NMK40403 ARTIFICIAL INTELLIGENCE

SVM Optimization Problem

Mohamed Elshaikh

SVM Math 2: The Optimal Hyperplane

Previous slide

- How can we find What is the goal of SVM?
- How to compute the margin?
- However, even if it separating the data it was not the optimal hyperplane.
- ***the optimal hyperplane ?***
- ***How do we calculate the distance between two hyperplanes ?***
- ***What is the SVM optimization problem ?***

SVM Math 2: The Optimal Hyperplane

How to find the optimal hyperplane ?

The optimal hyperplane is the one which maximizes the margin of the training data.

In Figure 1, we can see that the margin M1, delimited by the two blue lines, is not the biggest margin separating perfectly the data.

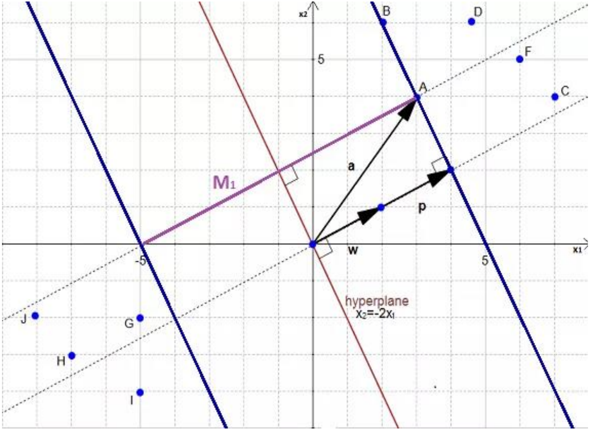


Figure 1

SVM Math 2: The Optimal Hyperplane

How to find the optimal hyperplane ?

The biggest margin is the margin M2 shown in Figure 2 below.

The optima ... our initial hyperplane.

How did I ... M2 in its middle.

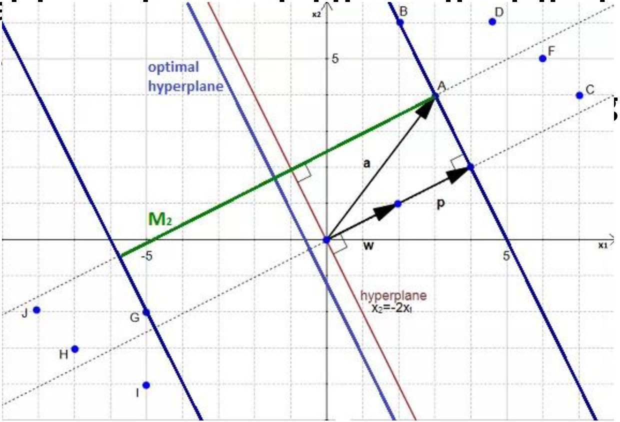


Figure 2

SVM Math 2: The Optimal Hyperplane

How to find the optimal hyperplane ?

Hyperplanes and margins are closely related.

If I have an hyperplane I can compute its margin with respect to some data point. If I have a margin delimited by two hyperplanes (the dark blue lines in Figure 2), I can find a third hyperplane passing right in the middle of the margin.

Finding the biggest margin, is the same thing as finding the optimal hyperplane.

SVM Math 2: The Optimal Hyperplane

How can we find the biggest margin ?

It is rather simple:

You have a dataset

select two hyperplanes which separate the data with no points between them

maximize their distance (the margin)

The region bounded by the two hyperplanes will be the biggest possible margin.

If it is so simple why does everybody have so much pain understanding SVM ?

It is because as always the simplicity requires some abstraction and mathematical terminology to be well understood.

SVM Math 2: The Optimal Hyperplane

How can we find the biggest margin ?

So we will now go through this recipe step by step:

Step 1: You have a dataset D and you want to classify it

Step 2: You need to select two hyperplanes separating the data with no points between them

Step 3: Maximize the distance between the two hyperplanes

SVM Math 2: The Optimal Hyperplane

Step 1: You have a dataset D and you want to classify it

Most of the time your data will be composed of n vectors xi.

Each xi will also be associated with a value yi indicating if the element belongs to the class (+1) or not (-1).

Note that yi can only have two possible values -1 or +1.

Moreover, most of the time, for instance when you do text classification, your vector xi ends up having a lot of dimensions. We can say that xi is a p-dimensional vector if it has p dimensions.

So your dataset D is the set of n couples of element (xi,yi)

The more formal definition of an initial dataset in set theory is :

$$D = \{(x_i, y_i) | x_i \in R^P, y_i \in \{-1,1\}\}_{i=1}^{n}$$

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

Finding two hyperplanes separating some data is easy when you have a pencil and a paper. But with some p-dimensional data it becomes more difficult because you can't draw it.

Moreover, even if your data is only 2-dimensional it might not be possible to find a separating hyperplane !

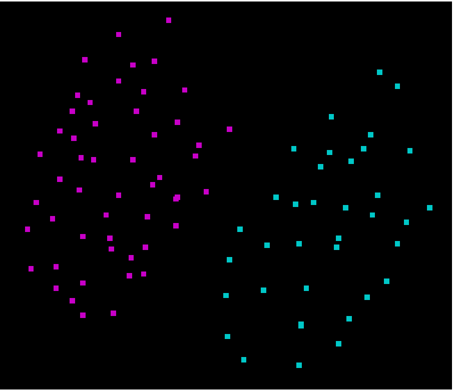
You can only do that if your data is linearly separable

So let's assume that our dataset D is linearly separable. We now want to find two hyperplanes with no points between them, but we don't have a way to visualize them.

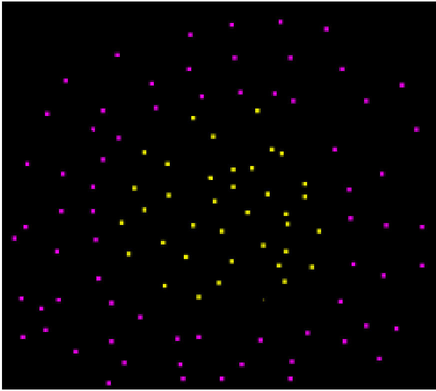
What do we know about hyperplanes that could help us ?

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them



Linearly separable data



Non linearly separable data

Figure 3: Data on the left can be separated by an hyperplane, while data on the right can't

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

Taking another look at the hyperplane equation

We saw previously, that the equation of a hyperplane can be written wTx=0

However, in the Wikipedia article about Support Vector Machine it is said that :

> Any hyperplane can be written as the set of points x satisfying w·x+b=0 .

First, we recognize another notation for the dot product, the article uses w·x instead of wTx.

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

You might wonder… Where does the +b comes from ? Is our previous definition incorrect ?

Not quite. Once again it is a question of notation. In our definition the vectors w and x have three dimensions, while in the Wikipedia definition they have two dimensions:

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

Given two 3-dimensional vectors $w(b, -a, 1)$ **and** $x(1, x, y)$

$$w.x = b \times (1) + (-a) \times x + 1 \times y$$

$$w.x = y - ax + b \tag{1}$$

Given two 2-dimensional vectors $w'(-\mathrm{a}, 1)$ **and** $x'(x, y)$

$$w'.x' = (-a) \times x + 1 \times y$$

$$w'.x' = y - ax \tag{2}$$

Now if we add b on both side of the equation (2) we got :

$$w'.x' + b = y - ax + b$$

$$w'.x' + b = w.x \tag{3}$$

For the rest of this article we will use 2-dimensional vectors (as in equation (2)).

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them
Given a hyperplane H0 separating the dataset and satisfying:

$$w \cdot x + b = 0$$

We can select two others hyperplanes H1 and H2 which also separate the data and have the following equations :

$$\mathbf{w \cdot x + b = \delta}$$

and

$$\mathbf{w \cdot x + b = -\delta}$$

so that H0 is equidistant from H1 and H2.

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

However, here the variable δ is not necessary. So we can set δ=1 to simplify the problem.

$$w \cdot x + b = 1$$

and

$$w \cdot x + b = -1$$

Now we want to be sure that they have no points between them.

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them
We won't select any hyperplane, we will only select those who meet the two following constraints:

For each vector xi either :

$$w \cdot x_i + b \geq 1 \text{ for } x_i \text{ having the class } 1 \qquad (4)$$

or

$$w \cdot x_i + b \leq -1 \text{ for } x_i \text{ having the class } -1 \qquad (5)$$

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

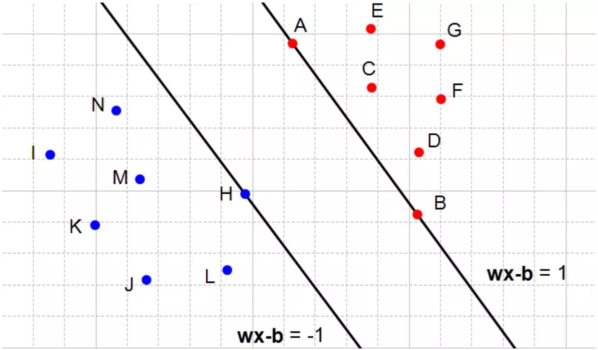


Figure 4: Two hyperplanes satisfying the constraints

Understanding the constraints

On the following figures, all red points have the class 1 and all blue points have the class −1.

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

So let's look at Figure 4 and consider the point A. It is red so it has the class 1 and we need to verify it does not violate the constraint $w \cdot x_i + b \geq 1$

When $x_i = A$ we see that the point is on the hyperplane so $w \cdot x_i + b = 1$ and the constraint is respected. The same applies for B.

When $x_i = C$ we see that the point is above the hyperplane so $w \cdot x_i + b > 1$ and the constraint is respected. The same applies for D, E, F and G.

With an analogous reasoning you should find that the second constraint is respected for the class −1.

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

On Figure 5, we see another couple of hyperplanes respecting the constraints:

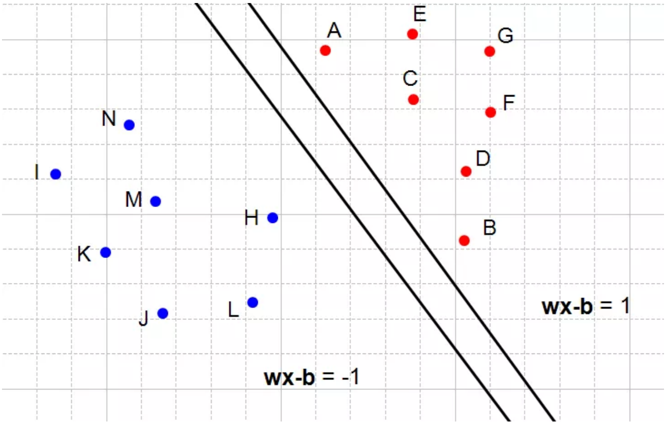


Figure 5: Two hyperplanes also satisfying the constraints

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

And now we will examine cases where the constraints are not respected:

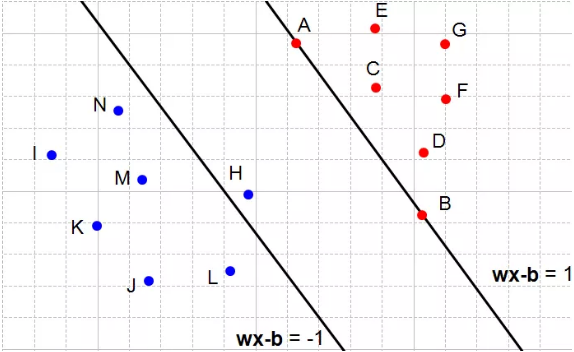


Figure 7: The left hyperplane does not satisfy the second constraint

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

What does it means when a constraint is not respected ? It means that we cannot select these two hyperplanes. You can see that every time the constraints are not satisfied (Figure 6, 7 and 8) there are points between the two hyperplanes.

By defining these constraints, we found a way to reach our initial goal of selecting two hyperplanes without points between them. And it works not only in our examples but also in p-dimensions !

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

Combining both constraints

In mathematics, people like things to be expressed concisely.

Equations (4) and (5) can be combined into a single constraint:

We start with equation (5)

for xi having the class −1

$$w \cdot x_i + b \leq -1$$

And multiply both sides by yi (which is always -1 in this equation)

$$y_i(w \cdot x_i + b) \geq y_i(-1)$$

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them

Which means equation (5) can also be written:

$$y_i(w \cdot x_i+b) \geq 1 \text{ for } x_i \text{ having the class } -1 \qquad (6)$$

In equation (4), as $y_i=1$ it doesn't change the sign of the inequation.

$$y_i(w \cdot x_i+b) \geq 1 \text{ for } x_i \text{ having the class } 1 \qquad (7)$$

We combine equations (6) and (7) :

$$y_i(w \cdot x_i+b) \geq 1 \text{ for all } 1 \leq i \leq n \qquad (8)$$

SVM Math 2: The Optimal Hyperplane

Step 2: You need to select two hyperplanes separating the data with no points between them
We now have a unique constraint (equation 8) instead of two (equations 4 and 5) , but they are mathematically equivalent. So their effect is the same (there will be no points between the two hyperplanes).

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

a) What is the distance between our two hyperplanes ?

b) How to maximize the distance between our two hyperplanes

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

a) What is the distance between our two hyperplanes ?

Before trying to maximize the distance between the two hyperplane, we will first ask ourselves: how do we compute it ?

Let:

H0 be the hyperplane having the equation $w \cdot x + b = -1$

H1 be the hyperplane having the equation $w \cdot x + b = 1$

x0 be a point in the hyperplane H0.

We will call m the perpendicular distance from x0 to the hyperplane H1 . By definition, m is what we are used to call the margin.

As x0 is in H0, m is the distance between hyperplanes H0 and H1 .

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

We will now try to find the value of m.

You might be tempted to think that if we add m to x0 we will get another point, and this point will be on the other hyperplane !

But it does not work, because m is a scalar, and x0 is a vector and adding a scalar with a vector is not possible. However, we know that adding two vectors is possible, so if we transform m into a vector we will be able to do an addition.

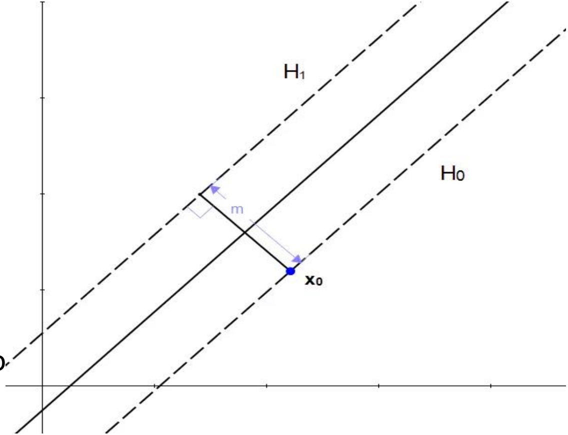


Figure 9: m is the distance between the two hyperplanes

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

We can find the set of all points which are at a distance m from x0. It can be represented as a circle :

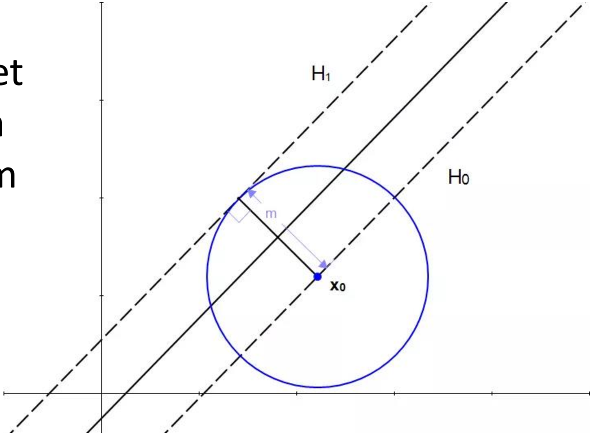


Figure 10: All points on the circle are at the distance m from x_0

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

Looking at the picture, the necessity of a vector become clear. With just the length m we don't have one crucial information : the direction. (a vector has a magnitude and a direction).

We can't add a scalar to a vector, but we know if we multiply a scalar with a vector we will get another vector.

From our initial statement, we want this vector:

to have a magnitude of m

to be perpendicular to the hyperplane H1

Fortunately, we already know a vector perpendicular to H1, that is w (because H1=w·x+b=1)

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

w is a vector perpendicular to H1 (because H1=w·x+b=1)

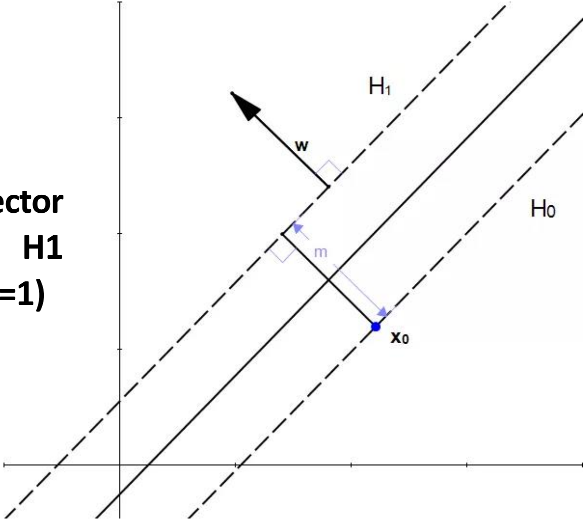


Figure 11: w is perpendicular to H_1

SVM Math 2: The Optimal Hyperplane

Let's define $u = \frac{w}{\|w\|}$ the unit vector of w.
As it is a unit vector ‖u‖=1 and it has the same direction as w so it is also perpendicular to the hyperplane.

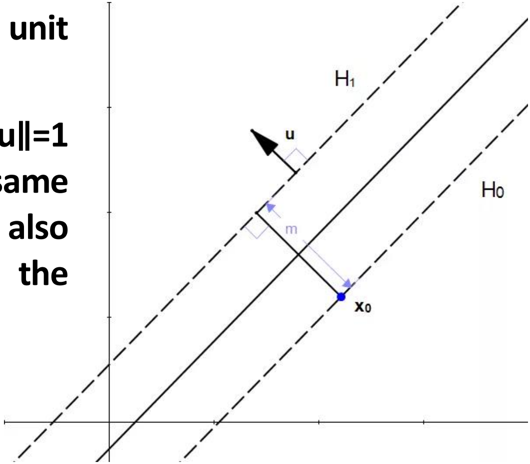


Figure 12: u is also is perpendicular to H_1

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

If we multiply u by m we get the vector k=mu and :

1) ‖k‖=m
2) k is perpendicular to H1 (because it has the same direction as u)

From these properties we can see that k is the vector we were looking for.

$$k = mu = m\frac{w}{\|w\|}$$

(9)

- We did it ! We transformed our scalar m into a vector k which we can use to perform an addition with the vector x_0.

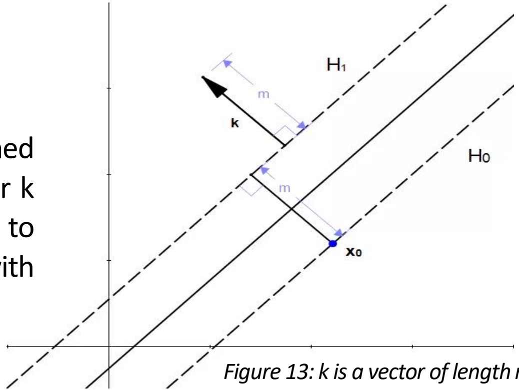


Figure 13: k is a vector of length m perpendicular to H_1

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

If we start from the point x_0 and add k we find that the point $z_0 = x_0 + k$ is in the hyperplane H_1 as shown on Figure 14.

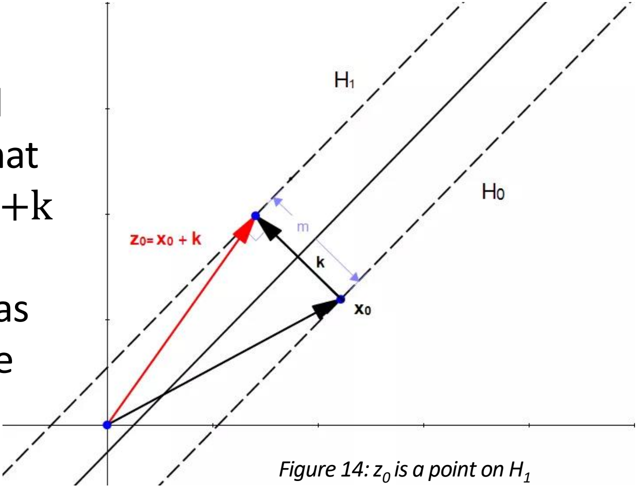


Figure 14: z_0 is a point on H_1

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

The fact that z0 is in H1 means that

$$\mathbf{w \cdot z0 + b = 1} \qquad (10)$$

We can replace z0 by x0+k because that is how we constructed it.

$$\mathbf{w \cdot (x0 + k) + b = 1} \qquad (11)$$

We can now replace k using equation (9)

$$\mathbf{w \cdot (x0 +} \, m \frac{w}{\|w\|} \mathbf{) + b = 1} \qquad (12)$$

We now expand equation (12)

$$\mathbf{w \cdot x0 +} \, m \frac{w.w}{\|w\|} \mathbf{+ b = 1} \qquad (13)$$

***The dot product of a vector with itself is the square of its norm**

SVM Math 2: The Optimal Hyperplane

Step 3: Maximize the distance between the two hyperplanes

The dot product of a vector with itself is the square of its norm so :

$$\mathbf{w \cdot x0 +} \, m\frac{\|w\|^2}{\|w\|}\mathbf{+ b = 1} \qquad \textbf{(14)}$$

$$\mathbf{w \cdot x0 +} \, m\|w\|\mathbf{+ b = 1} \qquad \textbf{(15)}$$

$$\mathbf{w \cdot x0 + b = 1 -} \, m\|w\| \qquad \textbf{(16)}$$

As x0 is in H0 then w·x0+b=−1

$$\mathbf{-1 = 1 -} \, m\|w\| \qquad \textbf{(17)}$$

$$m\|w\|\mathbf{= 2} \qquad \textbf{(18)}$$

$$m\mathbf{=}\frac{2}{\|w\|} \qquad \textbf{(19)}$$

This is it ! We found a way to compute m.

SVM Math 2: The Optimal Hyperplane

Maximizing the margin is the same thing as minimizing the norm of w. Our goal is to maximize the margin. Among all possible hyperplanes meeting the constraints, we will choose the hyperplane with the smallest $\|w\|$ because it is the one which will have the biggest margin.

This give us the following optimization problem:

Minimize in (w,b) $\|w\|$

subject to $y_i(w \cdot x_i + b) \geq 1$ (for any $i=1,\ldots,n$)

Solving this problem is like solving and equation. Once we have solved it, we will have found the couple (w,b) for which $\|w\|$ is the smallest possible and the constraints we fixed are met. Which means we will have the equation of the optimal hyperplane !

Q&A