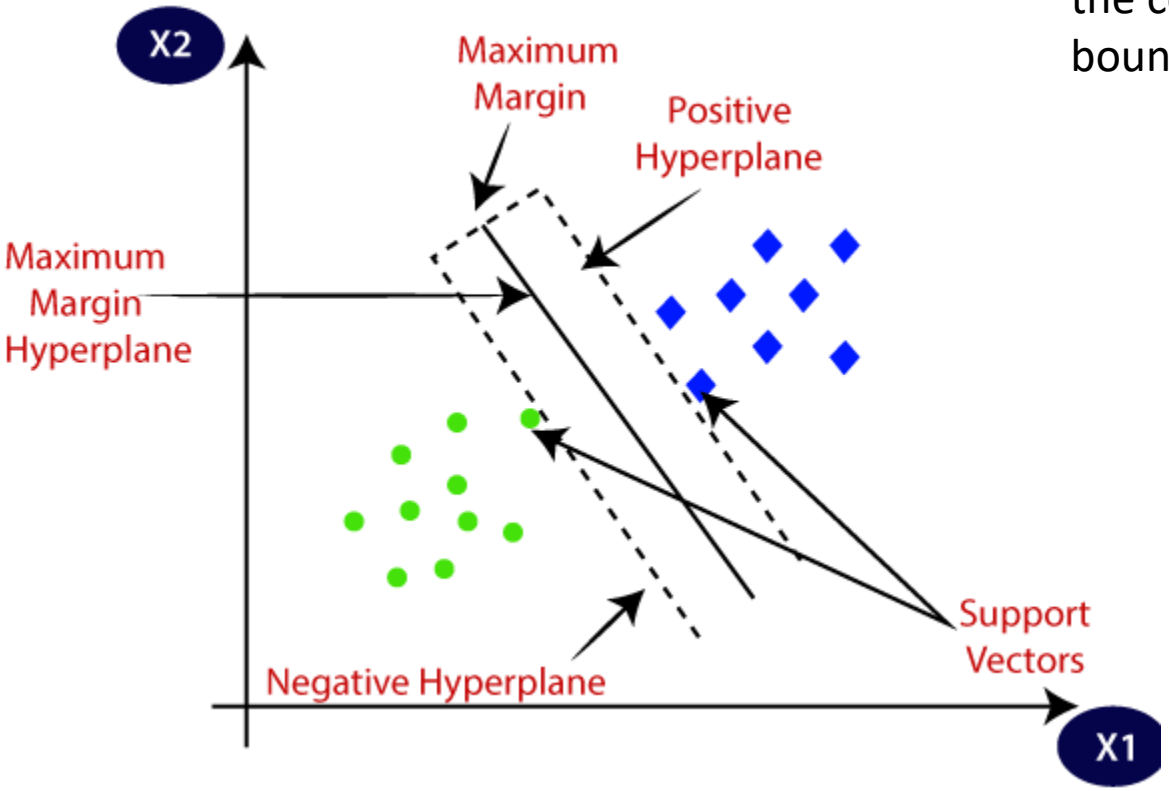
Do It Skills
do it. enjoy it.

SUPPORT VECTOR MACHINE (SVM)

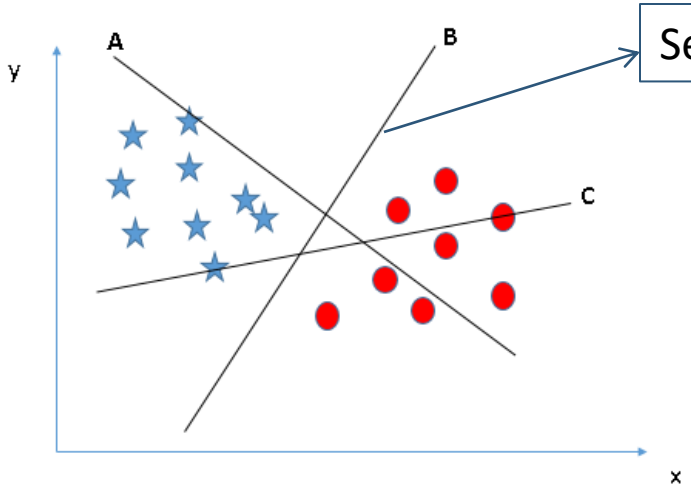
MACHINE LEARNING ALGORITHM

SVM



The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

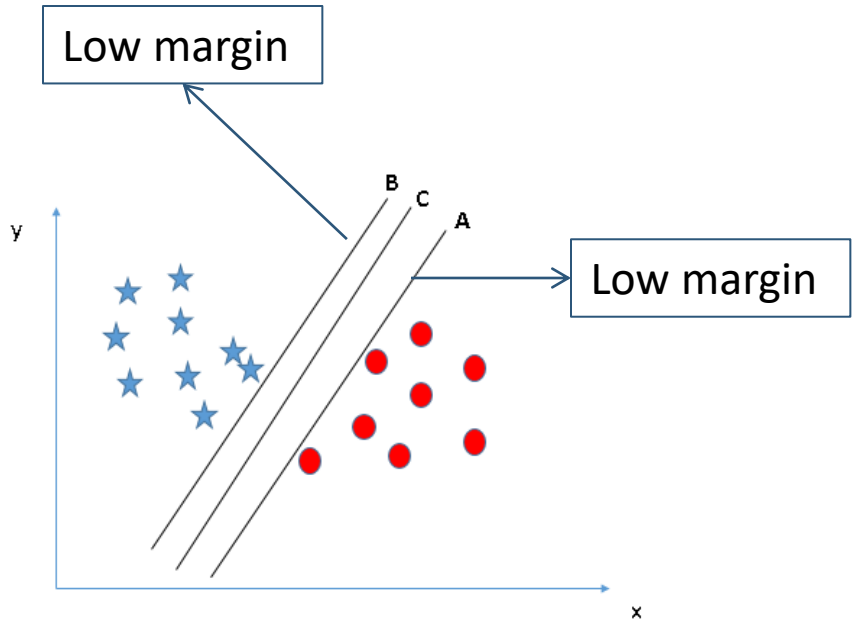
IDENTIFY THE RIGHT HYPER-PLANE (SCENARIO-1):

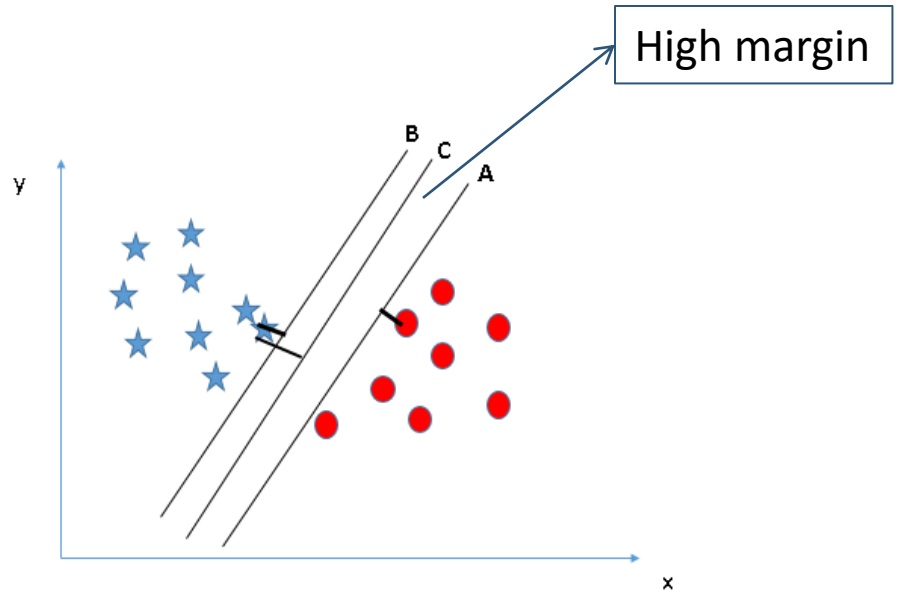


Separating both class correctly

From this plot, hyperplane A and C is not separating both the classes.
B is separating every data point correctly. No data point is at the wrong side of class.

IDENTIFY THE RIGHT HYPER-PLANE (SCENARIO-2):

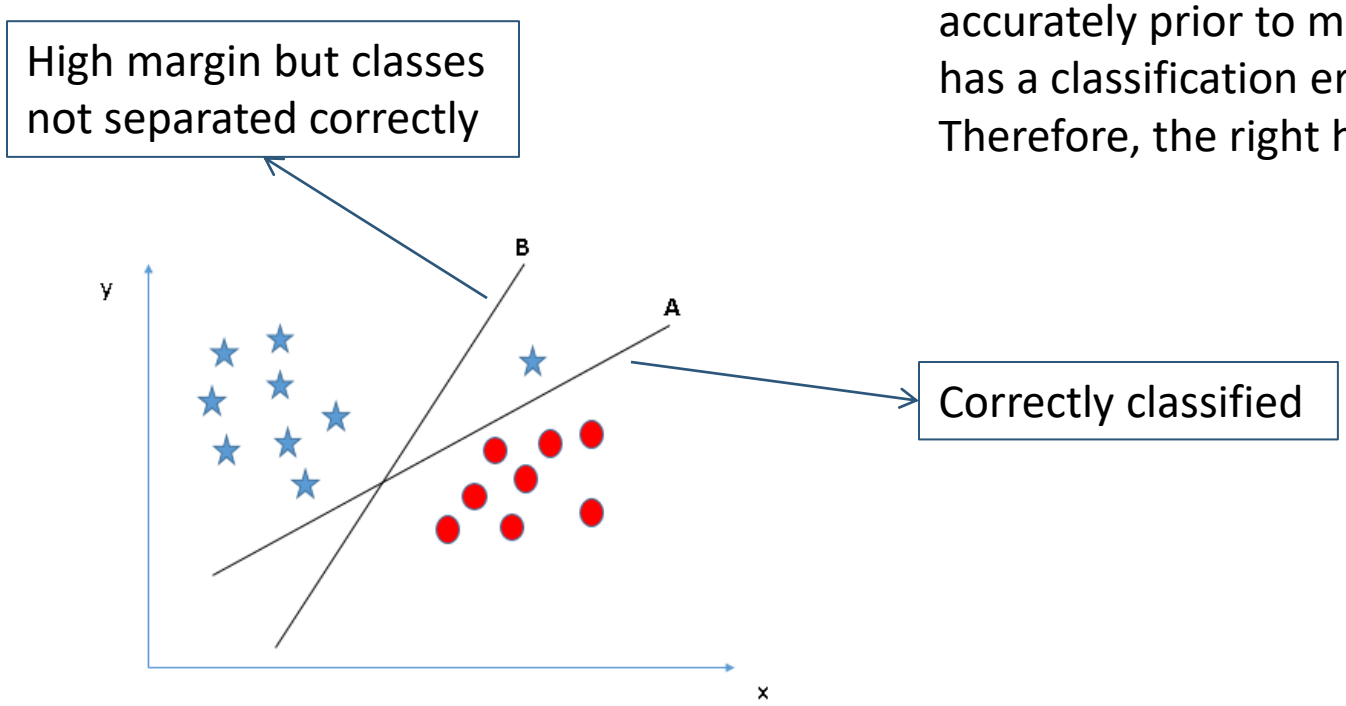




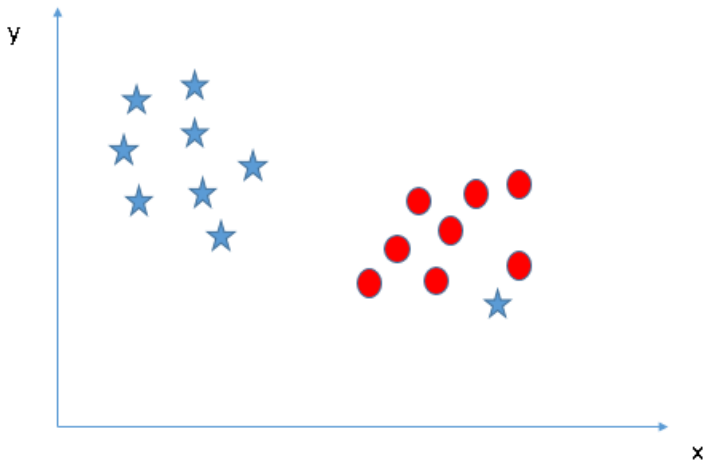
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

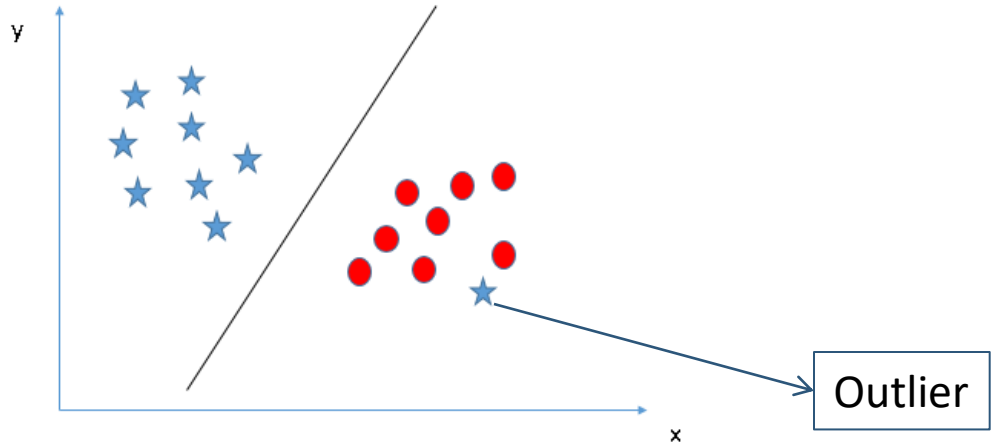
IDENTIFY THE RIGHT HYPER-PLANE (SCENARIO-3)

The hyper-plane **B** has higher margin compared to **A.** But SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is **A.**



CAN WE CLASSIFY TWO CLASSES (SCENARIO-4)?





One star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.

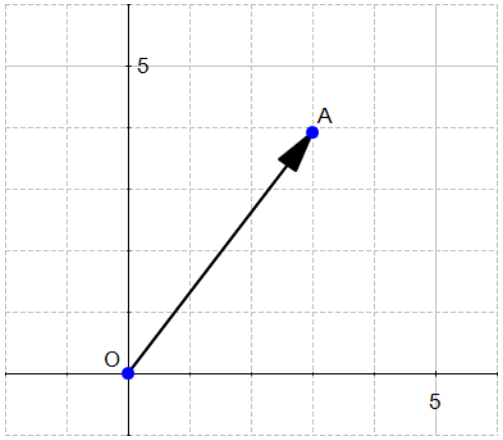
Do It Skills
do it. enjoy it.

LINEAR ALGEBRA FOR SVM

VECTOR

WHAT IS VECTOR?

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



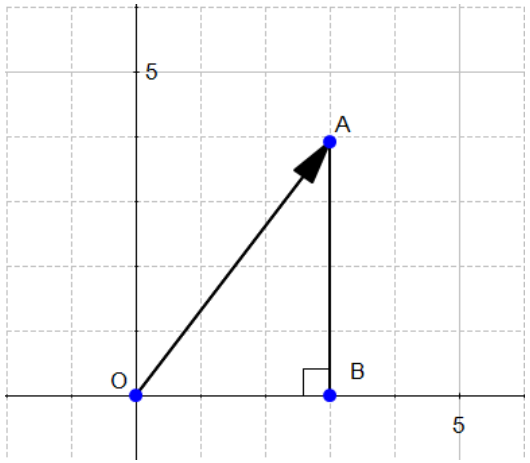
the vector above is the vector $\overrightarrow{OA}$. We could also give it an arbitrary name such as u.

Math behind SVM link : https://www.svm-tutorial.com/2014/11/svm-understanding-math-part-2/

Do It Skills
do it. enjoy it.

MAGNITUDE OF VECTOR

The magnitude or length of a vector x is written ‖x‖ (norm of x) and is called its norm.



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

$OA^2 = OB^2 + AB^2$

$OA^2 = 3^2 + 4^2$

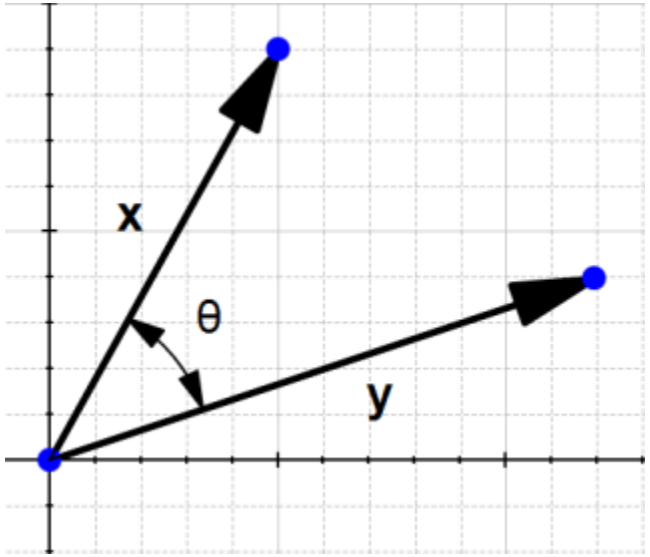
$OA^2 = 25$

$OA = \sqrt{25}$

$\|OA\| = OA = 5$

Norm of OA = ‖OA‖ is magnitude of vector OA.

THE DOT PRODUCT

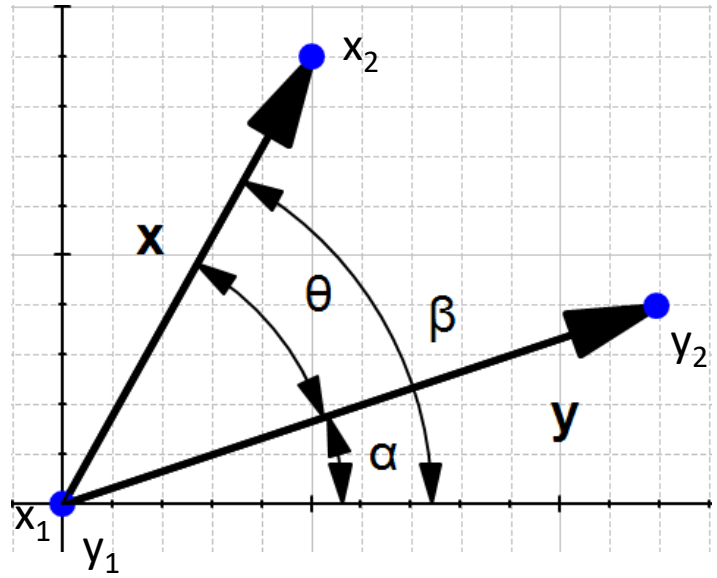


We know that in a right-angled triangle

$$cos(\theta) = \frac{adjacent}{hypotenuse}$$

$\theta = \beta - \alpha$

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



$$cos(\theta) = \frac{x_1}{\|x\|}\frac{y_1}{\|y\|} + \frac{x_2}{\|x\|}\frac{y_2}{\|y\|}$$

$$cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{\|x\|\|y\|}$$

$$\|x\|\|y\|cos(\theta) = x_1 y_1 + x_2 y_2$$

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

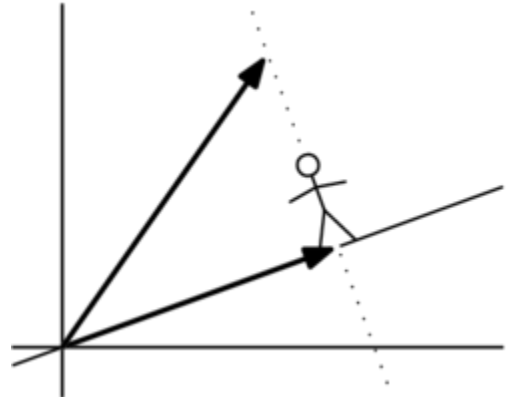
$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 = \sum_{i=1}^{2}(x_i y_i)$$

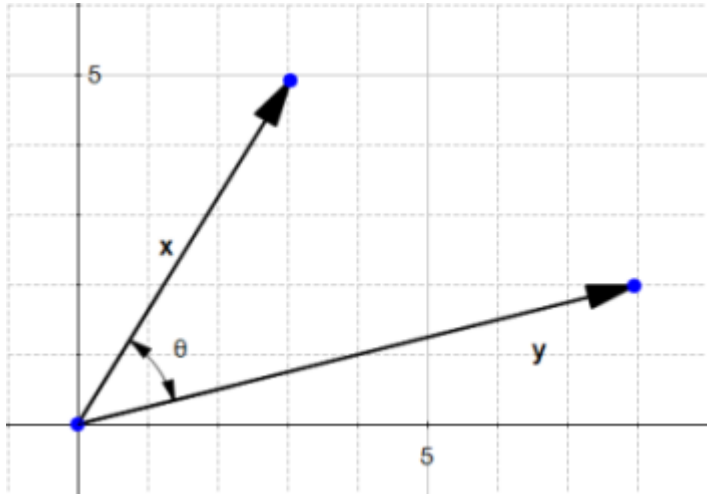
Do It Skills
do it. enjoy it.

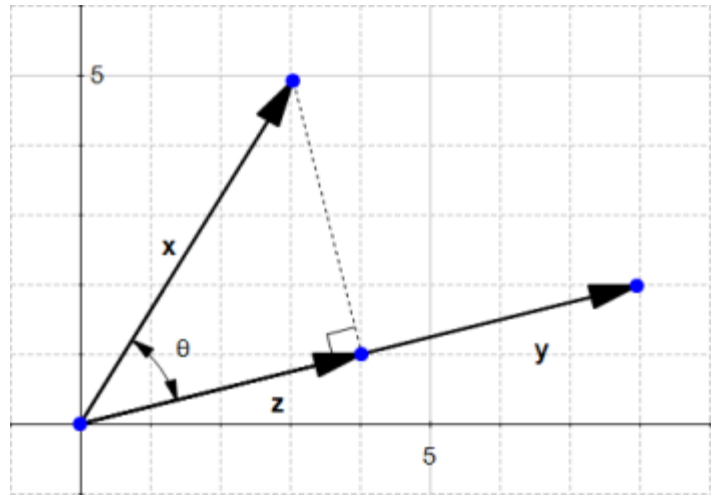
30 SEP 2020

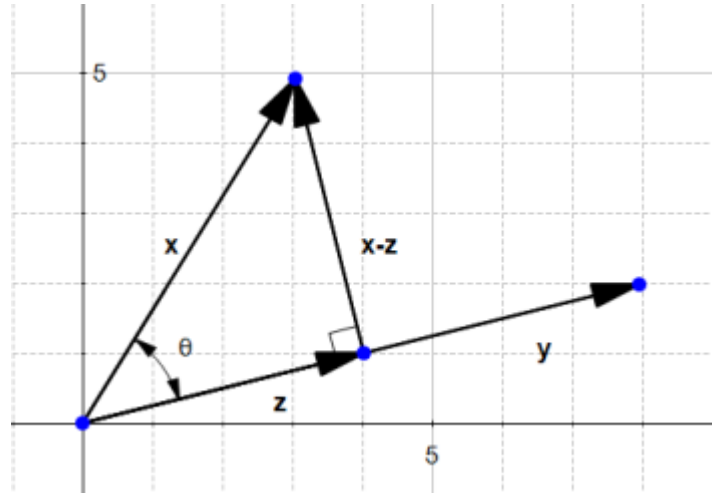
THE ORTHOGONAL PROJECTION OF A VECTOR

This is orthogonal projection of x on y









Do It Skills
do it. enjoy it.

Do It Skills
do it. enjoy it.

SVM MARGIN CALCULATION

THE EQUATION OF THE HYPERPLANE

Note that

$$y = ax + b$$

is the same thing as

$$y - ax - b = 0$$

Given two vectors $\mathbf{w}\begin{pmatrix} -b \\ -a \\ 1 \end{pmatrix}$ and $\mathbf{x}\begin{pmatrix} 1 \\ x \\ y \end{pmatrix}$

$$\mathbf{w}^T\mathbf{x} = -b \times (1) + (-a) \times x + 1 \times y$$

$$\mathbf{w}^T\mathbf{x} = y - ax - b$$

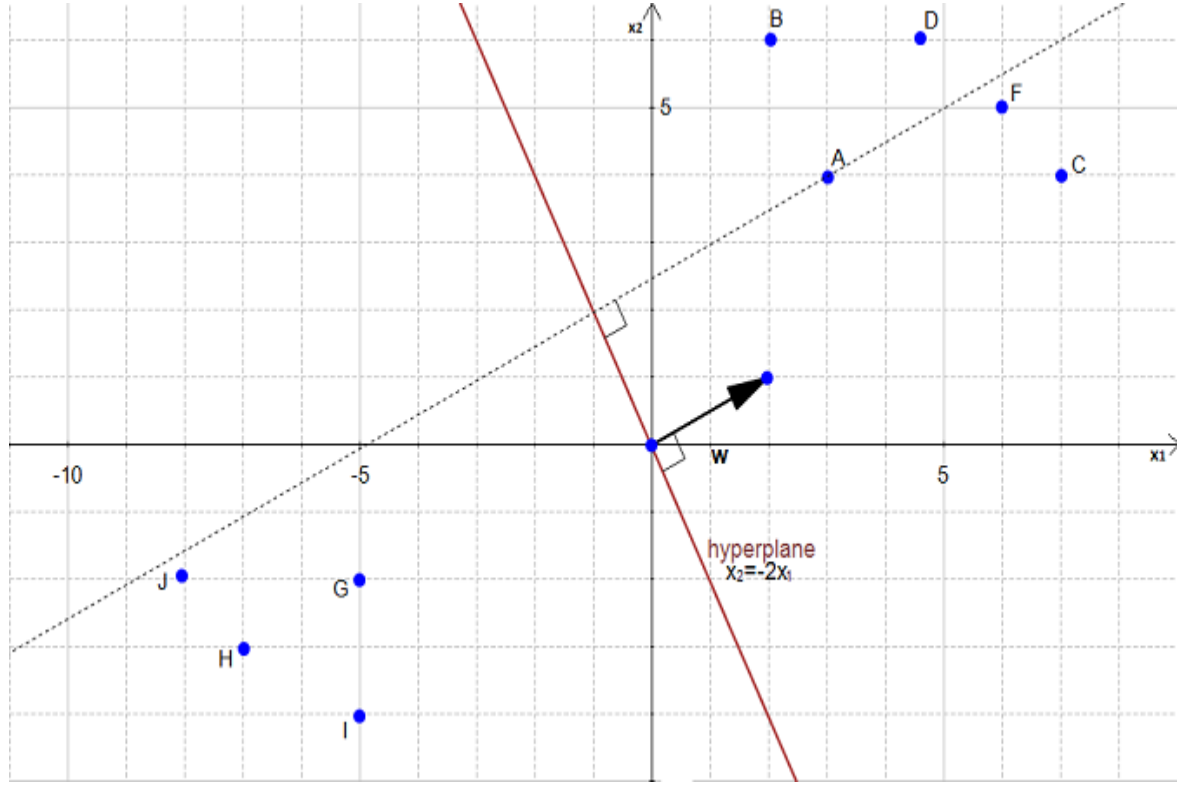
COMPUTE THE DISTANCE FROM A POINT TO THE HYPERPLANE

Equation of the hyperplane is :
$x_2 = -2x_1$ which is equivalent to
$w^T x = 0$

with $w \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $x \begin{pmatrix} x1 \\ x2 \end{pmatrix}$

$W(2\ 1) * \begin{pmatrix} x1 \\ x2 \end{pmatrix}$

$2 * x_1 + 1 * X_2 = 0$
$x_2 = -2x_1$
Note that the vector w is shown on the Figure
(w is not a data point)

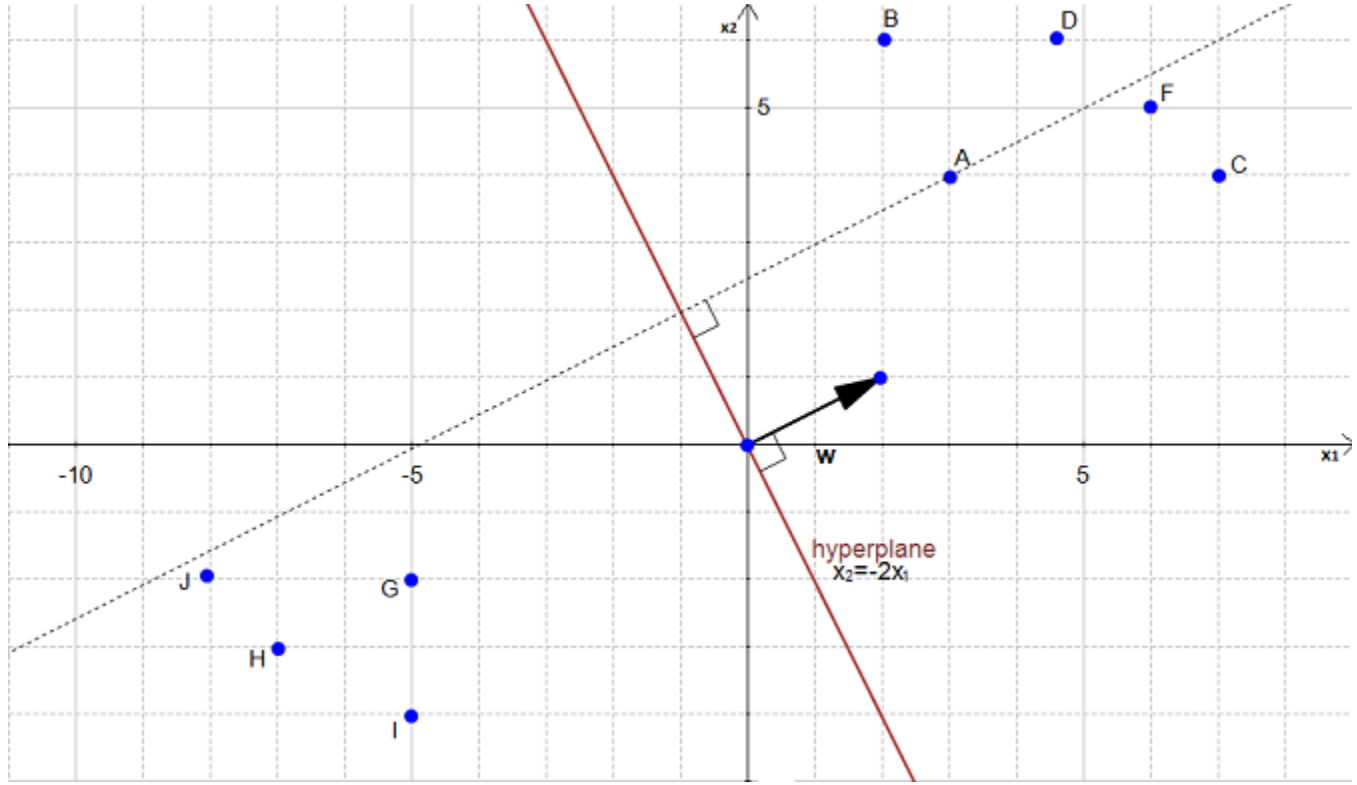


STEP2

We would like to compute the distance between the point A(3,4)

and the hyperplane.

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

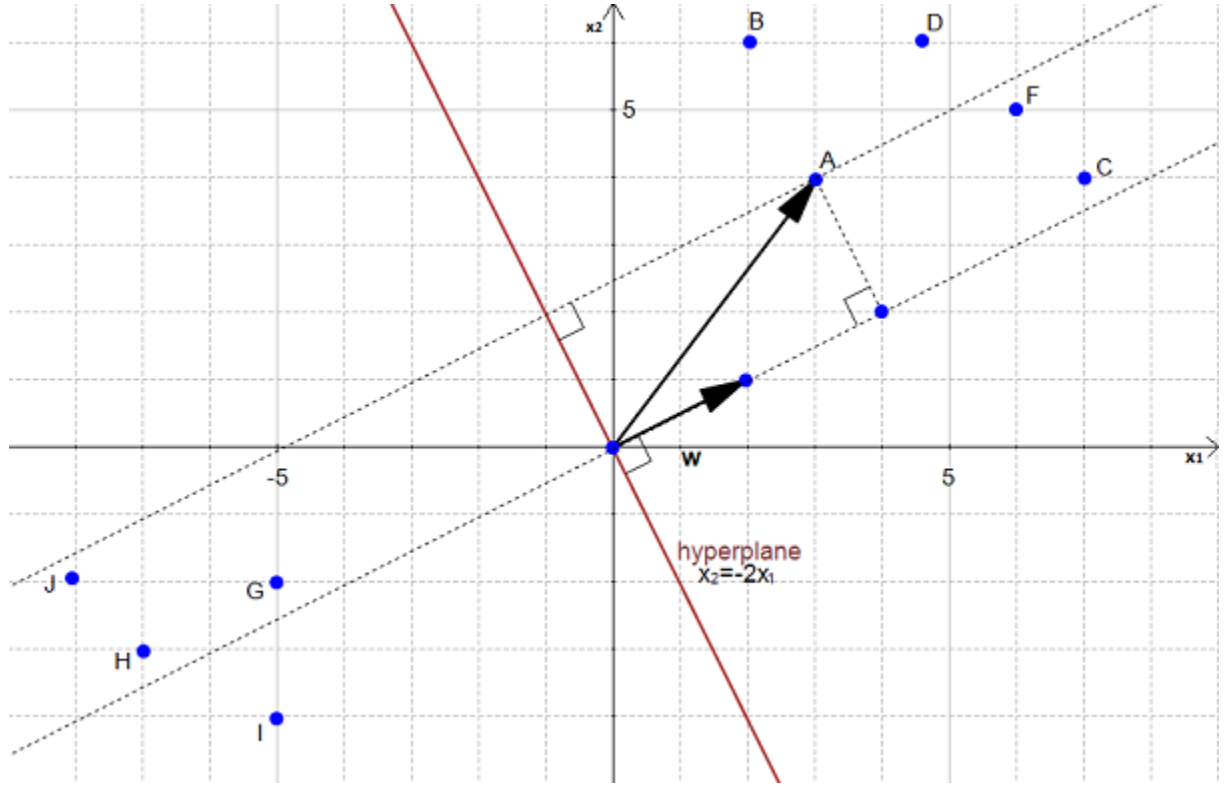


Do It Skills
do it. enjoy it.

STEP3

We can view the point A as a vector from the origin to A.
If we project it onto the normal vector w

A **normal** vector is any vector whose direction is **perpendicular** to that of the plane

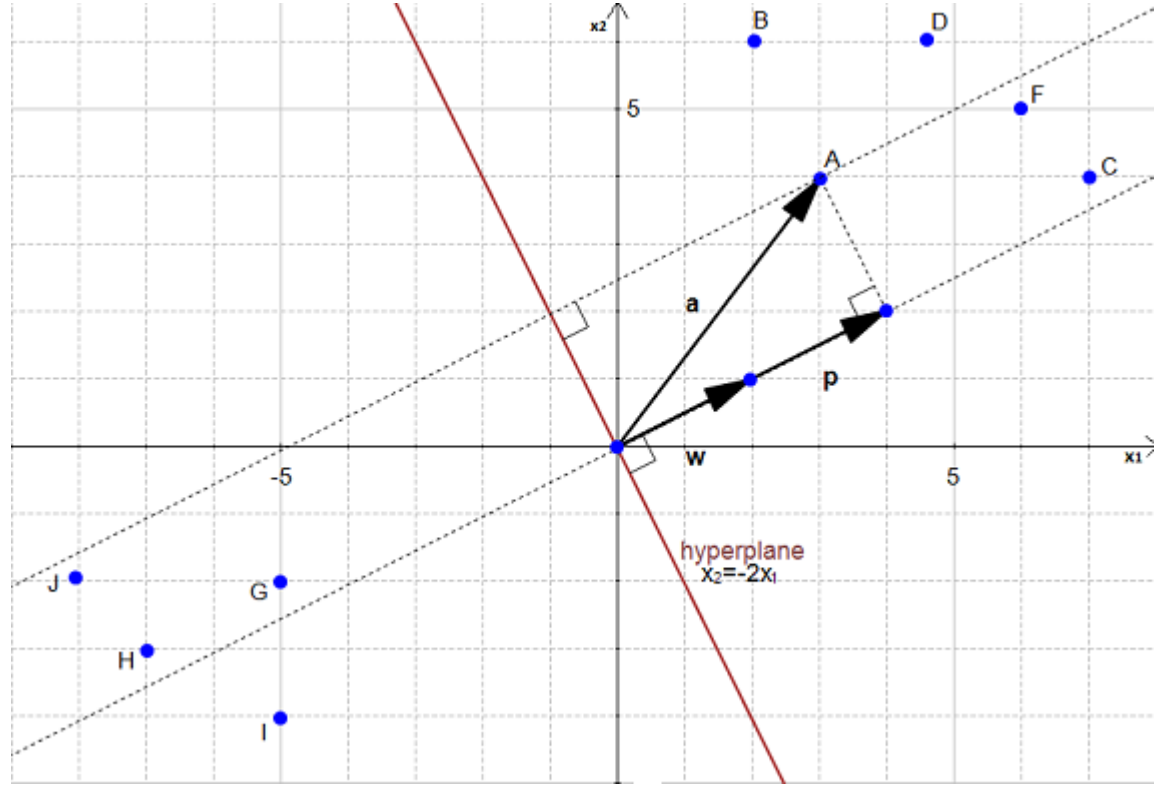


STEP4

Our goal is to find the distance between the point $A(3,4)$ and the hyperplane.
We can see in Figure 23 that this distance is the same thing as $\|p\|$.
Let's compute this value.

We start with two vectors, $\mathbf{w} = (2,1)$ which is normal to the hyperplane, and $\mathbf{a} = (3,4)$ which is the vector between the origin and A.

$$\|w\| = \sqrt{2^2 + 1^2} = \sqrt{5}$$



Do It Skills
do it. enjoy it.

STEP5

We start with two vectors, w=(2,1) which is normal to the hyperplane, and a=(3,4) which is the vector between the origin and A.

$$\|w\| = \sqrt{2^2 + 1^2} = \sqrt{5}$$

Let the vector u be the direction of w

$$u = \left(\frac{W_1}{||w_1||} \; \frac{W_2}{||w_2||} \right)$$

$$u = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right)$$

p is the orthogonal projection of **a** onto **w** so :

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

$$\mathbf{p} = (3 \times \frac{2}{\sqrt{5}} + 4 \times \frac{1}{\sqrt{5}})\mathbf{u}$$

$$\mathbf{p} = (\frac{6}{\sqrt{5}} + \frac{4}{\sqrt{5}})\mathbf{u}$$

$$\mathbf{p} = \frac{10}{\sqrt{5}}\mathbf{u}$$

$$\mathbf{p} = (\frac{10}{\sqrt{5}} \times \frac{2}{\sqrt{5}}, \frac{10}{\sqrt{5}} \times \frac{1}{\sqrt{5}})$$

$$\mathbf{p} = (\frac{20}{5}, \frac{10}{5})$$

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

$$\|p\| = \sqrt{4^2 + 2^2} = 2\sqrt{5}$$

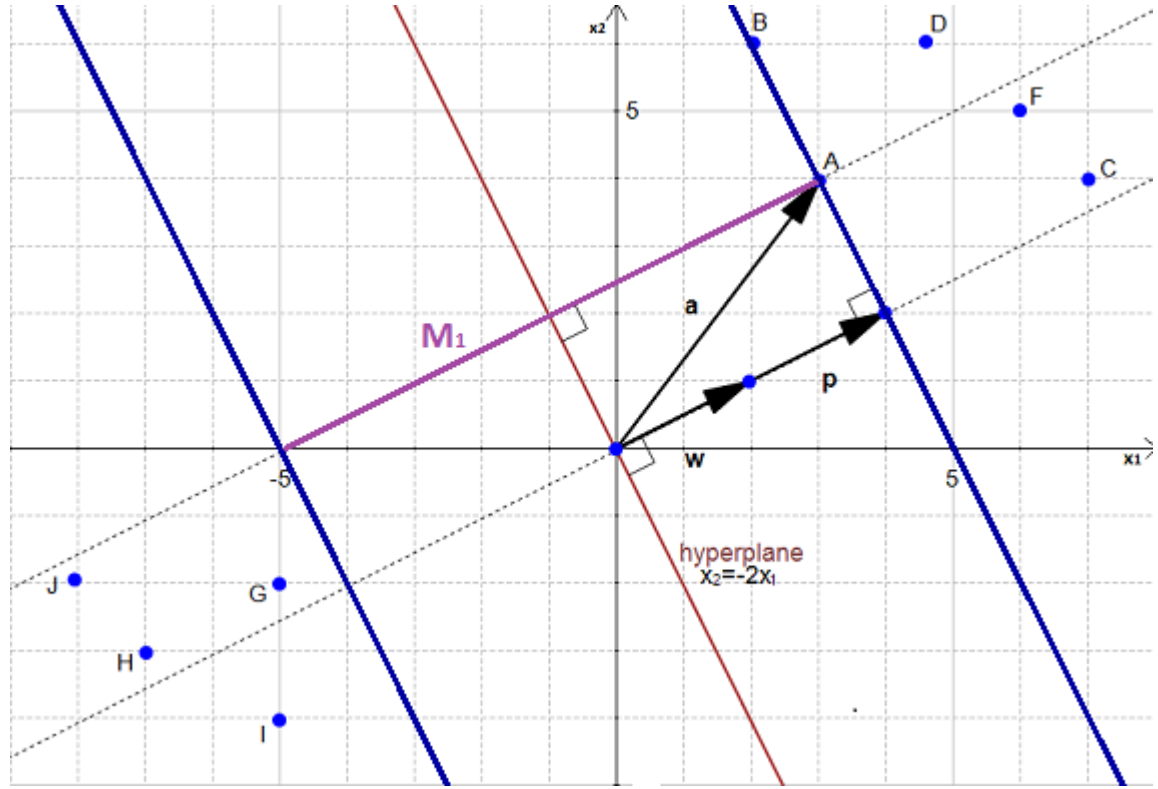
STEP6

Now that we have the distance ‖p‖ between A

and the hyperplane, the margin is defined by :

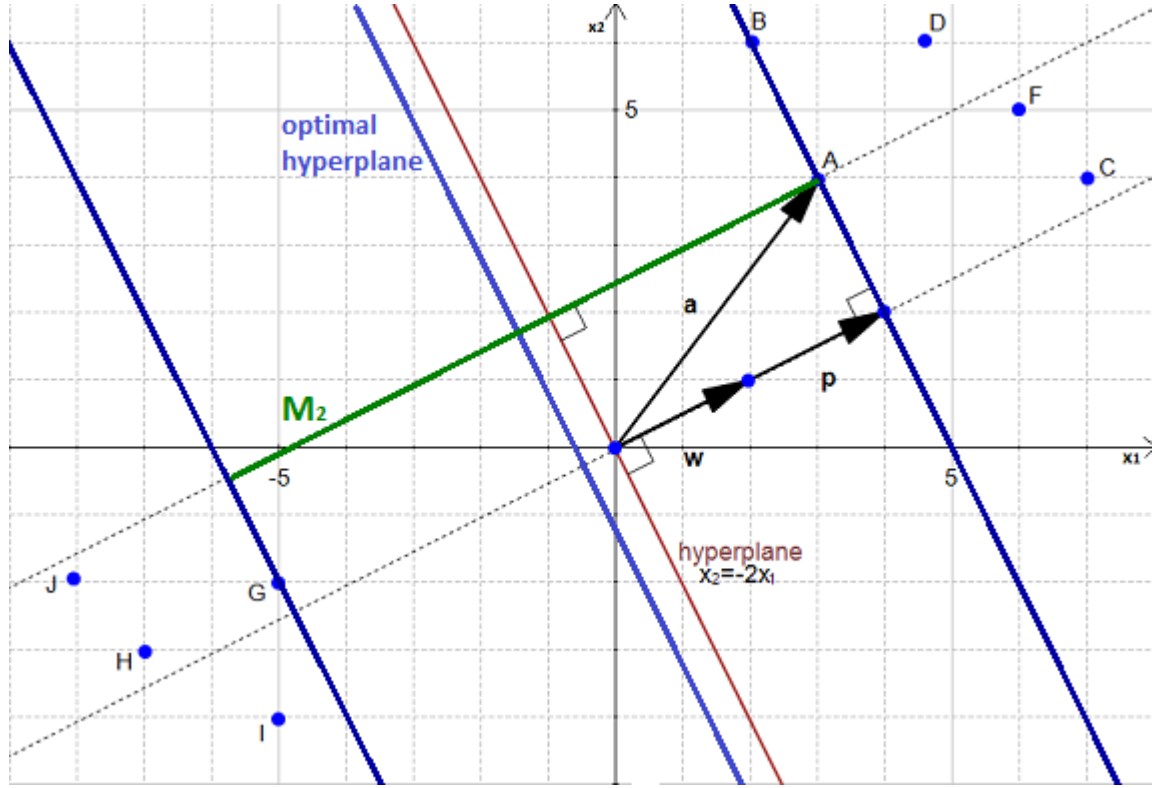
margin=2 * ‖p‖=4√5

However it is not an optimal hyperplane



OPTIMAL HYPERPLANE

We can see that the margin M1, delimited by the two blue lines, is not the biggest margin separating perfectly the data. The biggest margin is the margin M2 shown in this Figure



Do It Skills
do it. enjoy it.

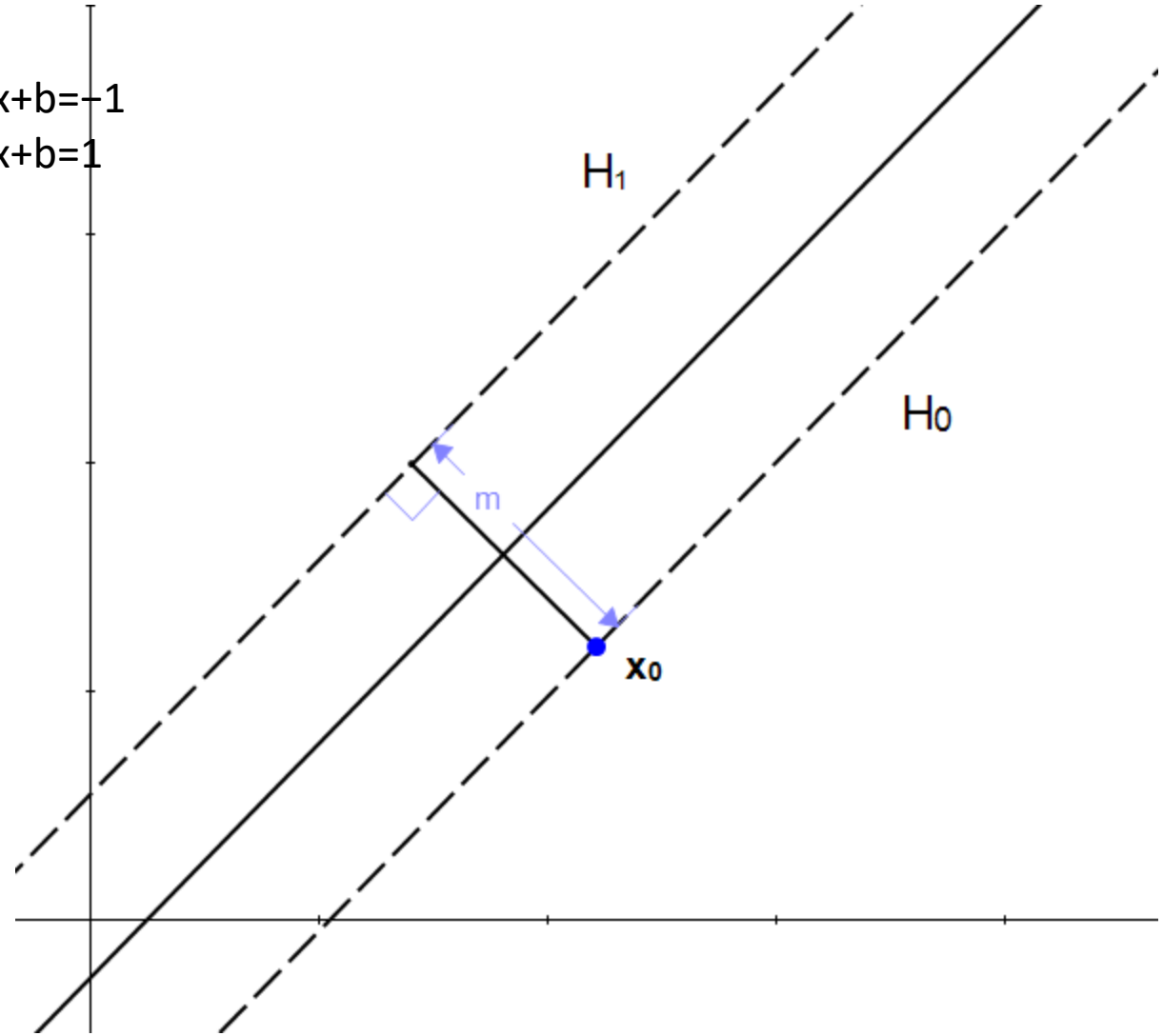
Do It Skills
do it. enjoy it.

MAXIMIZE THE MARGIN

MAXIMUM MARGIN

Let:
H_0 : be the hyperplane having the equation $w \cdot x+b=-1$
H_1 : be the hyperplane having the equation $w \cdot x+b=1$
X_0 : be a point in the hyperplane H0.

m: margin between the hyperplane H_0 and H_1
X_0: Data point on hyperplane H_0

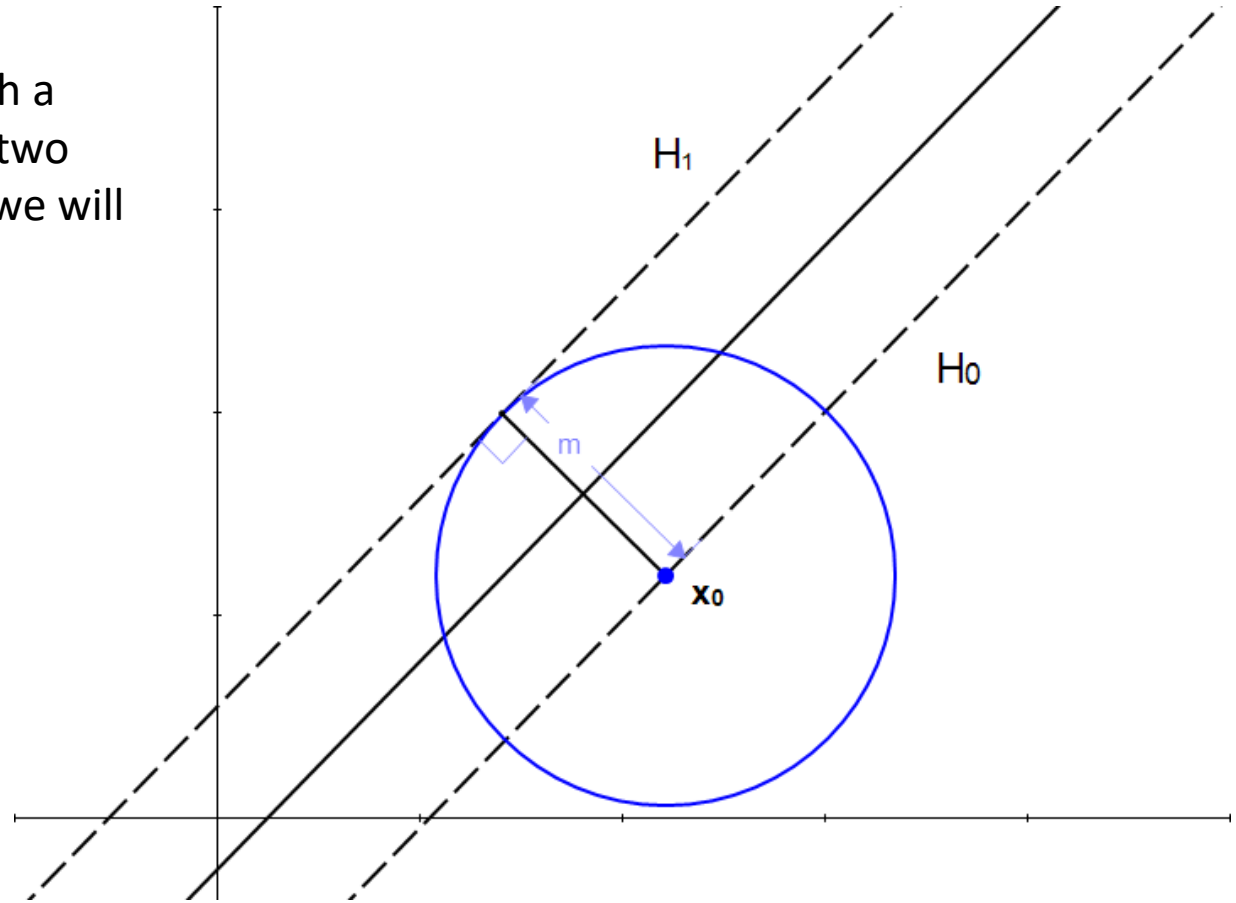


Do It Skills
do it. enjoy it.

STEP2

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

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.

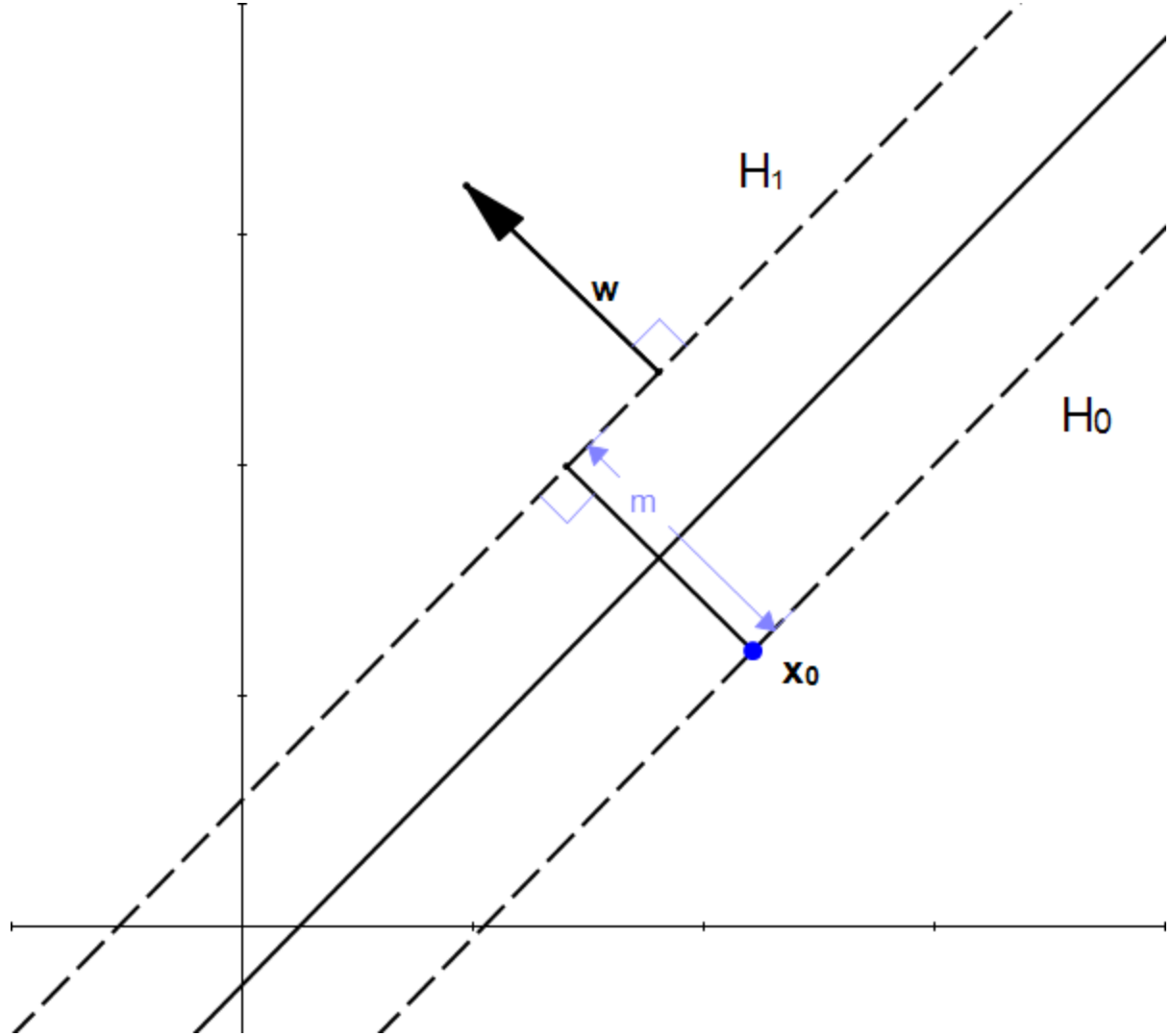


Do It Skills
do it. enjoy it.

STEP3

Fortunately, we already know a vector perpendicular to H_1, that is w (because $H_1 = w \cdot x + b = 1$)

Let's define $\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$ 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.

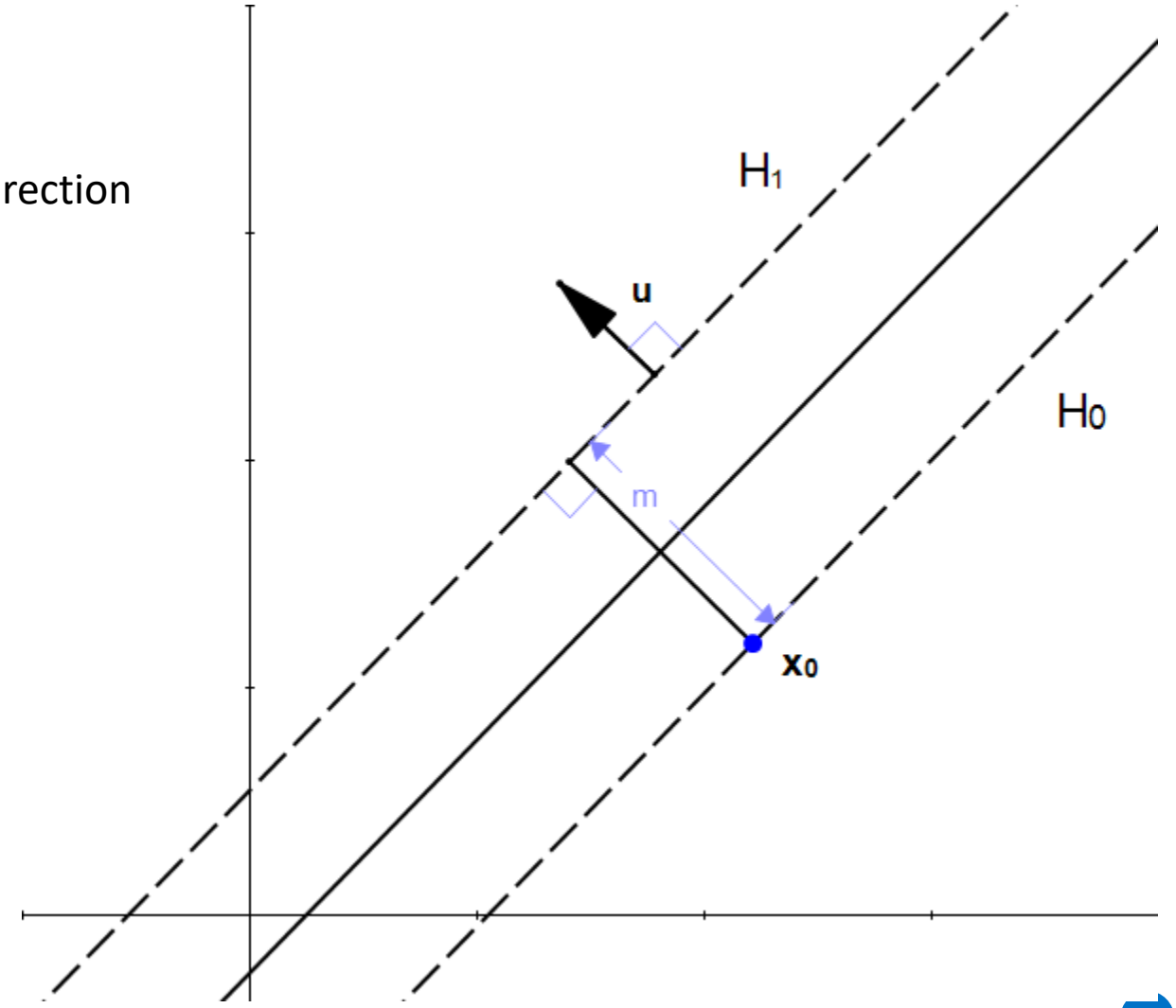


Do It Skills
do it. enjoy it.

STEP4

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

- ||k||=m
- k is perpendicular to H1 (because it has the same direction as u)



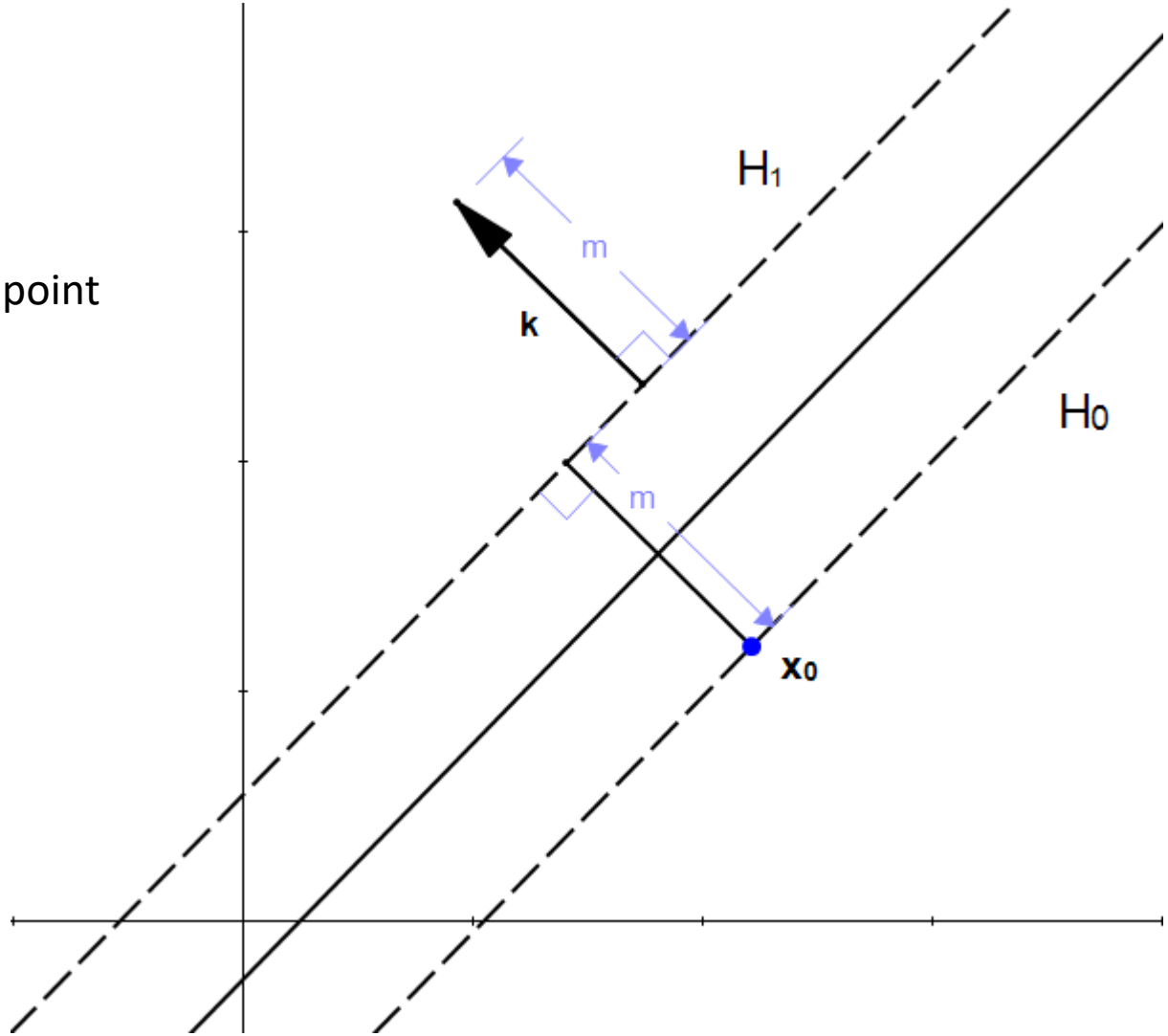
Do It Skills
do it. enjoy it.

STEP5

$$\mathbf{k} = m\mathbf{u} = m\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

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 in figure

The fact that z_0 is in H_1
means that $w \cdot z_0 + b = 1$



Do It Skills
do it. enjoy it.

STEP5 CONTINUES

We can replace z_0 by x_0 +k

because that is how we constructed it.

w·(x0+k)+b=1(11)

We can now replace k using $\mathbf{k} = m\mathbf{u} = m\frac{\mathbf{w}}{\|\mathbf{w}\|}$

w·(x_0 + $m\frac{\mathbf{w}}{\|\mathbf{w}\|}$) +b=1(12)

We now expand equation (12)

w·x_0 + $m\frac{\mathbf{w}\cdot\mathbf{w}}{\|\mathbf{w}\|}$ +b=1(13)

STEP5 CONTINUES

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

$$\mathbf{w} \cdot \mathbf{x}_0 + m\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} + b = 1$$

$$\mathbf{w} \cdot \mathbf{x}_0 + m\|\mathbf{w}\| + b = 1$$

$$\mathbf{w} \cdot \mathbf{x}_0 + b = 1 - m\|\mathbf{w}\|$$

As $\mathbf{x}_0$ is in $\mathcal{H}_0$ then $\mathbf{w} \cdot \mathbf{x}_0 + b = -1$

$$-1 = 1 - m\|\mathbf{w}\|$$

$$m\|\mathbf{w}\| = 2$$

$$m = \frac{2}{\|\mathbf{w}\|}$$

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

Do It Skills
do it. enjoy it.

HOW TO MAXIMIZE THE DISTANCE BETWEEN OUR TWO HYPERPLANES

We now have a formula to compute the margin:

$$m = \frac{2}{\|\mathbf{w}\|}$$

The only variable we can change in this formula is the norm of w.

Let's try to give it different values:

When ‖w‖=1 then m=2

When ‖w‖=2 then m=1

When ‖w‖=4 then m=12

Maximizing the margin is the same thing as minimizing the norm of w

COST FUNCTION

Our objective function is then to minimize this function

$$\max \frac{2}{\|\mathbf{w}\|} \rightarrow \max \frac{1}{\|\mathbf{w}\|} \rightarrow \min \|\mathbf{w}\| \rightarrow \min \frac{1}{2}\|\mathbf{w}\|^2$$

$$min\ \frac{1}{2}\|w\|^2 \quad \text{since} \quad \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{2}x^2 = x$$

Now, in most machine learning algorithms, we'd use something like gradient descent to minimize said function, however, for support vector machines, we use the Lagrange Multiplier.

$$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left[\frac{1}{N}\sum_{i}^{n} \max\left(0,\ 1 - y_i * (\mathbf{w} \cdot x_i + b)\right)\right]$$

Do It Skills
do it. enjoy it.

Do It Skills
do it. enjoy it.

KERNEL TRICK

KERNEL SVM

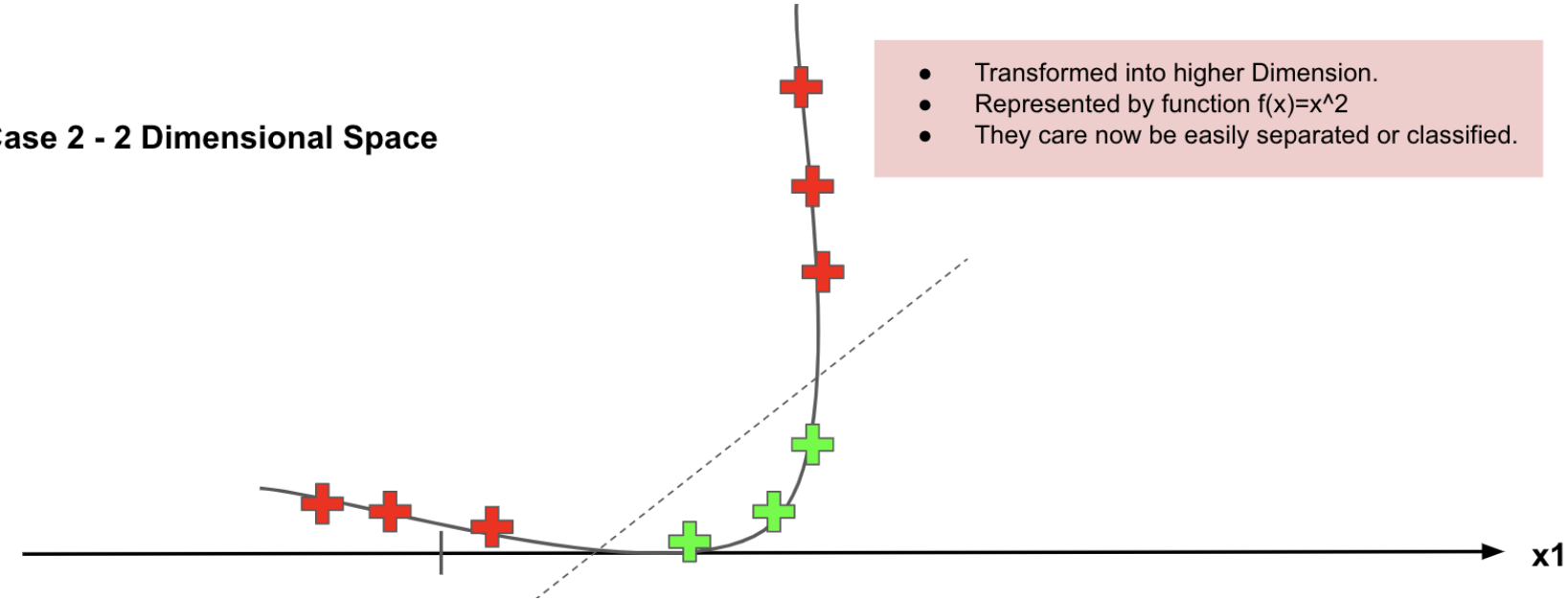
NON LINEAR SVM

Case 1 - 1 Dimensional Space

- Points in 1 Dimension Plan.
- Represented by function f(x)=x
- They cannot be separated or classified.

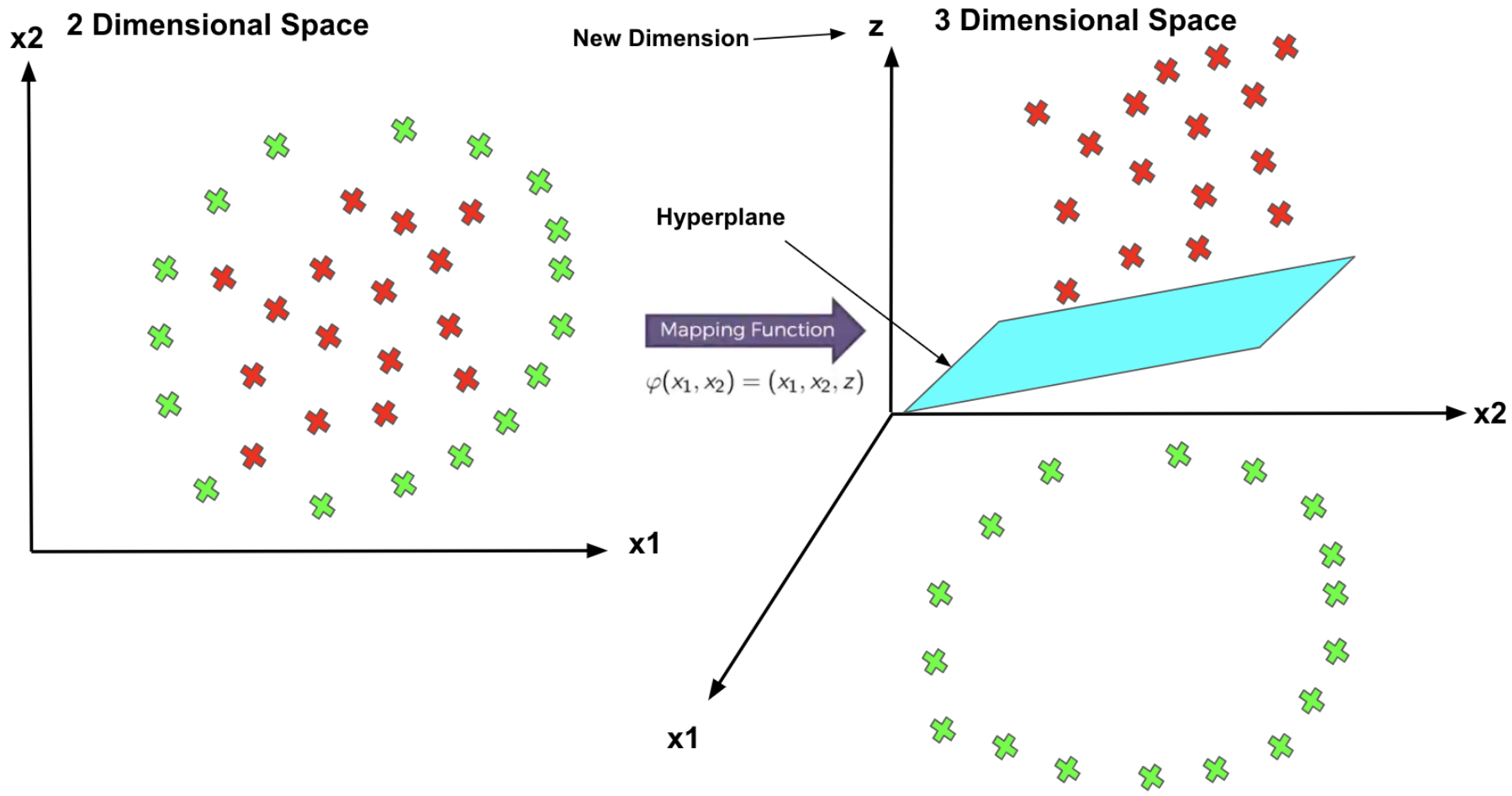
Case 2 - 2 Dimensional Space

- Transformed into higher Dimension.
- Represented by function f(x)=x^2
- They care now be easily separated or classified.



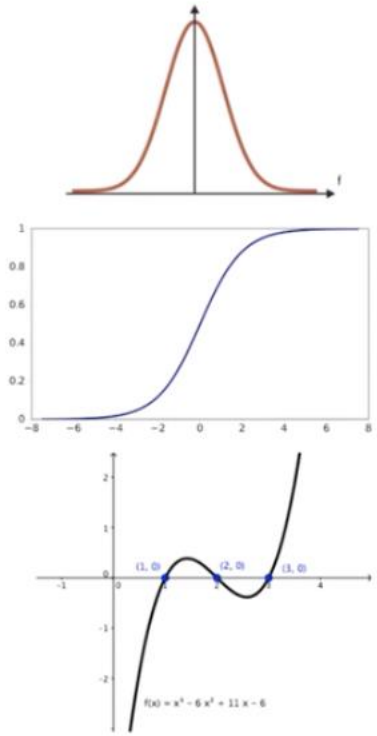
Do It Skills
do it. enjoy it.

MAPPING TO A HIGHER DIMENSION



Do It Skills
do it. enjoy it.

FREQUENTLY USED KERNELS



Gaussian RBF Kernel

$$K(\vec{x}, \vec{l}^i) = e^{-\frac{\|\vec{x}-\vec{l}^i\|^2}{2\sigma^2}}$$

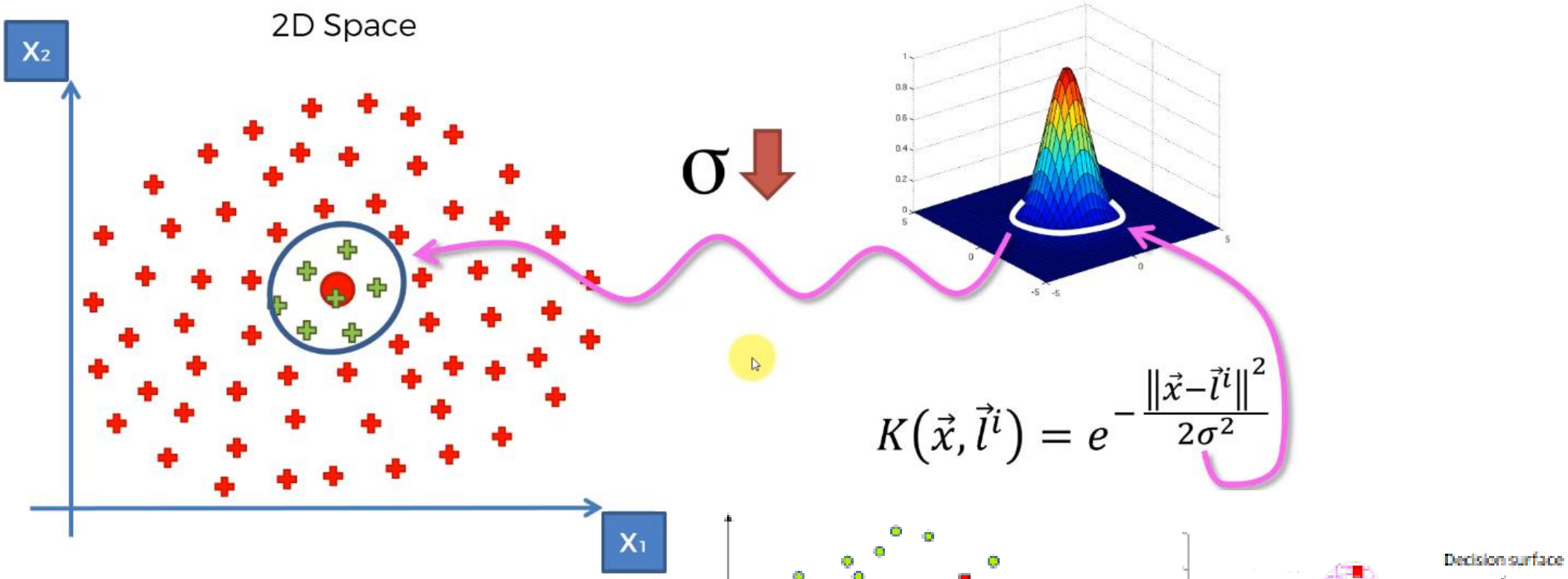
Sigmoid Kernel

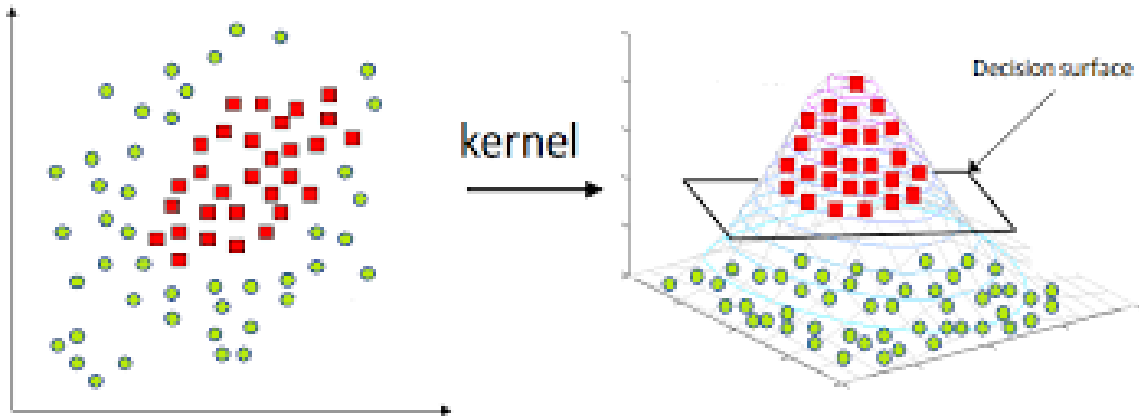
$$K(X, Y) = \tanh(\gamma \cdot X^T Y + r)$$

Polynomial Kernel

$$K(X, Y) = (\gamma \cdot X^T Y + r)^d, \gamma > 0$$

GAUSSIAN KERNEL





Do It Skills
do it. enjoy it.

LINEAR AND POLYNOMIAL KERNEL

A linear kernel can be used as normal dot product any two given observations. The product between two vectors is the sum of the multiplication of each pair of input values.

$K(x, xi) = sum(x * xi)$

A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or nonlinear input space.

$K(x,xi) = 1 + sum(x * xi)^d$

RADIAL BASIS FUNCTION KERNEL(RBF)

The Radial basis function kernel is a popular kernel function commonly used in support vector machine classification. RBF can map an input space in infinite dimensional space.

$$K(x,xi) = \exp(-gamma * sum((x - xi^2))$$

Here gamma is a parameter, which ranges from 0 to 1. A higher value of gamma will perfectly fit the training dataset, which causes over-fitting. Gamma=0.1 is considered to be a good default value. The value of gamma needs to be manually specified in the learning algorithm.

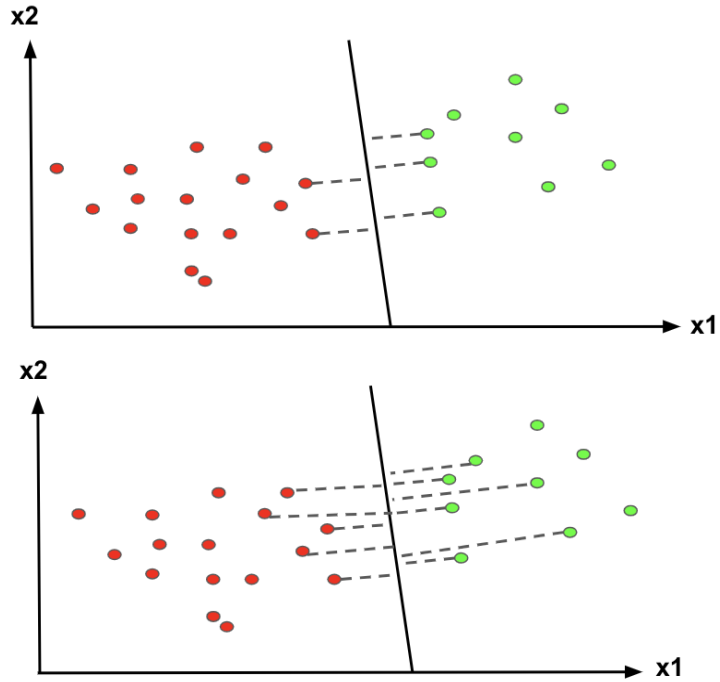
GAMMA VS C PARAMETER

For a linear kernel, we just need to optimize the c parameter. However, if we want to use an RBF kernel, both c and gamma parameter need to optimized simultaneously.

	Large Gamma	Small Gamma	Large C	Small C
Variance	Low	High	High	Low
Bias	High	Low	Low	High

GAMMA IN HYPER PARAMETER

Gamma: It defines how far influences the calculation of plausible line of separation.



High Gamma

- only near points are considered.

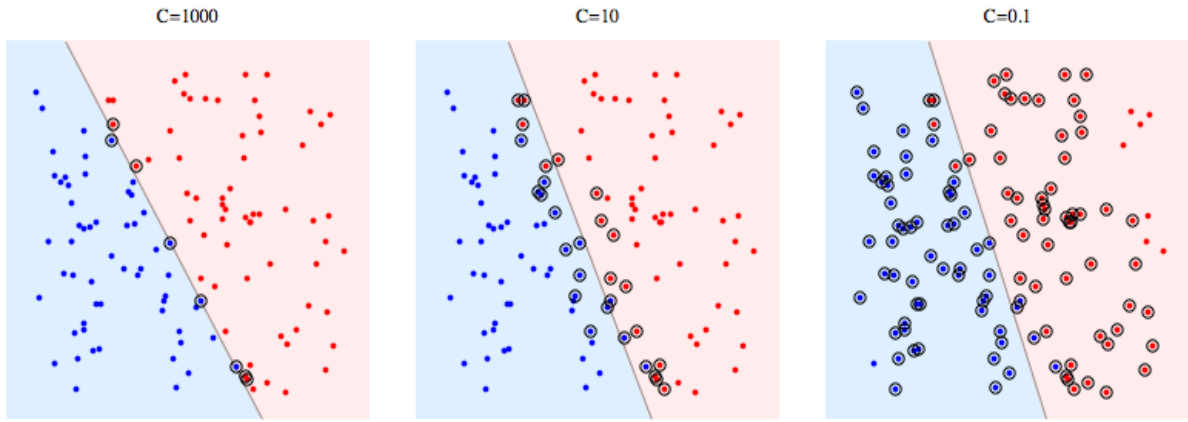
Low Gamma

- far away points are also considered

Do It Skills
do it. enjoy it.

C IN HYPER PARAMETER

C (Regularization): C is the penalty parameter, which represents misclassification or error term. The misclassification or error term tells the SVM optimization how much error is bearable. This is how you can control the trade-off between decision boundary and misclassification term.



TUNING THE HYPER-PARAMETERS OF AN ESTIMATOR

Hyper-parameters are parameters that are not directly learnt within estimators. In scikit-learn, they are passed as arguments to the constructor of the estimator classes. **Grid search** is commonly used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

Do It Skills
do it. enjoy it.

THANK YOU

ARUNKG99@GMAIL.COM

WWW.DOITSKILLS.COM