Lecture 2 : SVMs

By : Khalil idrissi

Recap

Recap

Linear regression

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Recap

Linear regression

$$y(x) = w^T x + \varepsilon \quad \text{where } w, \varepsilon, x, y \text{ are vectors}$$

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Using GD ,SGD,...

Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Using GD ,SGD,…

$$w = w - \eta \frac{\partial J(w)}{\partial w}$$

Recap

Linear regression

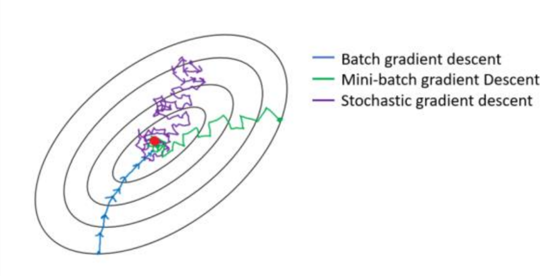
$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Using GD ,SGD,...

$$w = w - \eta \frac{\partial J(w)}{\partial w}$$



Recap

Linear regression

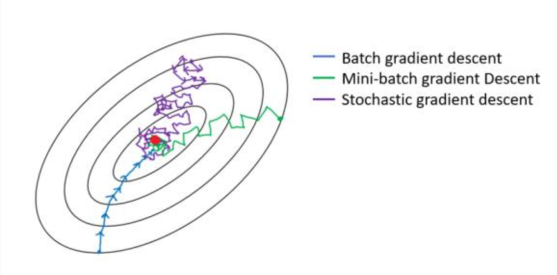
$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

Using GD ,SGD,...

$$w = w - \eta \frac{\partial J(w)}{\partial w}$$



Recap

Linear regression

$y(x) = w^T x + \varepsilon$ where w, ε, x, y are vectors

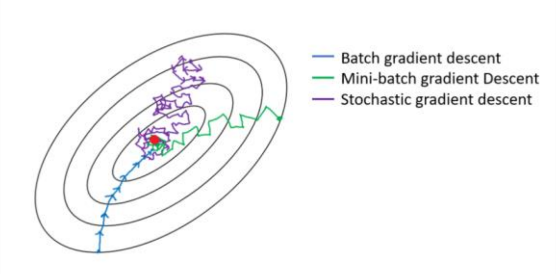
Goal : find the parameters w to obtain the best fit hyperplane

Using OLS

MLE ,MAP, full bayesian approach

Using GD ,SGD,...

$$w = w - \eta \frac{\partial J(w)}{\partial w}$$



SVMs =Support Vector Machines

From Wikipedia, the free encyclopedia

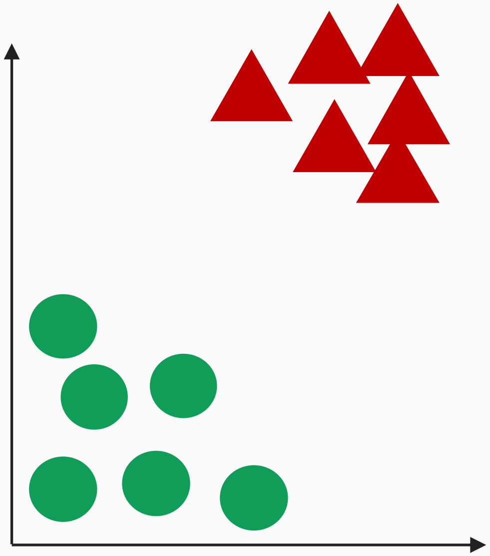
In machine learning, **support-vector machines** (**SVMs**, also **support-vector networks**[1]) are supervised learning models with associated learning algorithms

Quiz

Quiz

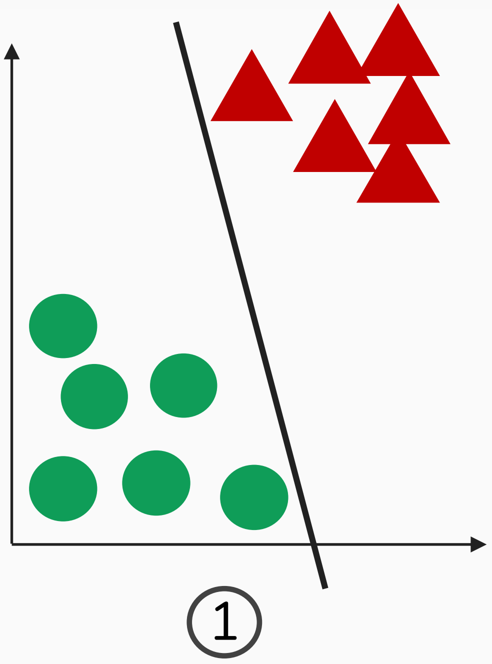
①

Quiz

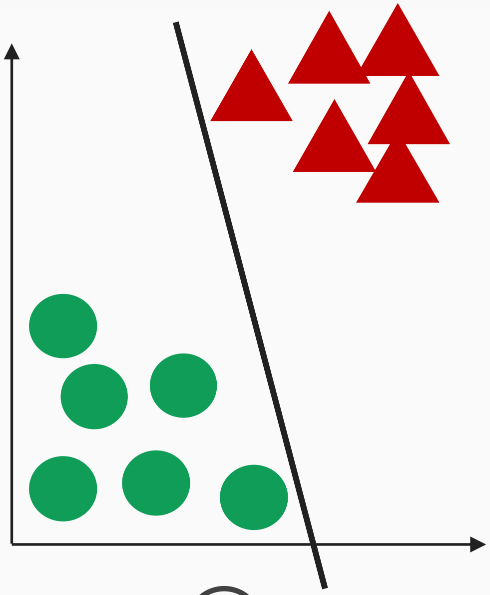


①

Quiz



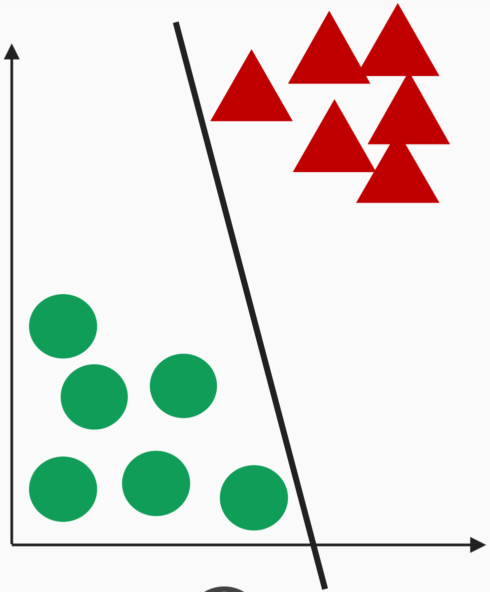
Quiz



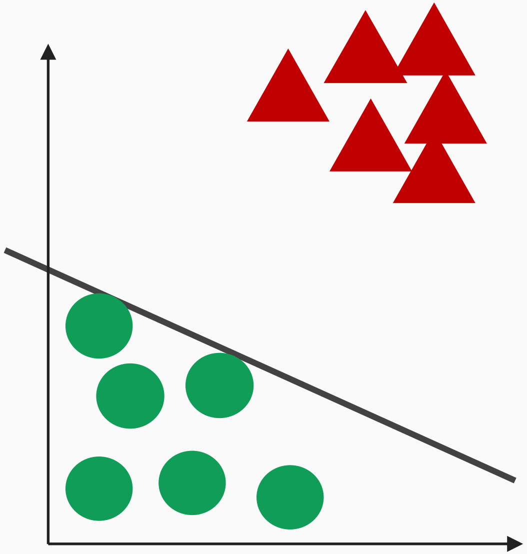
①

②

Quiz

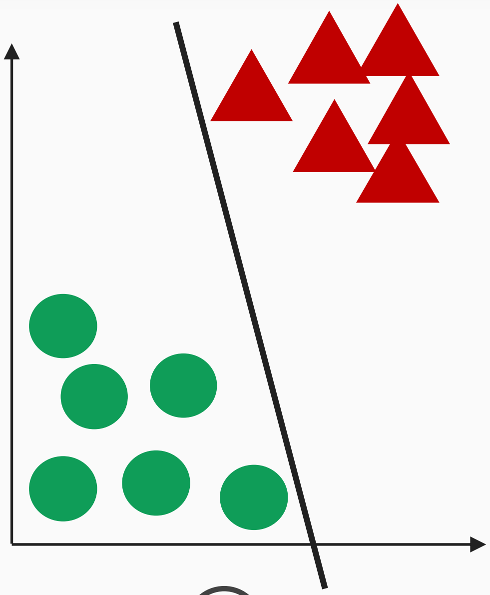


1

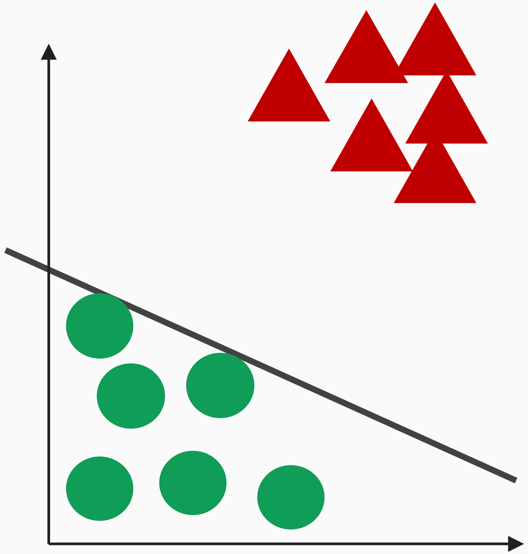


2

Quiz



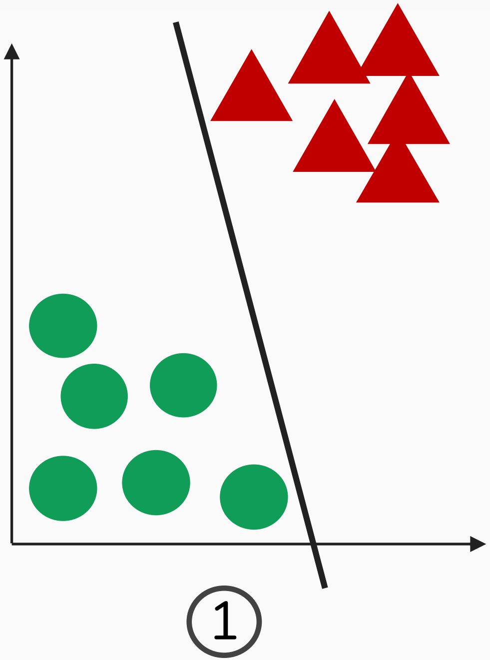
①

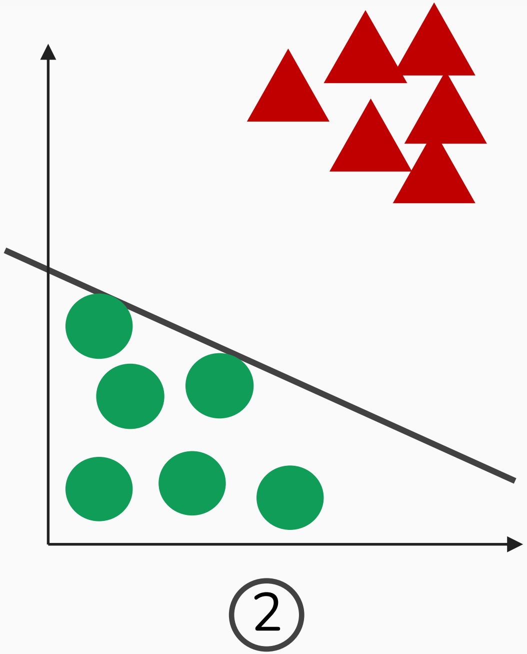


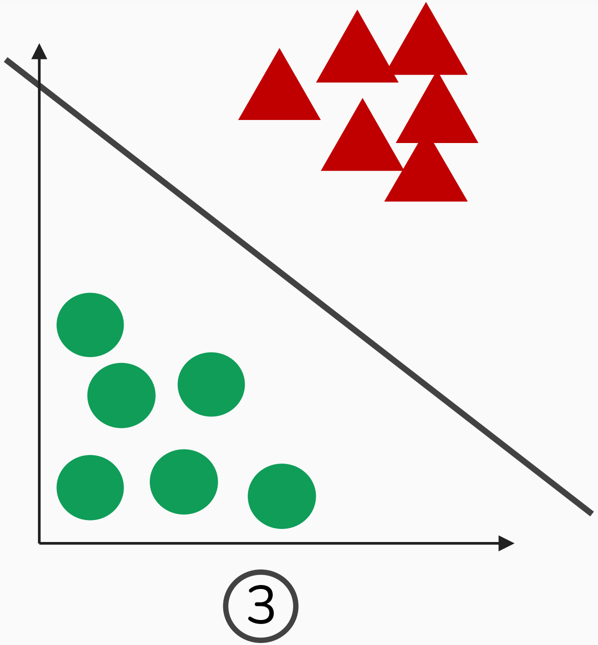
②

③

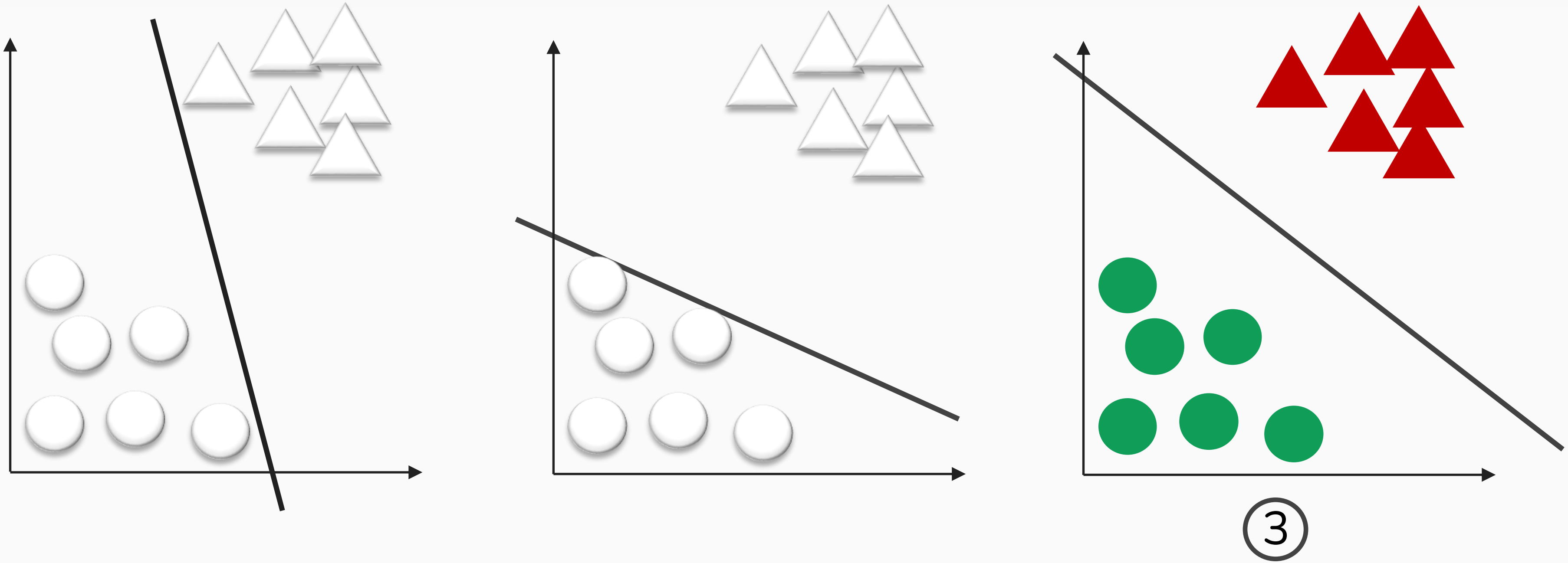
Quiz





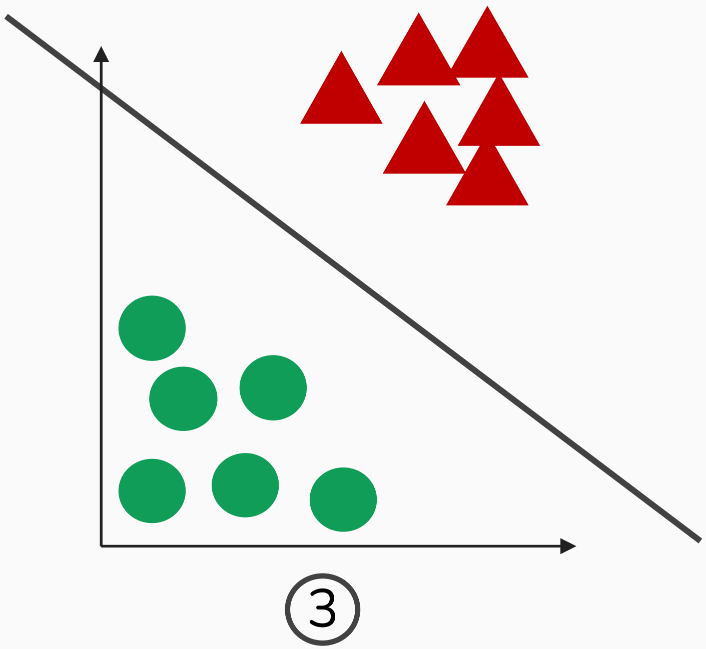


Quiz

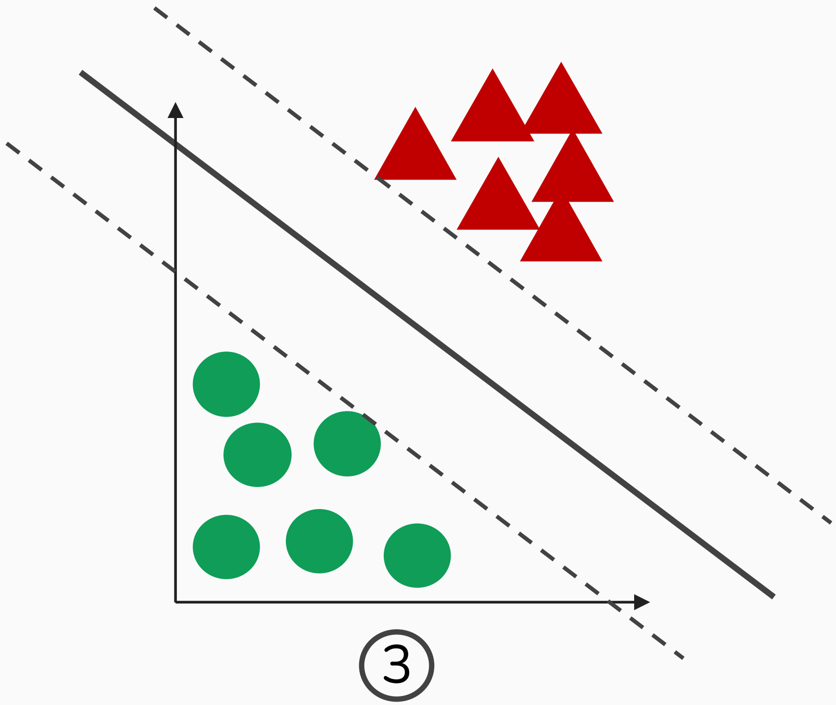


What did we do ?

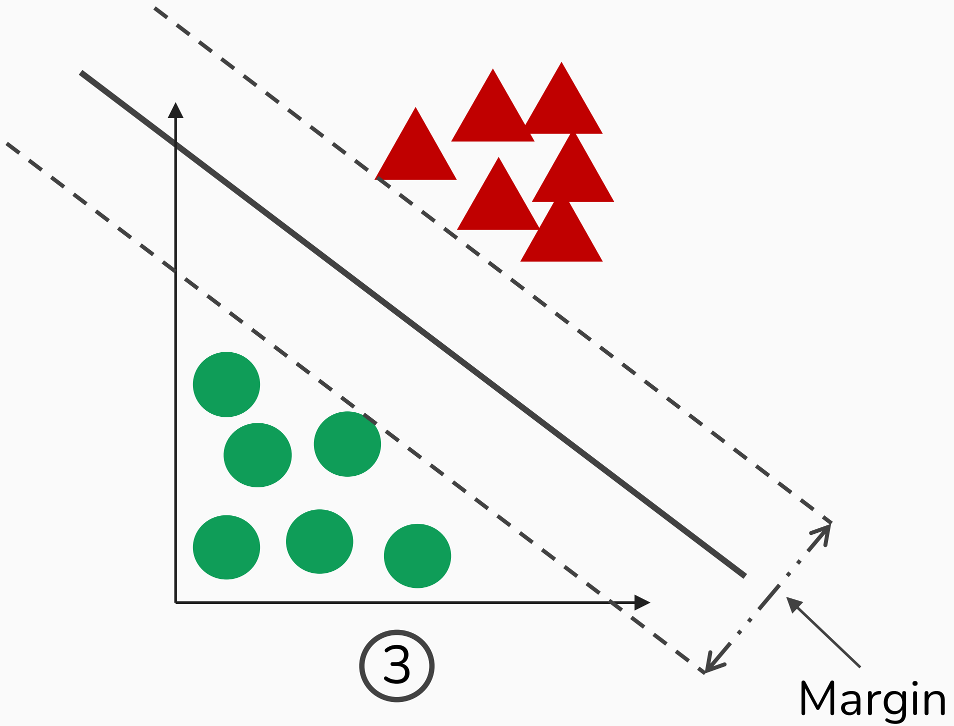
What did we do ?



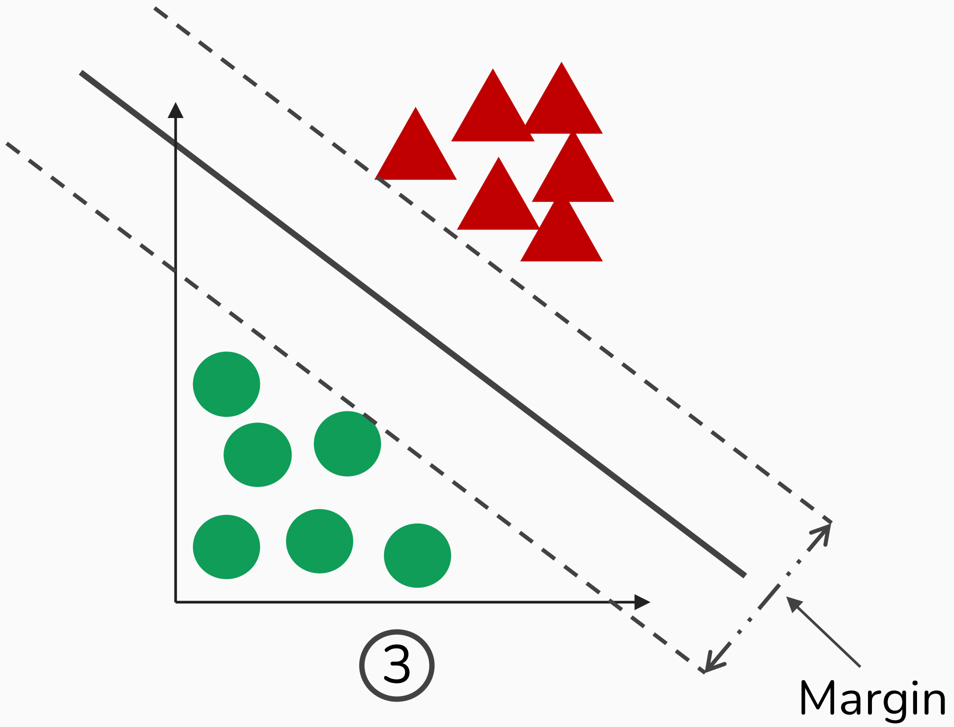
What did we do ?



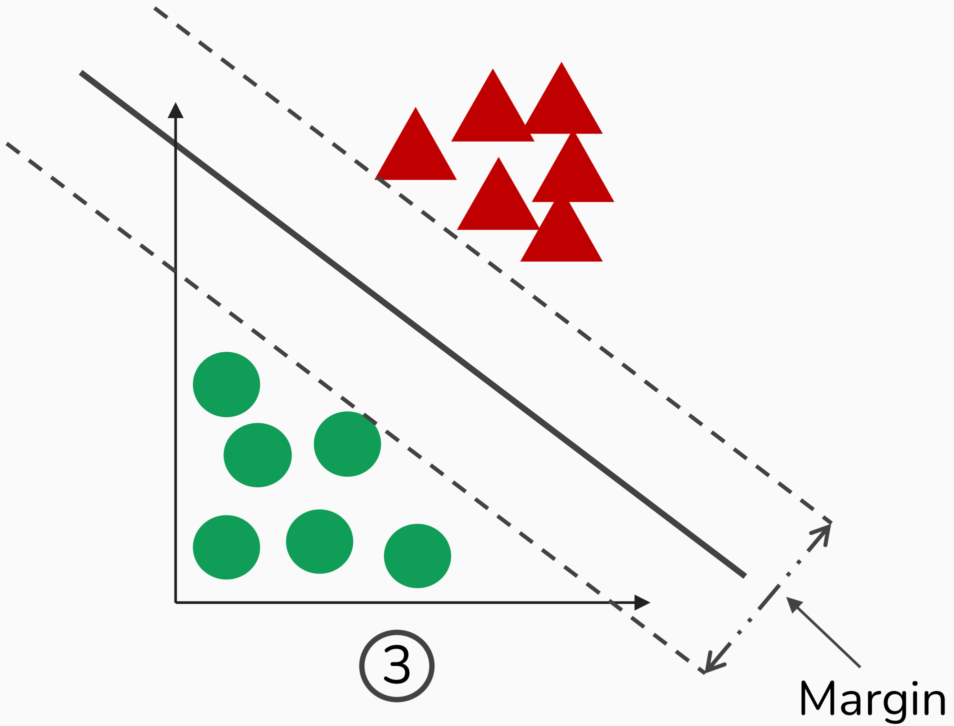
We maximized the margin

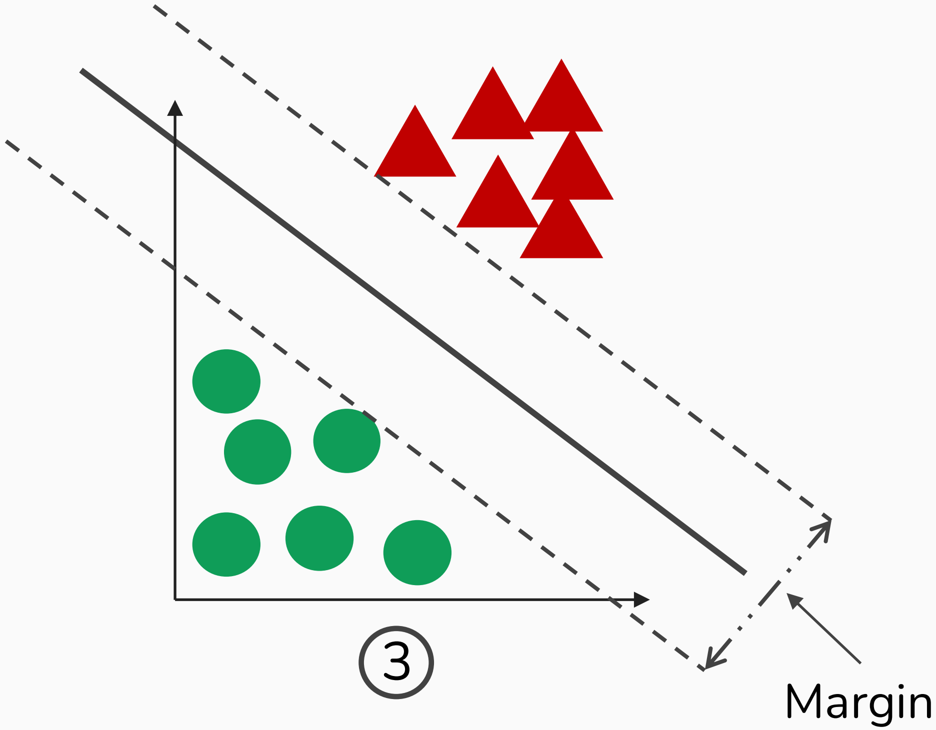


And … This is ….

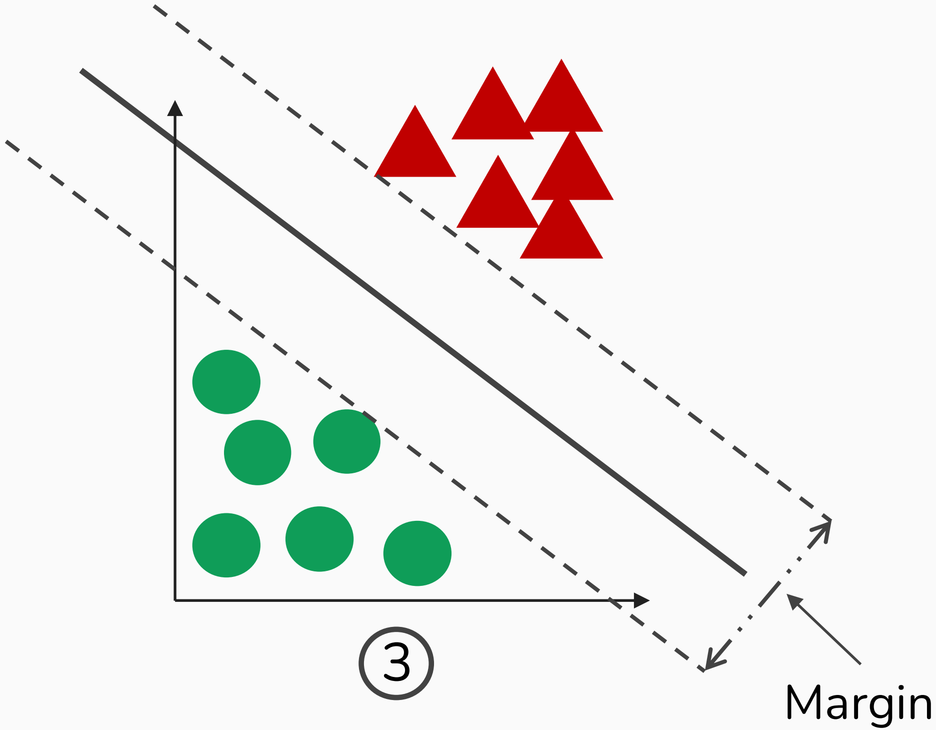


And … This is …. SVMs

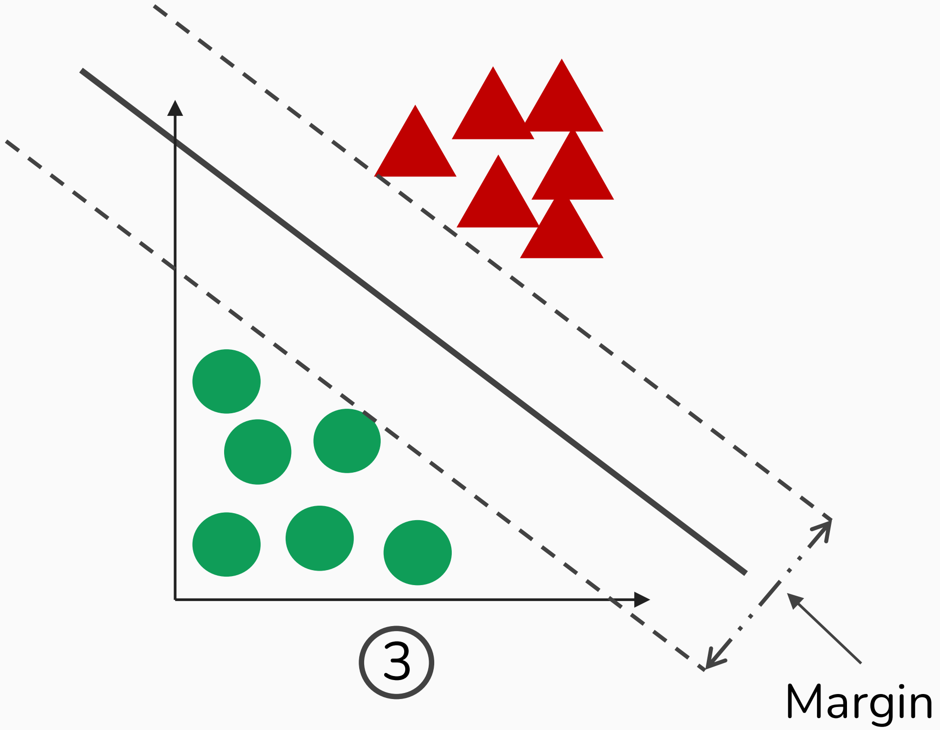



3
Margin

SVMs algorithm learns a linear model (hyperplane)

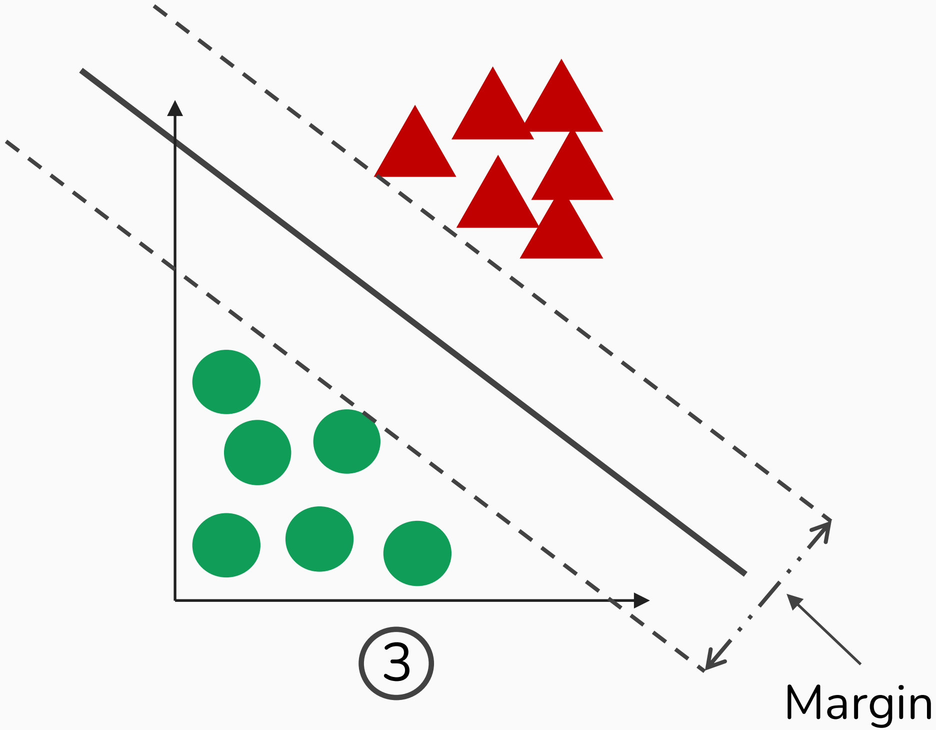


SVMs algorithm learns a linear model (hyperplane)
SVMs algorithm maximizes the margin around the hyperplane



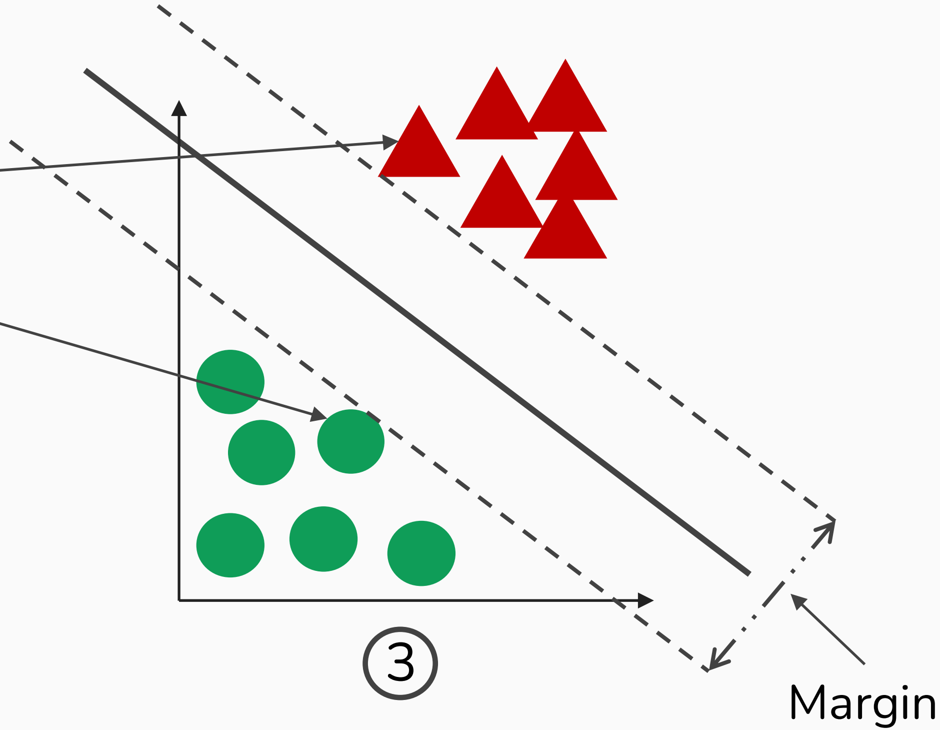
SVMs algorithm learns a linear model (hyperplane)
SVMs algorithm maximizes the margin around the hyperplane

These points are called
support vectors



SVMs algorithm learns a linear model (hyperplane)
SVMs algorithm maximizes the margin around the hyperplane

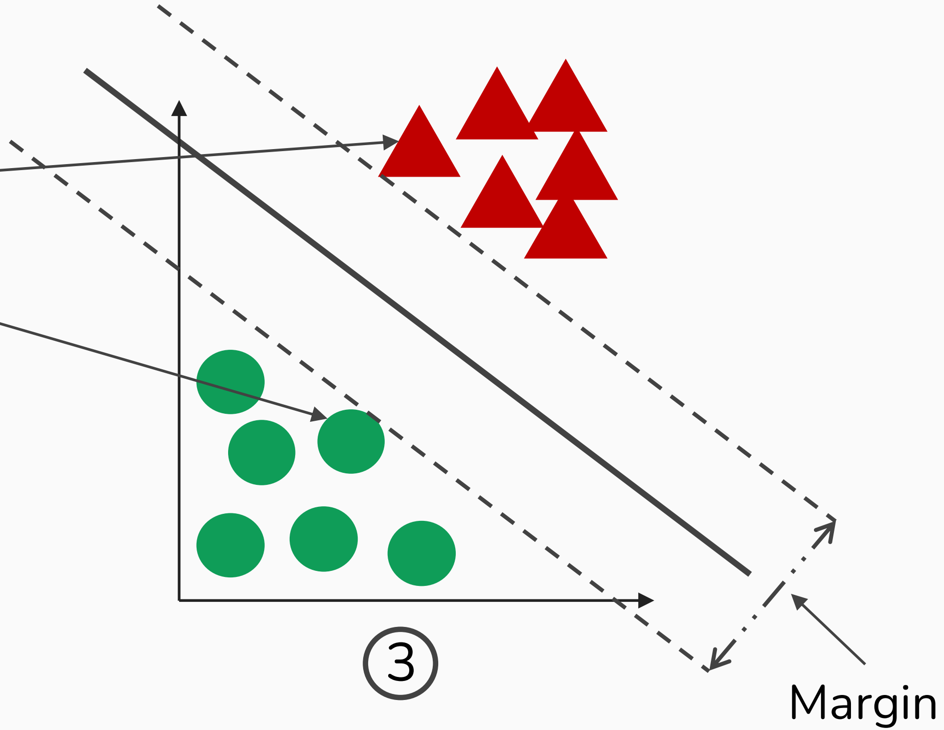
These points are called
support vectors



SVMs algorithm learns a linear model (hyperplane)
SVMs algorithm maximizes the margin around the hyperplane

These points are called support vectors

These points are the data points that lie closest to the hyperplane

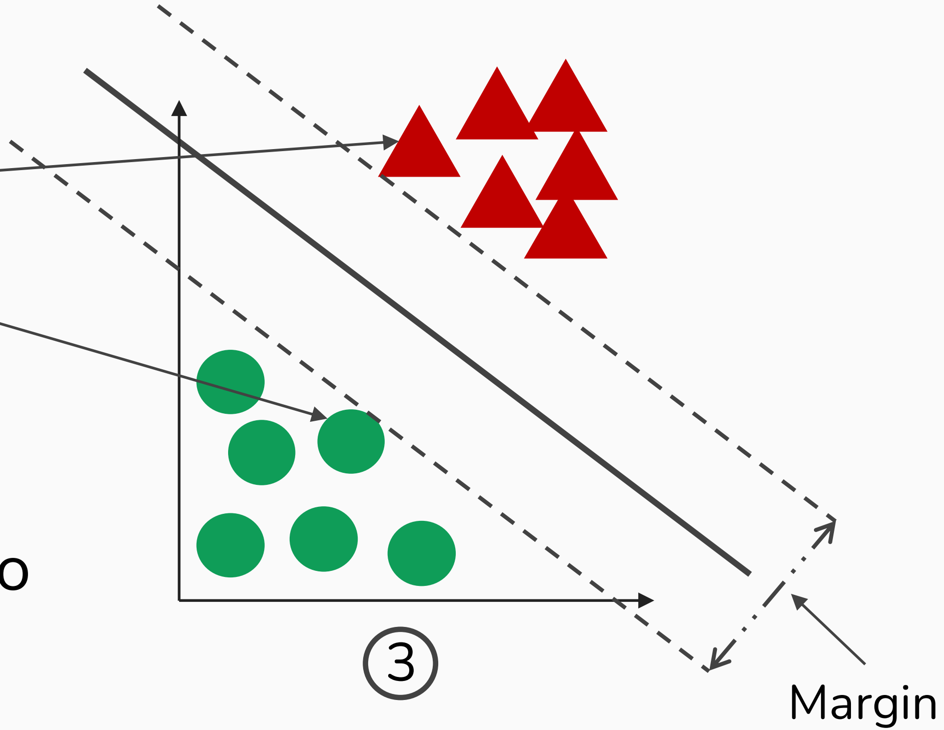


SVMs algorithm learns a linear model (hyperplane)
SVMs algorithm maximizes the margin around the hyperplane

These points are called support vectors

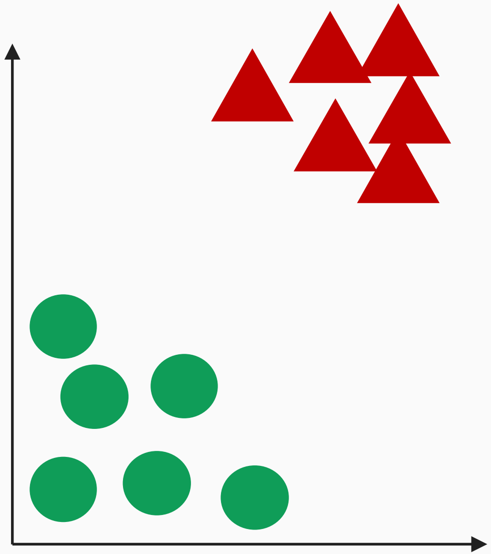
These points are the data points that lie closest to the hyperplane

The data points most difficult to classify



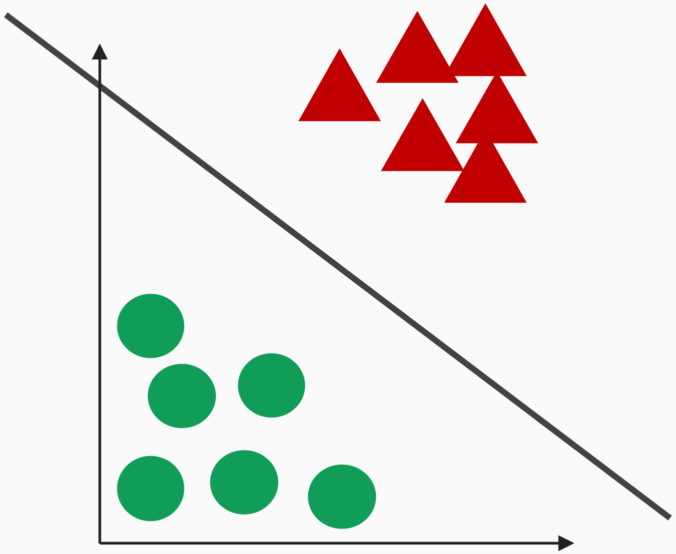
2 types of data :

2 types of data :



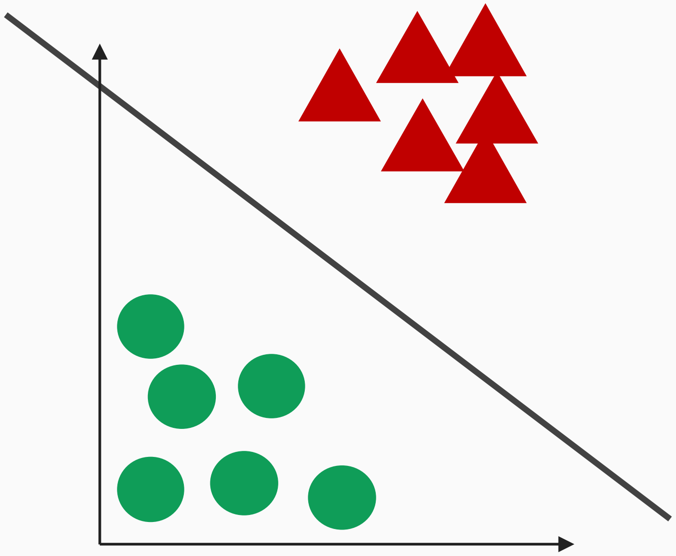
2 types of data :

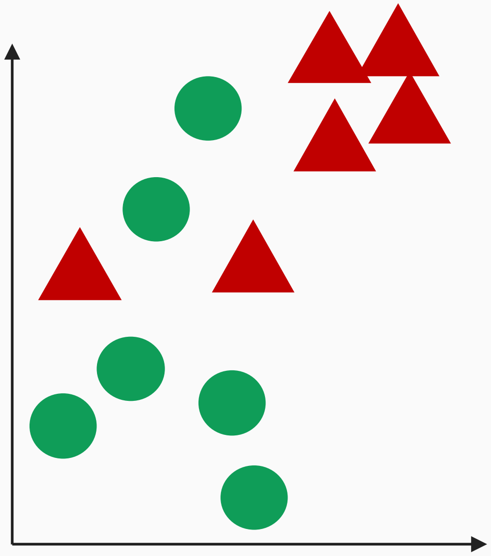
Linearly separable data



2 types of data :

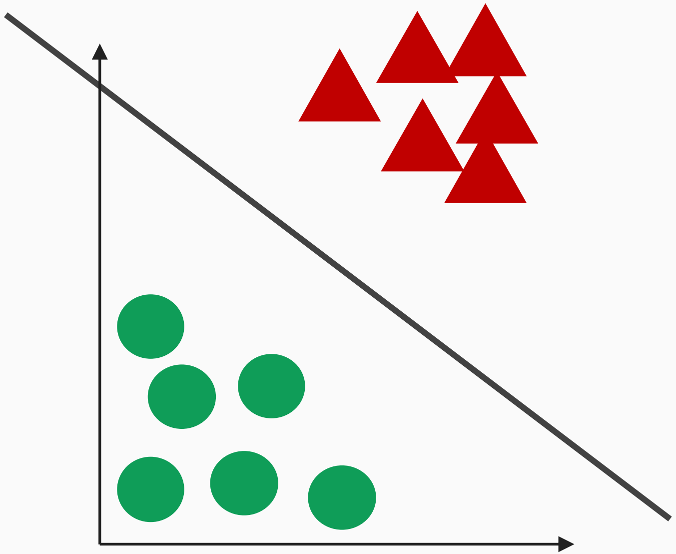
Linearly separable data



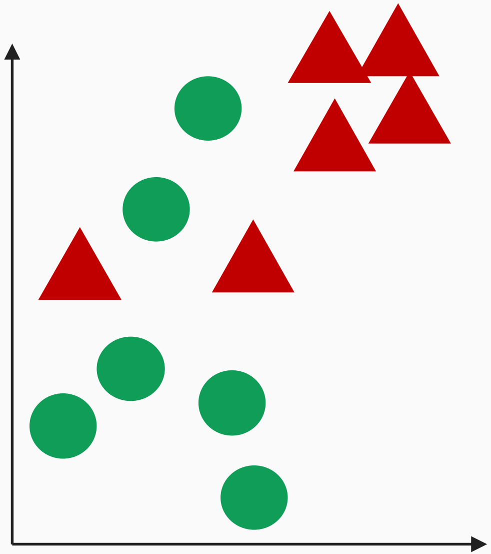


2 types of data :

Linearly separable data

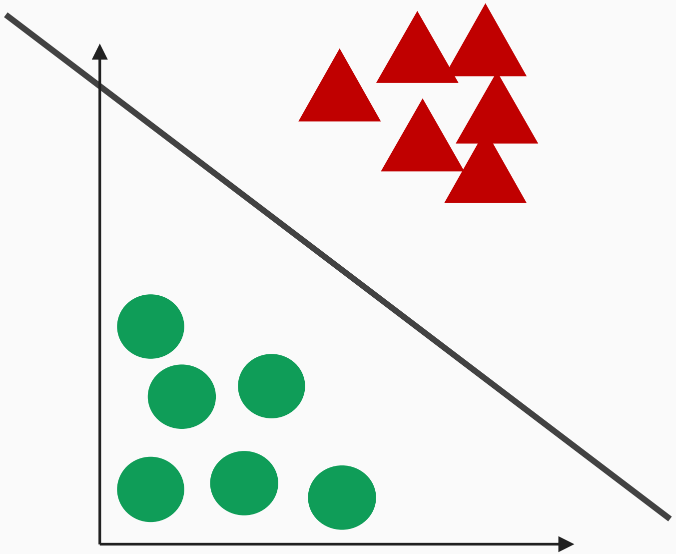


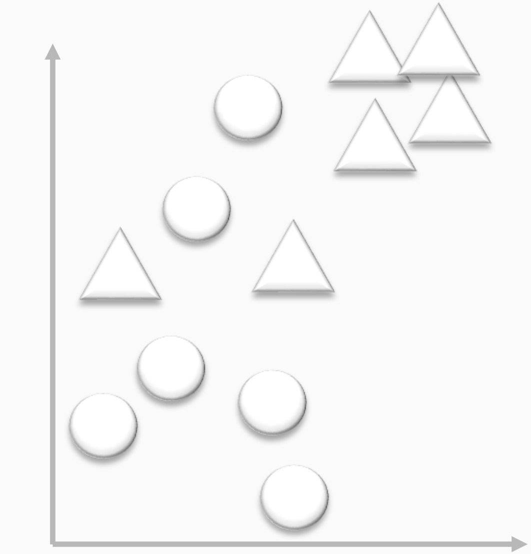
Non linearly separable data



2 types of data :

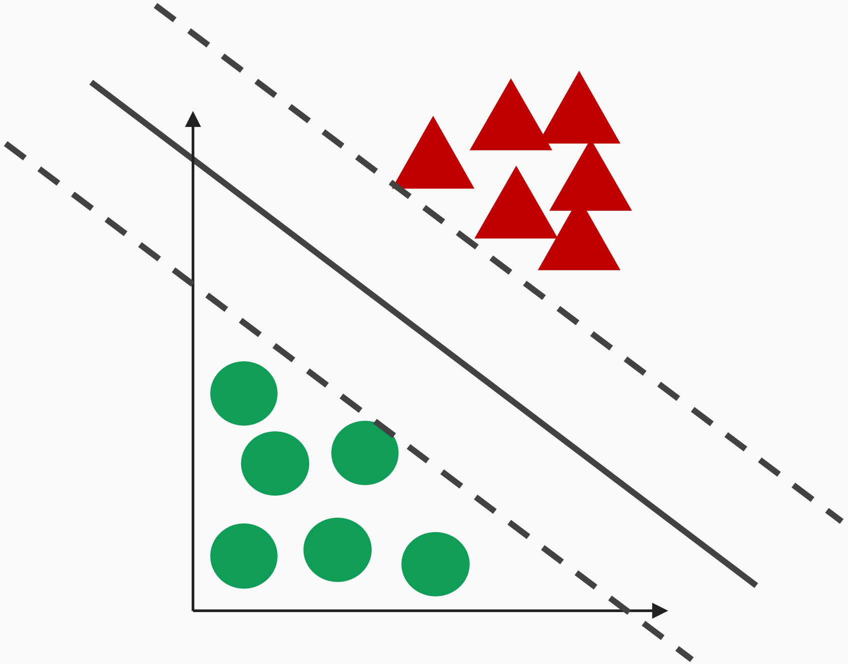
Linearly separable data





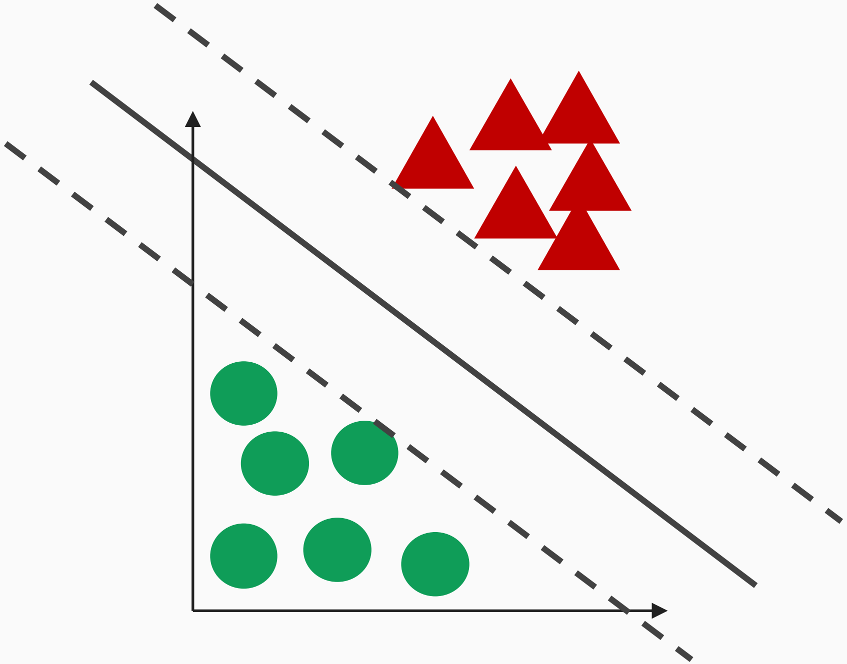
Linearly separable data

Linearly separable data



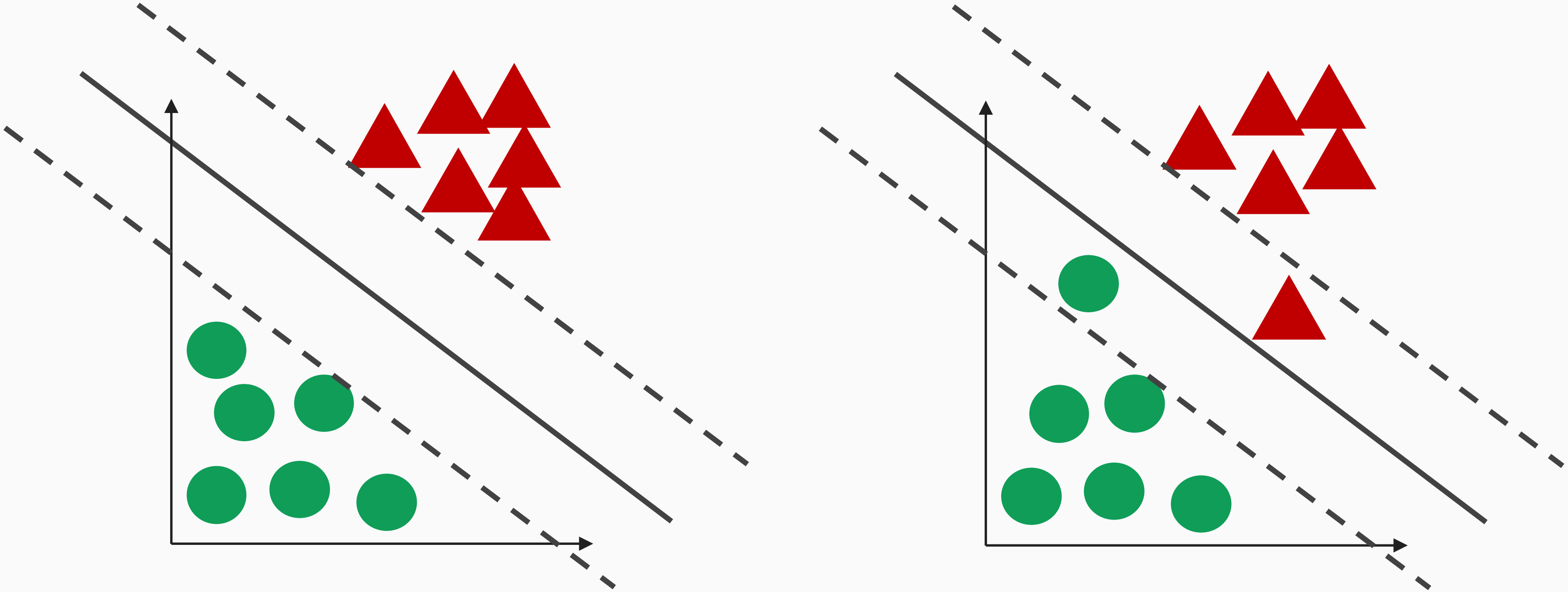
Linearly separable data

Hard margin SVM



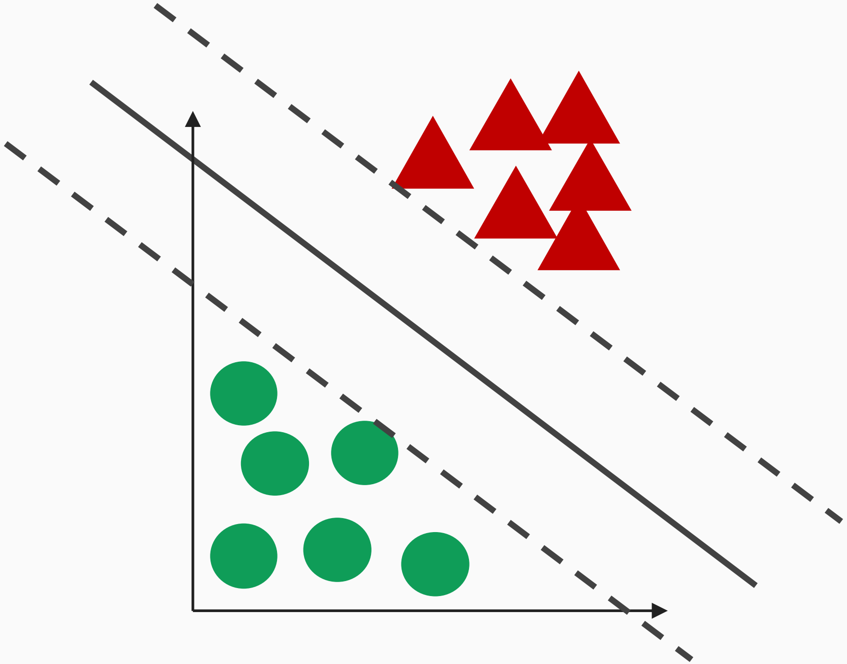
Linearly separable data

Hard margin SVM

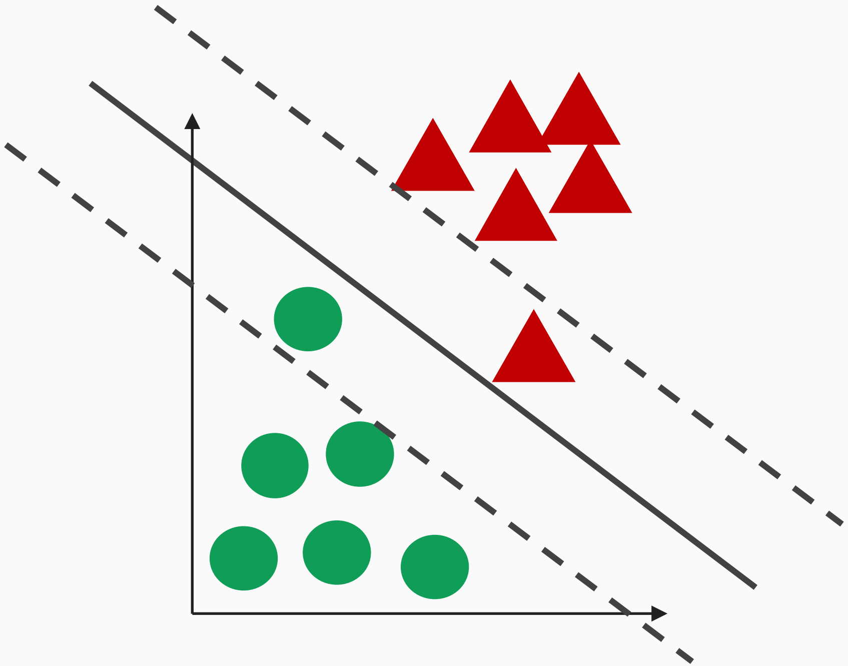


Linearly separable data

Hard margin SVM



Soft margin SVM



NOTE : hard margin SVM is only applicable when the dataset is linearly separable

WARNING !

The rest of the slides
are math heavy !!

(LINEAR) HARD MARGIN SVMs

Equation of a line : $y = ax + b \Rightarrow y - ax - b = 0$

Equation of an hyperplane: $w^T x = 0 \Leftrightarrow w^T x + b = 0$

$assuming\ that\ x_0 = 1$

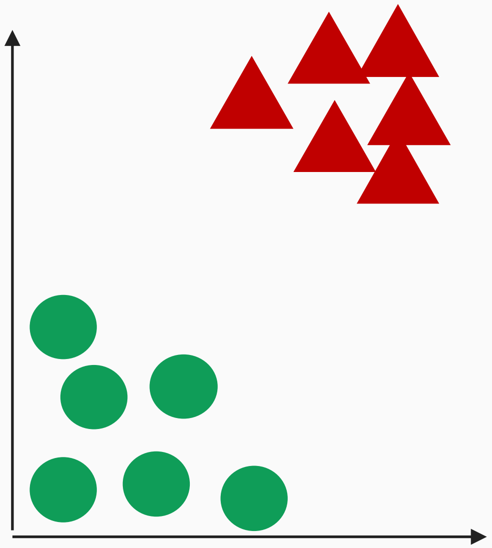
First step : prepare the dataset

First step : prepare the dataset

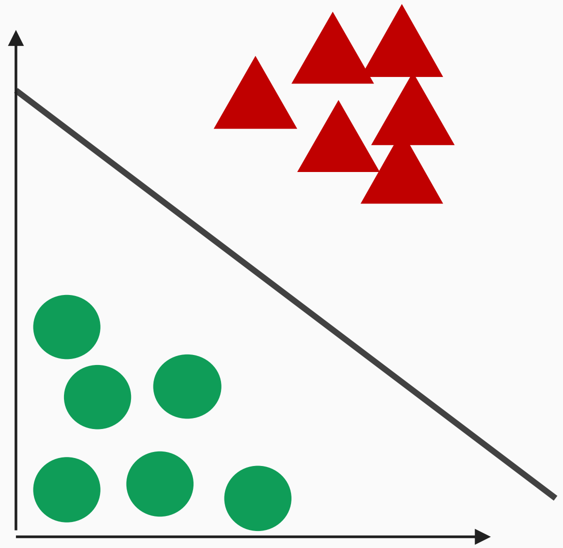
$$D = \{(x_i, y_i) \,|\, x_i \in \mathbb{R}^p, y_i \in \{\bullet, \blacktriangle\}\}_{i=1}^{N}$$

Second Step : select 2 hyperplanes that separates the data

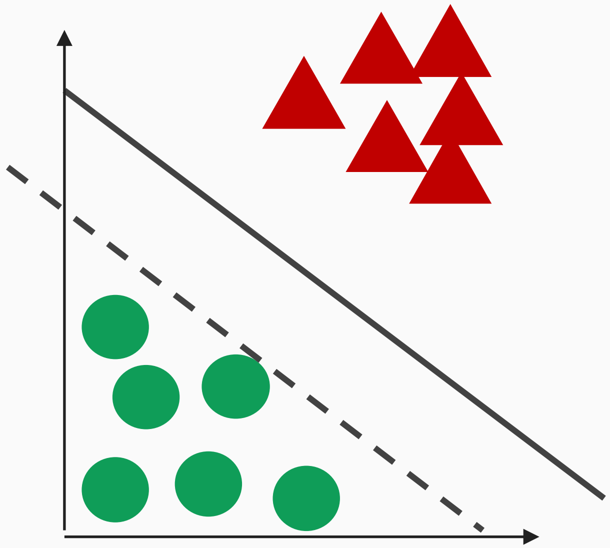
Second Step : select 2 hyperplanes that separates the data



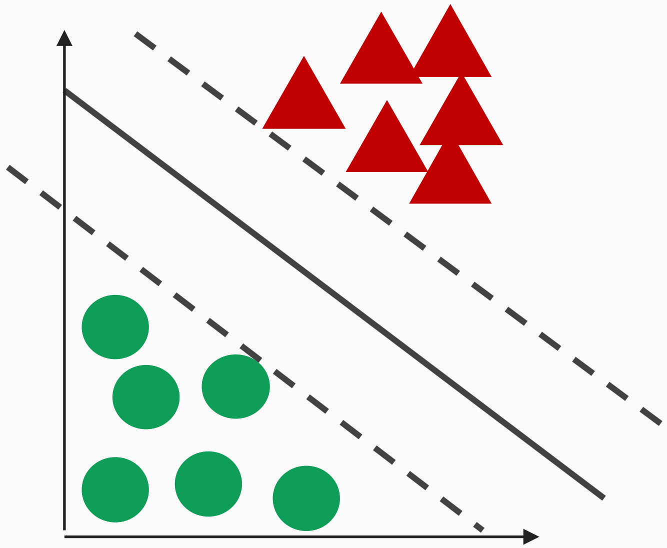
Second Step : select 2 hyperplanes that separates the data



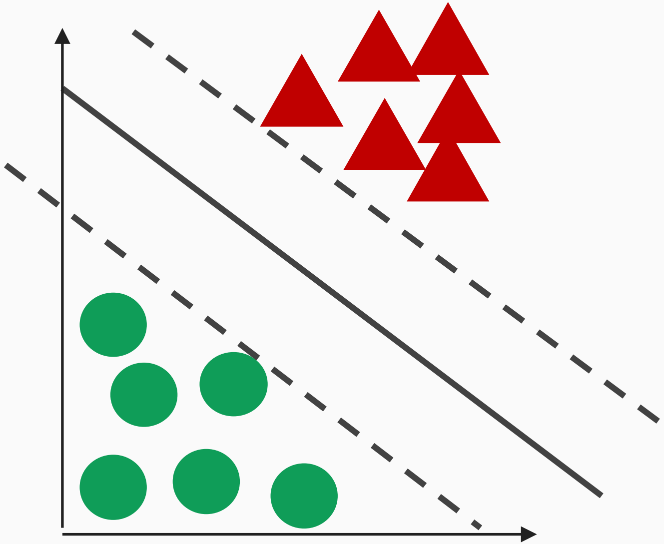
Second Step : select 2 hyperplanes that separates the data



Second Step : select 2 hyperplanes that separates the data

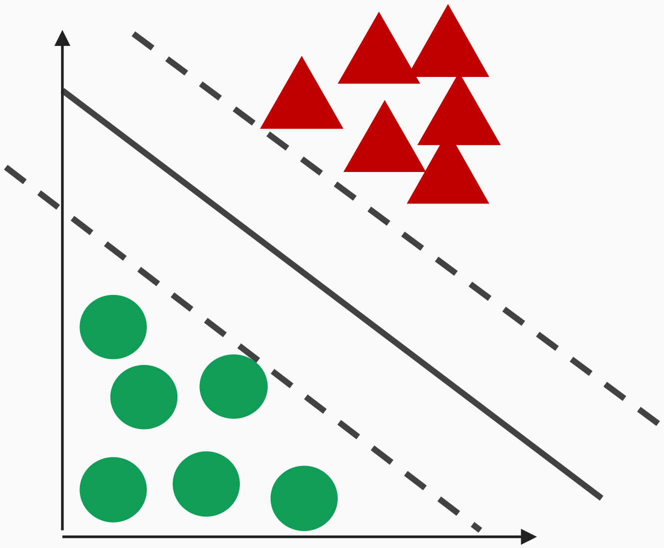


Second Step : select 2 hyperplanes that separates the data



1-Assume the equation of the hyperplane is :

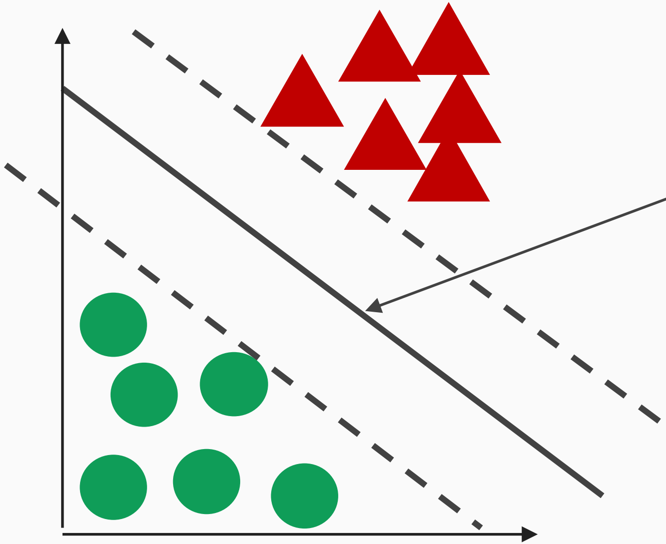
Second Step : select 2 hyperplanes that separates the data



1-Assume the equation of the hyperplane is :

$$H_0 : w^T x + b = 0$$

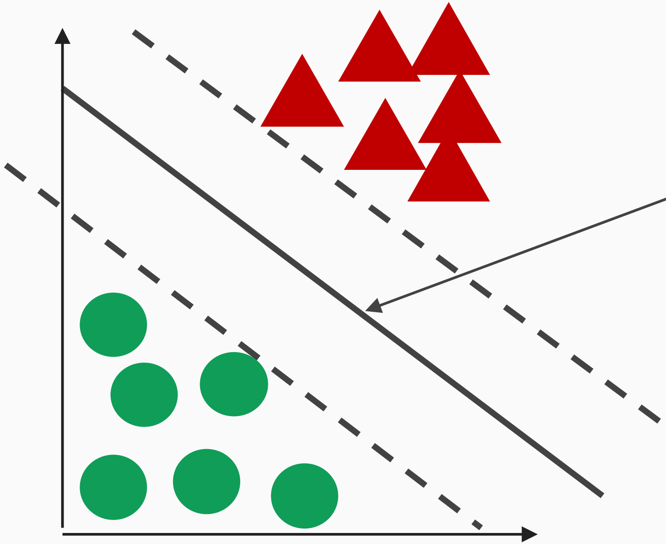
Second Step : select 2 hyperplanes that separates the data



1-Assume the equation of the hyperplane is :

$$H_0 : w^T x + b = 0$$

Second Step : select 2 hyperplanes that separates the data

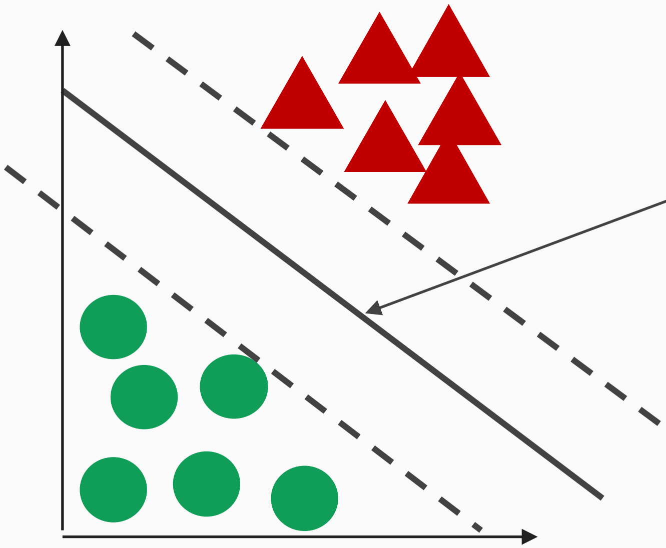


1-Assume the equation of the hyperplane is :

$$H_0 : w^T x + b = 0$$

2-We can select 2 other hyperplanes that separates the data :

Second Step : select 2 hyperplanes that separates the data



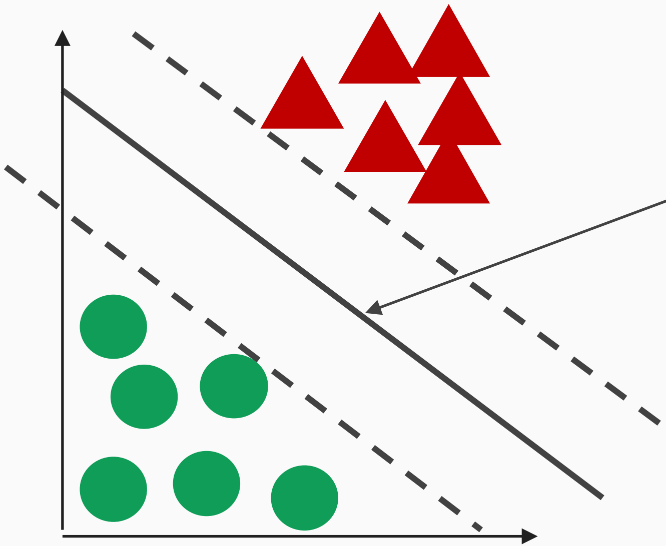
1-Assume the equation of the hyperplane is :

$$H_0 : w^T x + b = 0$$

2-We can select 2 other hyperplanes that separates the data :

$$\begin{cases} H_1 : wx + b = \delta \\ H_2 : wx + b = -\delta \end{cases}$$

Second Step : select 2 hyperplanes that separates the data



1-Assume the equation of the hyperplane is :

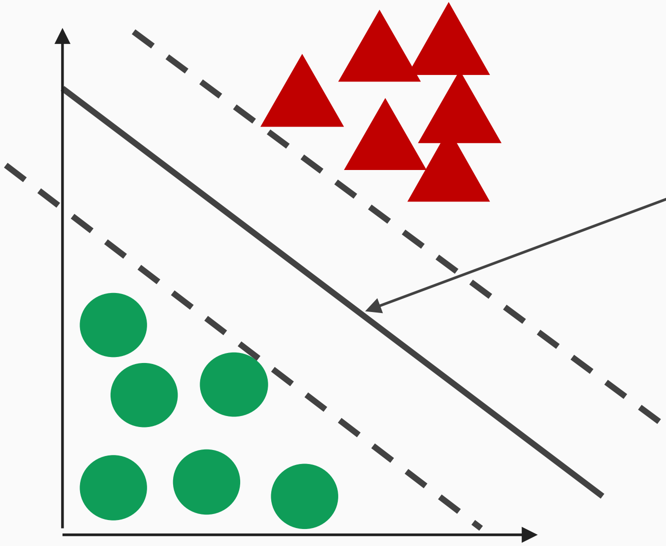
$$H_0 : w^T x + b = 0$$

2-We can select 2 other hyperplanes that separates the data :

$$\begin{cases} H_1 : wx + b = \delta \\ H_2 : wx + b = -\delta \end{cases}$$

To simplify the problem we choose $\delta = 1$

Second Step : select 2 hyperplanes that separates the data



1-Assume the equation of the hyperplane is :

$$H_0 : w^T x + b = 0$$

2-We can select 2 other hyperplanes that separates the data :

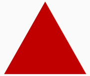
$$\begin{cases} H_1 : wx + b = \delta \\ H_2 : wx + b = -\delta \end{cases}$$

To simplify the problem we choose $\delta = 1$

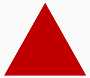
$$\begin{cases} H_1 : wx + b = 1 \\ H_2 : wx + b = -1 \end{cases}$$

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class  :

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

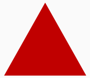
For x_i having the class  : $$wx_i + b \geq 1$$

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class ▲ : $$wx_i + b \geq 1$$

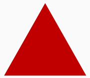
For x_i having the class ● :

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class  : $$wx_i + b \geq 1$$

For x_i having the class  : $$wx_i + b \leq -1$$

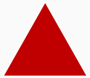
3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class  : $$wx_i + b \geq 1$$

i goes from 1 to N

For x_i having the class  : $$wx_i + b \leq -1$$

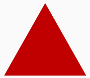
3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class  : $$wx_i + b \geq 1$$

For x_i having the class  : $$wx_i + b \leq -1$$

i goes from 1 to N

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

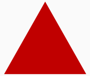
For x_i having the class  : $$wx_i + b \geq 1$$

For x_i having the class  : $$wx_i + b \leq -1$$

i goes from 1 to N

N constraints

3-Since we are dealing with Hard margin , we need to be sure that there is no margin violations by imposing the constraints :

For x_i having the class  : $$wx_i + b \geq 1$$

For x_i having the class  : $$wx_i + b \leq -1$$

i goes from 1 to N

N constraints

REMEMBER : our main goal is to maximize the margin

Geometrically , the distance between 2 hyperplanes is

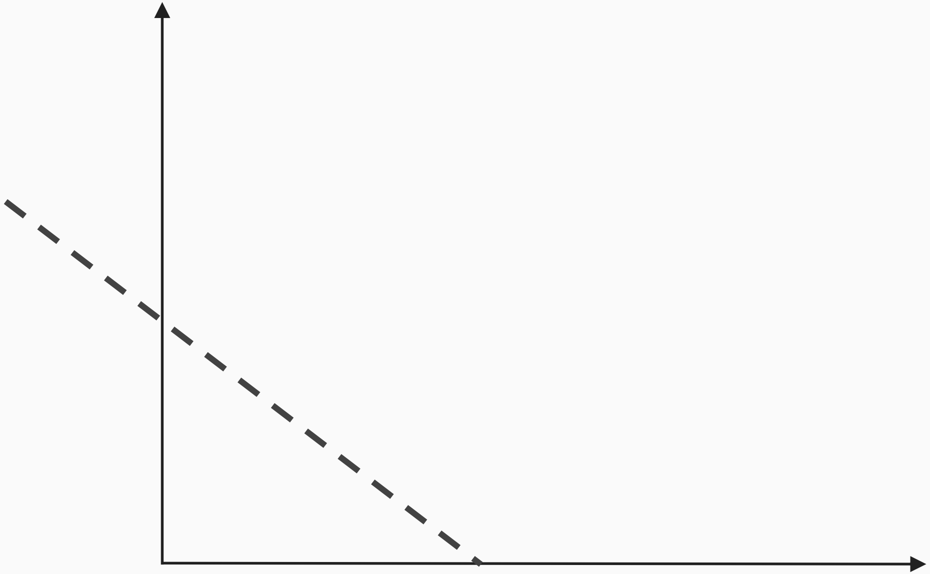
Geometrically , the distance between 2 hyperplanes is

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

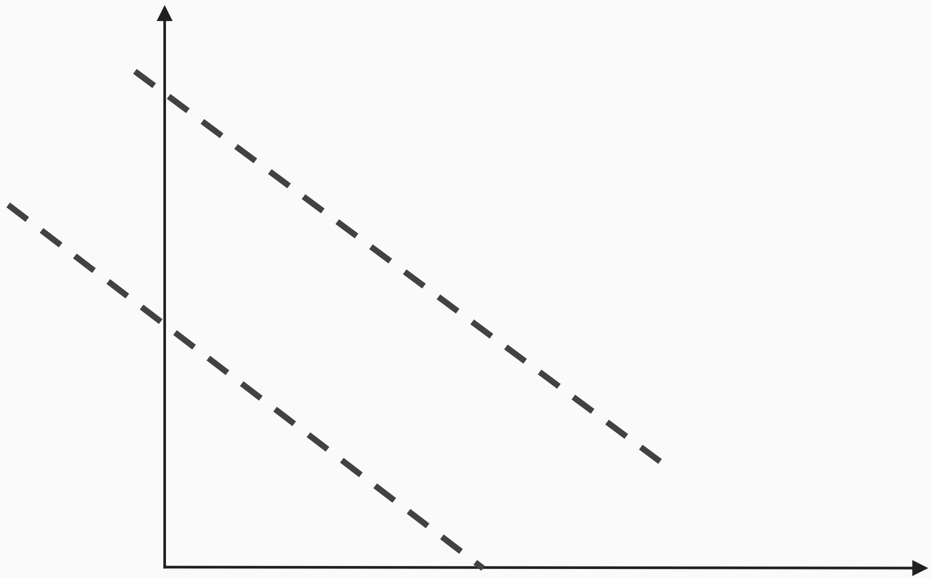
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .

Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .

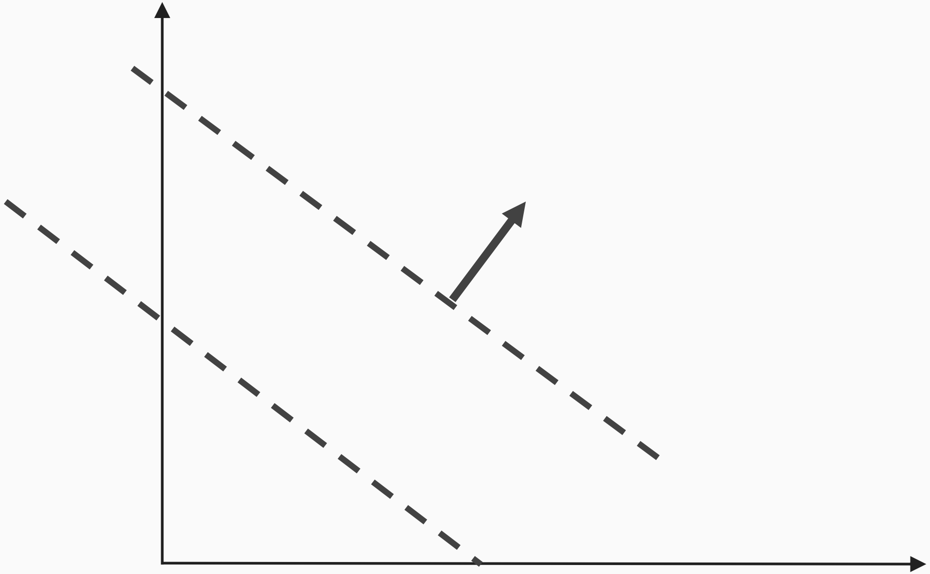
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



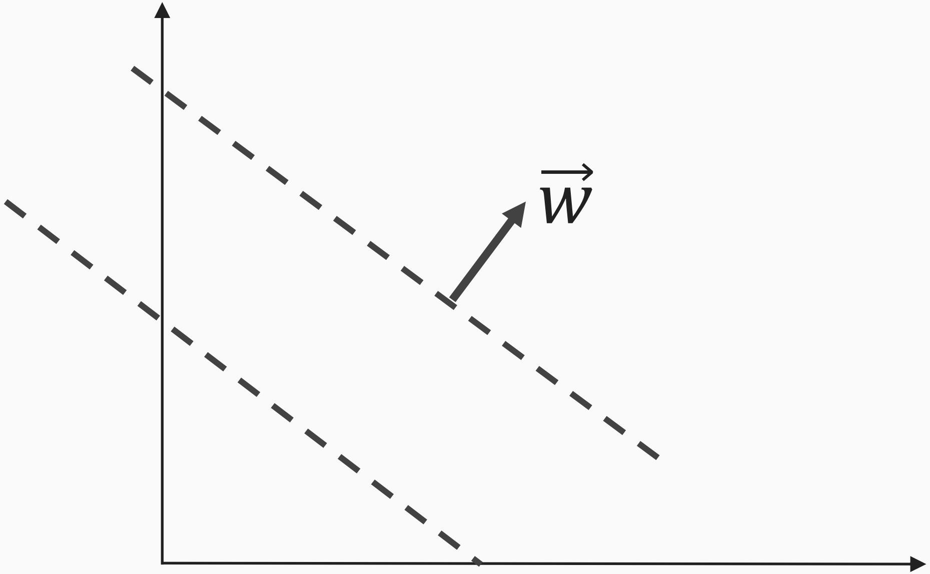
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



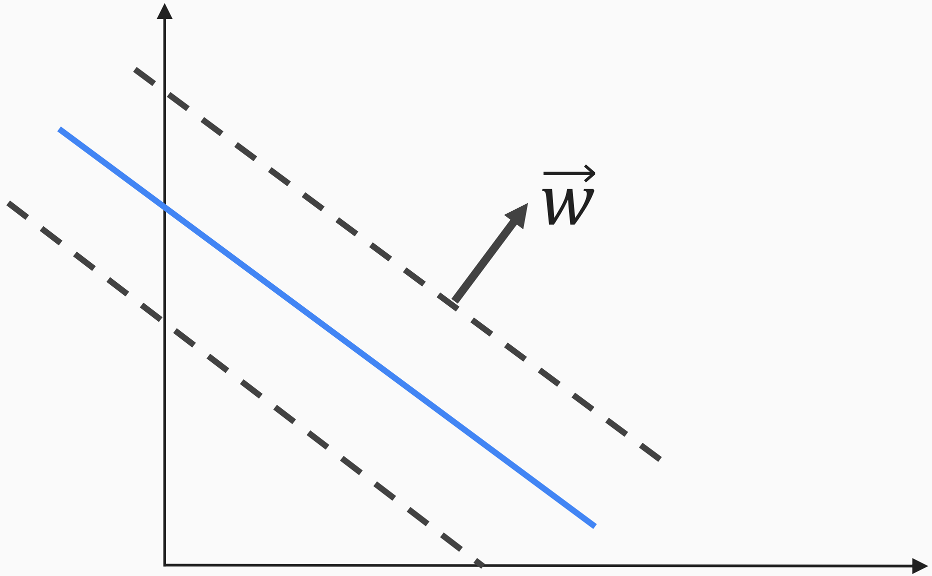
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



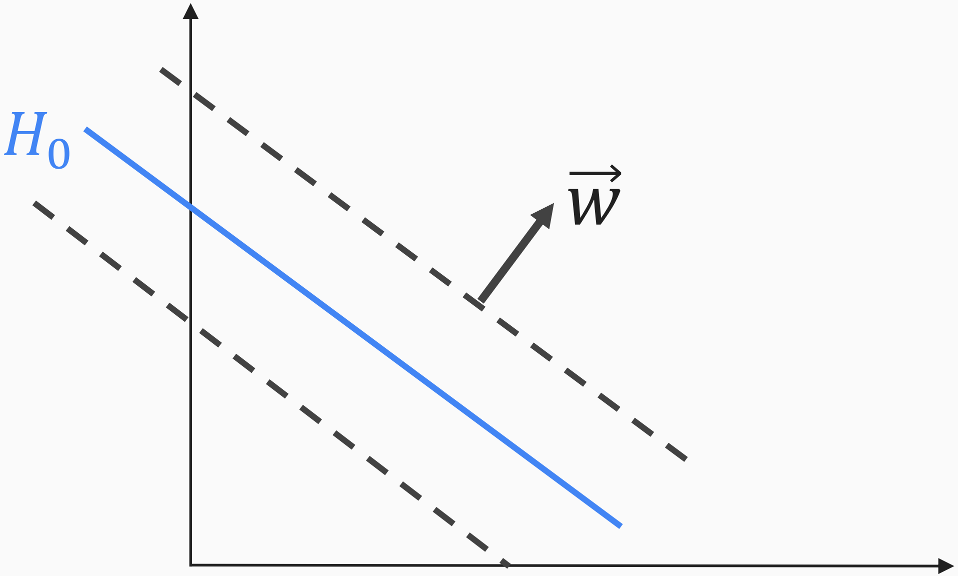
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



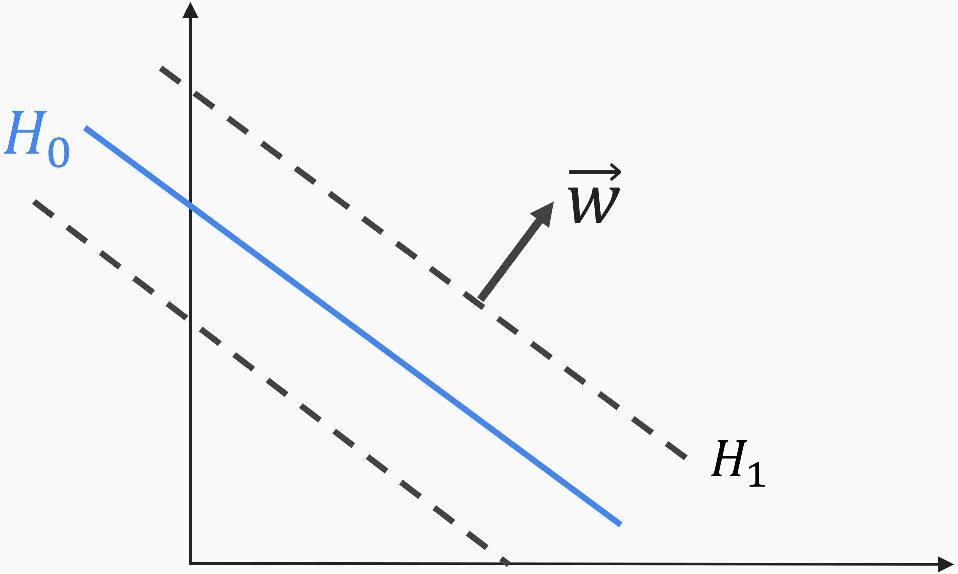
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



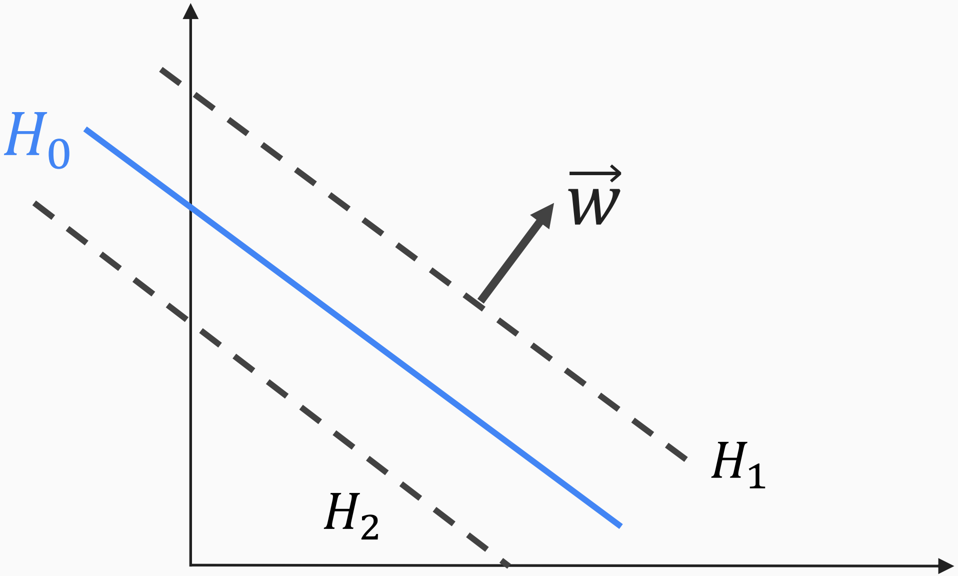
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



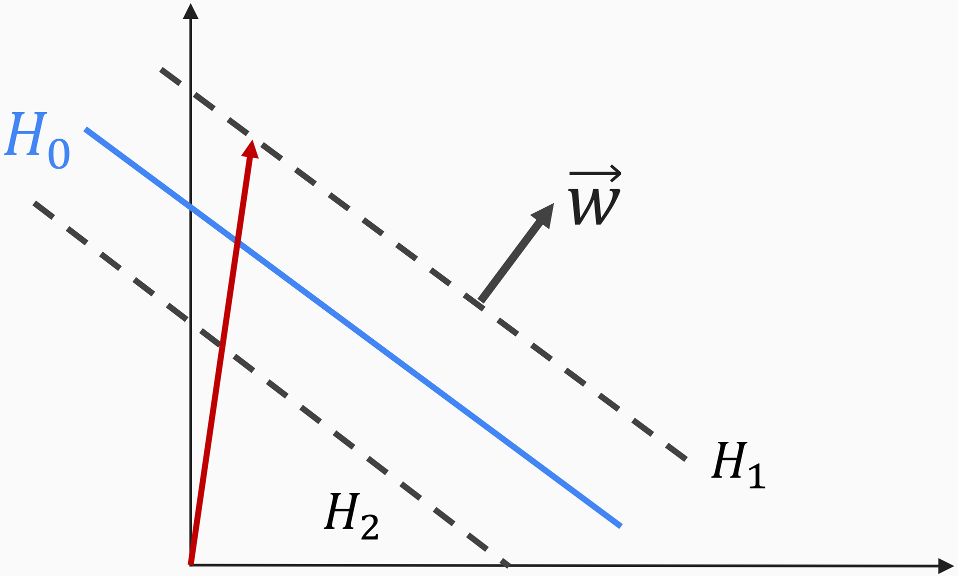
Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



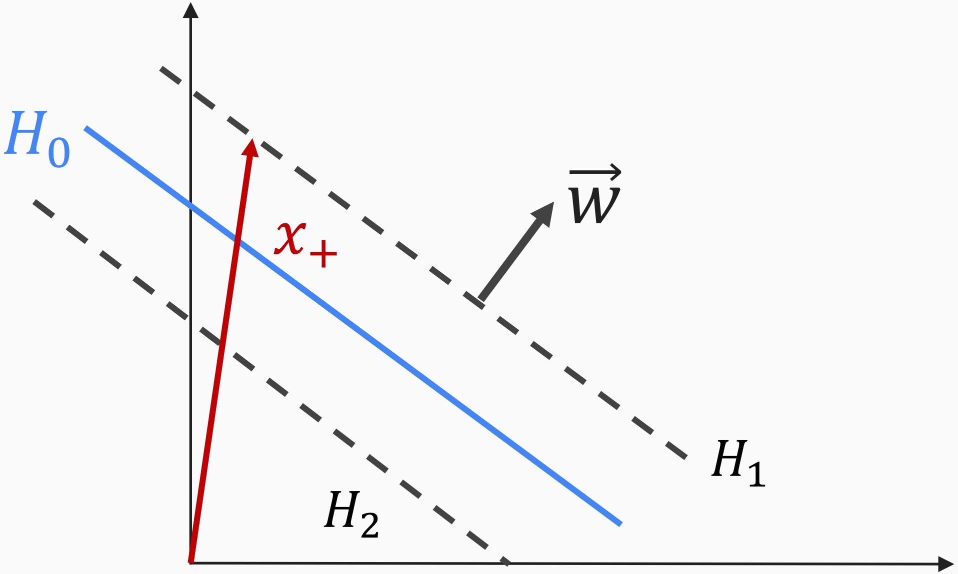
Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



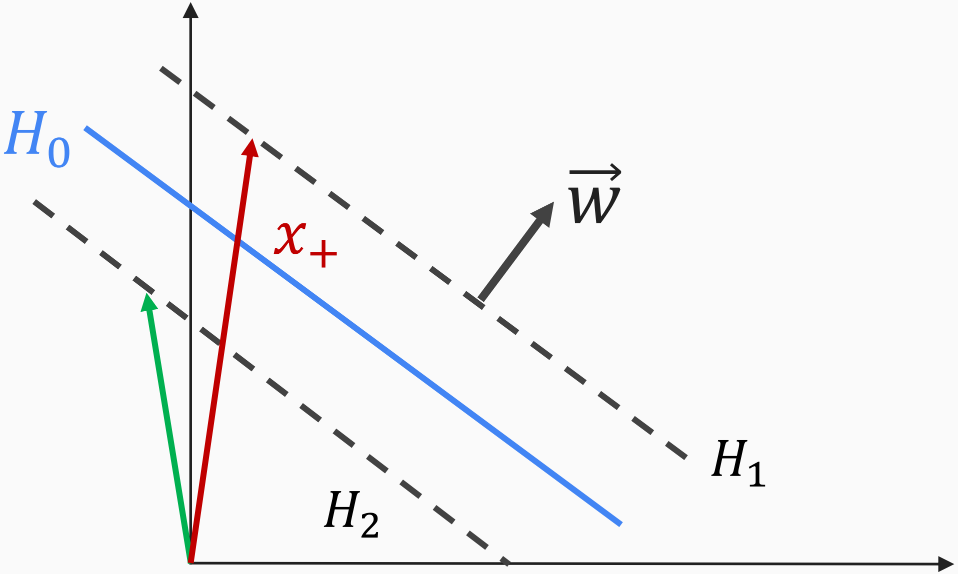
Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



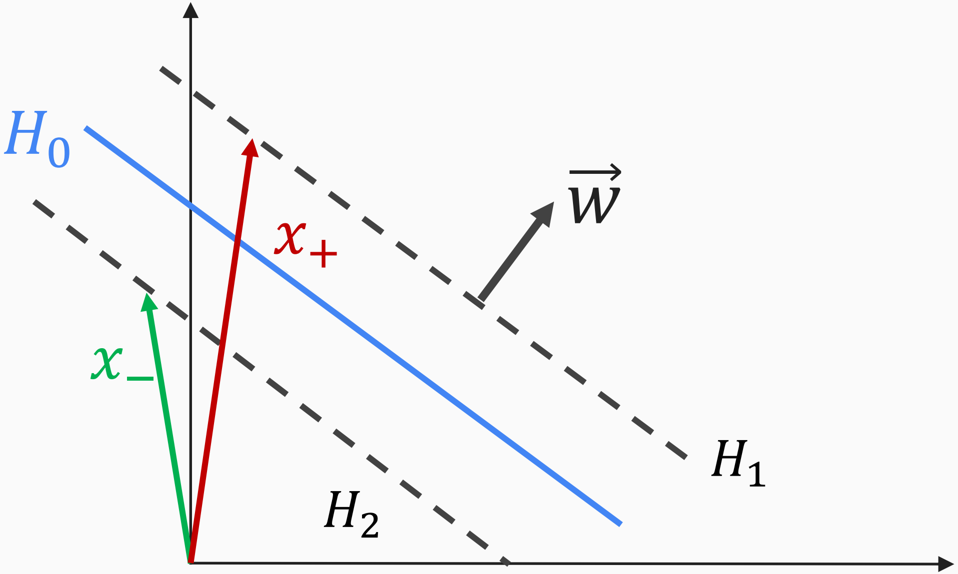
Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



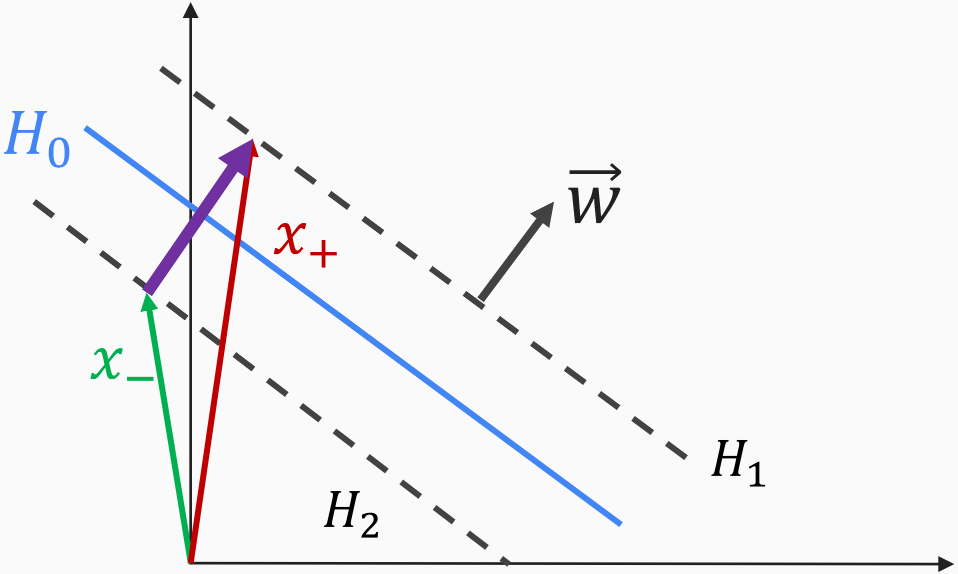
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+ x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



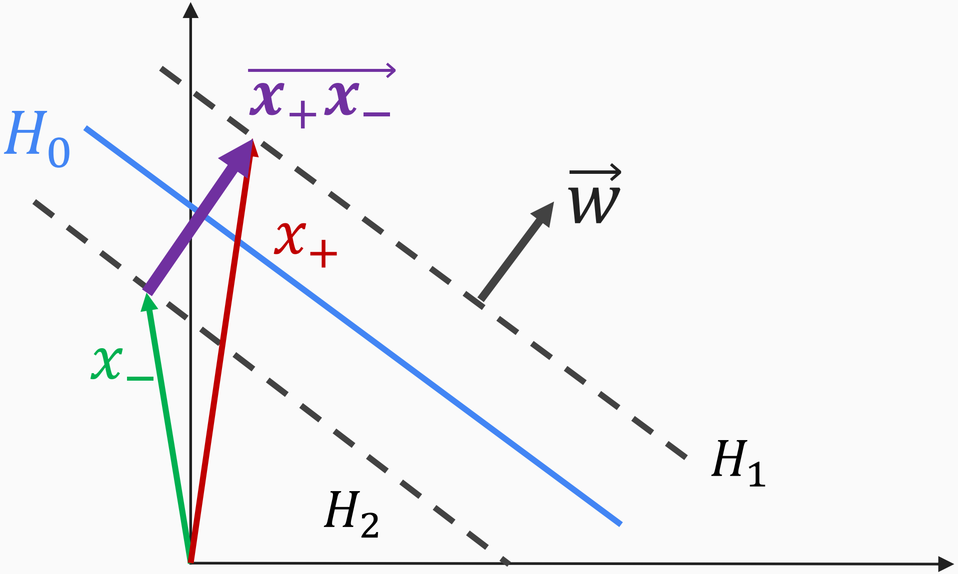
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



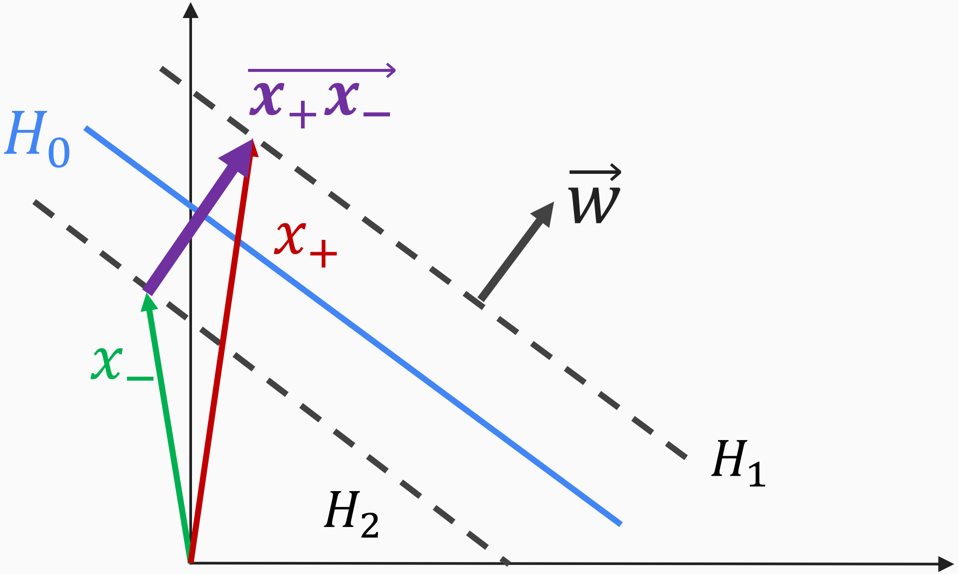
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .

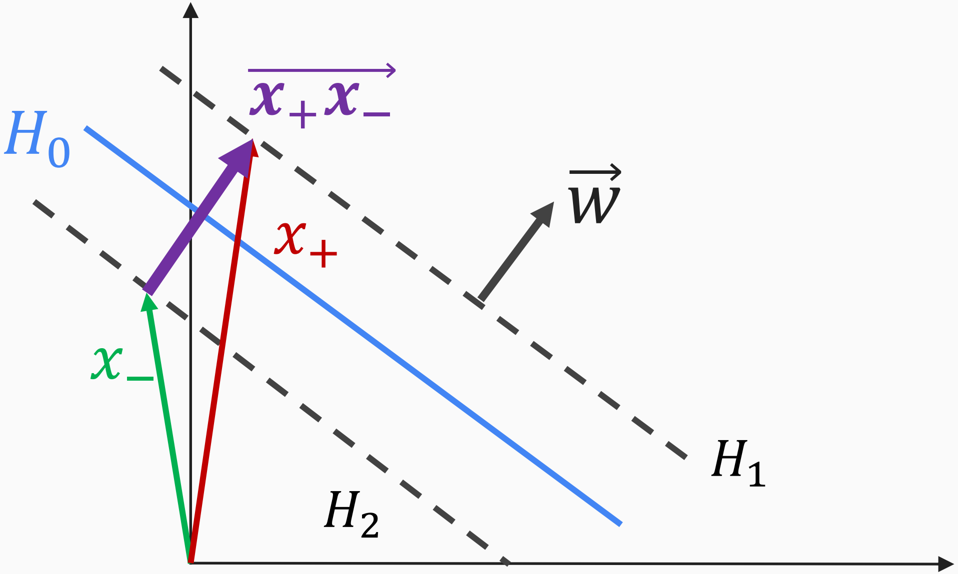


Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

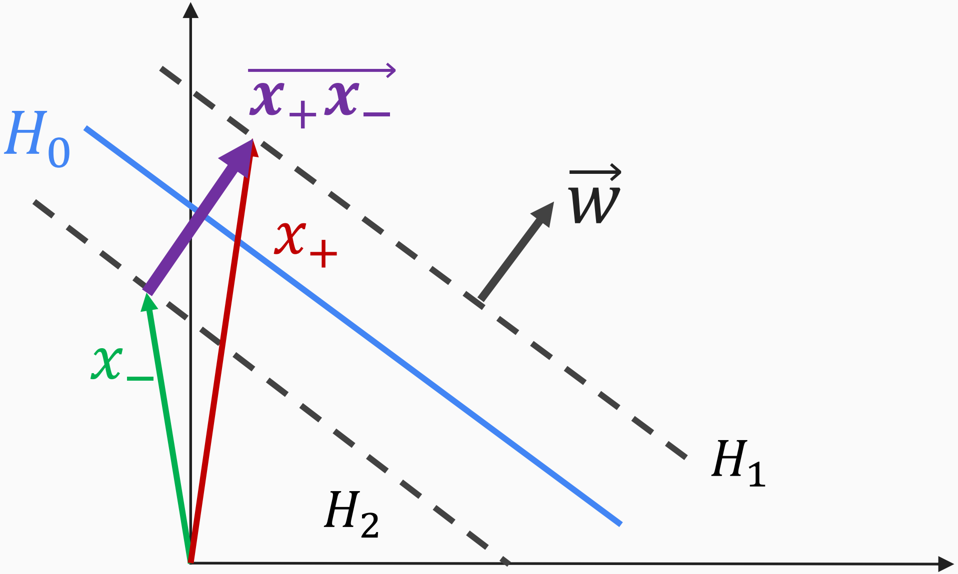
Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$\downarrow$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .

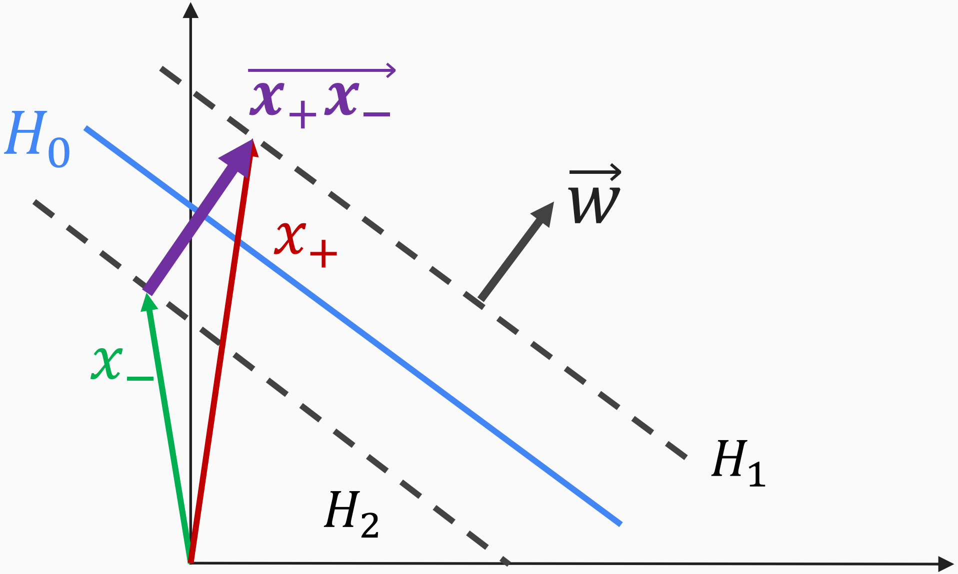


$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



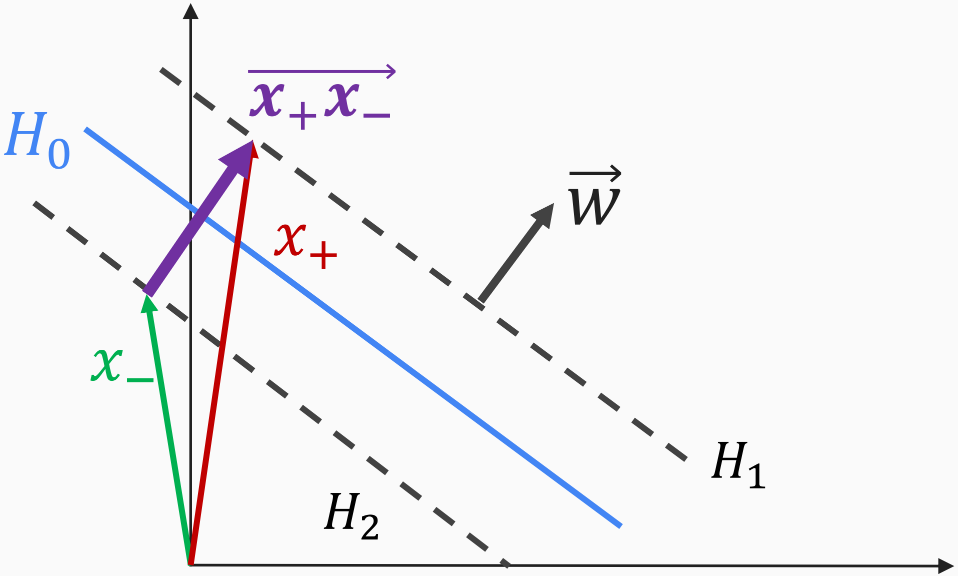
$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\downarrow$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .

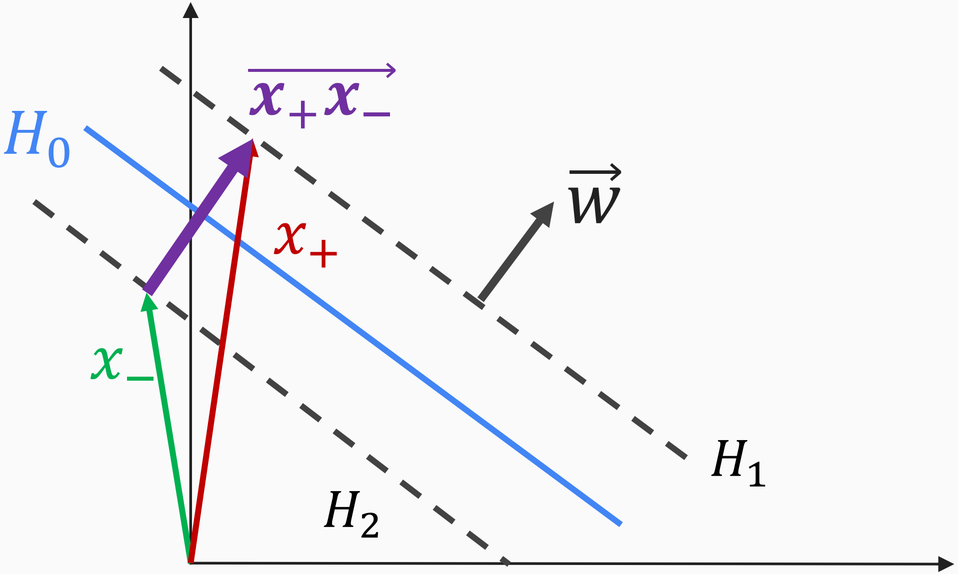


$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\|w(x_+ - x_-)\| = 2$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

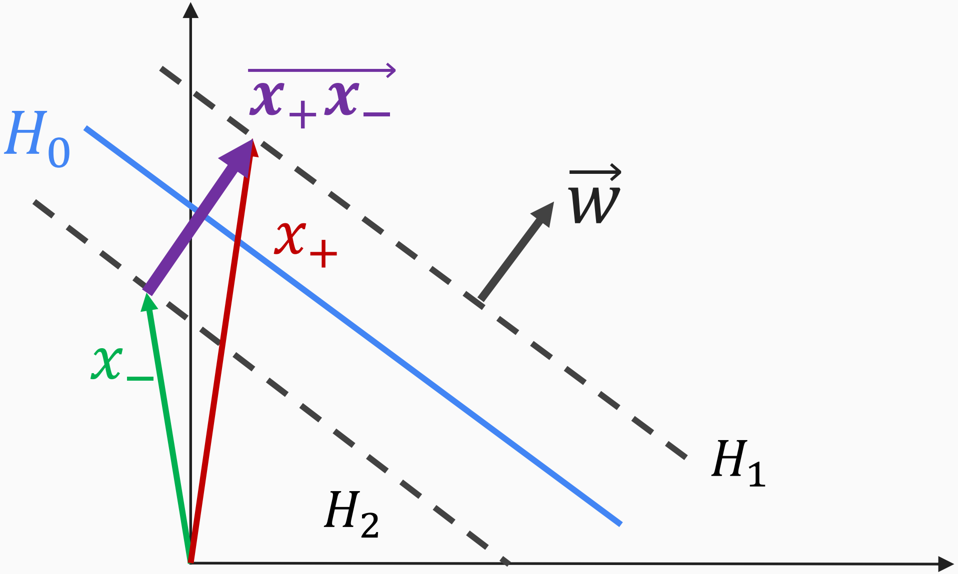
$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

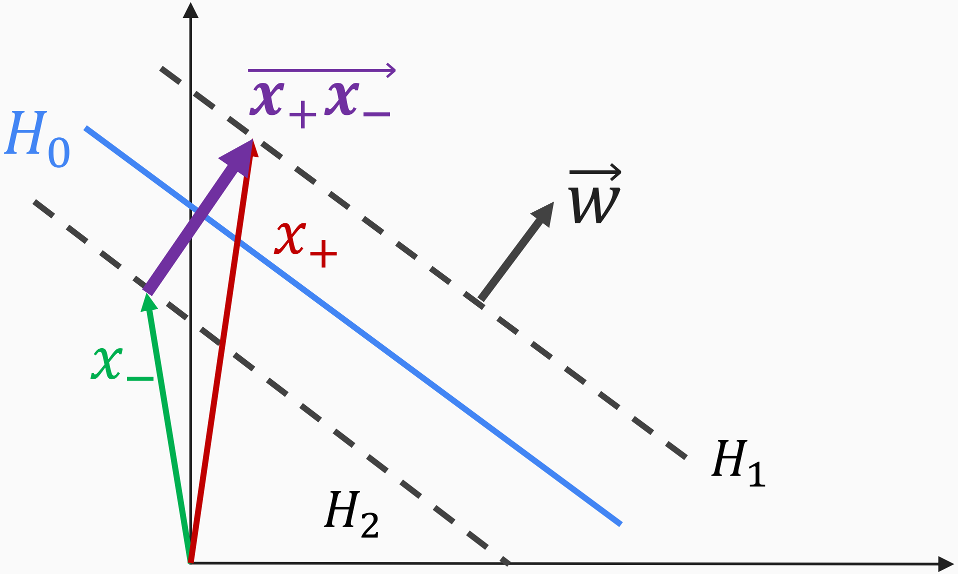
$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\|\|(x_+ - x_-)\| = 2$$

Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

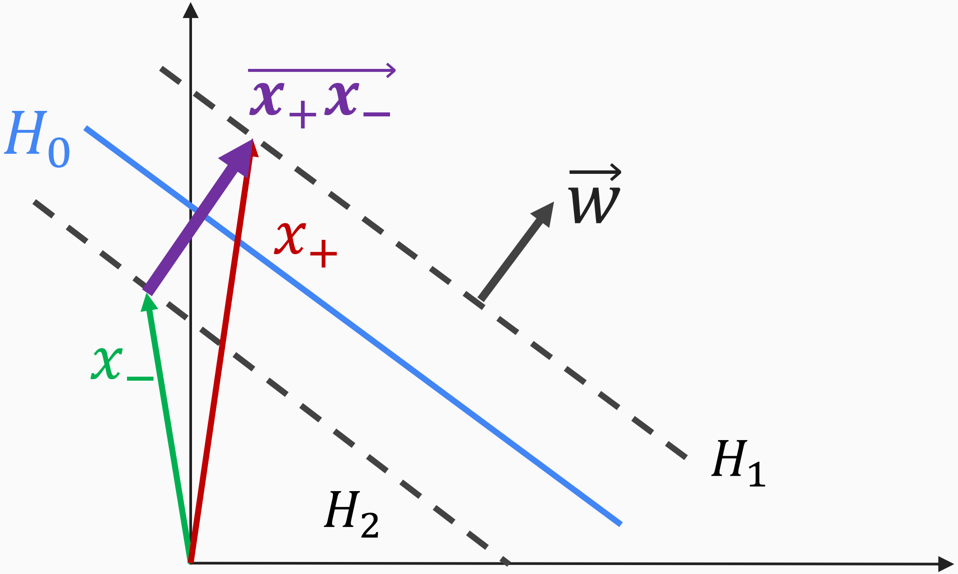
$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\|\underbrace{\|(x_+ - x_-)\|}_{} = 2$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

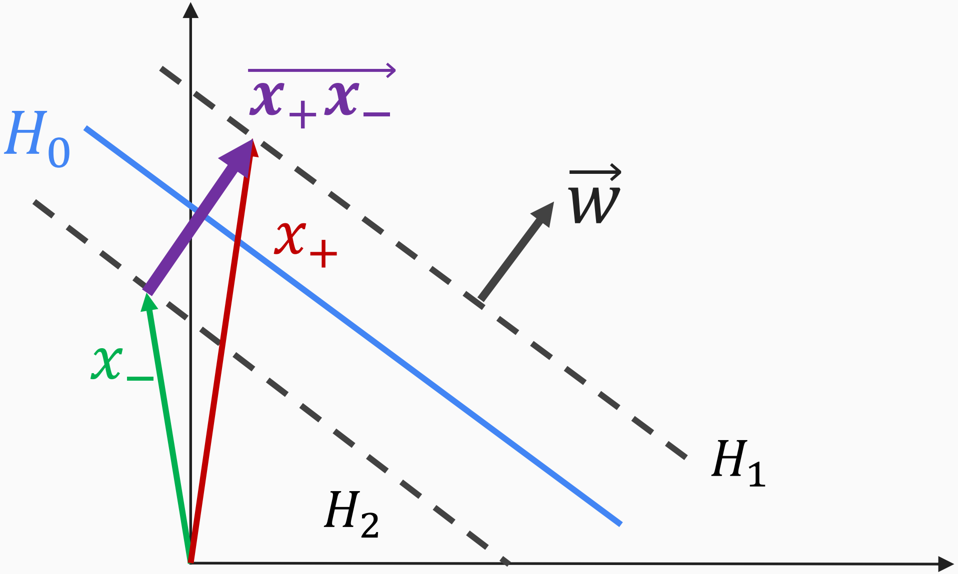
$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\| \underbrace{\|(x_+ - x_-)\|}_{m = margin} = 2$$

Let $x_+ \in H_1$ and $x_- \in H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

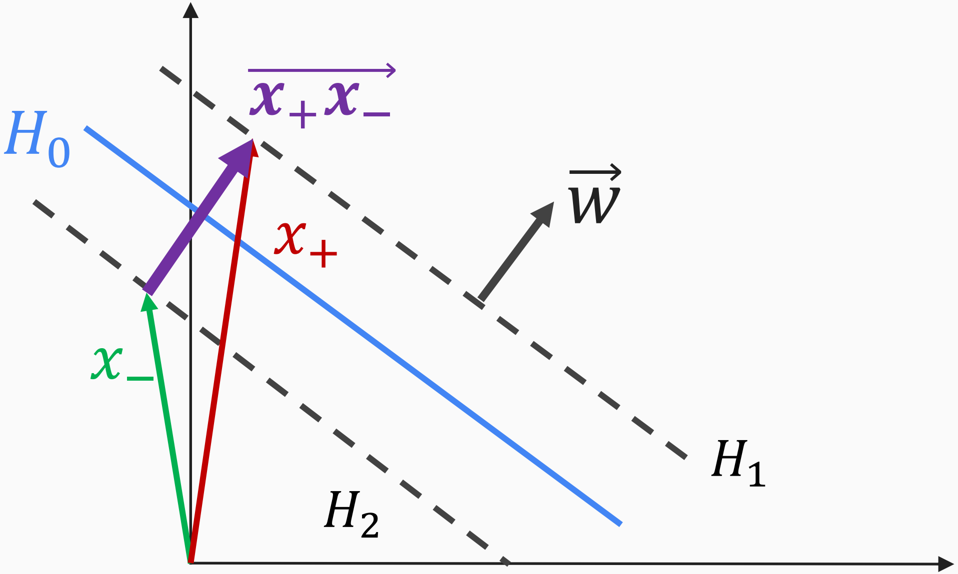
$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\| \underbrace{\|(x_+ - x_-)\|}_{m \,=\, margin} = 2$$

Let $x_+ \,\epsilon\, H_1$ and $x_- \,\epsilon\, H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\downarrow$$

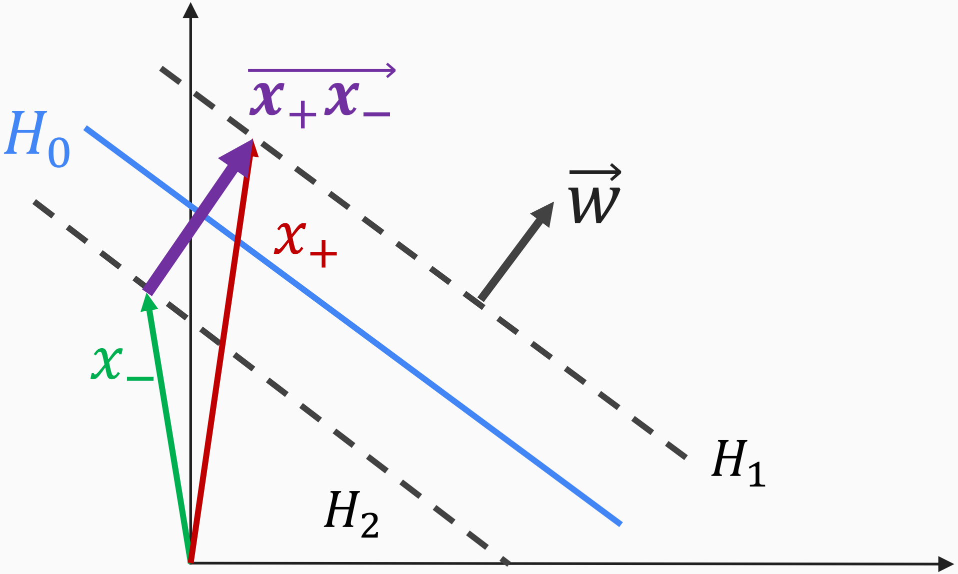
$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\|\underbrace{\|(x_+ - x_-)\|}_{m \,=\, margin} = 2$$

$$\|w\|.m = 2 \longleftarrow m = margin$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

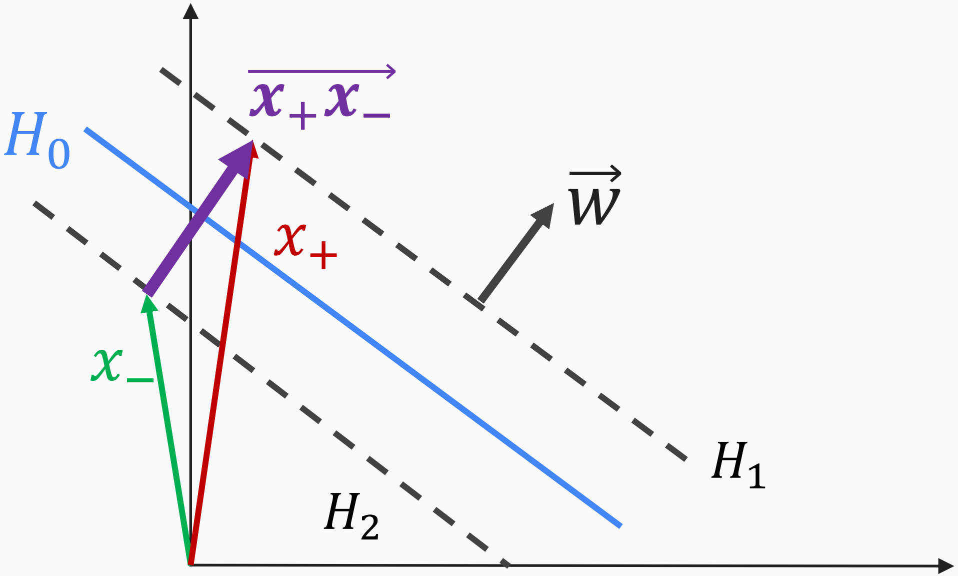
$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\|w(x_+ - x_-)\| = 2$$

$$\|w\|\underbrace{\|(x_+ - x_-)\|}_{m = margin} = 2$$

$$\|w\|.m = 2 \longleftarrow m = margin$$

Let $x_+ \epsilon H_1$ and $x_- \epsilon H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\downarrow$$

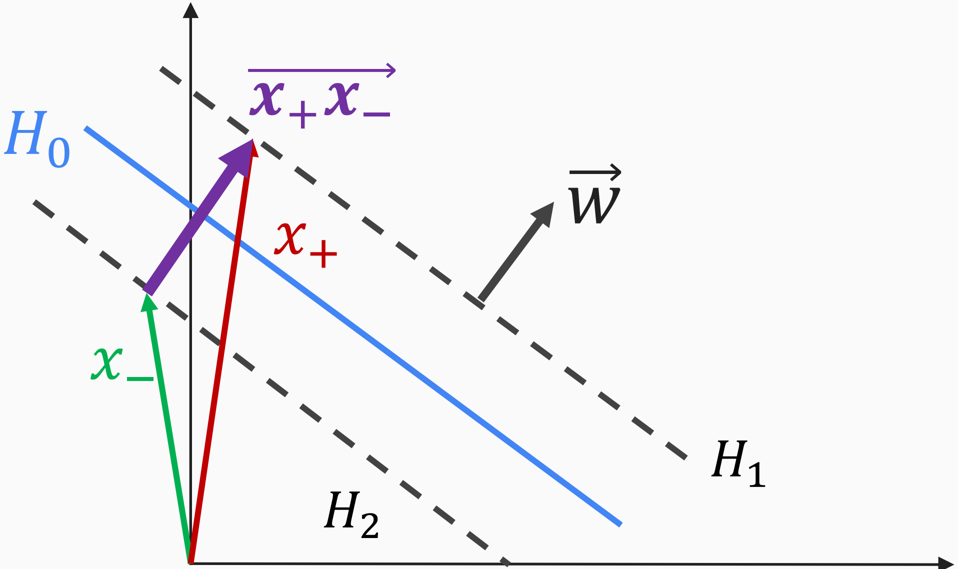
$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\|\underbrace{\|(x_+ - x_-)\|}_{m\,=\,margin} = 2$$

$$\longleftarrow \|w\|.m = 2 \longleftarrow m = margin$$

Let $x_+ \; \epsilon \; H_1$ and $x_- \; \epsilon \; H_2$ such that : $\overrightarrow{x_+x_-}$ has the same direction as $\vec{w}$ where w is the normal vector on H_1 . x_+ and x_- are vectors .



$$\begin{cases} H_1 : wx_+ + b = 1 \\ H_2 : wx_- + b = -1 \end{cases}$$

$$\downarrow$$

$$H_1 - H_2 : w(x_+ - x_-) = 2$$

$$\downarrow$$

$$\|w(x_+ - x_-)\| = 2$$

$$\downarrow$$

$$\|w\| \underbrace{\|(x_+ - x_-)\|}_{m = margin} = 2$$

$$\boxed{\mathbf{m} = \frac{\mathbf{2}}{\|\boldsymbol{w}\|}} \longleftarrow \|w\|.m = 2 \longleftarrow m = margin$$

We need to maximize the margin :

We need to maximize the margin :

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

We need to maximize the margin :

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

Subject to the constraints

We need to maximize the margin :

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

Subject to the constraints

$$wx_i + b \geq 1$$
$$wx_i + b \leq -1$$

We need to maximize the margin :

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

Subject to the constraints

$$wx_i + b \geq 1$$
$$wx_i + b \leq -1$$

i goes from 1 to N

Maximizing the margin:

Maximizing the margin:

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

Maximizing the margin:

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

$$\Leftrightarrow$$

Maximizing the margin:

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

$$\Leftrightarrow$$

Minimizing $\|w\|$

Our linear program becomes

Our linear program becomes

$$\text{minimize} \quad \|w\|$$

Our linear program becomes

$$\textit{minimize} \qquad \|w\|$$

$$\textit{subject to}:$$

Our linear program becomes

$$minimize \qquad \|w\|$$

$$subject\ to: \begin{cases} wx_i + b \geq 1 \\ wx_i + b \leq -1 \end{cases} \qquad i\ goes\ from\ 1\ to\ N$$

Suppose we have

$$y_i = 1 = \blacktriangle$$

$$y_i = -1 = \bullet$$

Our linear program becomes

$$minimize \qquad \|w\|$$

$$subject\ to: \begin{cases} (wx_i + b \geq 1) \\ (wx_i + b \leq -1) \end{cases} \qquad i\ goes\ from\ 1\ to\ N$$

Our linear program becomes

$$minimize \qquad \|w\|$$

$$subject\ to: \begin{cases} (wx_i + b \geq 1)\ y_i & i\ goes\ from\ 1\ to\ N \\ (wx_i + b \leq -1)\ y_i \end{cases}$$

Our linear program becomes

$$minimize \qquad \|w\|$$

$$subject\ to: \begin{cases} y_i(wx_i + b) \geq y_i \\ y_i(wx_i + b) \leq -y_i \end{cases} \ i\ goes\ from\ 1\ to\ N$$

Our linear program becomes

$$minimize \qquad ||w||$$

$$subject\ to: \qquad\qquad i\ goes\ from\ 1\ to\ N$$

Our linear program becomes

$$\mathit{minimize} \qquad \|w\|$$

$$\mathit{subject\ to:} \quad y_i(wx_i + b) \geq 1 \qquad i\ \mathit{goes\ from}\ 1\ \mathit{to}\ N$$

When we solve this linear program we will find w and b that minimizes $\|w\|$

When we solve this linear program we will find w and b that minimizes $\|w\|$

When we solve this linear program we will find w and b that minimizes $\|w\|$

Problem solved ☺ : we found the equation of the optimal hyperplane

In reality, we don’t use the previous program to find the optimal hyperplane.

In reality, we don’t use the previous program to find the optimal hyperplane.

Instead ,we will minimize $\frac{1}{2}||w||^2$ rather than $||w||$ since $||w||$ is not differentiable at $w = 0$

In reality, we don't use the previous program to find the optimal hyperplane.

Instead ,we will minimize $\frac{1}{2}||w||^2$ rather than $||w||$ since $||w||$ is not differentiable at $w = 0$

Optimization algorithms work much better on differentiable functions

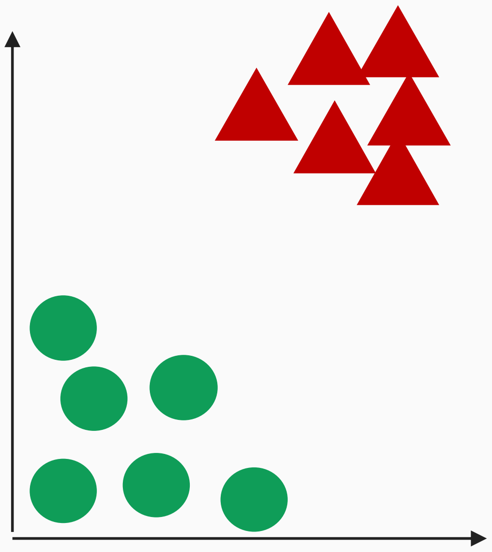
So our program becomes a quadratic program

$$minimize \quad \frac{1}{2}\|w\|^2$$

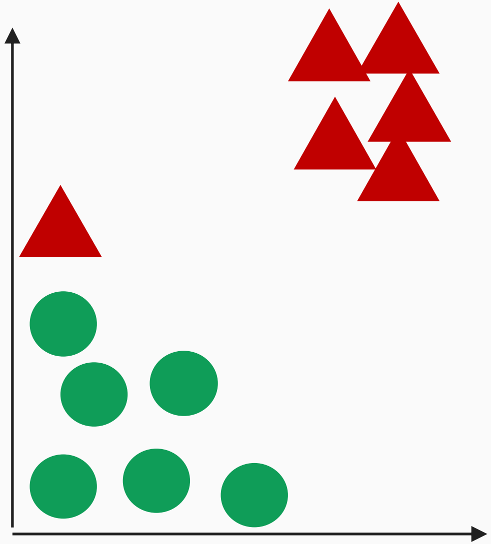
$$subject\ to: \quad y_i(wx_i + b) \geq 1 \qquad i\ goes\ from\ 1\ to\ N$$

Problem

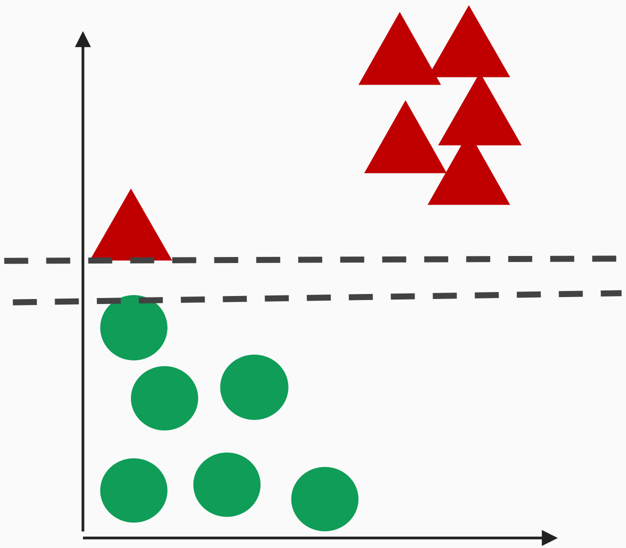
Problem



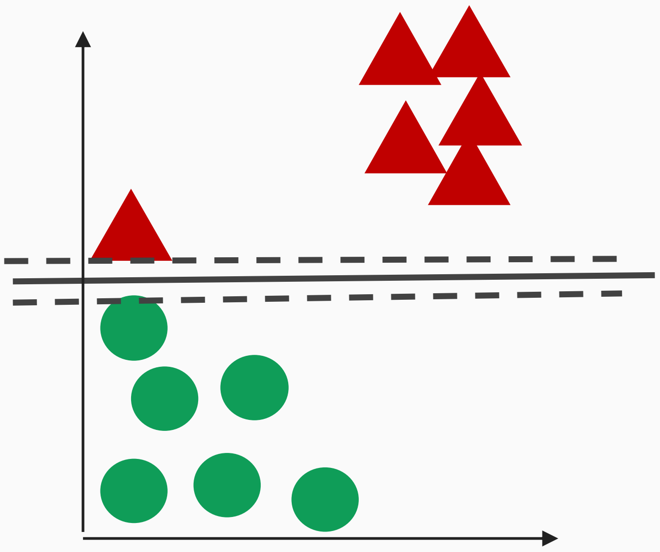
Problem



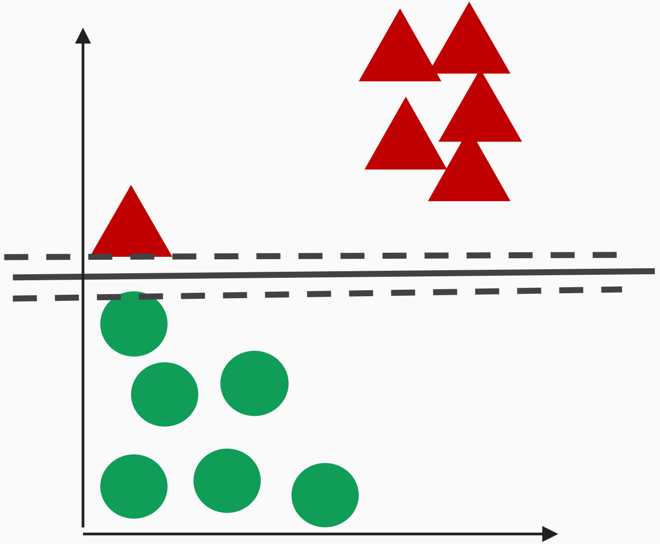
Problem



Problem



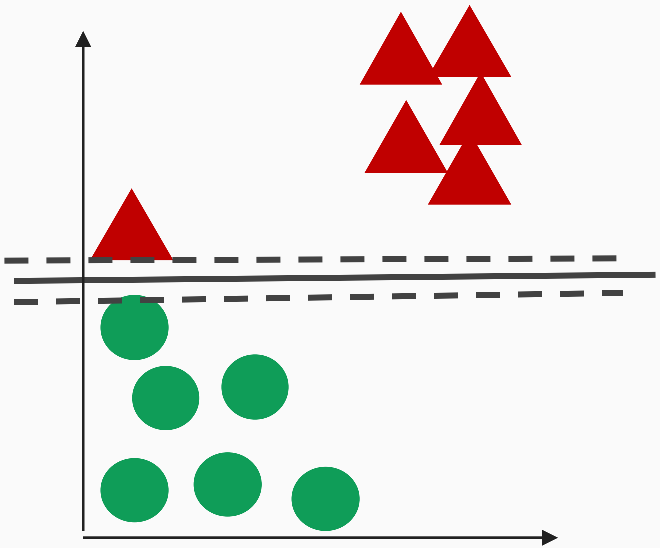
Problem

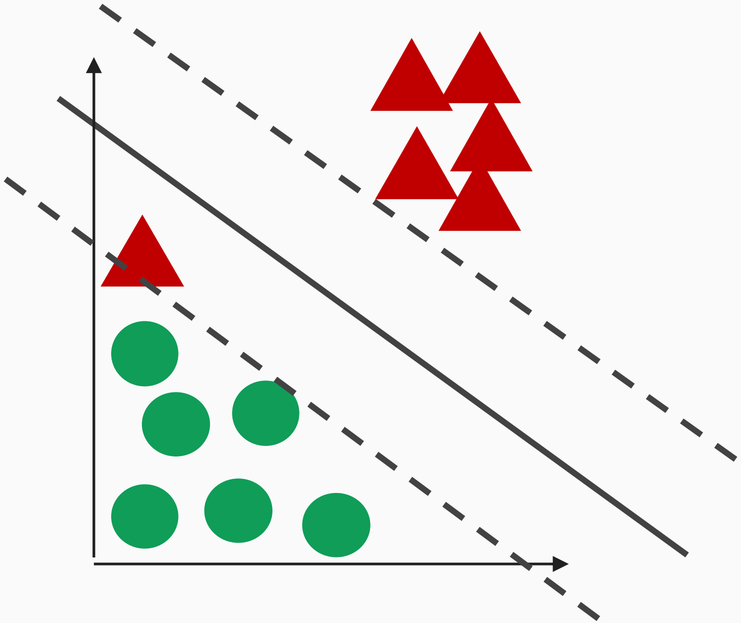


Hard SVM is sensitive to outliers !

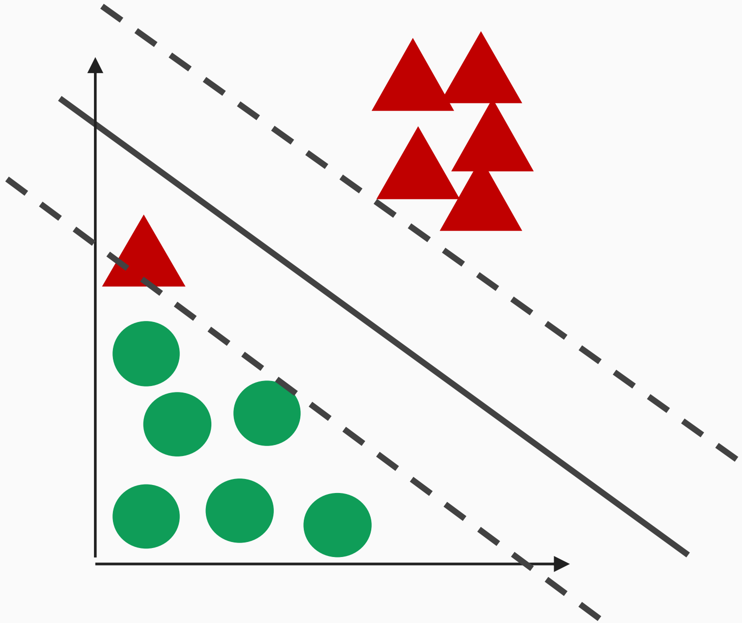
Solution is : ....

(linear) Soft margin SVMs





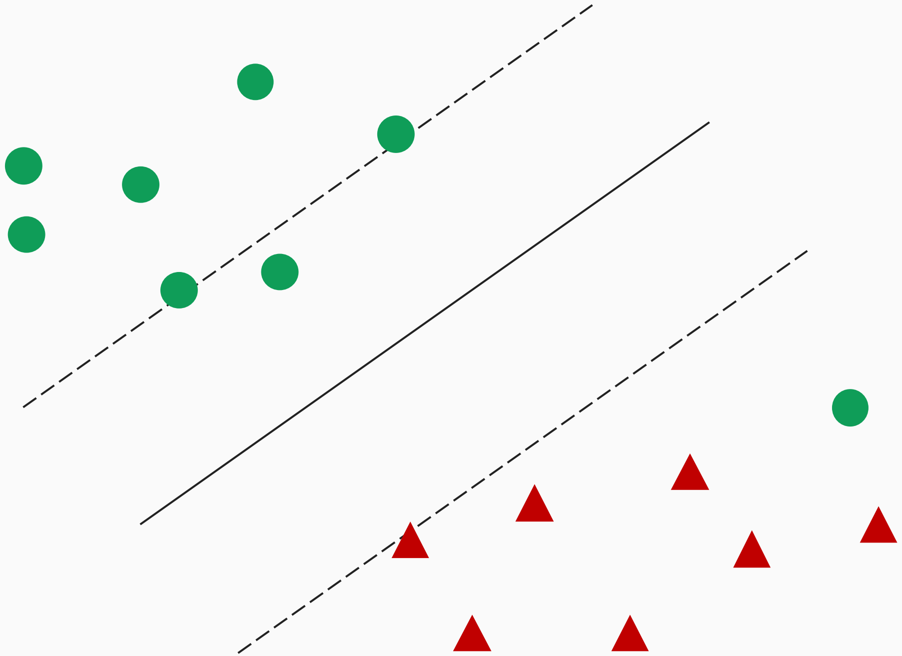
This is Soft margin SVMs



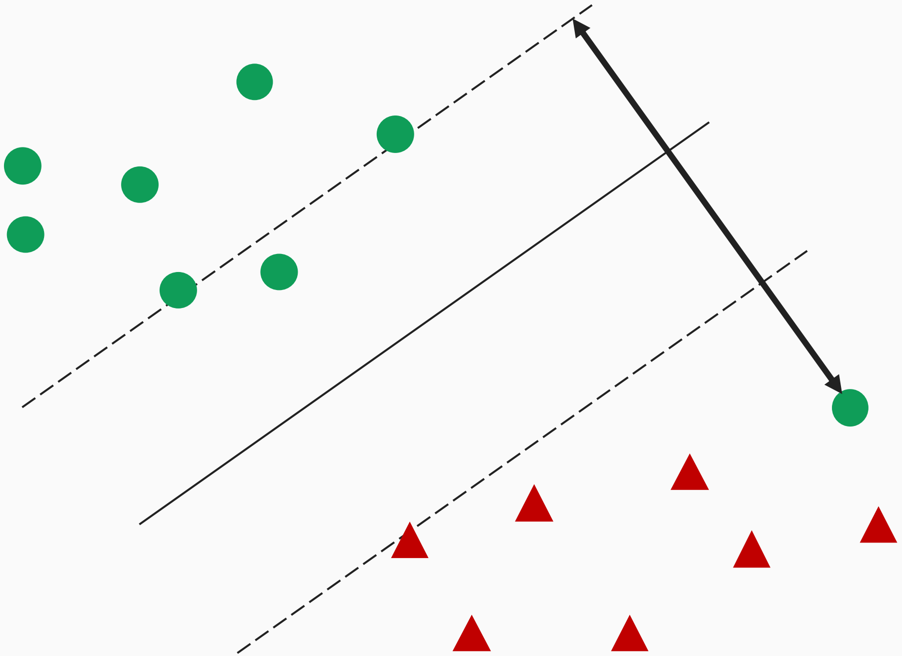
mathematically ☹ :

We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .

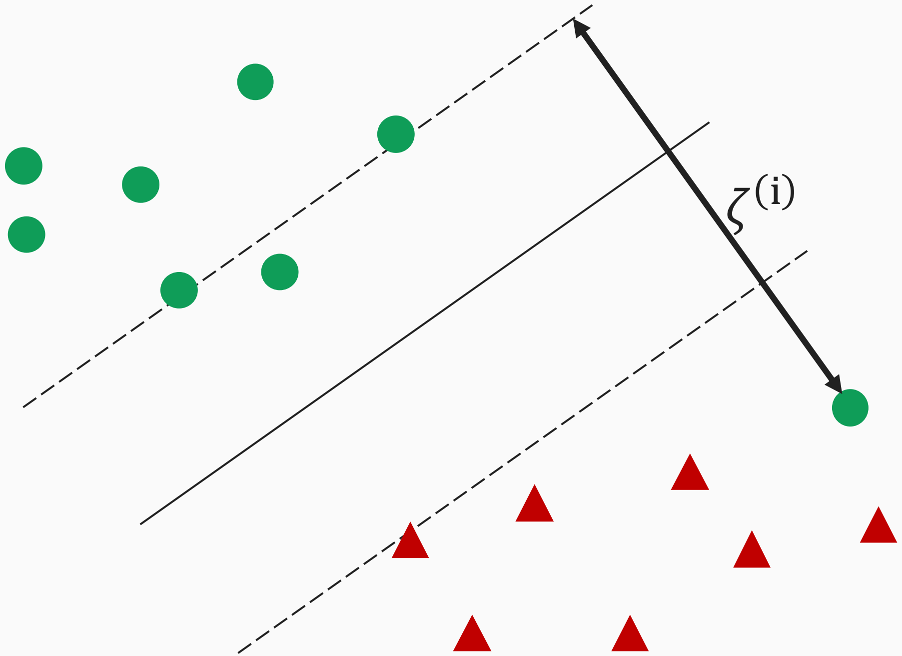
We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



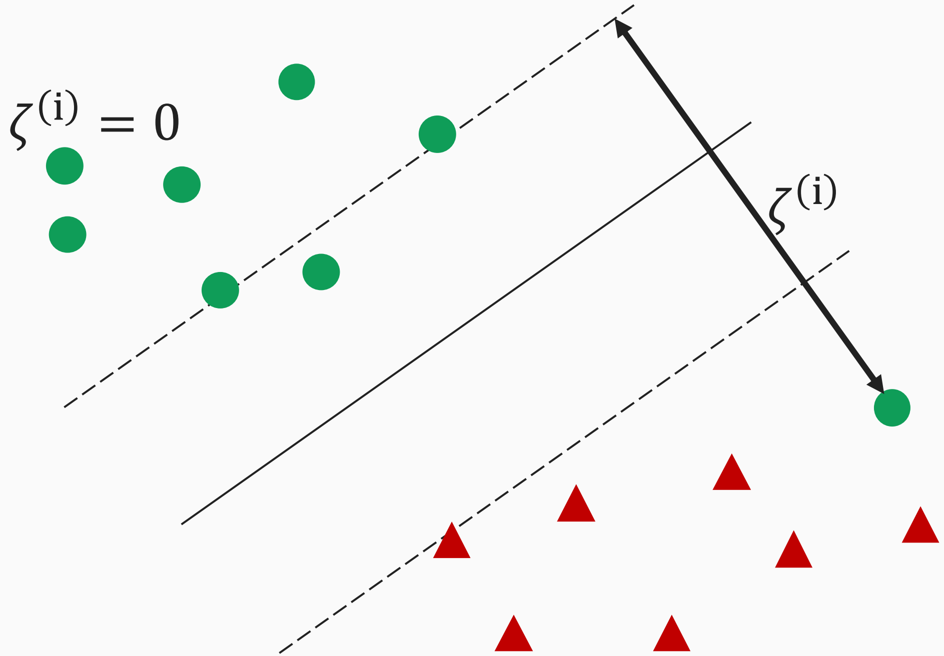
We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



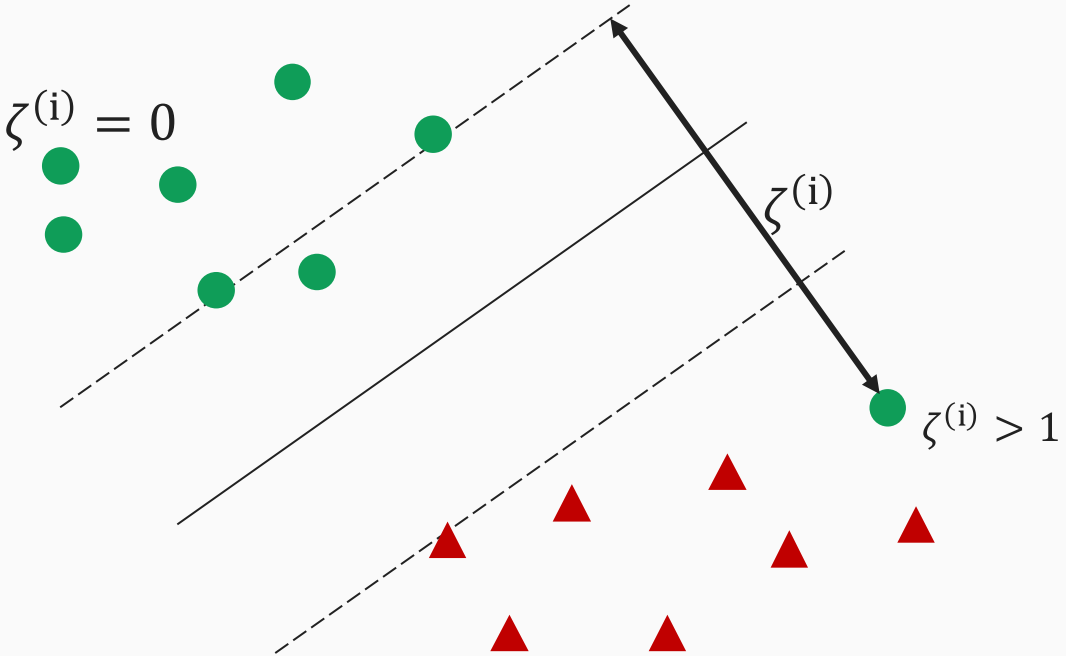
We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



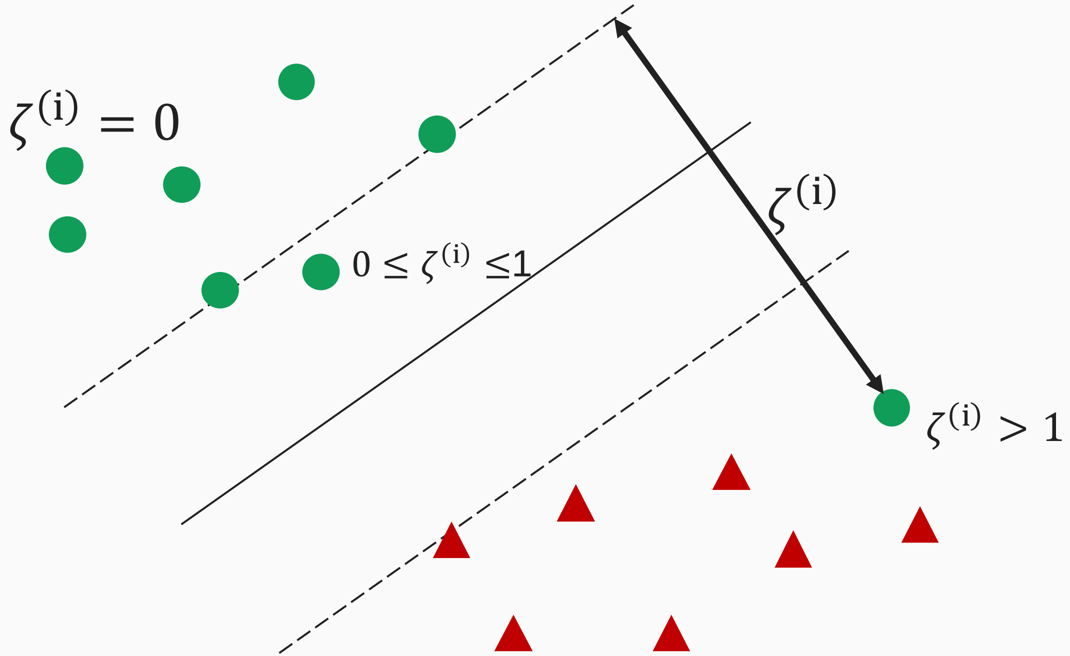
We introduce a slack variable $\zeta^{(i)} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



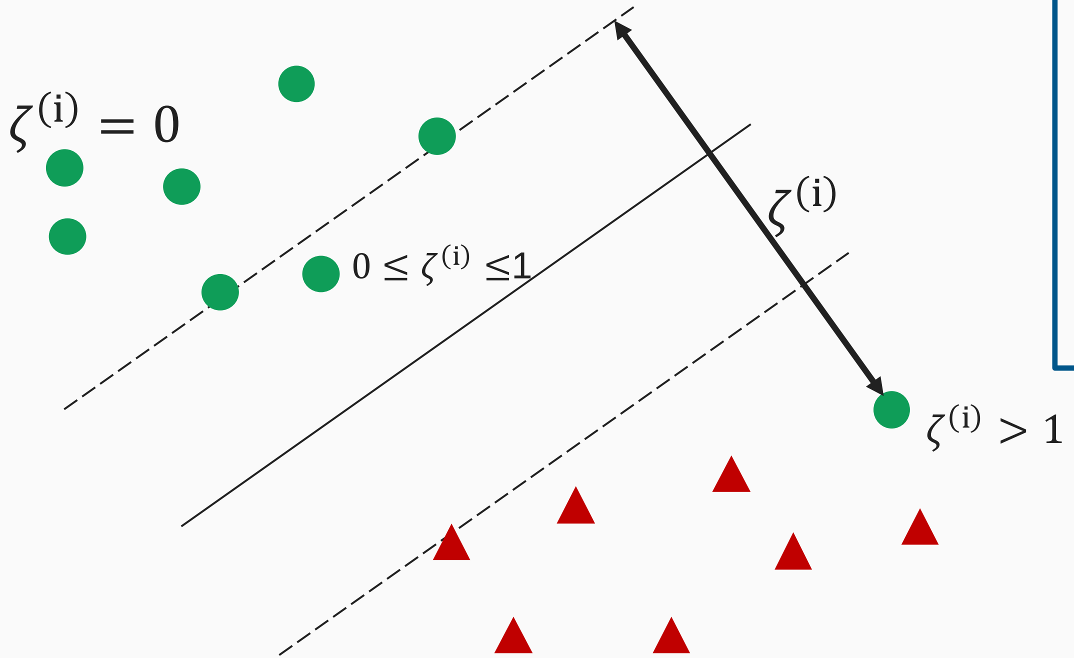
We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



We introduce a slack variable $\zeta^{(i)} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



We introduce a slack variable $\zeta^{(\mathrm{i})} \geq 0$ that tells us how much the i^{th} data point is allowed to violate the margin .



$\zeta^{(\mathrm{i})} > 1$ incorrectly classified

$0 \leq \zeta^{(\mathrm{i})} \leq 1$ correctly classified but inside the margin

$\zeta^{(\mathrm{i})} = 0$ correctly classified

Trade off :

Maximize the margin

Minimize the violations

$$\begin{aligned} &minimize \quad \frac{1}{2}\|w\|^2 \\ &subject\ to: \quad y_i(wx_i + b) \geq 1 \qquad i \in \{1, \cdots, N\} \end{aligned}$$

$$\textit{minimize} \qquad \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \zeta^{(i)}$$

$$\textit{subject to :} \qquad y_i(wx_i + b) \geq 1 \qquad i \in \{1, \cdots, N\}$$

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \zeta^{(i)}$$

$$\text{subject to :} \quad y_i(wx_i + b) + \zeta^{(i)} \geq 1 \quad i \in \{1, \cdots, N\}$$

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \zeta^{(i)}$$

$$\text{subject to}: \quad y_i(wx_i + b) + \zeta^{(i)} \geq 1 \quad i \in \{1, \cdots, N\}$$

$$\zeta^{(i)} \geq 0$$

Soft margin

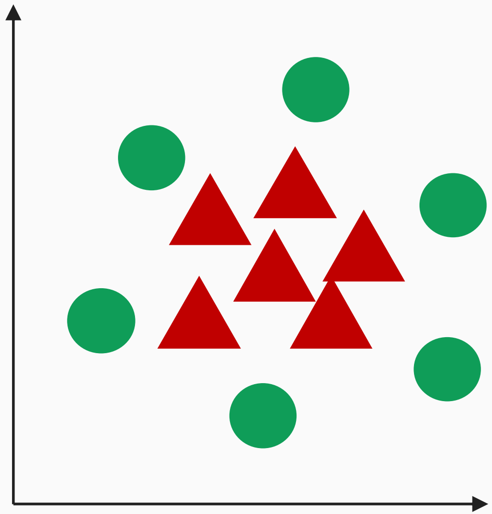
$$minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\zeta^{(i)}$$

$$subject\ to: \quad y_i(wx_i + b) + \zeta^{(i)} \geq 1 \quad i \in \{1, \cdots, N\}$$

$$\zeta^{(i)} \geq 0$$

What if our data is not linearly separable?

What if our data is not linearly separable?



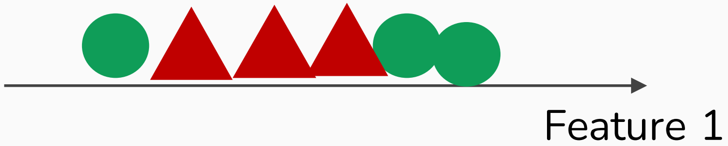
Solution is : ....

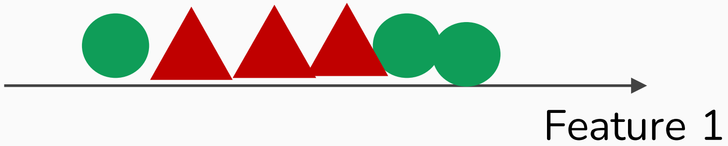
Or Solutions are : ....

We can find a non linear decision boundary to separate data using multiple ideas :

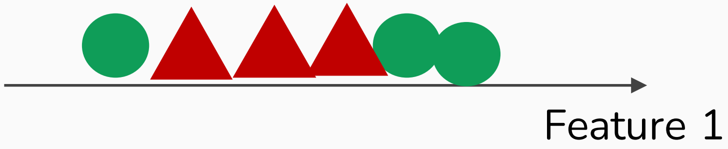
1. Adding features
2. Adding features using similarity function and deleting the old ones
3. Kernel trick

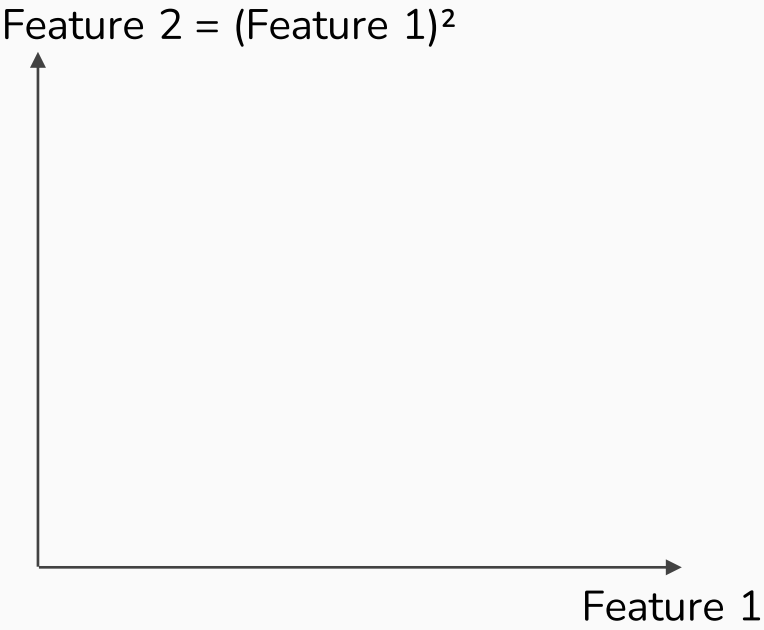
Adding features

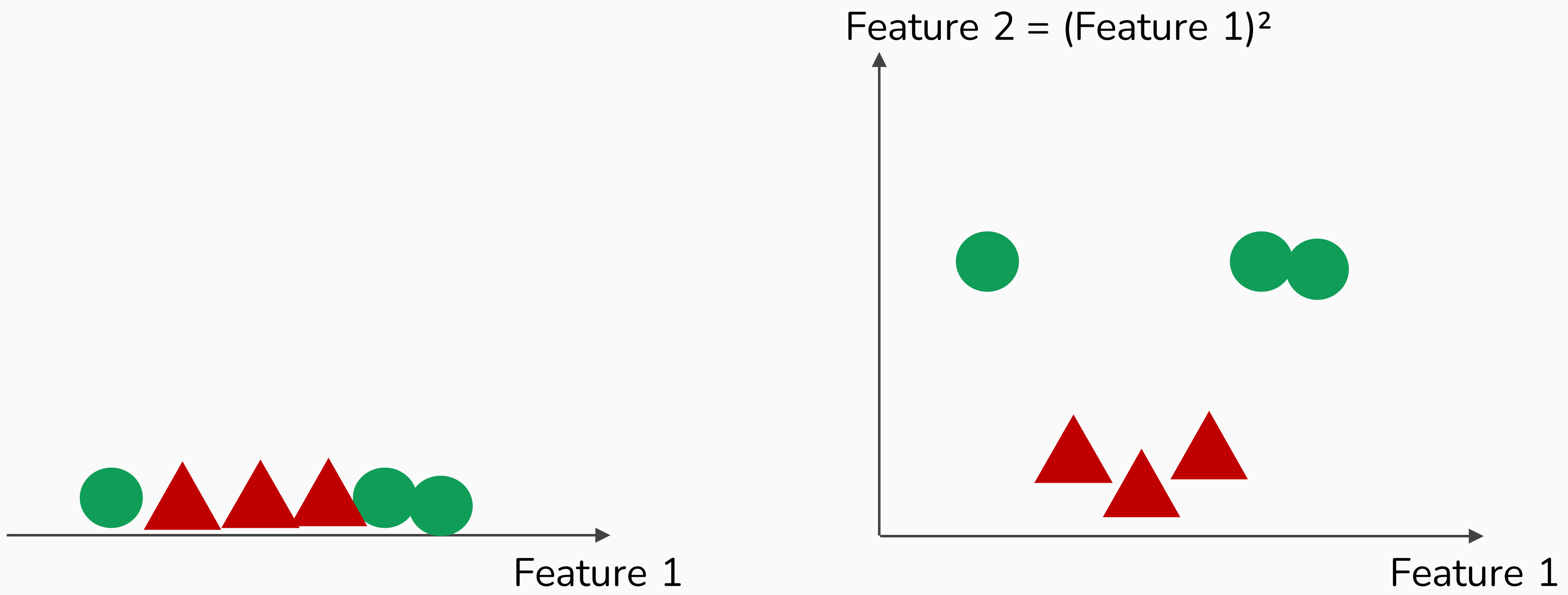

Feature 1


Feature 1


Feature 1

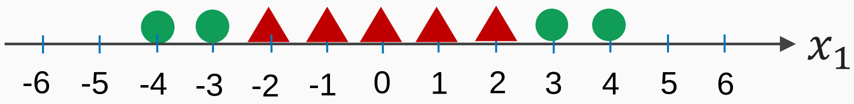

Feature 1


Feature 2 = (Feature 1)²
Feature 1

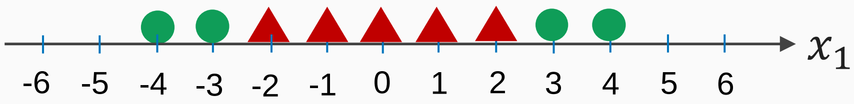


Feature 1
Feature 2 = (Feature 1)²
Feature 1

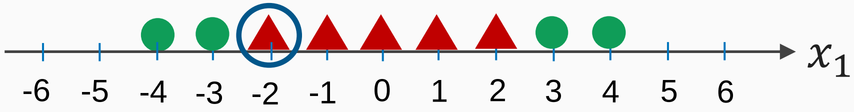
Adding features using similarity functions


-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
x_1

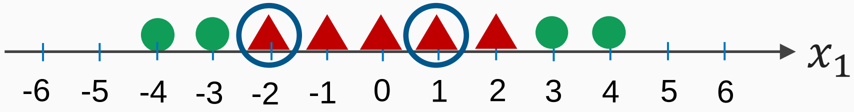
Choose landmarks



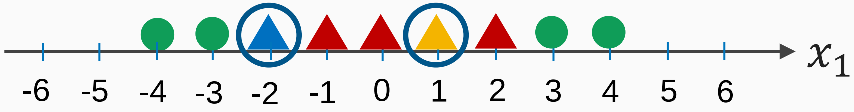
Choose landmarks



Choose landmarks

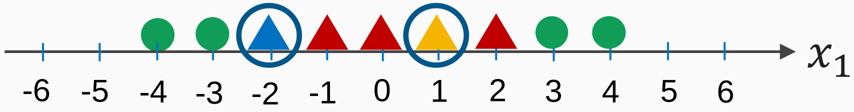


Choose landmarks



Choose landmarks

Our landmarks are :

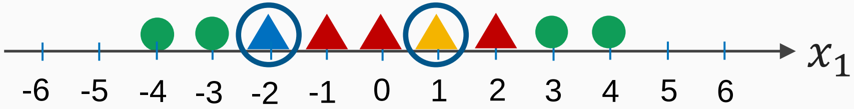


Choose landmarks

Our landmarks are :

 $= -2$

 $= 1$

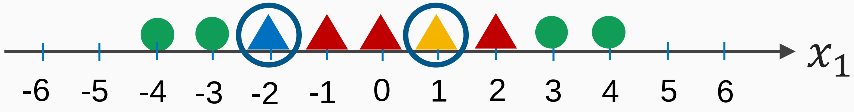


Next we measure the similarity between each point and the landmarks

Our landmarks are :

$= -2$

$= 1$

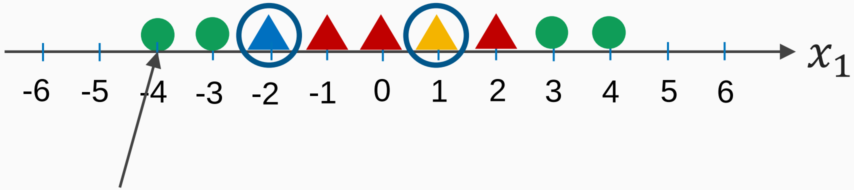


Next we measure the similarity between each point and the landmarks

Our landmarks are :

Blue triangle $= -2$

Yellow triangle $= 1$

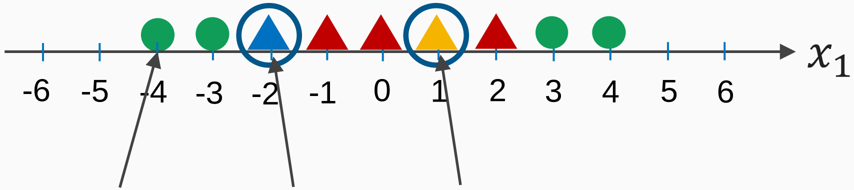


Next we measure the similarity between each point and the landmarks

Our landmarks are :

▲ $= -2$

▲ $= 1$

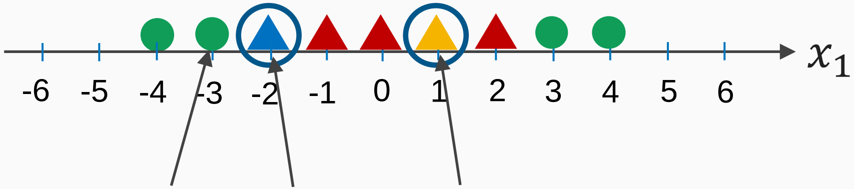


Next we measure the similarity between each point and the landmarks

Our landmarks are :

 $= -2$

 $= 1$

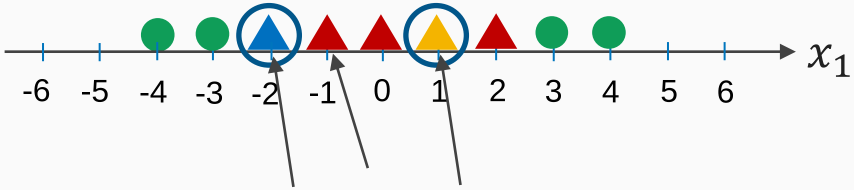


Next we measure the similarity between each point and the landmarks

Our landmarks are :

$\blacktriangle = -2$

$\blacktriangle = 1$

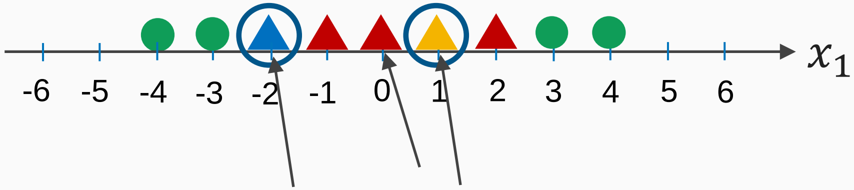


Next we measure the similarity between each point and the landmarks

Our landmarks are :

 $= -2$

 $= 1$

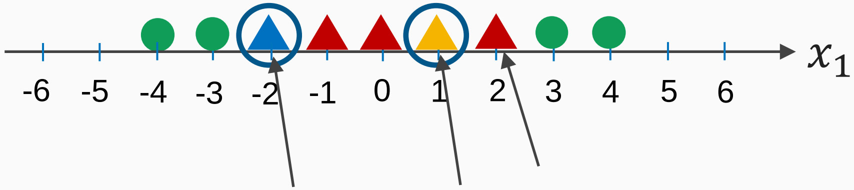


Next we measure the similarity between each point and the landmarks

Our landmarks are :

▲ $= -2$

▲ $= 1$

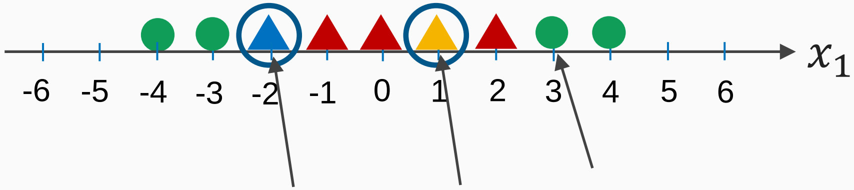


Next we measure the similarity between each point and the landmarks

Our landmarks are :

▲ $= -2$

▲ $= 1$

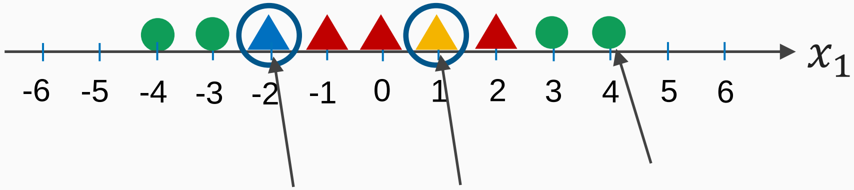


Next we measure the similarity between each point and the landmarks

Our landmarks are :

▲ $= -2$

▲ $= 1$

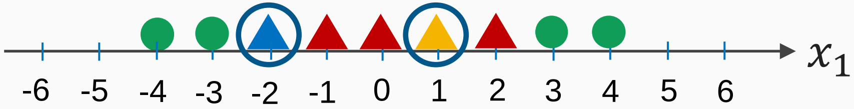


Next we measure the similarity between each point and the landmarks

Our landmarks are :

 $= -2$

 $= 1$

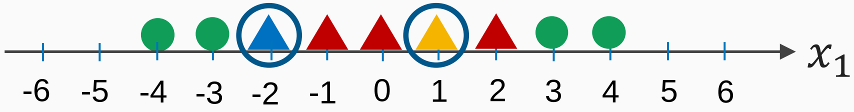


Next we measure the similarity between each point and the landmarks

Our landmarks are :

▲ $= -2$ ⟶

▲ $= 1$

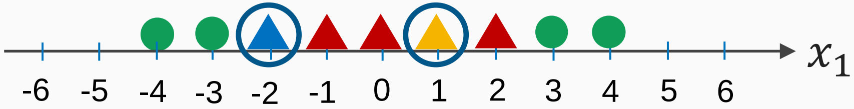


Next we measure the similarity between each point and the landmarks

Our landmarks are :

 $= -2 \longrightarrow x_2$

 $= 1$

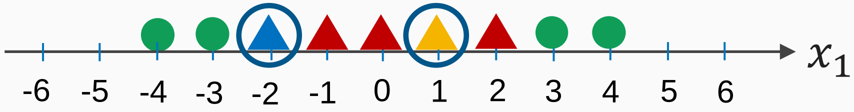


Next we measure the similarity between each point and the landmarks

Our landmarks are :

 $= -2 \longrightarrow x_2$

 $= 1 \longrightarrow$

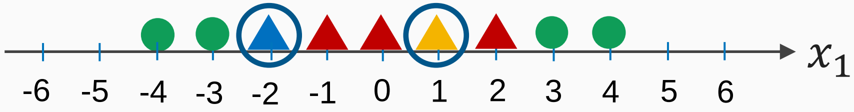


Next we measure the similarity between each point and the landmarks

Our landmarks are :

$= -2 \longrightarrow x_2$

$= 1 \longrightarrow x_3$

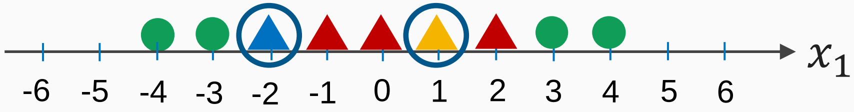


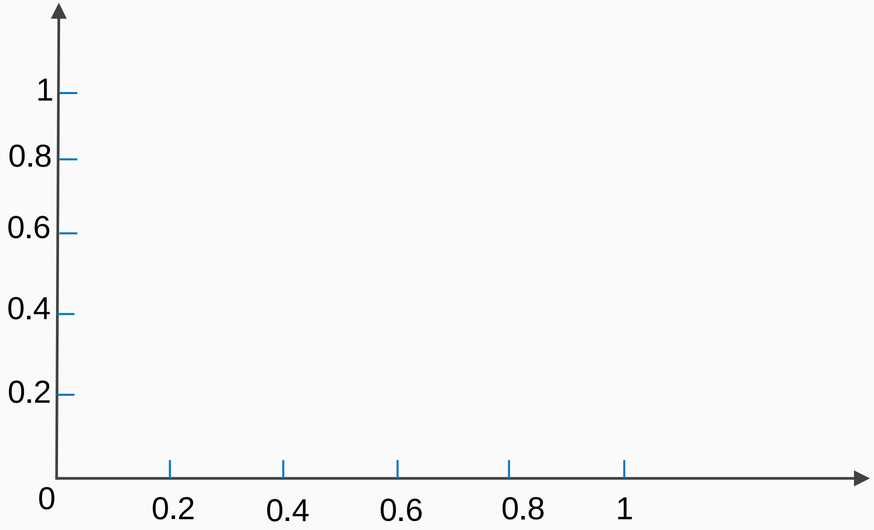
Next we measure the similarity between each point and the landmarks

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

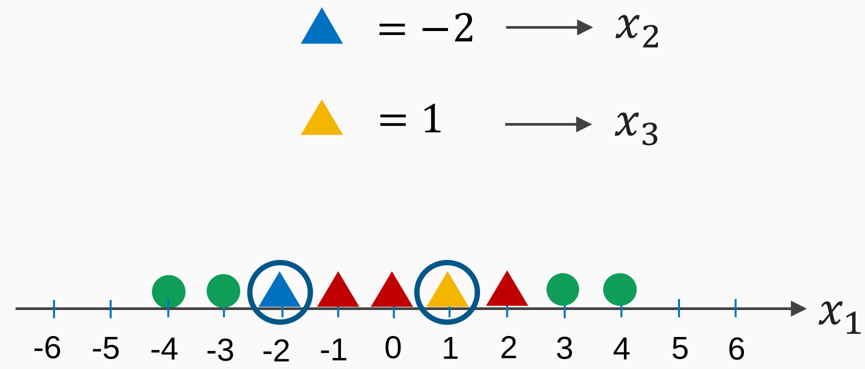
$\blacktriangle = 1 \longrightarrow x_3$

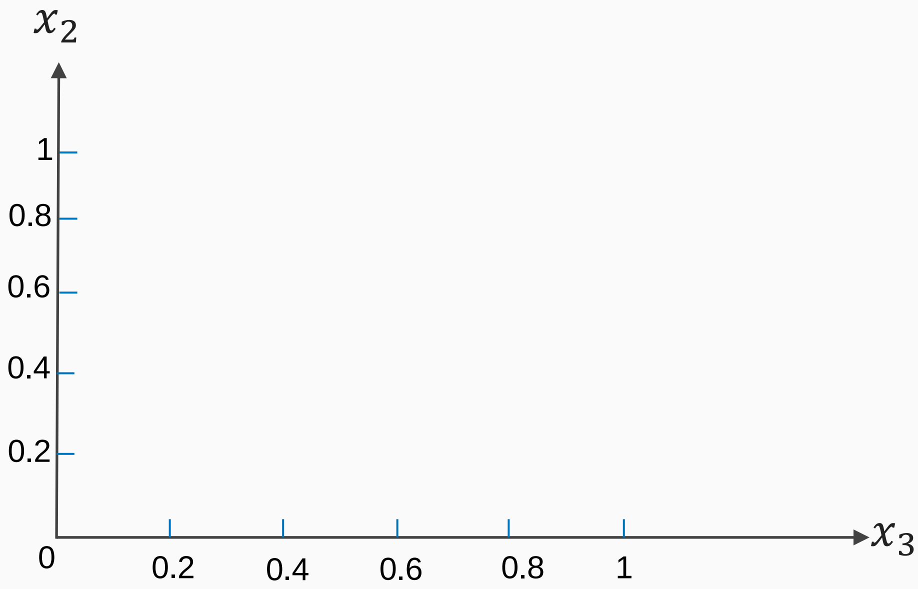




Next we measure the similarity between each point and the landmarks

Our landmarks are :





The question is : How do we measure the similarity ?

The answer is :
using a **kernel function**

The kernel function measures the similarity between each data point and a specific landmark

The kernel function measures the similarity between each data point and a specific landmark

We will use as a kernel function the (Gaussian) Radial Basis function (RBF)

The kernel function measures the similarity between each data point and a specific landmark

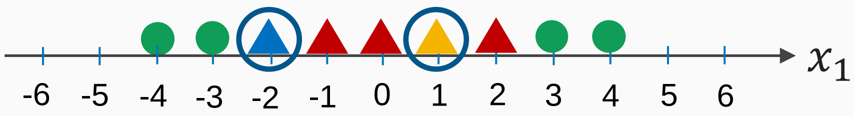
We will use as a kernel function the (Gaussian) Radial Basis function (RBF)

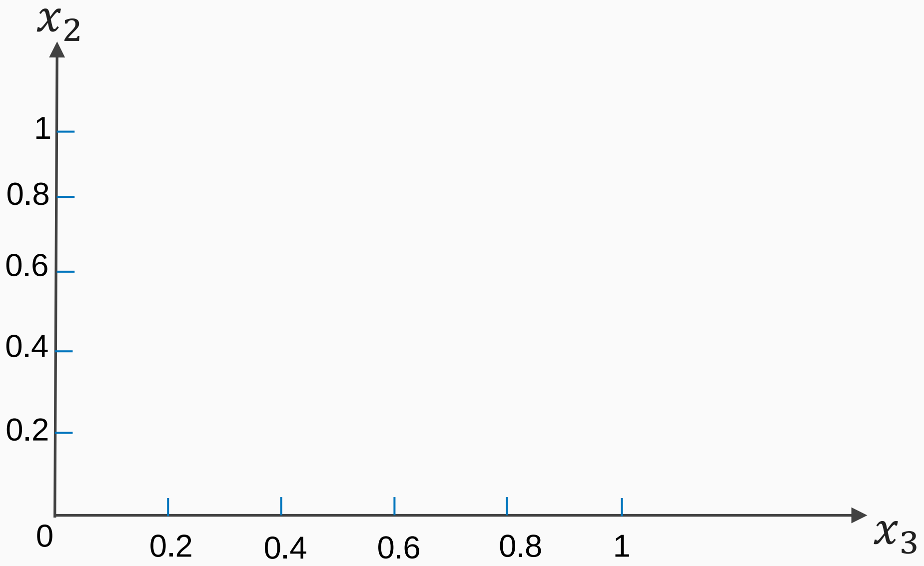
$$K\left(x_i, l_j\right) = e^{\left(-\gamma \left\| x_i - l_j \right\|^2\right)}$$

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

$\blacktriangle = 1 \longrightarrow x_3$



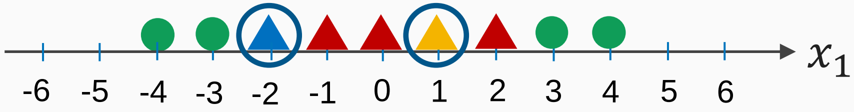


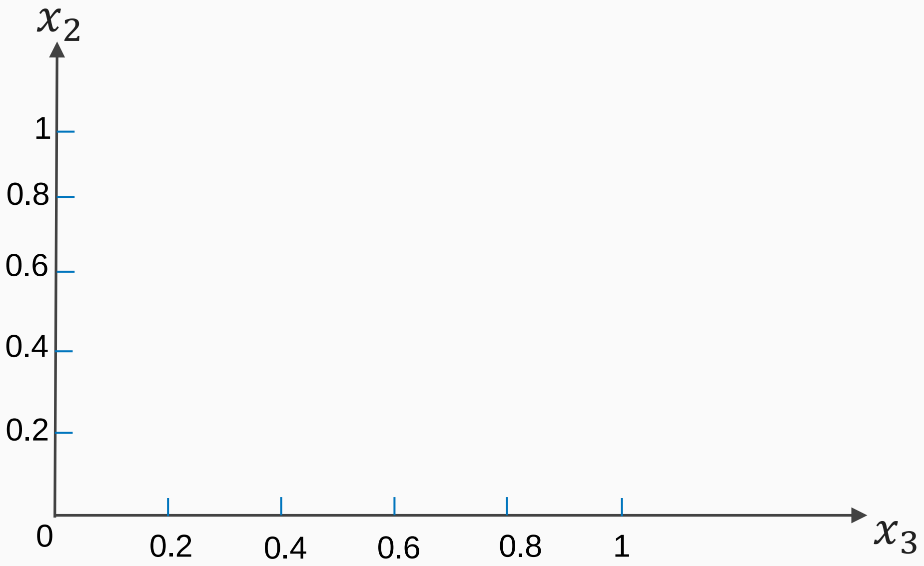
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$

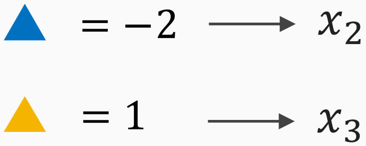


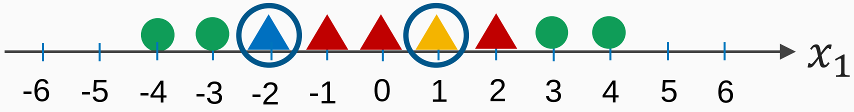


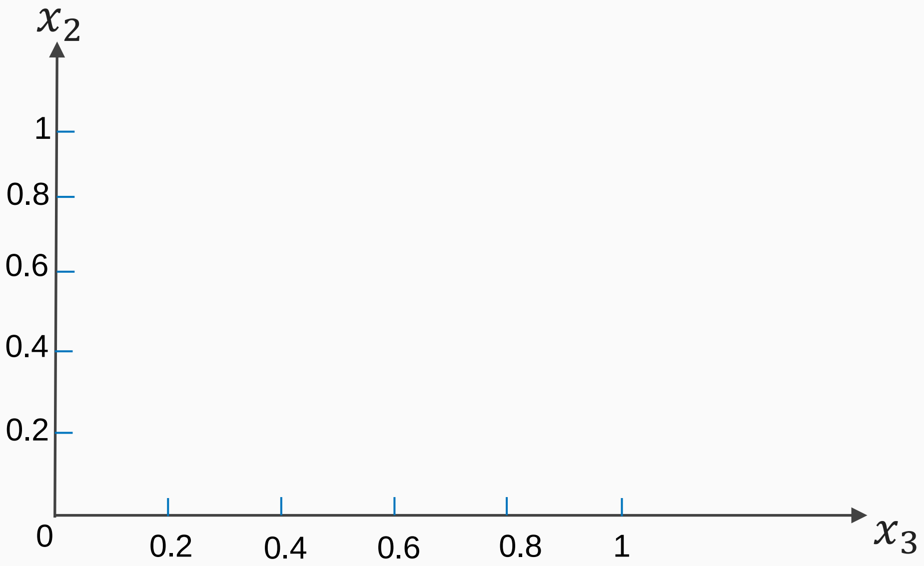
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

Let $\gamma = 0.3$

Our landmarks are :







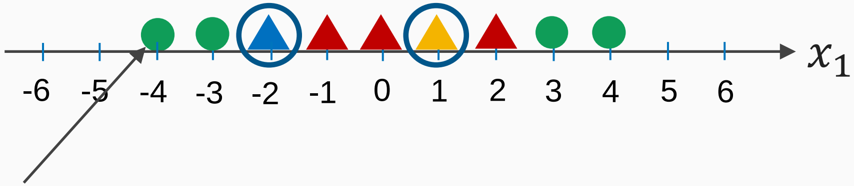
$$K(x_i, l_j) = e^{\left(-\gamma\|x_i - l_j\|^2\right)}$$

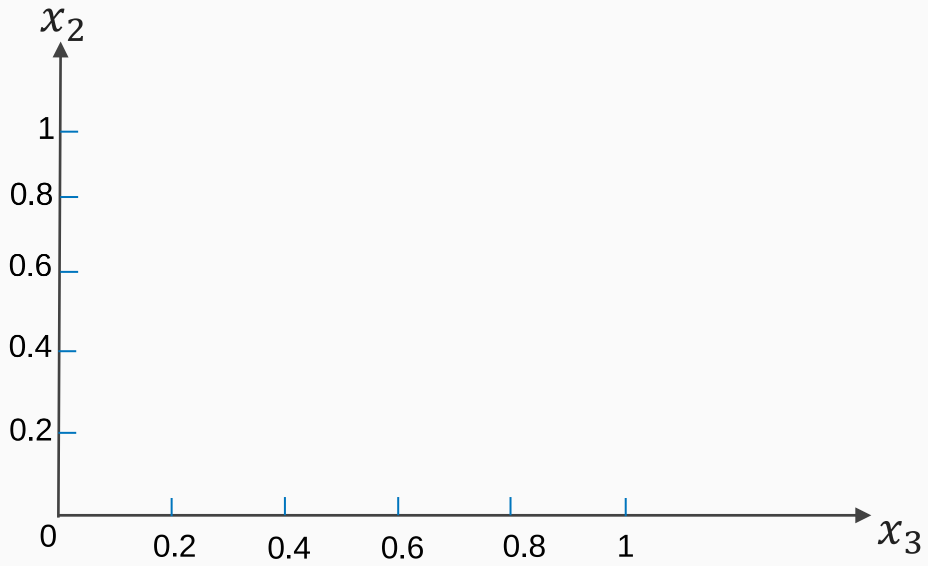
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$





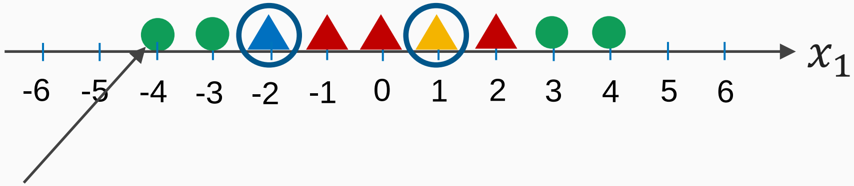
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

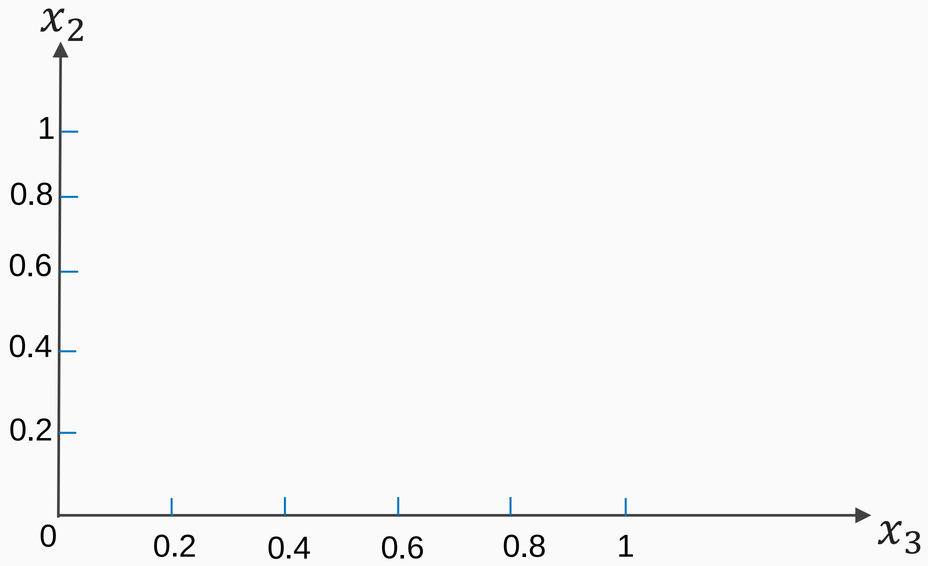
Let $\gamma = 0.3$

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

$\blacktriangle = 1 \longrightarrow x_3$





$$x_2 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(-2)\|^2\right)}$$

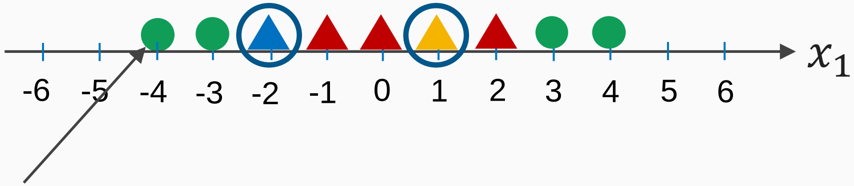
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

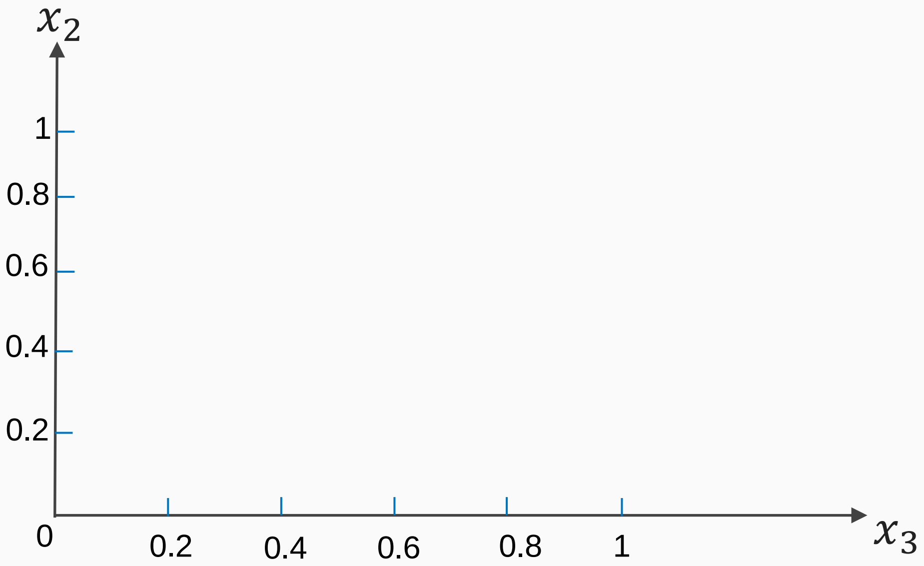
Let $\gamma = 0.3$

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

$\blacktriangle = 1 \longrightarrow x_3$





$$x_2 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(-2)\|^2\right)} = 0.30$$

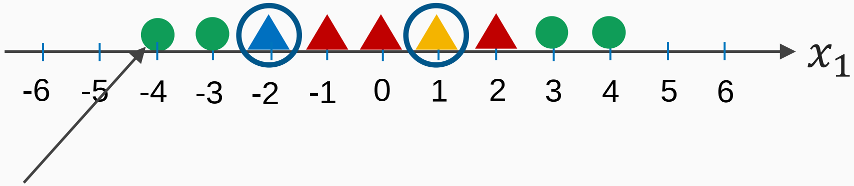
$$K(x_i, l_j) = e^{\left(-\gamma\|x_i - l_j\|^2\right)}$$

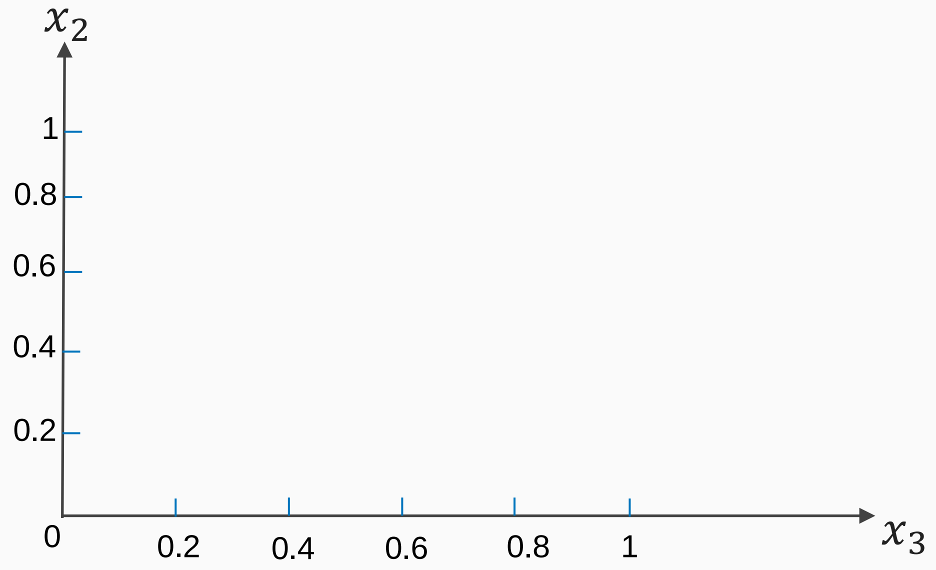
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$





$$x_2 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(-2)\|^2\right)} = 0.30$$

$$x_3 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(1)\|^2\right)}$$

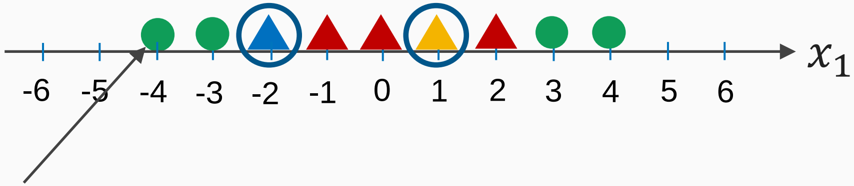
$$K(x_i, l_j) = e^{\left(-\gamma\|x_i - l_j\|^2\right)}$$

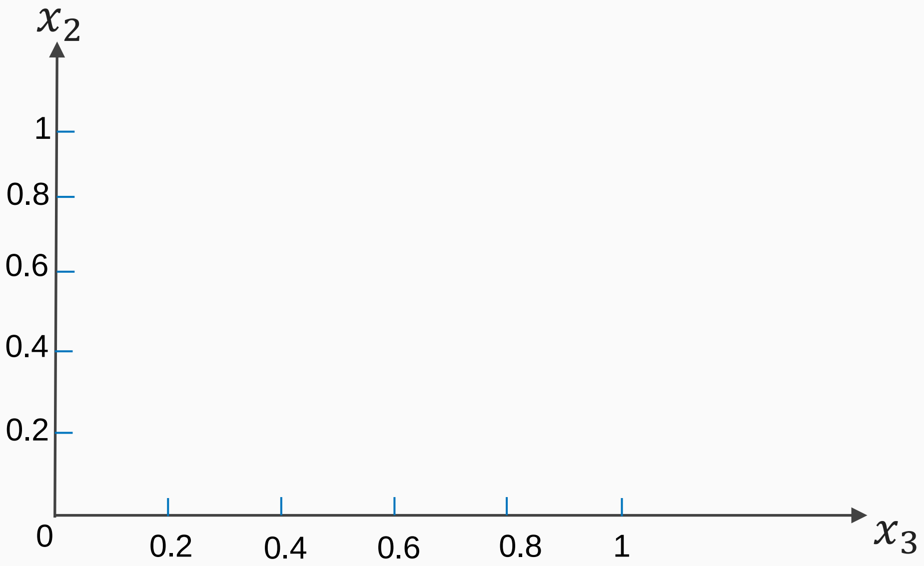
Let $\gamma = 0.3$

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

$\blacktriangle = 1 \longrightarrow x_3$





$$x_2 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(-2)\|^2\right)} = 0.30$$

$$x_3 = K(\bullet, \blacktriangle) = e^{\left(-0.3\|-4-(1)\|^2\right)} = 0.0005$$

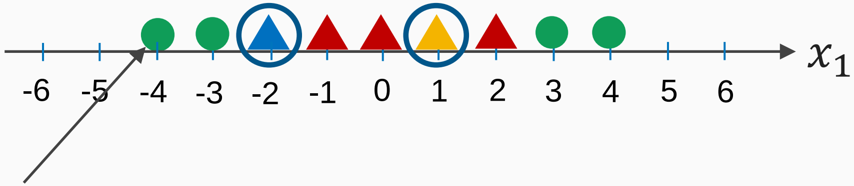
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

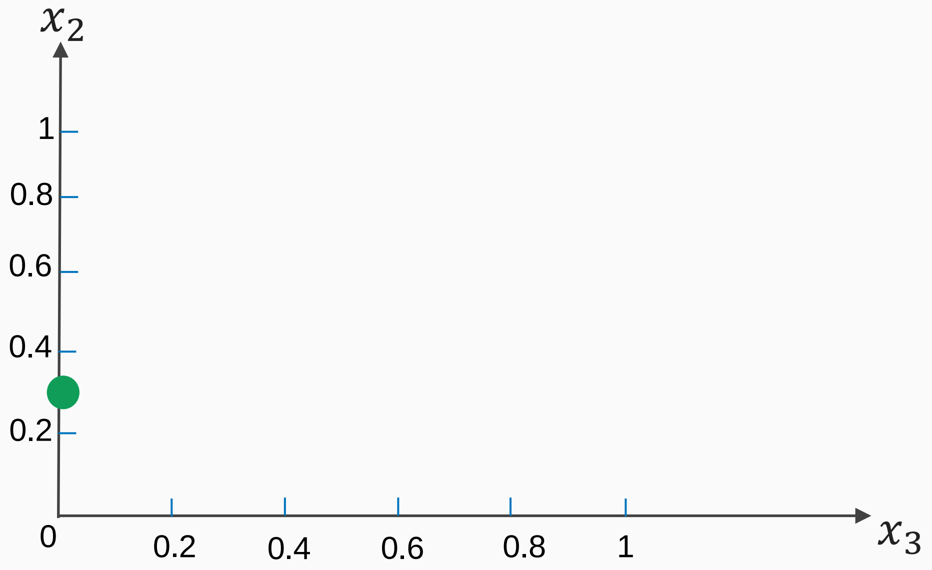
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$





$$x_2 = K(\,●\,,\,▲\,) = e^{\left(-0.3\|-4-(-2)\|^2\right)} = 0.30$$

$$x_3 = K(\,●\,,\,▲\,) = e^{\left(-0.3\|-4-(1)\|^2\right)} = 0.0005$$

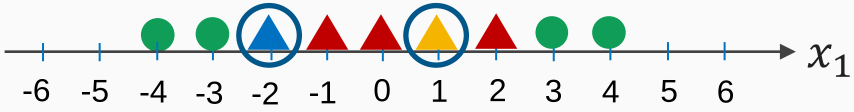
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

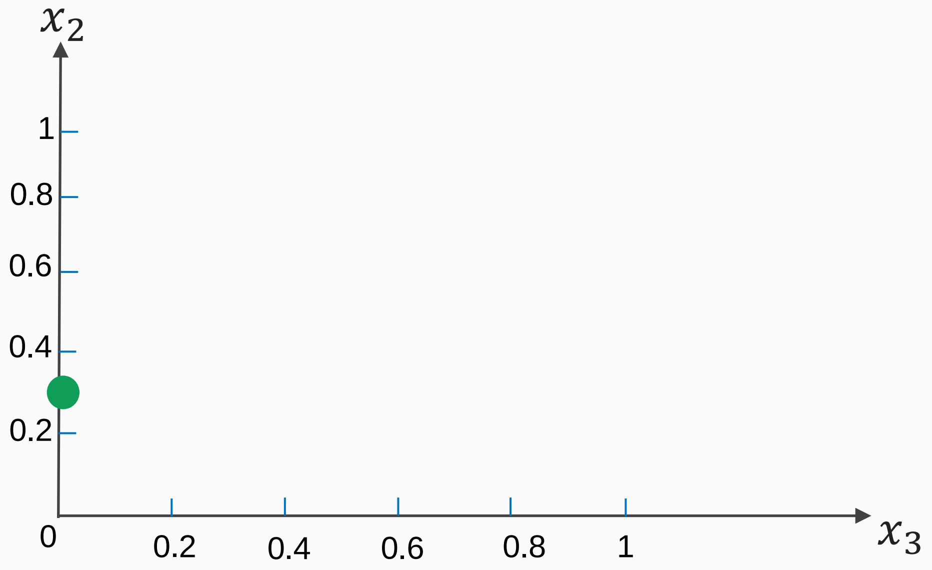
Let $\gamma = 0.3$

Our landmarks are :

$= -2 \longrightarrow x_2$

$= 1 \longrightarrow x_3$





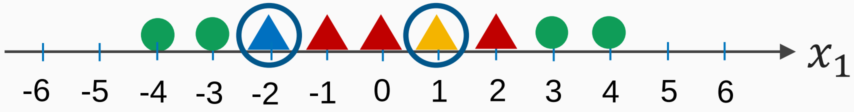
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

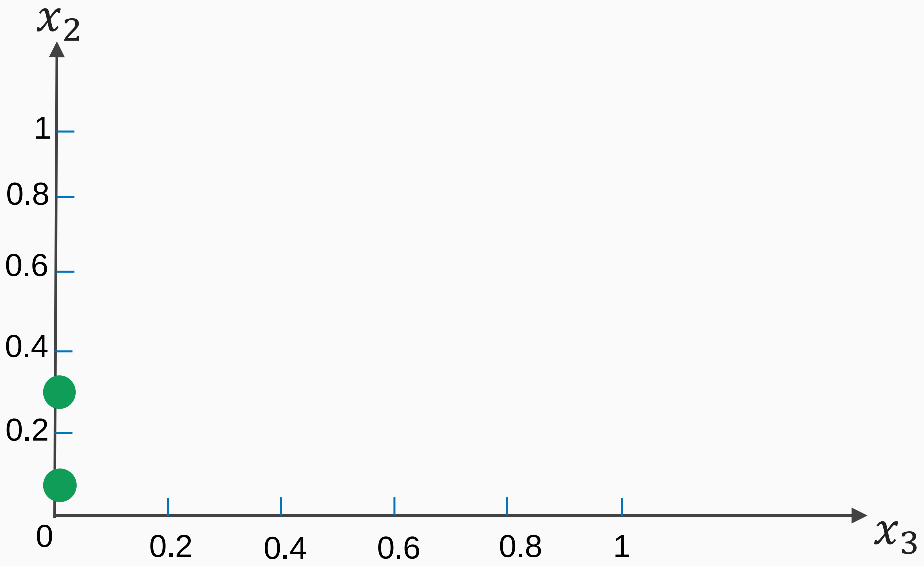
Let $\gamma = 0.3$

Our landmarks are :

$\blacktriangle = -2 \longrightarrow x_2$

$\blacktriangle = 1 \longrightarrow x_3$





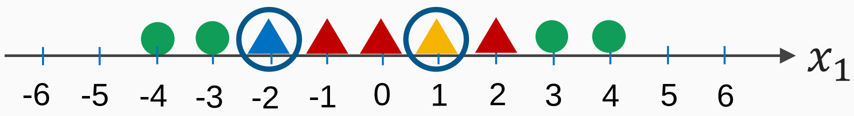
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

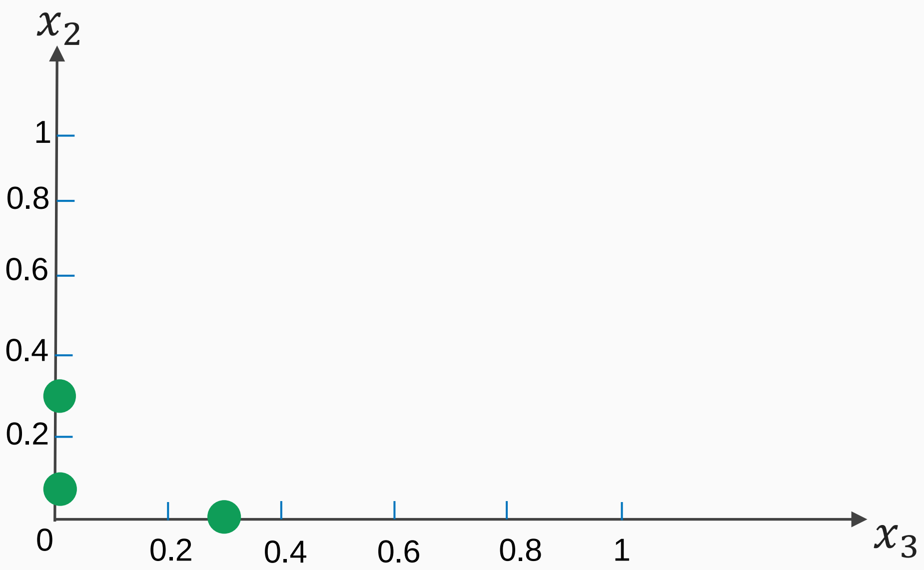
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$





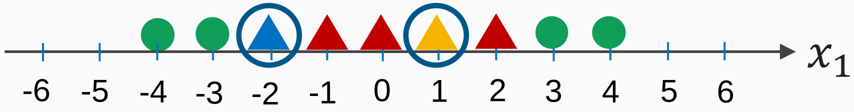
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

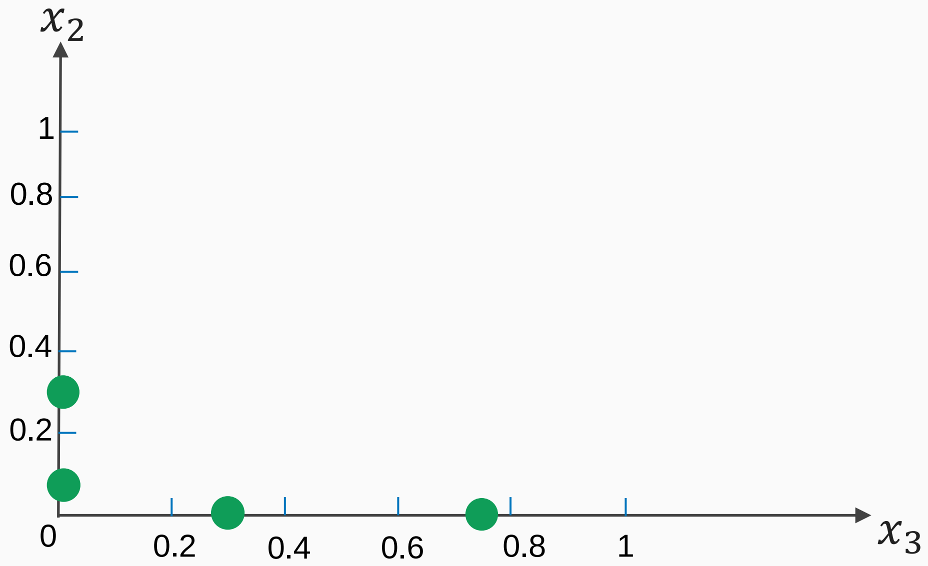
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$

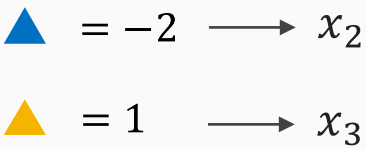


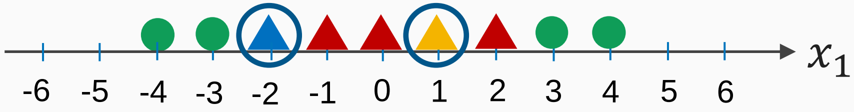


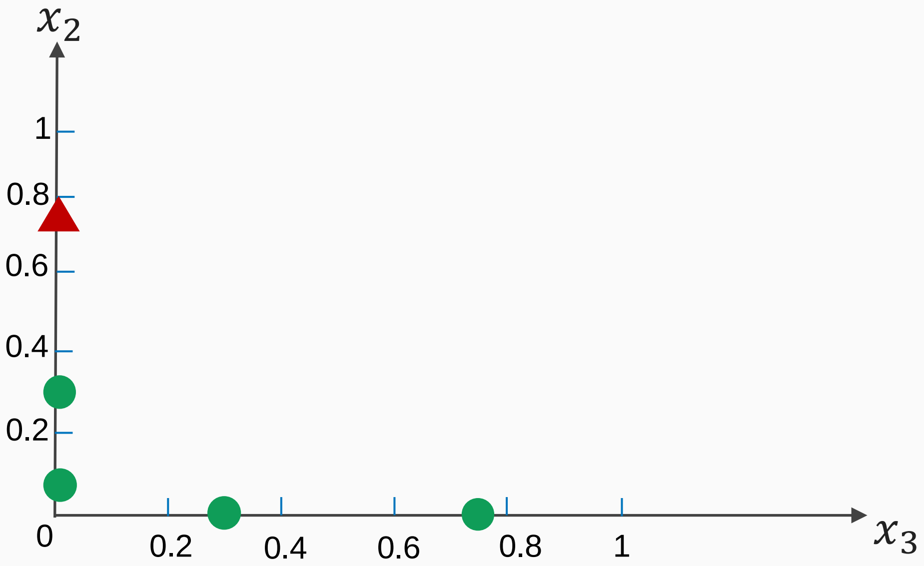
$$K(x_i, l_j) = e^{\left(-\gamma\|x_i - l_j\|^2\right)}$$

Let $\gamma = 0.3$

Our landmarks are :







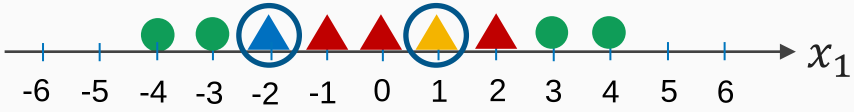
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

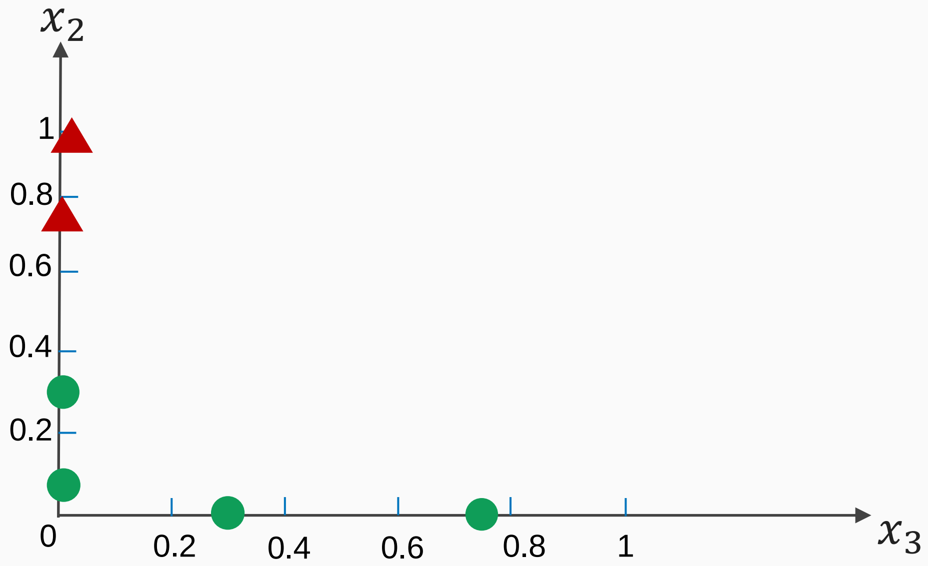
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$

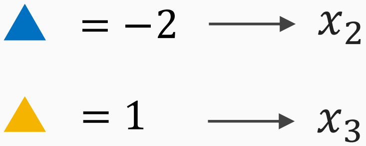


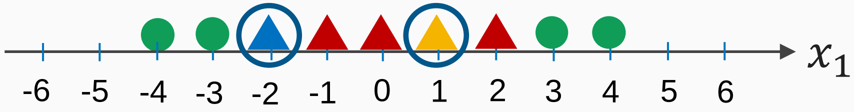


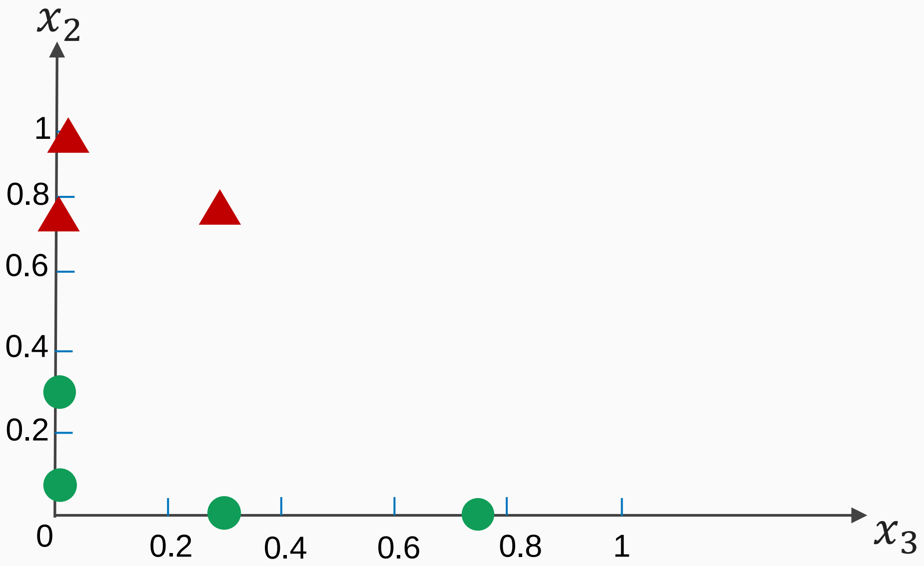
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

Let $\gamma = 0.3$

Our landmarks are :







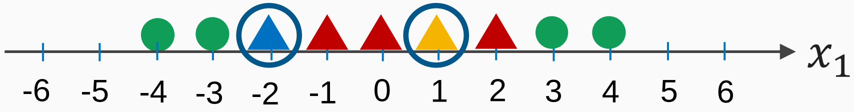
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

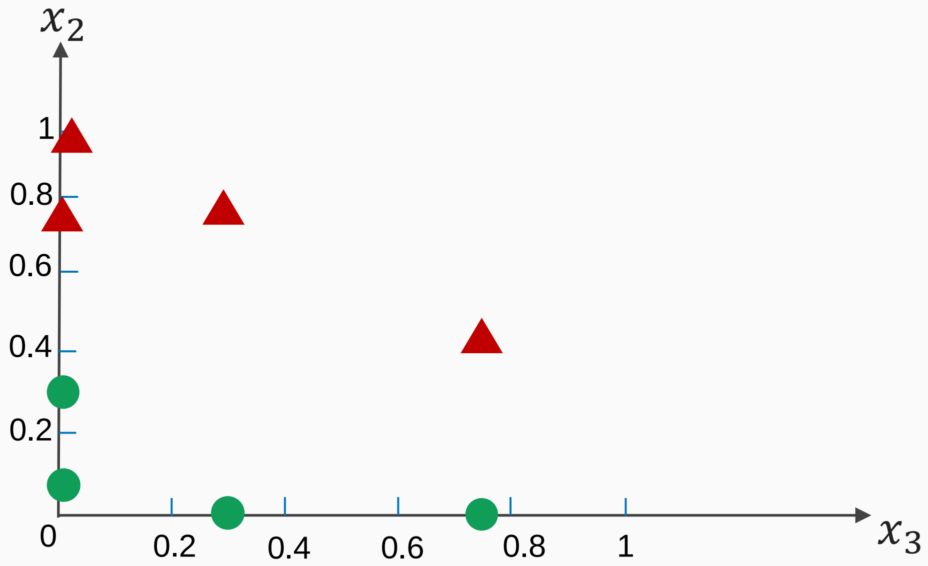
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$





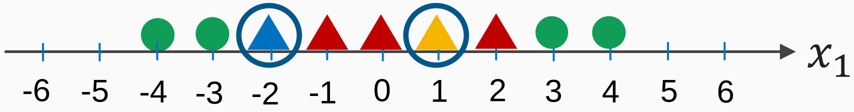
$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

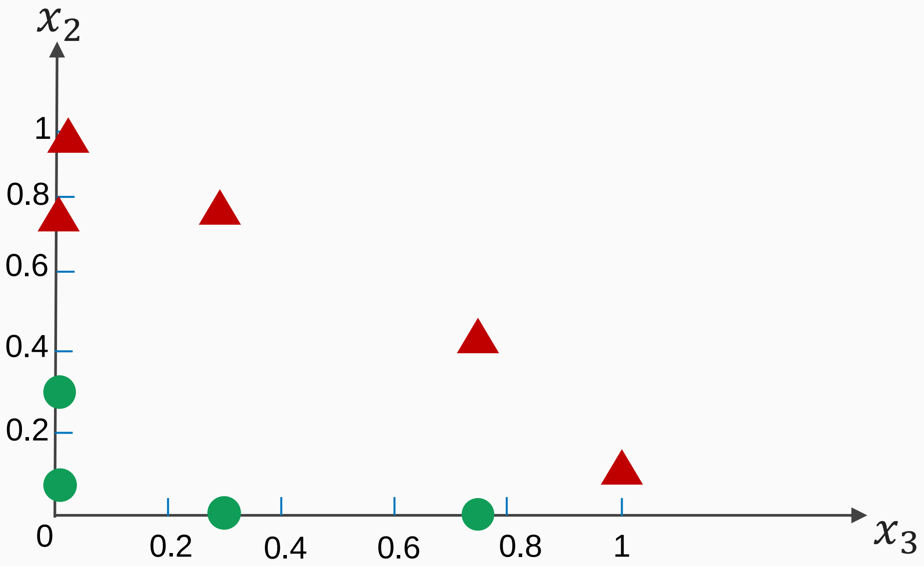
Let $\gamma = 0.3$

Our landmarks are :

▲ $= -2 \longrightarrow x_2$

▲ $= 1 \longrightarrow x_3$

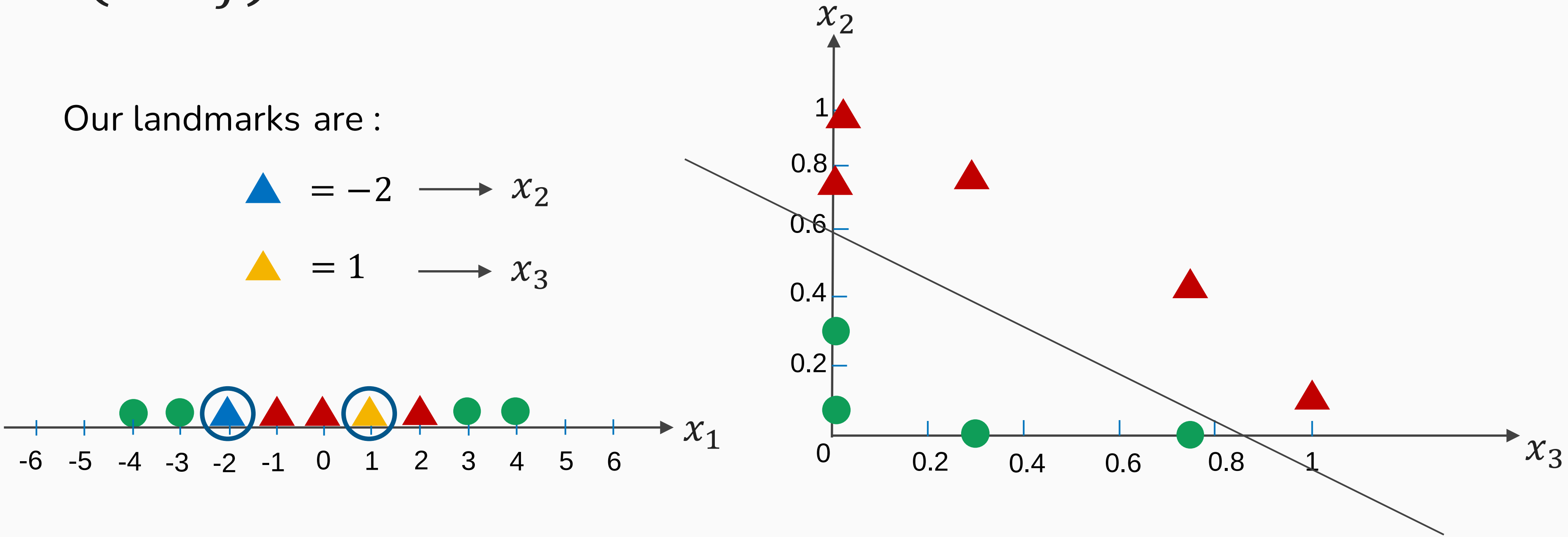




$$K(x_i, l_j) = e^{\left(-\gamma \|x_i - l_j\|^2\right)}$$

Let $\gamma = 0.3$

Our landmarks are :



Linearly separable

How do we select the landmarks ?

We choose every data point as a landmark

Downsides ?

Computationally expensive

Kernel trick

Instead of the primal problem :

$$\textit{minimize} \quad \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N}\zeta^{(i)}$$

$$\textit{subject to}: \quad y_i(wx_i + b) + \zeta^{(i)} \geq 1 \quad i \in \{1, \cdots, N\}$$

$$\zeta^{(i)} \geq 0$$

Using KKT conditions we obtain the dual problem

Using KKT conditions we obtain the dual problem

$$\text{maximize} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle {x_i}^T, x_j \rangle$$

$$\text{subject to :} \qquad 0 \le \alpha_i \le \mathrm{C} \qquad i = 1, \dots, \mathrm{N}$$

$$\textit{and} \sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

Now we can transform our features x_i using some transformation to overcome the non linear data and SVM can learn using some new features

Now we can transform our features x_i using some transformation to overcome the non linear data and SVM can learn using some new features

$$\text{maximize} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \phi({x_i}^T), \phi(x_j) \rangle$$

$$\text{subject to :} \qquad 0 \leq \alpha_i \leq \mathrm{C} \qquad i = 1, \ldots, \mathrm{N}$$

$$and \sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

Now we can transform our features x_i using some transformation to overcome the non linear data and SVM can learn using some new features

$$\text{maximize} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \phi({x_i}^T), \phi(x_j) \rangle$$

$$\text{subject to}: \qquad 0 \leq \alpha_i \leq \mathrm{C} \qquad i = 1, \ldots, \mathrm{N}$$

$$and \sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

$$\langle \phi({x_i}^T), \phi(x_j) \rangle$$

$$\langle \phi({x_i}^T), \phi(x_j) \rangle$$

$$\langle \phi({x_i}^T), \phi(x_j) \rangle$$

We could find ϕ and then calculate the inner product

$$\langle \phi({x_i}^T), \phi(x_j) \rangle$$

We could find ϕ and then calculate the inner product

$$\langle \phi({x_i}^T), \phi(x_j) \rangle$$

We could find ϕ and then calculate the inner product

Or we could use a kernel function to calculate the inner product without even visiting the high dimensional space

For example

Suppose we have a 2nd transformation ϕ :

For example

Suppose we have a 2nd transformation ϕ :

$$\phi(x_i) = \begin{pmatrix} {x_{i1}}^2 \\ x_{i1}x_{i2} \\ x_{i2}x_{i1} \\ {x_{i2}}^2 \end{pmatrix} \quad \phi(x_j) = \begin{pmatrix} {x_{j1}}^2 \\ x_{j1}x_{j2} \\ x_{j2}x_{j1} \\ {x_{j2}}^2 \end{pmatrix}$$

$$x_i = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad , \quad x_j = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$x_i = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad , \quad x_j = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\phi(x_i) = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}$$

$$x_i = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad , \quad x_j = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\phi(x_i) = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix} \qquad \phi(x_j) = \begin{pmatrix} 9 \\ 15 \\ 15 \\ 25 \end{pmatrix}$$

$$x_i = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad , \quad x_j = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\phi(x_i) = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix} \qquad \phi(x_j) = \begin{pmatrix} 9 \\ 15 \\ 15 \\ 25 \end{pmatrix}$$

$$\langle \phi({x_i}^T), \phi(x_j) \rangle = 9 + 30 + 30 + 100 = 169$$

$$\langle \phi({x_i}^T), \phi(x_j) \rangle = K\left(x_j, x_i\right) = \left(\begin{pmatrix}1\\2\end{pmatrix}^T \cdot \begin{pmatrix}3\\5\end{pmatrix}\right)^2$$

$$= (3 + 10)^2 = 169$$

$$\left\langle \phi({x_i}^T), \phi(x_j) \right\rangle = K\left(x_j, x_i\right) = \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}^T \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix}\right)^2$$

$$= (3 + 10)^2 = 169$$

This kernel is called : linear kernel

And finally our program becomes :

And finally our program becomes :

$$\text{maximize} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot K(x_j, x_i)$$

$$\text{subject to :} \qquad 0 \leq \alpha_i \leq C \qquad i = 1, \ldots, N$$

$$\textit{and} \sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

Once we find the corresponding α_i using a QP solver we use the following equation to find w and b that minimizes the primal problem

Once we find the corresponding α_i using a QP solver we use the following equation to find w and b that minimizes the primal problem

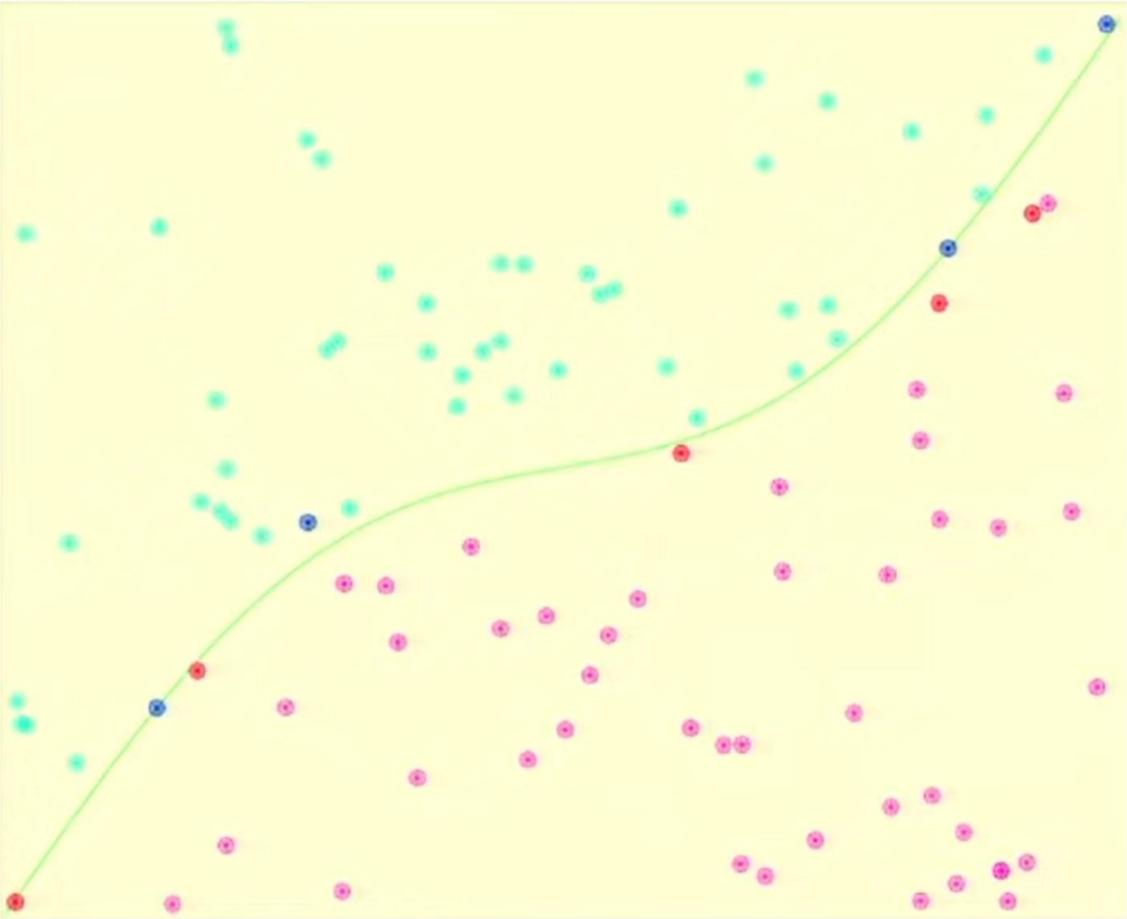
$$w = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot x_i$$

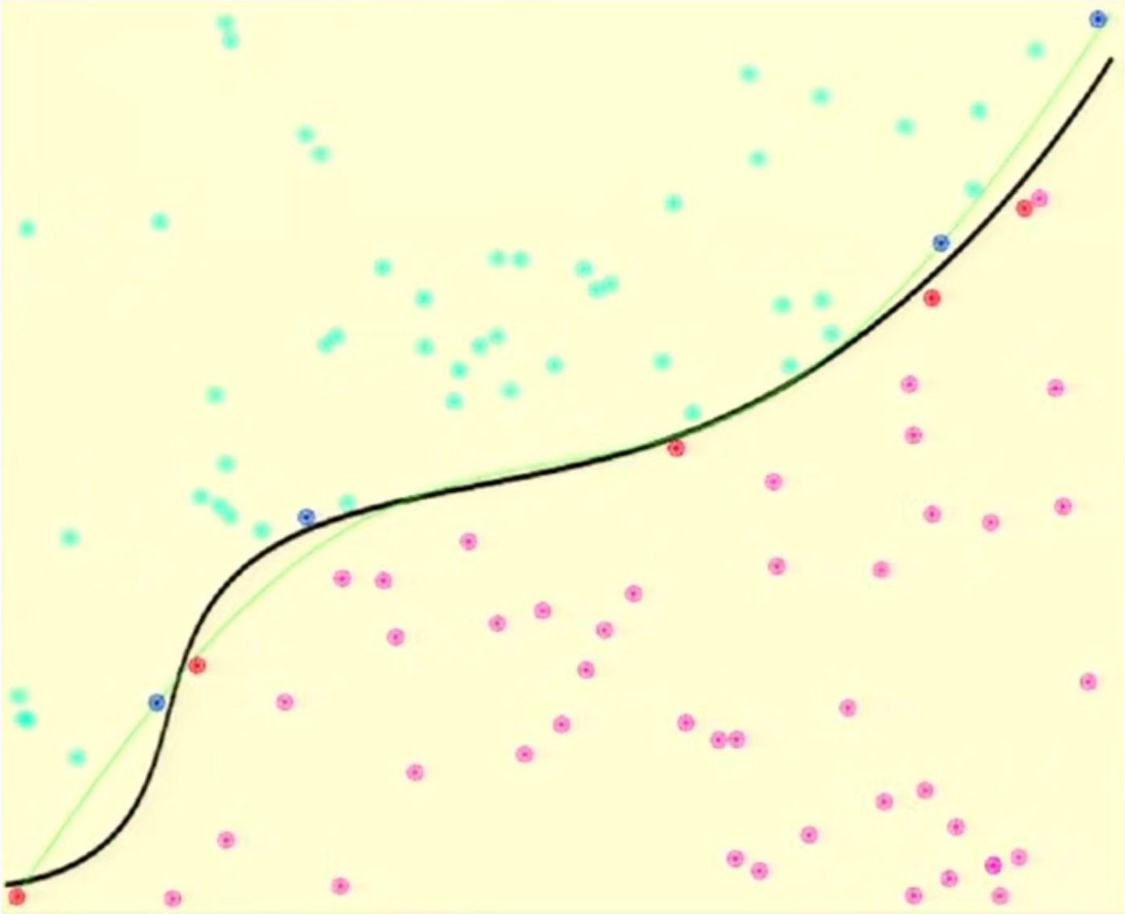
$$b = \frac{1}{N} \sum_{i=1}^{N} (y_i - w^T x_i)$$

One special kernel function is the RBF kernel :

One special kernel function is the RBF kernel :

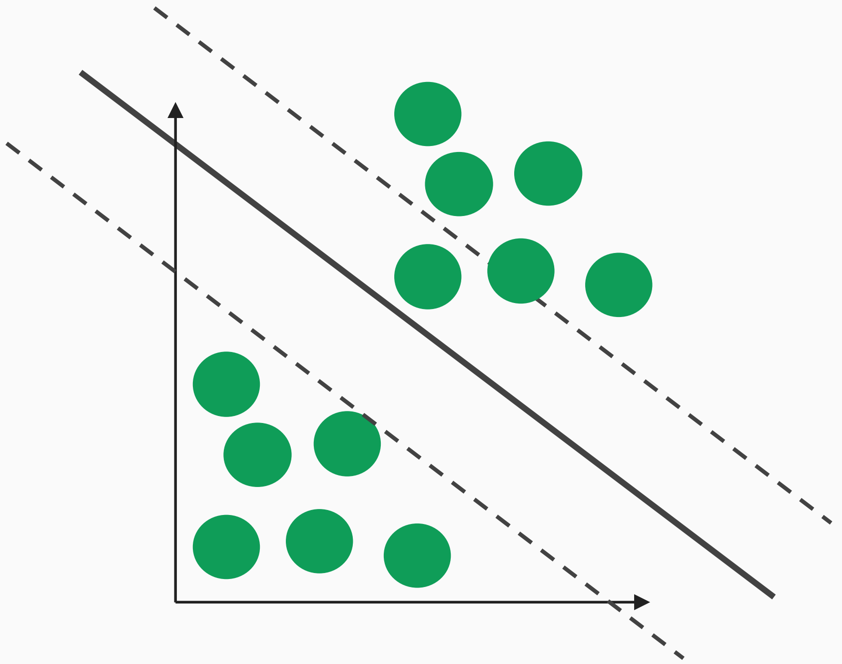
$$K\left(x_i, x_j\right) = e^{\left(-\gamma\|x_i - x_j\|^2\right)}$$



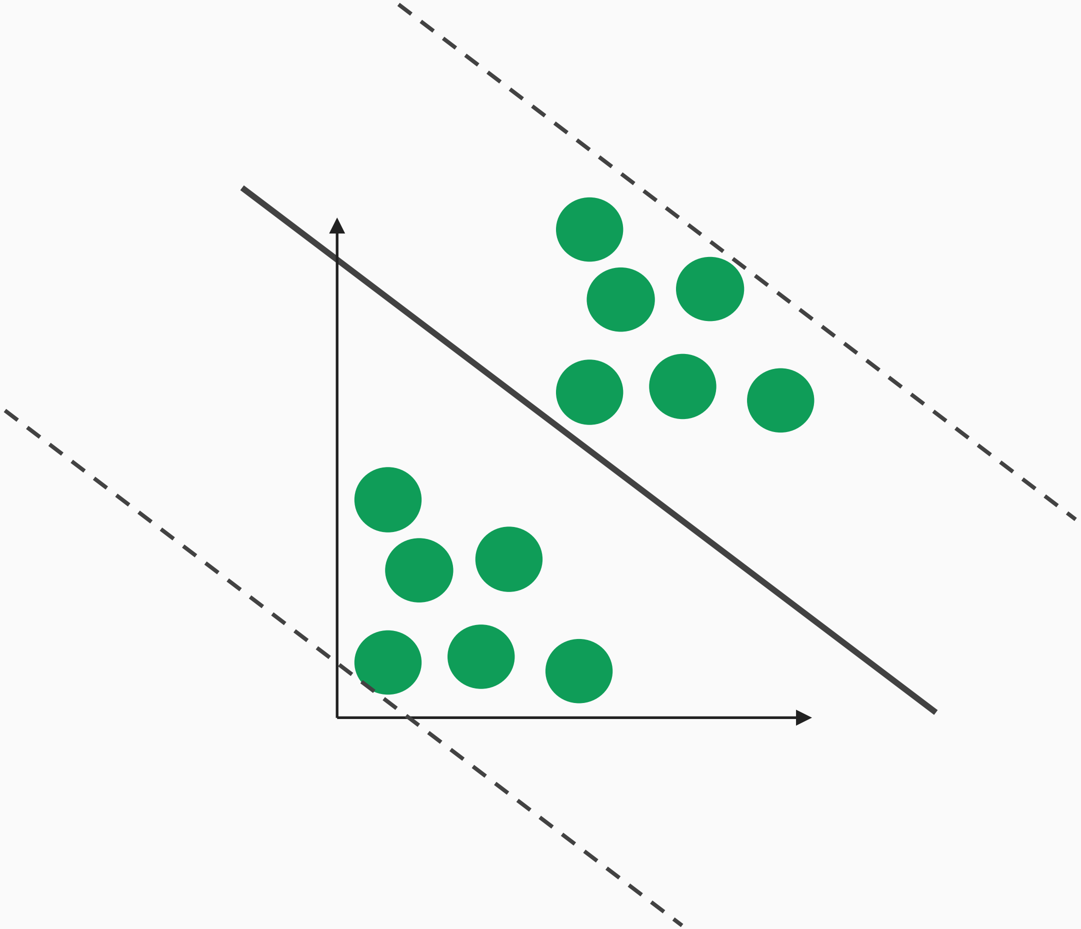


SVMs for regression

Instead of :



we maximize the instances on the street :



END