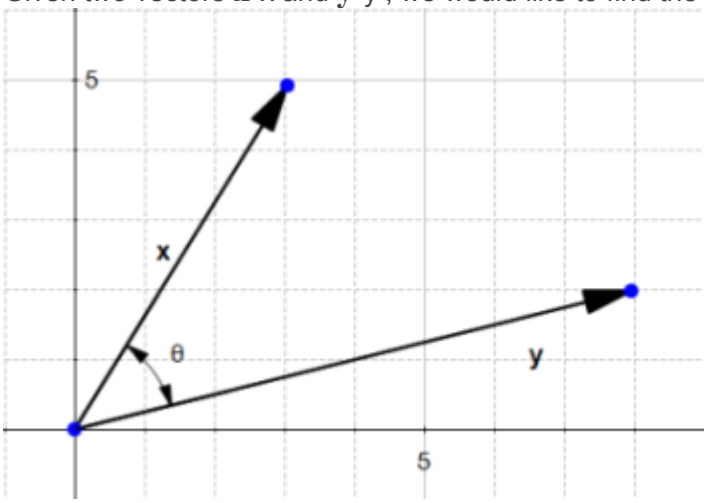


Figure 16

To do this we project the vector $\mathbf{x}$ x onto $\mathbf{y}$ y

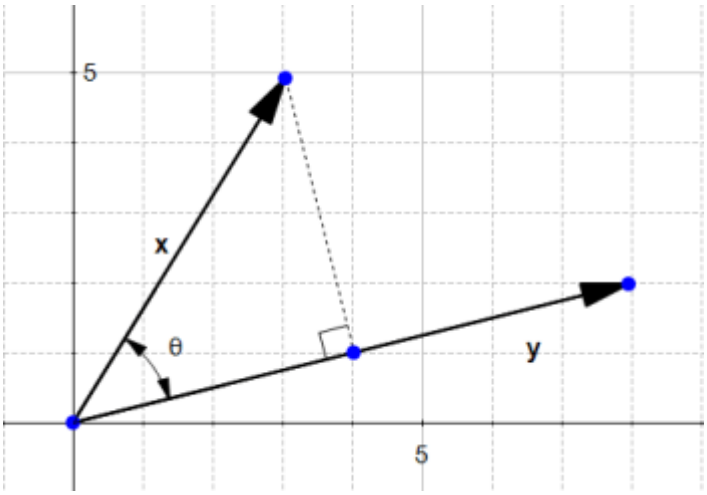


Figure 17

This give us the vector $\mathbf{z}$ z

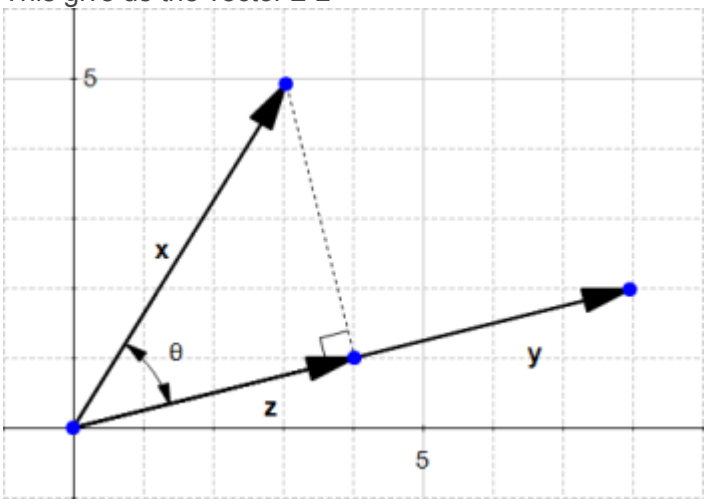


Figure 18 : z is the projection of x onto y

By definition :

$$cos(\theta)=\|z\|\|x\| \quad \cos(\theta)=\|z\|\|x\|$$

$$\|z\|=\|x\|cos(\theta) \quad \|z\|=\|x\|\cos(\theta)$$

We saw in the section about the dot product that

$$cos(\theta)=\mathbf{x}\cdot\mathbf{y}\|x\|\|y\| \quad \cos(\theta)=x\cdot y\|x\|\|y\|$$

So we replace $cos(\theta)$ cos(θ) in our equation:

$$\|z\|=\|x\|\mathbf{x}\cdot\mathbf{y}\|x\|\|y\| \quad \|z\|=\|x\|x\cdot y\|x\|\|y\|$$

$$\|z\|=\mathbf{x}\cdot\mathbf{y}\|y\| \quad \|z\|=x\cdot y\|y\|$$

If we define the vector $\mathbf{u}$ u as the **direction** of $\mathbf{y}$ y then

$$\mathbf{u}=\mathbf{y}\|y\| \quad u=y\|y\|$$

and

$$\|z\| = \mathbf{u} \cdot \mathbf{x} \quad \|z\| = u \cdot x$$

We now have a simple way to compute the norm of the vector $\mathbf{z}$ z .
Since this vector is in the same direction as $\mathbf{y}$ y it has the direction $\mathbf{u}$ u

$$\mathbf{u} = \frac{\mathbf{z}}{\|z\|} \quad u = \frac{z}{\|z\|}$$

$$\mathbf{z} = \|z\| \mathbf{u} \quad z = \|z\| u$$

And we can say :

The vector $\mathbf{z} = (\mathbf{u} \cdot \mathbf{x})\mathbf{u}$ z=(u·x)u *is the orthogonal projection of* $\mathbf{x}$ x *onto* $\mathbf{y}$ y .

Why are we interested by the orthogonal projection ? Well in our example, it allows us to compute the distance between $\mathbf{x}$ x and the line which goes through $\mathbf{y}$ y .

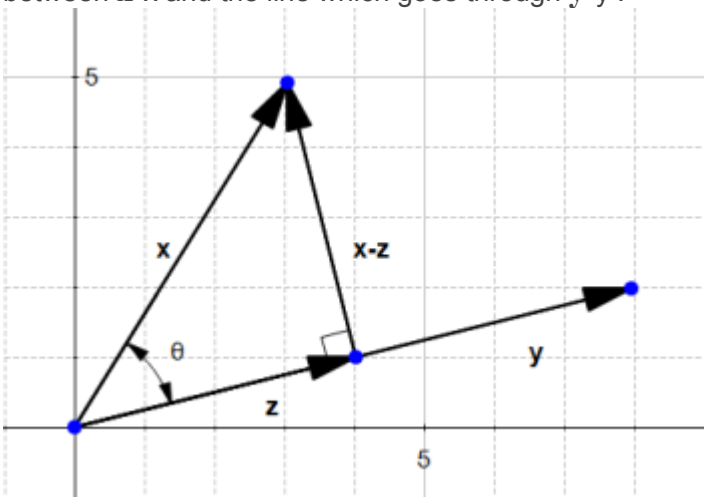


Figure 19

We see that this distance is $\|x - z\|$ ‖x−z‖

$$\|x - z\| = \sqrt{(3-4)^2 + (5-1)^2} = \sqrt{17} \quad \|x-z\| = \sqrt{(3-4)2+(5-1)2} = \sqrt{17}$$

The SVM hyperplane

Understanding the equation of the hyperplane

You probably learnt that an equation of a line is : $y = ax + b$ y=ax+b . However when reading about hyperplane, you will often find that the equation of an hyperplane is defined by :

$$\mathbf{w}^T \mathbf{x} = 0 \quad wTx=0$$

How does these two forms relate ?
In the hyperplane equation you can see that the name of the variables are in bold. Which means that they are vectors ! Moreover, $\mathbf{w}^T \mathbf{x}$ wTx is how we compute the inner product of two vectors, and if you recall, the inner product is just another name for the dot product !
Note that

$$y = ax + b \quad y=ax+b$$

is the same thing as

$$y - ax - b = 0 \quad y-ax-b=0$$

Given two vectors $\mathbf{w}(-b,-a,1)$ w(−b,−a,1) and $\mathbf{x}(1,x,y)$ x(1,x,y)

$$\mathbf{w}^T\mathbf{x}=-b\times(1)+(-a)\times x+1\times y \quad \text{wTx=−b×(1)+(−a)×x+1×y}$$

$$\mathbf{w}^T\mathbf{x}=y-ax-b \quad \text{wTx=y−ax−b}$$

The two equations are just different ways of expressing the same thing.

It is interesting to note that w_0 w0 is $-b$ −b , which means that this value determines the intersection of the line with the vertical axis.
Why do we use the hyperplane equation $\mathbf{w}^T\mathbf{x}$ wTx instead of $y=ax+b$ y=ax+b ?
For two reasons:

- it is easier to work in more than two dimensions with this notation,
- the vector $\mathbf{w}$ w will always be normal to the hyperplane

And this last property will come in handy to compute the distance from a point to the hyperplane.

Compute the distance from a point to the hyperplane

In Figure 20 we have an hyperplane, which separates two group of data.

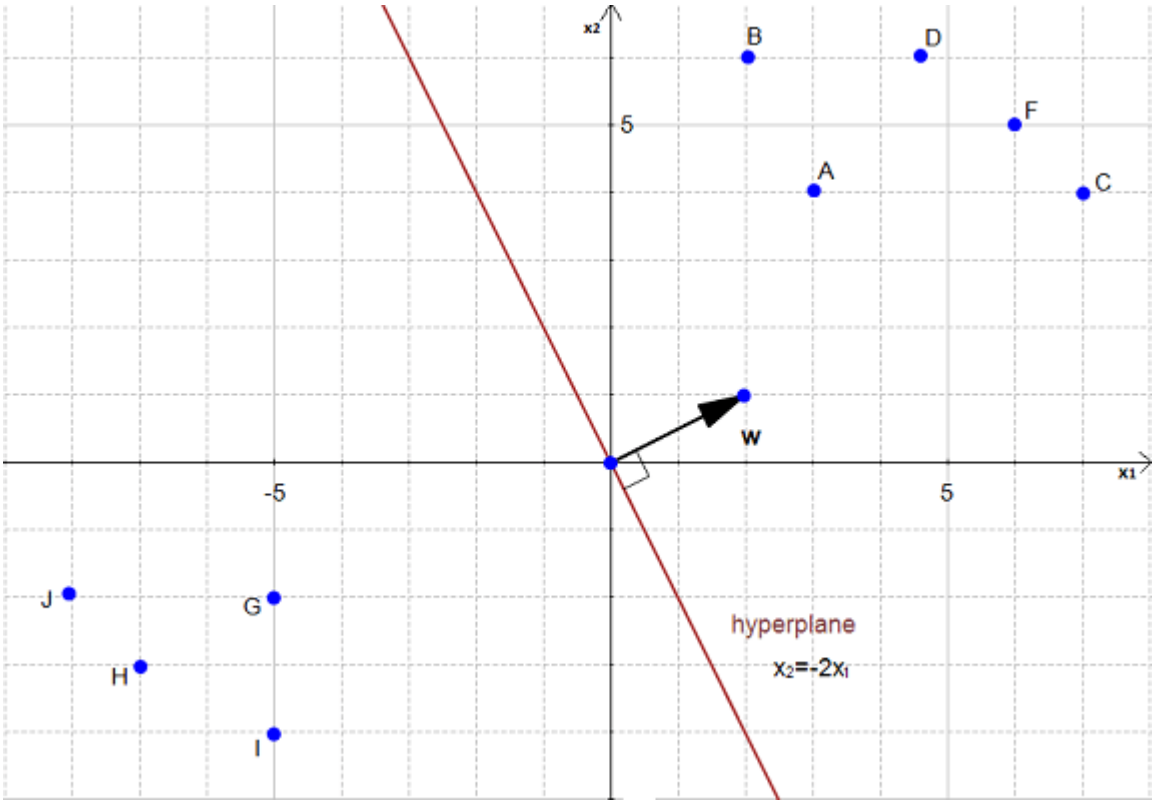


Figure 20

To simplify this example, we have set $w_0=0$ w0=0 .
As you can see on the Figure 20, the equation of the hyperplane is :

$$x_2=-2x_1 \quad \text{x2=−2x1}$$

which is equivalent to

$\mathbf{w}^T\mathbf{x}=0$ wTx=0

with $\mathbf{w}(2,1)$ w(2,1) and $\mathbf{x}(x_1, x_2)$ x(x1,x2)

Note that the vector $\mathbf{w}$ w is shown on the Figure 20. (w is not a data point)

We would like to compute the distance between the point $A(3,4)$ A(3,4) and the hyperplane.

This is the distance between A A and its projection onto the hyperplane

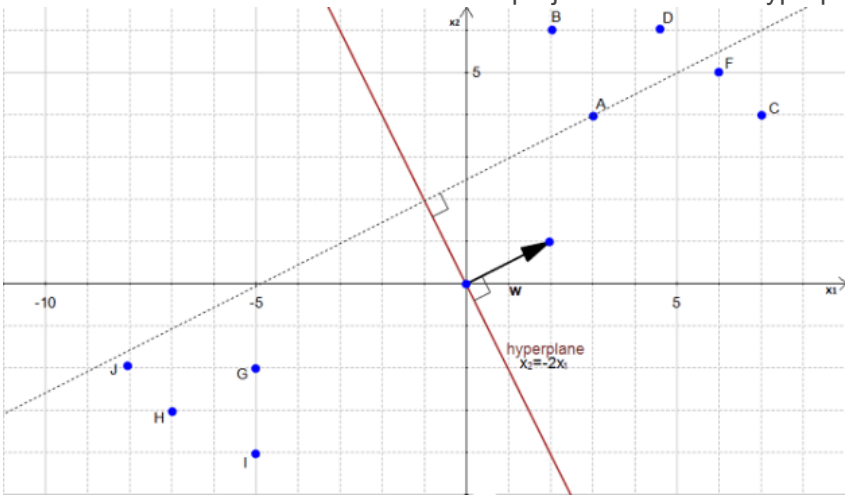


Figure 21

We can view the point A A as a vector from the origin to A A .

If we project it onto the normal vector $\mathbf{w}$ w

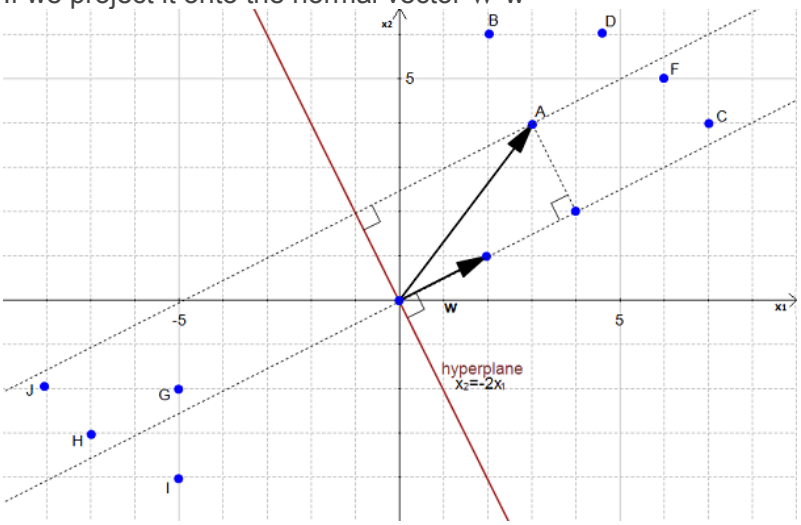


Figure 22 : projection of a onto w

We get the vector $\mathbf{p}$ p

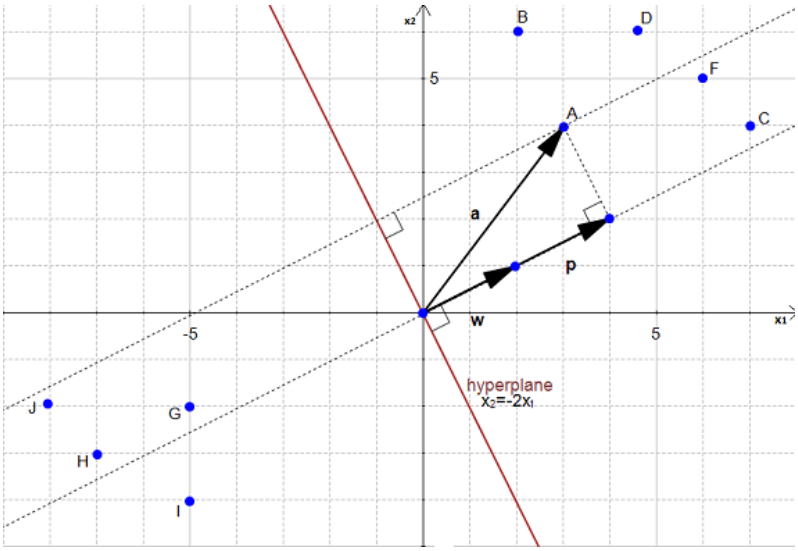


Figure 23: p is the projection of a onto w

Our goal is to find the distance between the point $A(3,4)$ A(3,4) and the hyperplane.
We can see in Figure 23 that this distance is the same thing as $\|p\|$ ‖p‖ .
Let's compute this value.
We start with two vectors, $\mathbf{w}=(2,1)$ w=(2,1) which is normal to the hyperplane, and $\mathbf{a}=(3,4)$ a=(3,4) which is the vector between the origin and A A .

$$\|w\|=2_{2}+1_{2}\quad\text{------}\quad\sqrt{}=5\quad\sqrt{}\quad \|w\|=22+12=5$$

Let the vector $\mathbf{u}$ u be the direction of $\mathbf{w}$ w

$$\mathbf{u}=(25\quad\sqrt{}\quad,15\quad\sqrt{}\quad)\quad u=(25,15)$$

$\mathbf{p}$ p is the orthogonal projection of $\mathbf{a}$ a onto $\mathbf{w}$ w so :

$$\mathbf{p}=(\mathbf{u}\cdot\mathbf{a})\mathbf{u}\quad p=(u\cdot a)u$$

$$\mathbf{p}=(3\times25\quad\sqrt{}\quad+4\times15\quad\sqrt{}\quad)\mathbf{u}\quad p=(3\times25+4\times15)u$$

$$\mathbf{p}=(65\quad\sqrt{}\quad+45\quad\sqrt{}\quad)\mathbf{u}\quad p=(65+45)u$$

$$\mathbf{p}=105\quad\sqrt{}\quad\mathbf{u}\quad p=105u$$

$$\mathbf{p}=(105\quad\sqrt{}\quad\times25\quad\sqrt{}\quad,105\quad\sqrt{}\quad\times15\quad\sqrt{}\quad)\quad p=(105\times25,105\times15)$$

$$\mathbf{p}=(205\ ,105\ )\quad p=(205,105)$$

$$\mathbf{p}=(4,2)\quad p=(4,2)$$

$$\|p\|=4_{2}+2_{2}\quad\text{------}\quad\sqrt{}=25\quad\sqrt{}\quad \|p\|=42+22=25$$

Compute the margin of the hyperplane

Now that we have the distance $\|p\|$ ‖p‖ between A A and the hyperplane, the margin is defined by :
$margin=2\|p\|=45\quad\sqrt{}$ margin=2‖p‖=45
We did it ! We computed the margin of the hyperplane !

Conclusion

This ends the Part 2 of this tutorial about the math behind SVM.
There was a lot more of math, but I hope you have been able to follow the article without problem.

http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-2/