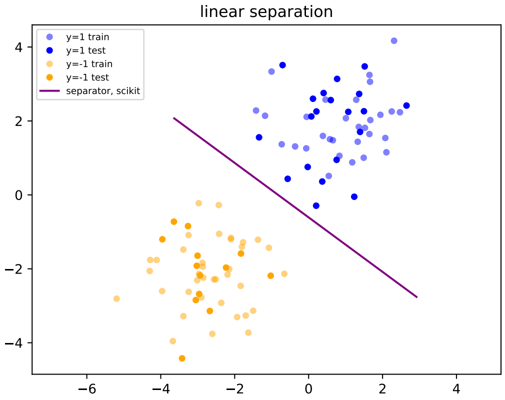
# Fondamentaux théoriques du machine learning



# Support vector machines

## Support vector machines

- Linear separation
- Optimization problem
- Link with empirical risk minimization

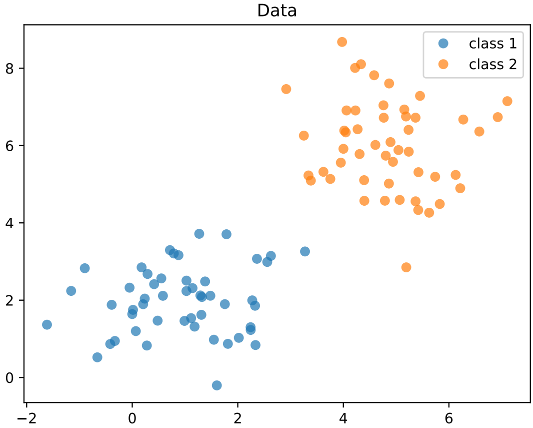



Figure – Linearly separable data

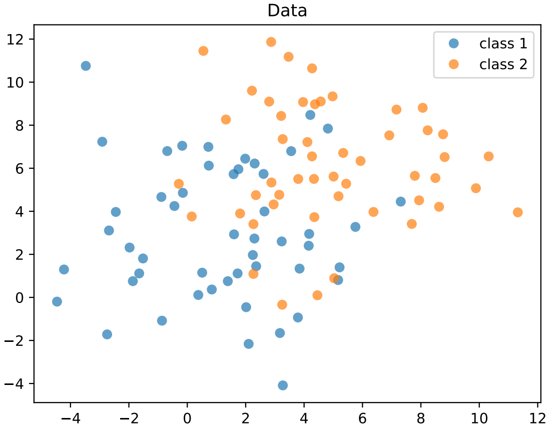



Figure – Non linearly-separable data

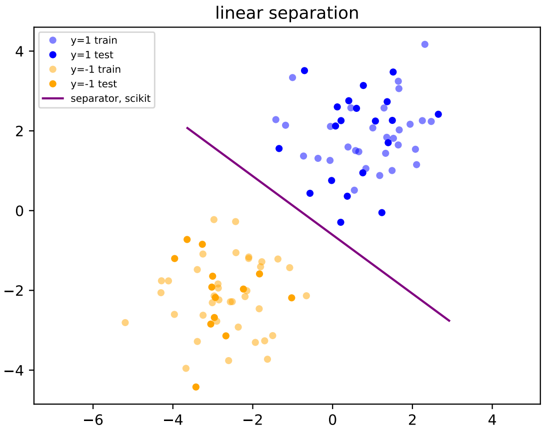



Figure – Linear separator

# Linear separator

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \{-1, 1\}$

Equation of a linear separator

$$\langle w, x \rangle + b = 0 \tag{1}$$

- $w \in \mathbb{R}^d$
- $x \in \mathbb{R}^d$
- $b \in \mathbb{R}$

Notation :

$$h_{w,b}(x) = \langle w, x \rangle + b \tag{2}$$

# Affine subspace

$$H = \{x \in \mathbb{R}^d, \langle w, x \rangle + b = 0\} \tag{3}$$

is an affine subspace.

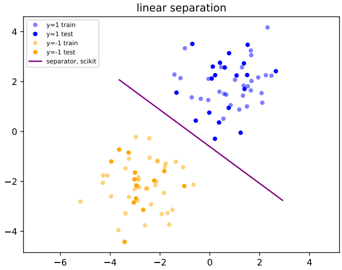
Any vector $x \in \mathbb{R}^d$ can uniquely be decomposed as

$$x = \lambda_w^x \frac{w}{||w||} + x_{w^\perp} \tag{4}$$

with $x_{w^\perp} \in \text{vect}(w)^\perp$. $x \in H$ if and only if

$$\begin{aligned} \langle w, x \rangle + b = 0 \\ \Leftrightarrow \langle w, \lambda_w^x \frac{w}{||w||} + x_{w^\perp} \rangle + b = 0 \\ \Leftrightarrow \langle w, \lambda_w^x \frac{w}{||w||} \rangle + b = 0 \\ \Leftrightarrow \lambda_w^x ||w|| + b = 0 \\ \Leftrightarrow \lambda_w^x = \frac{-b}{||w||} \end{aligned} \tag{5}$$

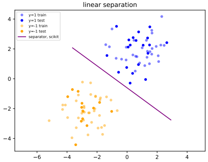
We first consider a linearly separable situation.




We recall the definition $h_{w,b}(x) = \langle w, x \rangle + b$. We look for separators that satisfy :

- $\forall x_i$ such that $y_i = 1$, $h_{w,b}(x) \geq 0$
- $\forall x_i$ such that $y_i = -1$, $h_{w,b}(x) \leq 0$

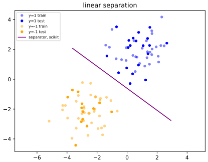
We first consider a linearly separable situation.




We note $h_{w,b}(x) = \langle w, x \rangle + b$. We look for separators that satisfy :

- $\forall x_i$ such that $y_i = 1$, $h_{w,b}(x) \geq 0$
- $\forall x_i$ such that $y_i = -1$, $h_{w,b}(x) \leq 0$

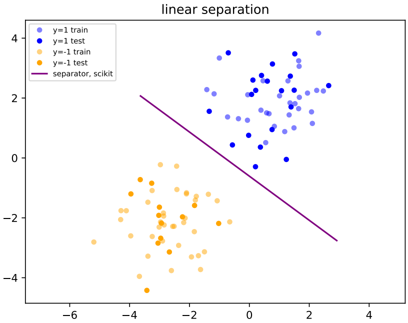
**However**, there exists an infinite number of such parameters. How could we choose the best one ?




- $\forall x_i$ such that $y_i = 1$, $h_{w,b}(x) \geq 0$
- $\forall x_i$ such that $y_i = -1$, $h_{w,b}(x) \leq 0$

The **margin** is the distance from $H$ to the dataset. We look for the separator with the largest margin, leading to **Support vector classification (SVC)**.

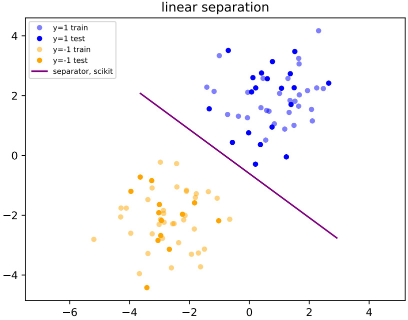
# Margin




Let $x$ be a point such that $h_{w,b}(x) = \langle w, x \rangle + b = c$, with $c \in \mathbb{R}$.

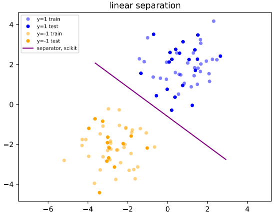
Exercice 1 : Compute the distance from $x$ to $H$.

# Margin




Let $x$ be a point such that $h_{w,b}(x) = \langle w, x \rangle + b = c$, with $c \in \mathbb{R}$.
The distance is $\frac{|c|}{||w||}$.

# Support vectors

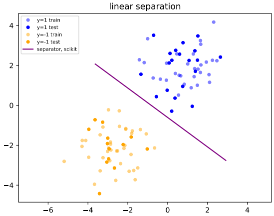



The **support vectors** are the vectors such that $|h_{w,b}(x)|$ is minimal among the dataset.

- the margin $M$ is the distance from $H$ to these vectors.
- if $H$ is the optimal separator, there has to be a vector $x_-$ and $x_+$ on each side, such that

$$M = d(x_-, H) = d(x_+, H) \tag{6}$$

# Support vectors




Exercice 2 : Show that if $H$ is optimal, then

$$M = d(x_-, H) = d(x_+, H) \tag{7}$$

# Rescaling

**Important remark** : multiplying $w$ and $b$ by a constant $\lambda \neq 0$ does not change $H$, as :

$$\begin{aligned} &\langle \lambda w, x\rangle + \lambda b = 0 \\ &\Leftrightarrow \lambda(\langle w, x\rangle + b) = 0 \\ &\Leftrightarrow \langle w, x\rangle + b = 0 \end{aligned} \tag{8}$$

# Rescaling

**Important remark** : multiplying $w$ and $b$ by a constant $\lambda \neq 0$ does not change $H$.
If the support vector $x$ is such that $h_{w,b}(x) = c$, we have seen that the margin is

$$\frac{|c|}{||w||} \tag{9}$$

When looking for the optimal $H$, we can impose, without loss of generality, that $|c| = 1$.
This means that we look for $w$ with minimal norm, such that $H$ separates the data (since the margin is $\frac{1}{||w||}$).

# Optimization problem

We can now formulate the optimization problem.

$$\arg\min_{w,b} \frac{1}{2}\langle w, w\rangle \tag{10}$$

subject to :

$$\forall i \in [1, n], y_i(\langle w, x_i\rangle + b) \geq 1 \tag{11}$$

# Slack variables

When the dataset is not linearly separable, the approach is to authorize some of the samples to have a margin smaller that 1. This means relaxing the constraint, from

$$y_i(\langle w, x_i \rangle + b) \geq 1 \tag{12}$$

to

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \tag{13}$$

The $\xi$ are called the *slack variables*, they are $\geq 0$. The smaller the slack variabes, the better.

# Optimization problem

In the general case, the optimization problem is :

$$\underset{w,b,\xi}{\arg\min} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}\xi_i \tag{14}$$

subject to :

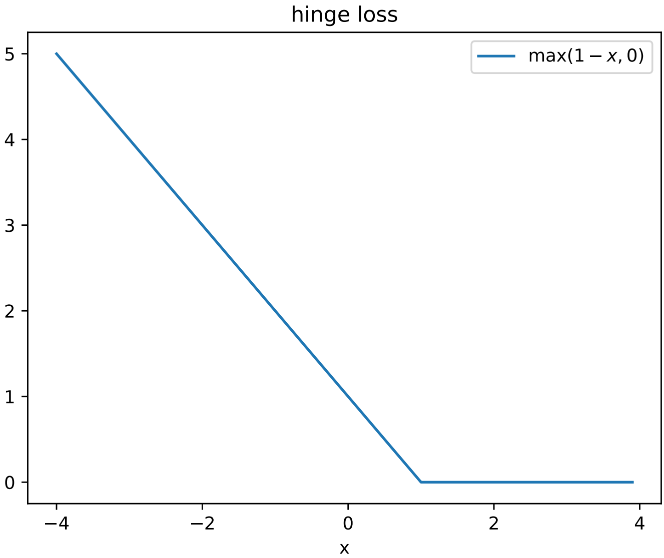
$$\forall i \in [1, n], y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \tag{15}$$

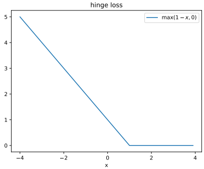
and

$$\forall i \in [1, n], \xi_i \geq 0 \tag{16}$$

# Margin vs ERM

The margin maximisation seems to differ from empirical risk minimization (ERM), which we have studied earlier.
However, with a specific loss function, we an show that margin maximisation is in fact an ERM.


hinge loss
max(1 − x, 0)
5
4
3
2
1
0
−4
−2
0
2
4
x




- estimation : $h(x) = \langle w, x \rangle + b$
- label : $y \in \{-1, 1\}$

**Hinge loss :**

$$L_{\text{hinge}}(h(x), y) = \max(0, 1 - yh(x)) \tag{17}$$

The hinge loss can be seen as an approximation of the binary loss.

# Problem reformulation

We recall the constraints on $\xi$

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \tag{18}$$

and

$$\xi_i \geq 0 \tag{19}$$

Equivalently,

$$\xi_i \geq max\big(0, 1 - y_i(\langle w, x_i \rangle + b)\big) \tag{20}$$

# Problem reformulation

The slack variables should be minimal. Hence, we can write that for the optimal solution, the inequality is in fact an equality ;

$$\xi_i = max\big(0, 1 - y_i(\langle w, x_i \rangle + b)\big) \tag{21}$$

# Problem reformulation

Finally, we can rewrite the problem as

$$\arg\min_{w,b} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n} max\big(0, 1 - y_i(\langle w, x_i\rangle + b)\big) \tag{22}$$

or equivalently

$$\arg\min_{w,b} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n} L_{\text{hinge}}(h(x_i)), y_i) \tag{23}$$

Which is an ERM problem with a $L2$ regularization.