SVM
Support Vector Machine
Machine à Vecteur Support

Catherine ACHARD
Institut des Systèmes Intelligents et de Robotique

catherine.achard@sorbonne-universite.fr

Bibliographie

Livres

C. Bishop, Pattern Recognition and Machine Learning, 2006
R. Duda, P. Hart et D. Stork, Pattern Classification, 2002

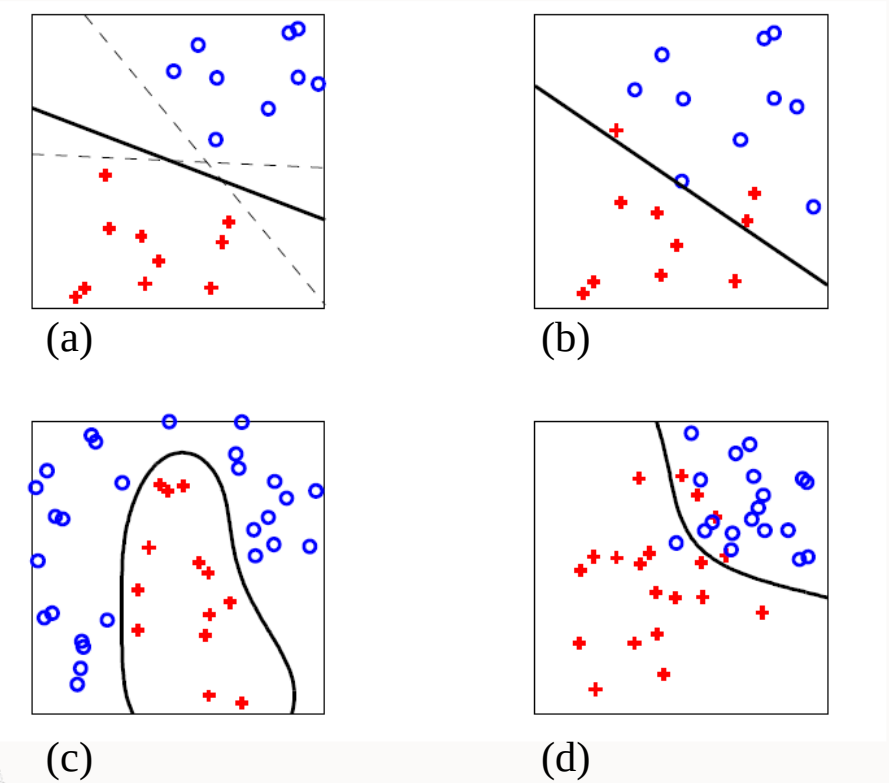
Internet

https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-svm.pdf
https://www.lri.fr/~antoine/Courses/Master-ISI/ISI-10/Tr-cours-SVM_2014_2x2.pdf

Introduction

But : trouver une fonction de décision $f(\boldsymbol{x})$ pour un problème à deux classes

Image issue de Wikistat/pdf/st-m-app-svm.pdf



Exemples d'apprentissage :
$\{(\boldsymbol{x}_i), i = 1, N\}$
Avec $\boldsymbol{x}_i \in \mathbb{R}^n$

Classes :
$\{(y_i), i = 1, N\}$
Avec $y_i \in \{+1, -1\}$

Deux cas:

- Linéairement séparable (a et b)
- Non linéairement séparable (c et d)

SVM linéaire

Les données sont linéairement séparables s'il existe une fonction de décision $f(\boldsymbol{x})$ linéaire (un hyperplan dans l'espace de dimension n) permettant de les classer

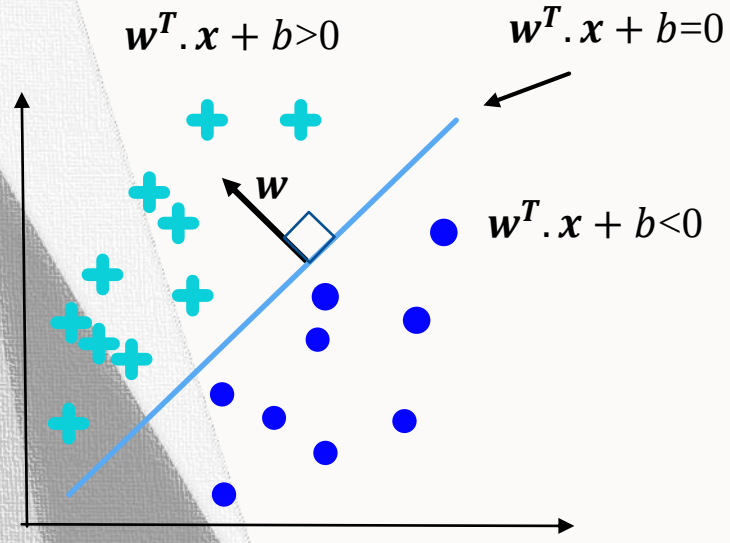
$$\boldsymbol{f(x) = w^T.x + b}$$

L'hyperplan séparateur vérifie

$$\boldsymbol{w^T.x_i + b > 0}\ si\ y_i = 1$$
$$\boldsymbol{w^T.x_i + b < 0}\ si\ y_i = -1$$

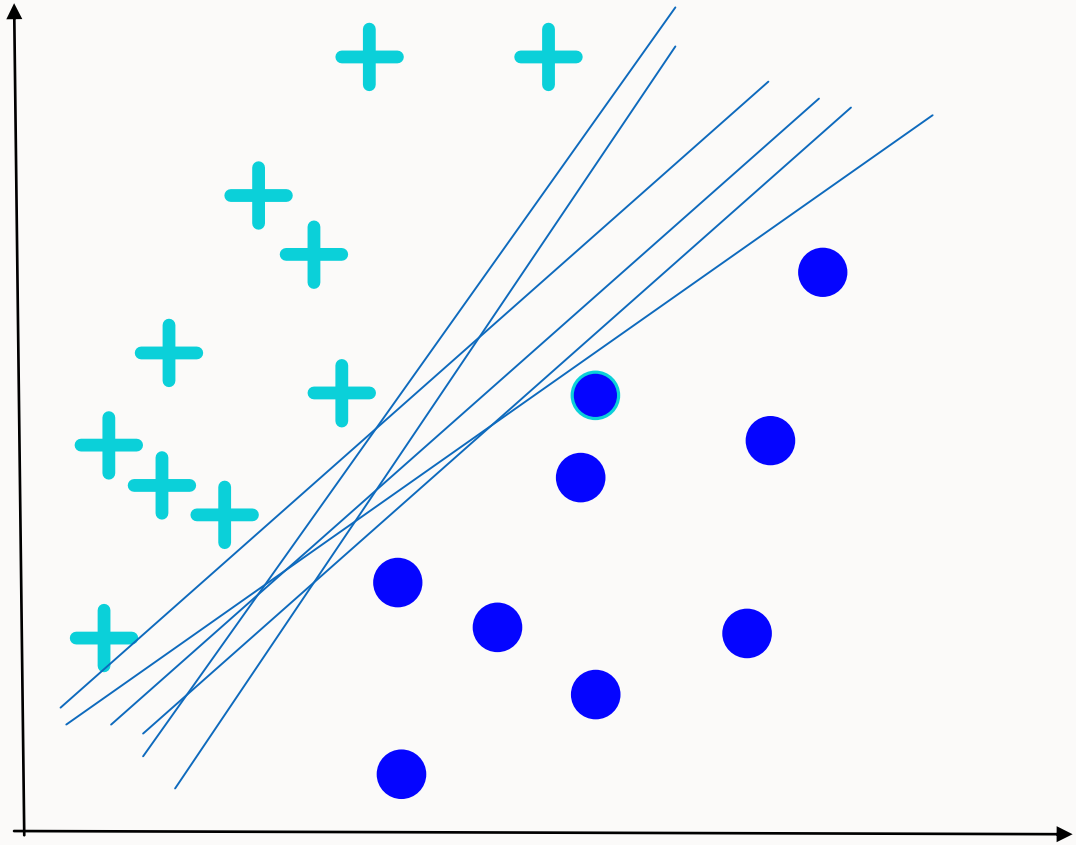
Ce qui revient à :

$$\forall i = 1, N \quad y_i\left(\boldsymbol{w^T.x_i + b}\right) > 0$$



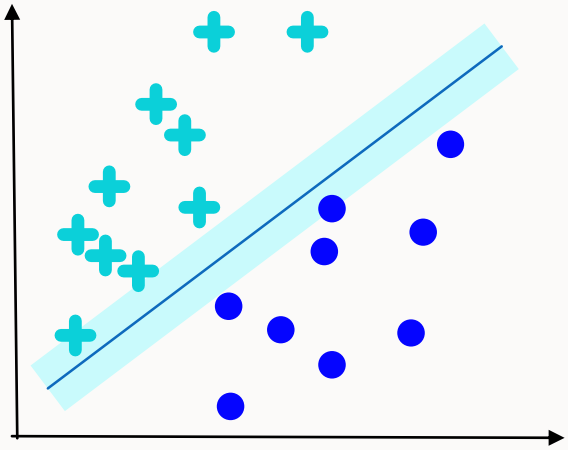
- $\boldsymbol{w}$ est un vecteur orthogonal à l'hyperplan
- b est une constante

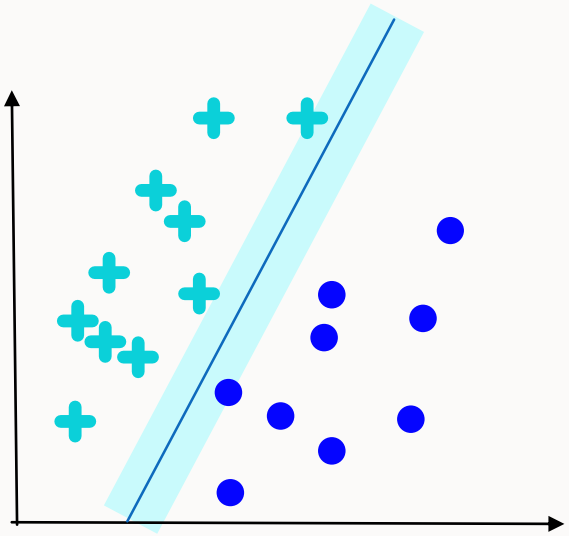
Il existe une infinité d'hyperplan, lequel choisir ?

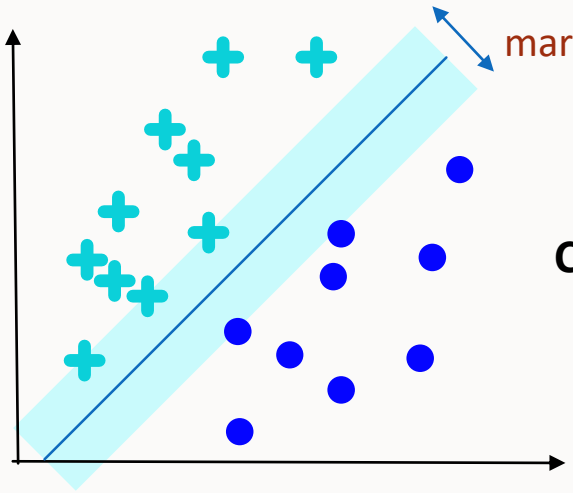


SVM linéaire

Il existe une infinité d'hyperplan, lequel choisir ?





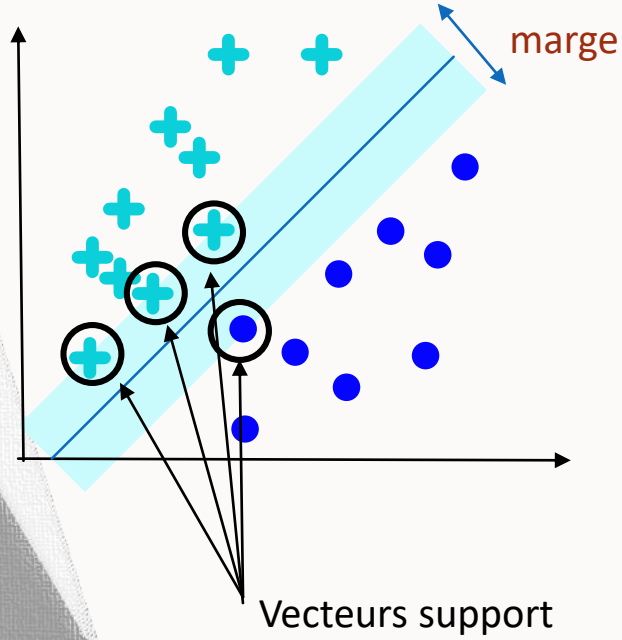


Celui qui maximise la marge

Toutes les équations $k\boldsymbol{w}^{\boldsymbol{T}}.\boldsymbol{x_i} + kb = 0$ où k est une constante correspondent au même hyperplan.

On diminue l'espace des possibles en choisissant $\boldsymbol{w}$ et b tq :

$$\min_{i=1,N} y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} + b\right) = 1$$



Les exemples tq

$$y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} + b\right) = 1$$

sont appelés **vecteurs support**

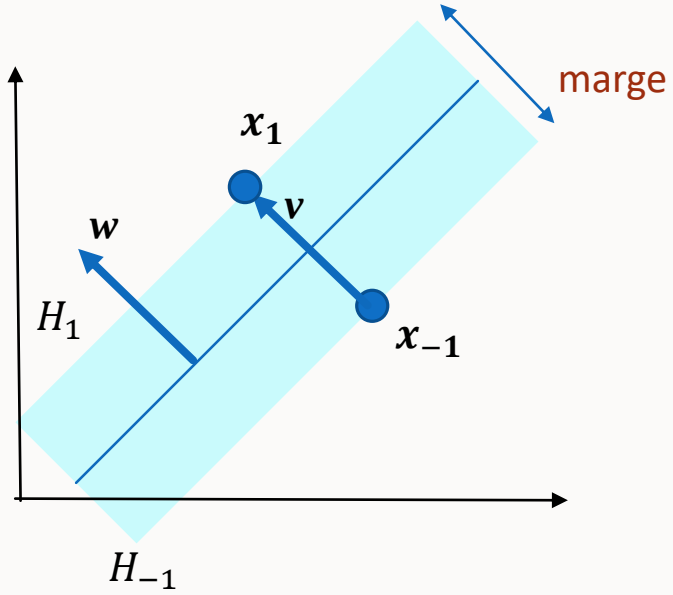
Mais comment calculer la marge m ?

Comment trouver la marge *m* ?

Considérons les points $\boldsymbol{x_{-1}}$ et $\boldsymbol{x_1}$, placés sur les hyperplans H_{-1} et H_1 tq

$$\boldsymbol{w^T x_1} + b = 1 \text{ (hyperplan } H_1)$$
$$\boldsymbol{w^T x_{-1}} + b = -1 \text{ (hyperplan } H_{-1})$$

$\boldsymbol{w}$ est perpendiculaire à l'hyperplan. Soit $\boldsymbol{v}$ le vecteur, colinéaire à $\boldsymbol{w}$, de norme *m* : $\boldsymbol{v} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} m$

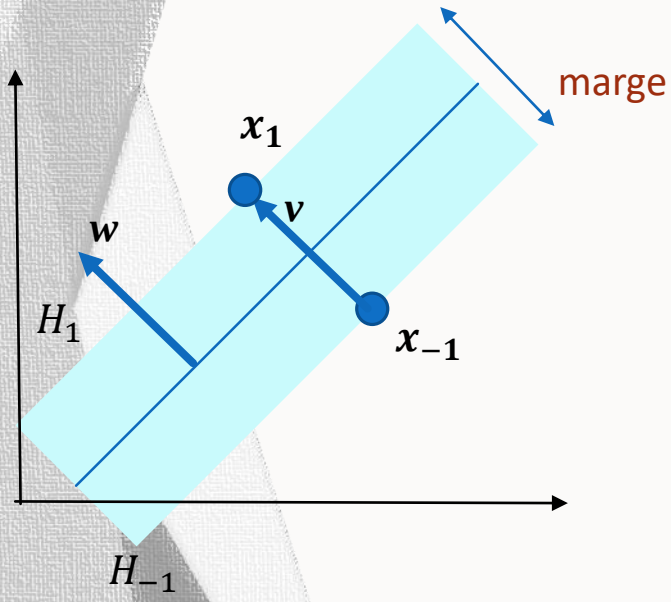


SVM linéaire

Considérons les points $\boldsymbol{x_{-1}}$ et $\boldsymbol{x_1}$, placés sur les hyperplans H_{-1} et H_1 tq

$\boldsymbol{w^T x_1} + b = 1$ et $\boldsymbol{w^T x_{-1}} + b = -1$

$\boldsymbol{w}$ est perpendiculaire à l'hyperplan. Soit $\boldsymbol{v}$ le vecteur, colinéaire à $\boldsymbol{w}$, de norme m : $\boldsymbol{v} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} m$



On a alors $\boldsymbol{x_1} = \boldsymbol{x_{-1}} + \boldsymbol{v}$ avec $\boldsymbol{v} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} m$

$$\boldsymbol{w^T x_1} = \boldsymbol{w^T x_{-1}} + \frac{\boldsymbol{w^T w}}{\|\boldsymbol{w}\|} m$$

$$\boldsymbol{w^T x_1} + \boldsymbol{b} = \boldsymbol{w^T x_{-1}} + \boldsymbol{b} + \frac{\boldsymbol{w^T w}}{\|\boldsymbol{w}\|} m$$

$$1 = -1 + \|\boldsymbol{w}\| m$$

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

Et donc la marge est :

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

Trouver l'hyperplan optimal revient donc au problème d'optimisation :

$$\max_{w,b} \frac{2}{\|w\|}$$

Sous contraintes : $\forall i = 1, N \quad y_i \left(\boldsymbol{w^T}.\boldsymbol{x_i} + b\right) \geq 1$

Ou encore,

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2 = \min_{w,b} \frac{1}{2}\boldsymbol{w^T}\,\boldsymbol{w} \quad sc \quad \forall i = 1, N \quad y_i \left(\boldsymbol{w^T}.\boldsymbol{x_i} + b\right) \geq 1$$

Ce problème d'optimisation sous contrainte est un problème quadratique de la forme :

$$\min_{\boldsymbol{z}} \tfrac{1}{2}\boldsymbol{z^T P z} + \boldsymbol{q^T z} \qquad \text{sc} \qquad \boldsymbol{Gz} \leq \mathbf{h}$$

Il existe de nombreux solvers pour le résoudre.

➔**On arrive donc aux valeurs (*w,b*) de l'hyperplan optimal.**

Pour classer un nouvel exemple $\boldsymbol{x}$, on utilise la **fonction de décision** :

$$\boldsymbol{f(x)} = \boldsymbol{w^T}.\boldsymbol{x} + b \quad \begin{cases} > 0 \Rightarrow y = 1 \\ < 0 \Rightarrow y = -1 \end{cases}$$

Résoudre

$$\min_{w,b} \frac{1}{2} \boldsymbol{w^T w}$$
$$sc \quad \forall i = 1, N \quad y_i \left(\boldsymbol{w^T}.\boldsymbol{x_i} + b \right) \geq 1$$

Équivaut à résoudre

$$\min_{\boldsymbol{z}} \frac{1}{2} \boldsymbol{z^T P z} + \boldsymbol{q^T z}$$
$$sc \; \boldsymbol{Gz} \leq \mathbf{h}$$

Trouver l'expression analytique de $\boldsymbol{z}$, $\boldsymbol{P}$, $\boldsymbol{G}$, $\boldsymbol{q}$ et $\mathbf{h}$

Résoudre

$$\min_{w,b} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}$$
$$sc \quad \forall i = 1, N \quad y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b \right) \geq 1$$

Équivaut à résoudre

$$\min_{\boldsymbol{z}} \frac{1}{2} \boldsymbol{z}^T \boldsymbol{P} \boldsymbol{z} + \boldsymbol{q}^T \boldsymbol{z}$$
$$sc \; \boldsymbol{G}\boldsymbol{z} \leq \mathbf{h}$$

Trouver l'expression analytique de $\boldsymbol{z}$, $\boldsymbol{P}$, $\boldsymbol{G}$, $\boldsymbol{q}$ et $\mathbf{h}$

$$\boldsymbol{z} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^{n+1}, \; \mathbf{q} = \mathbf{0}_{n+1,1}, \; \mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}_{1,n} \\ \mathbf{0}_{n,1} & \mathbf{I}_{n,n} \end{bmatrix}, \; \mathbf{G} = - \begin{bmatrix} y_1, y_1 \boldsymbol{x}_1^T \\ \vdots \\ y_N, y_N \boldsymbol{x}_N^T \end{bmatrix} \text{ et } \mathbf{h} = -\mathbf{1}_{N,1}$$

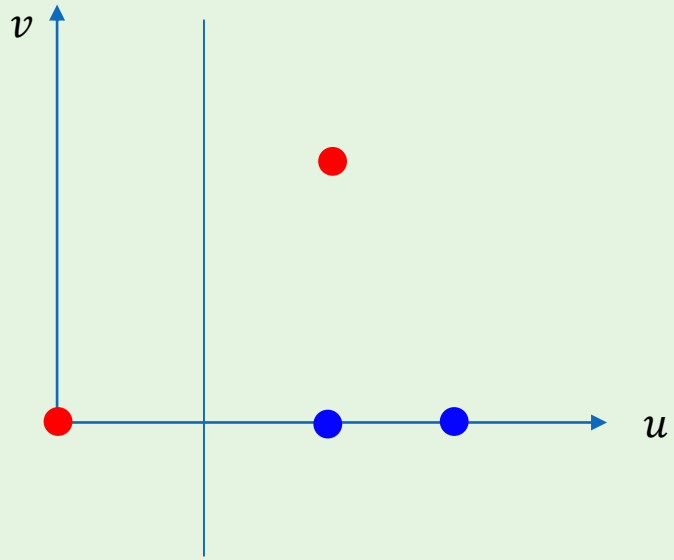
SVM linéaire

Considérons les points :

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \text{ et } y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

1. Représentez ces points, sont-ils linéairement séparable ?
2. Représenter l'hyperplan de paramètre $\boldsymbol{w}^T = [1,0]$ et $b = -1$. Est-il un bon classifieur pour ces données ?
3. Exprimer les contraintes de séparabilité pour chaque exemple
4. Calculer les paramètres **(*w*,*b*)** de l'hyperplan optimal. Que vaut la marge ? Quels sont les vecteurs supports ?
5. Dans le cas où on souhaiterait résoudre le problème des SVM par un solver de problème quadratique, déterminer les matrices P, G, q, h à passer au solver et la matrice *z* qu'il renverrait
6. Classer les points $\begin{bmatrix} 0 \\ 1.5 \end{bmatrix}$ et $\begin{bmatrix} 0 \\ 3.5 \end{bmatrix}$ en déterminant les valeurs de f(***x***).

SVM linéaire



1. Représentez ces points, sont-ils linéairement séparable ?

   Les points sont linéairement séparables

2. Représenter l'hyperplan de paramètre $\boldsymbol{w}^T = [1,0]$ et $b = -1$. Est-il un bon classifieur pour ces données ?

   L'hyperplan a pour équation $1.u + 0.v - 1 = 0$, soit $u = 1$. Ce n'est pas un hyperplan séparateur

3. Exprimer les contraintes de séparabilité pour chaque exemple

   Les contraintes de séparabilité sont

$$\forall i = 1, N \quad y_i \left(\boldsymbol{w^T}.\boldsymbol{x_i} + b\right) \geq 1$$

   Et donc

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \text{ et } y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{aligned} -b &\geq 1 && (1) \\ -2w_1 - 2w_2 - b &\geq 1 && (2) \\ 2w_1 + b &\geq 1 && (3) \\ 3w_1 + b &\geq 1 && (4) \end{aligned}$$

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \text{ et } y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \qquad \begin{aligned} -b &\geq 1 && (1) \\ -2w_1 - 2w_2 - b &\geq 1 && (2) \\ 2w_1 + b &\geq 1 && (3) \\ 3w_1 + b &\geq 1 && (4) \end{aligned}$$

4. Calculer les paramètres **(*w*,*b*)** de l'hyperplan optimal. Que vaut la marge ? Quels sont les vecteurs supports ?

En combinant (1) et (3), on a : $w_1 \geq 1$

En combinant (2) et (3), on a : $w_2 \leq -1$

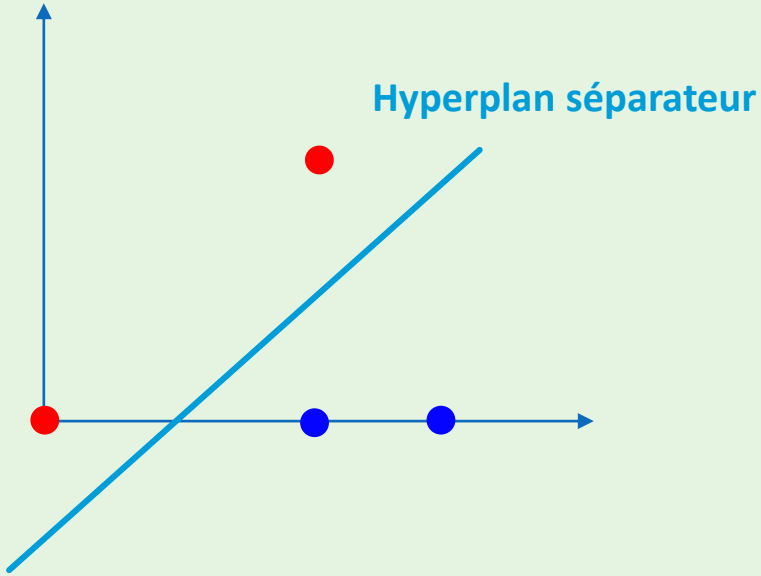
Or, on recherche $\min_{w,b} \frac{1}{2} \boldsymbol{w^T w} = \min_{w,b} \frac{1}{2}(w_1^2 + w_2^2)$ ➔ $w_1 = 1$ et $w_2 = -1$

On obtient b en combinant (1) et (3) :

(1) -> $b \leq -1$

(3) -> b ≥ -1

L'hyperplan est donc
$\boldsymbol{w}^T = [1, -1]$ et $b = -1$



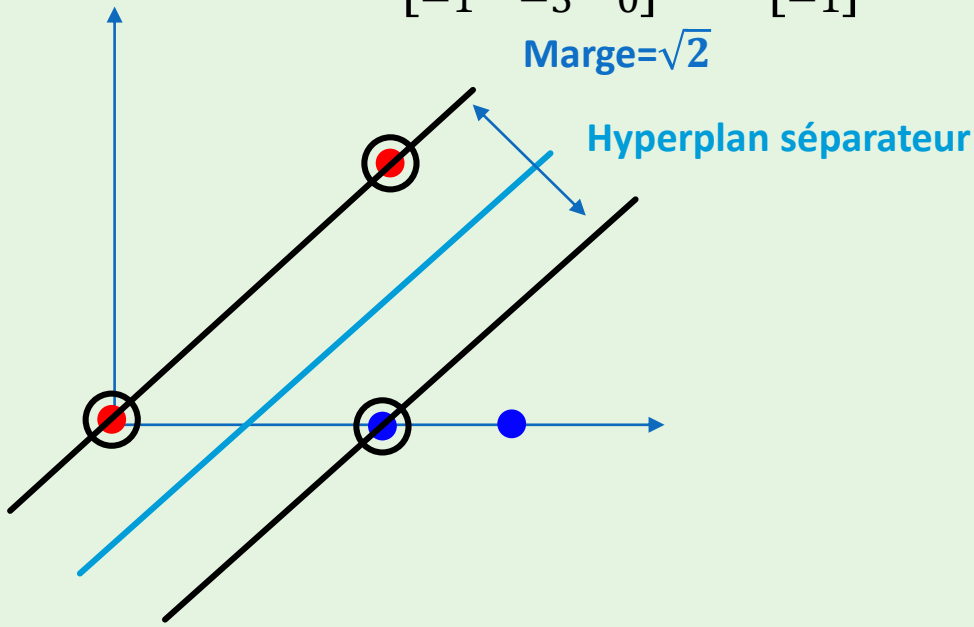
La marge vaut $\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{2}} = \sqrt{2}$

Les vecteurs supports sont tq $y_i(\boldsymbol{w}^T.\boldsymbol{x_i} + b) = 1$ soit les points 1, 2 et 3

5. Dans le cas où on souhaiterait résoudre le problème des SVM par un solver de problème quadratique, déterminer les matrices P, G, q, h à passer au solver et la matrice *z* qu'il renverrait

Pour la résolution avec un solveur de problème quadratique, on aurait

$$q = \begin{bmatrix}0\\0\\0\end{bmatrix}, P = \begin{bmatrix}0 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}, G = \begin{bmatrix}1 & 0 & 0\\1 & 2 & 2\\-1 & -2 & 0\\-1 & -3 & 0\end{bmatrix}, h = \begin{bmatrix}-1\\-1\\-1\\-1\end{bmatrix} \text{ et } z = \begin{bmatrix}-1\\1\\-1\end{bmatrix}$$

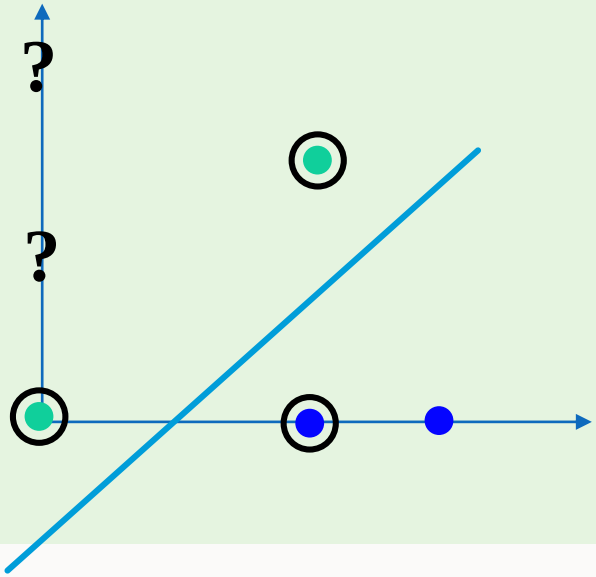


6. Classer les points $\begin{bmatrix} 0 \\ 1.5 \end{bmatrix}$ et $\begin{bmatrix} 0 \\ 3.5 \end{bmatrix}$ en déterminant les valeurs de f($\boldsymbol{x}$).

$\boldsymbol{f(x_1) = w^T . x_1 + b = -2.5}$

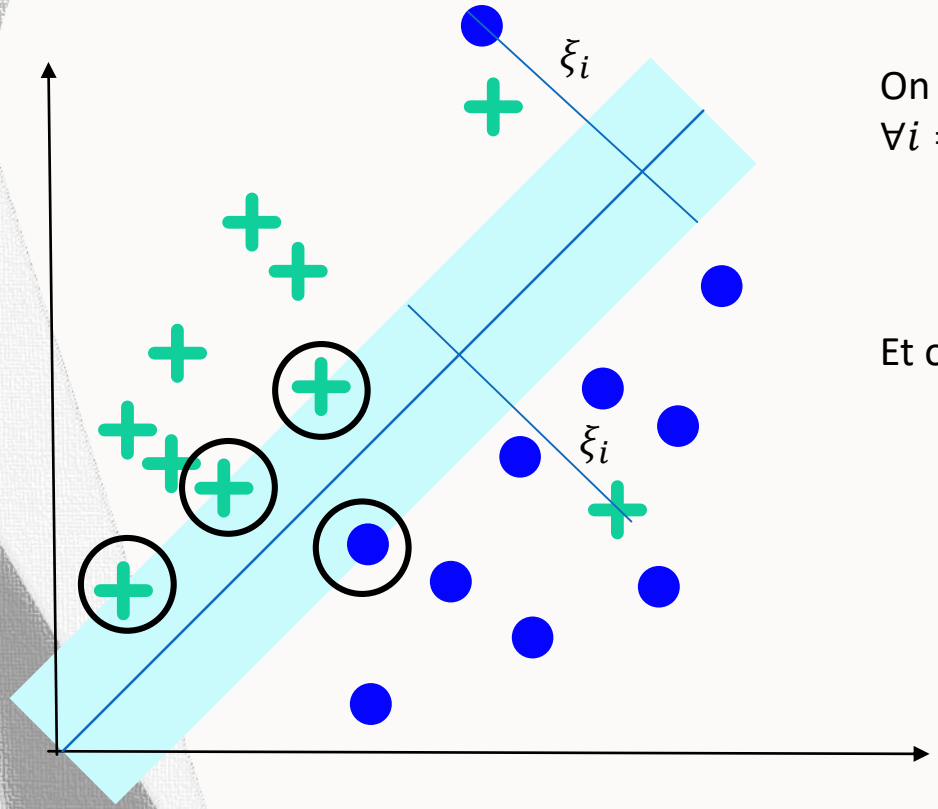
$\boldsymbol{f(x_2) = w^T . x_2 + b = -4.5}$

Les deux points appartiennent donc à la classe -1



SVM à marge souple

Il se peut que les données ne soient pas linéairement séparables à cause de quelques exemples



On ajoute des variables d'ajustement ξ_i tq :
$\forall i = 1, N$

$$y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b \right) \geq 1 - \xi_i$$

et $\xi_i \geq 0$

Et on souhaite que $\sum_{i=1}^{N} \xi_i$ soit minimal

SVM à marge souple

D'où la reformulation du problème :

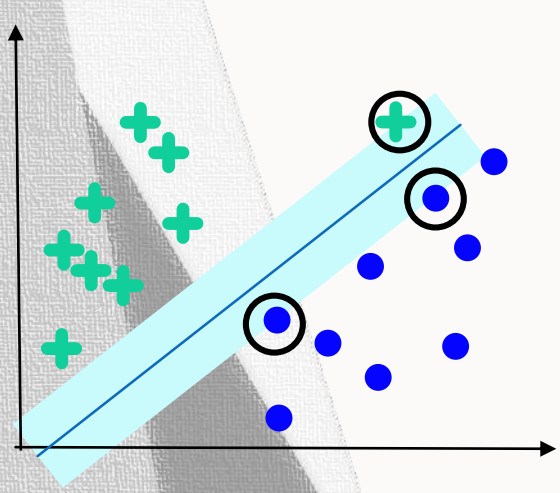
$$\min_{w,b,\xi_i} \left(\frac{1}{2} \boldsymbol{w^T w} + \boldsymbol{C} \sum_{i=1}^{N} \xi_i \right)$$

$sc\ \forall i = 1, N$

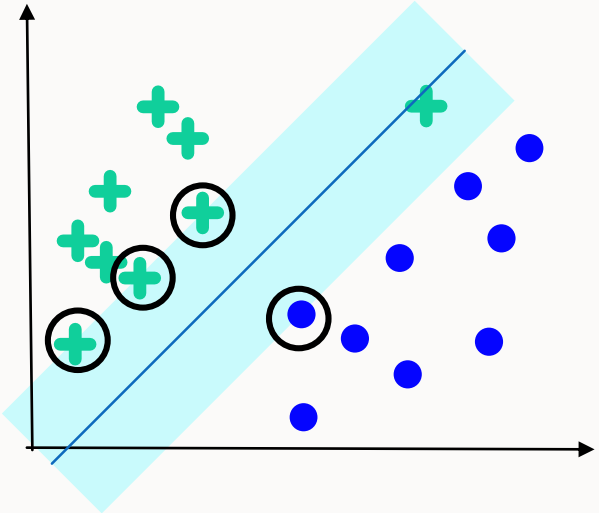
$y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b \right) \geq 1 - \xi_i$

et $\xi_i \geq 0$

- C règle le compromis entre une grande marge et le respect des contraintes
- Si C est grand, la marge sera petite pour que le maximum de point soit bien classé
- Si C est petit, la marge sera plus grande car on est plus laxiste sur les données mal classées
- C est à optimiser sur une base de validation



C grand



C petit

SVM à marge souple

D'où la reformulation du problème :

$$\min_{w,b,\xi_i} \left(\frac{1}{2} \boldsymbol{w^T w} + \boldsymbol{C} \sum_{i=1}^{N} \xi_i \right)$$

$sc\ \forall i = 1, N$

$y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b \right) \geq 1 - \xi_i$

et $\xi_i \geq 0$

De la même manière, le problème se met sous une forme quadratique :

$$\min_{\boldsymbol{z}} \frac{1}{2} \boldsymbol{z^T P z} + \boldsymbol{q^T z}$$

$$sc\ \boldsymbol{Gz} \leq \mathbf{h}$$

Exercice :

Déterminer les expressions analytiques de $\boldsymbol{z}$, $\boldsymbol{P}$, $\boldsymbol{G}$, $\boldsymbol{q}$ et $\mathbf{h}$

D'où la reformulation du problème :

$$\min_{w,b,\xi_i} \left(\frac{1}{2} \boldsymbol{w^T w} + \boldsymbol{C} \sum_{i=1}^{N} \xi_i \right)$$

$$sc \quad \forall i = 1, N \quad y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b \right) \geq 1 - \xi_i \;\; et \xi_i \geq 0$$

De la même manière, le problème se met sous une forme quadratique :

$\min_z \frac{1}{2} \boldsymbol{z^T P z} + \boldsymbol{q^T z} \;\; sc \; \boldsymbol{G z} \leq \mathbf{h}$

Et, en identifiant,

$$\boldsymbol{z} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_n \\ \xi_1 \\ \vdots \\ \xi_N \end{bmatrix}, \mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}_{1,n} & \mathbf{0}_{1,N} \\ \mathbf{0}_{n,1} & \mathbf{I}_{n,n} & \mathbf{0}_{n,N} \\ \mathbf{0}_{N,1} & \mathbf{0}_{N,n} & \mathbf{0}_{N,N} \end{bmatrix}, \mathbf{q} = \begin{bmatrix} \mathbf{0}_{n+1,1} \\ C.\mathbf{1}_{N,1} \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} -y_1, -y_1 \mathbf{x}_1^T & \\ \vdots & -\mathbf{I}_{N,N} \\ -y_N, -y_N \mathbf{x}_N^T & \\ \mathbf{0}_{N,n+1} & -\mathbf{I}_{N,N} \end{bmatrix}, \mathbf{h} = \begin{bmatrix} -\mathbf{1}_{N,1} \\ \mathbf{0}_{N,1} \end{bmatrix}$$

SVM à marge souple

En appliquant ces équations au problème précédant avec C=1, on trouve :

$$z = \begin{bmatrix} -1.0000 \\ 0.9999 \\ -0.9999 \\ 0.0000 \\ 0.0000 \\ 0.0003 \\ 0.0003 \end{bmatrix}$$

Que peut-on conclure ?

SVM à marge souple

En appliquant ces équations au problème précédant avec C=1, on trouve :

$$z = \begin{bmatrix} -1.0000 \\ 0.9999 \\ -0.9999 \\ 0.0000 \\ 0.0000 \\ 0.0003 \\ 0.0003 \end{bmatrix}$$

Que peut-on conclure ?

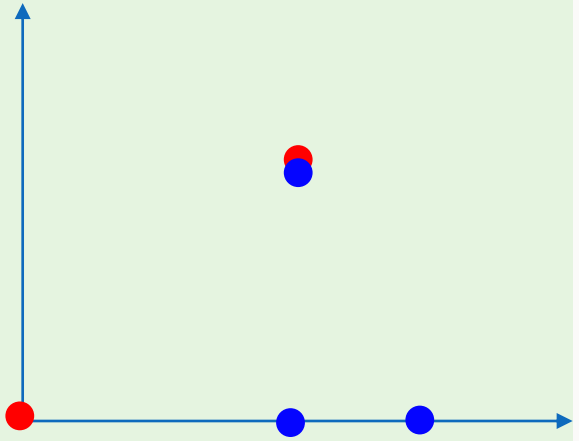
On retrouve presque le même hyperplan :$\boldsymbol{w}^T = [1, -1]$ et $b = -1$

Tous les ξ_i sont très petits ➔ les exemples respectent bien la contrainte

SVM à marge souple

On ajoute maintenant un point aux données :

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \\ 2 & 1.9 \end{bmatrix} \text{ et } y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$



En faisant tourner le programme, on obtient :

Avec C=1, $z = \begin{bmatrix} -1.0000 \\ 1 \\ -0.1 \\ 0.0000 \\ 1.8000 \\ 0.0000 \\ 0.1900 \end{bmatrix}$ et avec C=1000, $z = \begin{bmatrix} -1.0000 \\ 20 \\ -20 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{bmatrix}$

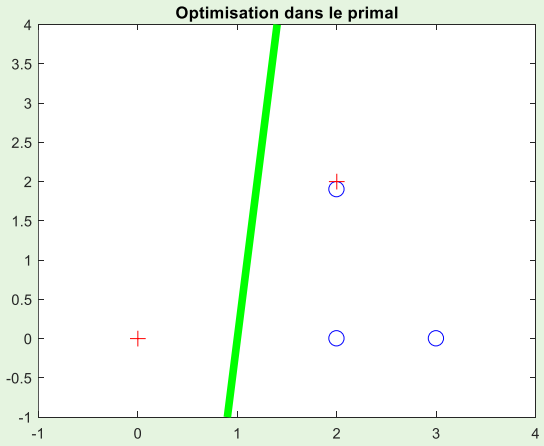
Commentez

SVM à marge souple

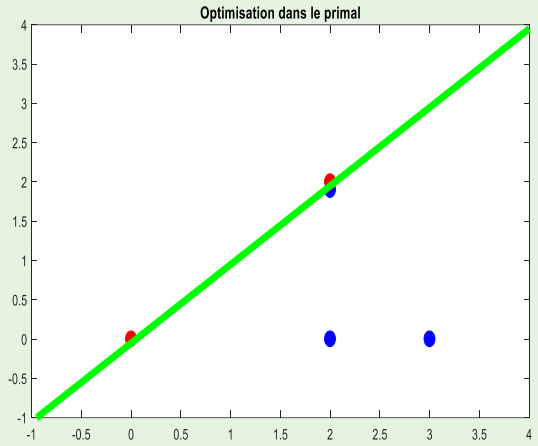
C=1 ➔ On donne moins de poids au respect des contraintes ➔ La marge est plus grande mais le second point viole la contrainte

C=1000 ➔ on donne un très fort poids au respect des contraintes ➔ Elles sont respectées mais la marge est très petite

C=1



C=1000



Formulation duale

Nous avons vu que l'expression primale des SVM est :

$$\min_{w,b} \frac{1}{2} \boldsymbol{w^T w}$$

$$sc \quad \forall i = 1, N \quad y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) \geq 1$$

On peut résoudre ce problème en utilisant le lagrangien :

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) \begin{cases} = \frac{1}{2} \boldsymbol{w^T w} - \sum_{i=1}^{N} \alpha_i \left[y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) - 1 \right] \\ = \frac{1}{2} \boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) \right] \end{cases}$$

Les α_i sont appelés **multiplicateurs de Lagrange et** $\forall i, \alpha_i \geq 0$.

Formulation duale

On recherche :

$$\min_{w,b} \frac{1}{2}\boldsymbol{w^T w} + \max_{\alpha_i} \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right]$$

Pourquoi ?

Pour des données linéairement séparables :

- Si l'hyperplan est tq la **contrainte n'est pas respectée en $\boldsymbol{x_i}$**
  - ➔ $\left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right] > 0$
  - ➔Le α_i correspondant devient très grand pour maximiser $\alpha_i\left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right]$
  - ➔ $\boldsymbol{w}$, b vont évoluer de manière à minimiser le Lagrangien et vérifier la contrainte

- Si la contrainte est **strictement respectée** en $\boldsymbol{x_i}$
  - ➔ $\left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right] < 0$
  - ➔$\alpha_i = 0$ pour maximiser $\alpha_i\left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right]$
- Si $\boldsymbol{x_i}$ est un vecteur support
  - ➔ $1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right) = \boldsymbol{0}$
  - ➔ $\alpha_i > 0$

Pour conclure, pour un cas linéairement séparable,
$\alpha_i = 0$ ➔ L'exemple $\boldsymbol{x_i}$ vérifie la contrainte
$\alpha_i > 0$ ➔ vecteur support

Formulation duale

D'autre part, résoudre:

$$\min_{w,b} \frac{1}{2} \boldsymbol{w^T w} + \max_{\alpha_i} \sum_{i=1}^{N} \alpha_i \left[1 - y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right)\right]$$

Revient à résoudre

$$\max_{\alpha_i} \min_{w,b} \underbrace{\frac{1}{2} \boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i \left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right)\right]}$$

$$\max_{\alpha_i} \min_{w,b} \; L(\mathbf{w}, b, \boldsymbol{\alpha})$$

Formulation duale

$$\max_{\alpha_i} \min_{w,b} \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right)\right]$$

En dérivant $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ par rapport à $\boldsymbol{w}$ et b,

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0$ ➔ ?

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{b}} = 0$ ➔ ?

Rappel sur les matrices

$(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T$

$(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a} = \mathbf{b}$

$\frac{\partial a^T \boldsymbol{B} a}{\partial a} = 2\mathbf{B}a$

Formulation duale

$$\max_{\alpha_i} \min_{w,b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right]$$

En dérivant $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ par rapport à $\boldsymbol{w}$ et b,

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x_i}$ **(1)**

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{b}} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i\, y_i = \mathbf{0}$ **(2)**

Rappel sur les matrices

$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

$(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a} = \mathrm{b}$

$\frac{\partial a^T \boldsymbol{B} a}{\partial a} = 2\mathbf{B}\mathrm{a}$

Formulation duale

Rappel sur les matrices
$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$
$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a} = b$
$\frac{\partial a^T \boldsymbol{B} a}{\partial a} = 2\boldsymbol{B}a$

En dérivant $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ par rapport à $\boldsymbol{w}$ et b,

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x_i}$ **(1)**

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{b}} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i\, y_i = \boldsymbol{0}$ **(2)**

Or,

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T.\boldsymbol{x_i} + b\right)\right]$$

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{w}^T.\boldsymbol{x_i} - \sum_{i=1}^{N} \alpha_i y_i\, b$$

On remplace $\boldsymbol{w}$ par sa valeur (eq 1)

On remplace $\boldsymbol{w}$ par sa valeur (eq 1)

=0 (eq 2)

Formulation duale

Rappel sur les matrices
$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$
$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a} = b$
$\frac{\partial a^T \boldsymbol{B} a}{\partial a} = 2\boldsymbol{B}a$

En dérivant $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ par rapport à $\boldsymbol{w}$ et b,

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x_i}$ **(1)**

$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{b}} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i\, y_i = \boldsymbol{0}$ **(2)**

Or, $L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i \left[1 - y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right)\right]$

$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{w}^T . \boldsymbol{x_i} - \sum_{i=1}^{N} \alpha_i y_i\, b$

Soit, en introduisant (1) et (2) dans le lagrangien,

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x_i^T}\, \boldsymbol{x_j} + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x_i^T}\, \boldsymbol{x_j}$$

$$\mathrm{L}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x_i^T}\, \boldsymbol{x_j} + \sum_{i=1}^{N} \alpha_i$$

On souhaite maximiser le lagrangien par rapport à $\boldsymbol{\alpha}$,

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j\, y_i y_j \boldsymbol{x_i^T}\, \boldsymbol{x_j} \quad sc \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ (eq 2) et } \alpha_i \geq 0$$

Formulation duale

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j \, y_i y_j \boldsymbol{x_i^T x_j} \qquad sc \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0$$

revient à

$$\min_{\alpha_i} \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j \, y_i y_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{N} \alpha_i \qquad sc \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0$$

Ce problème peut aussi être mis sous une forme quadratique.
➔ **On trouve les valeurs de $\boldsymbol{\alpha_i}$**

Mais alors, quel est l'intérêt ?

- L'astuce du noyau permet de résoudre des problèmes non linéaires

Formulation duale

Concrètement, la résolution du problème quadratique nous amène aux valeurs de α_i

Pour les problèmes linéairement séparable, on en déduit les valeurs de $\boldsymbol{w}$ avec :

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \, y_i \boldsymbol{x_i} = \sum_{\alpha_i > 0} \alpha_i \, y_i \boldsymbol{x_i}$$

Puis la valeur de b :

En se plaçant sur un des vecteurs supports (un des points tq α_i>0)

$$y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) = 1$$

Soit $$y_i y_i\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) = y_i$$

Ou $$\left(\boldsymbol{w}^T . \boldsymbol{x_i} + b\right) = y_i \;\; car \; y_i = \pm 1$$

Et donc, $$\boldsymbol{b} = y_i - \boldsymbol{w}^T . \boldsymbol{x_i}$$

$$\boldsymbol{b} = y_i - \left(\sum_{i=1}^{N} \alpha_i \, y_i \boldsymbol{x_i}\right)^T . \boldsymbol{x_i}$$

Pour classer un nouvel exemple $\boldsymbol{x}$, la fonction de décision est :

$$\boldsymbol{f(x)} = \boldsymbol{signe}\left(\boldsymbol{w^T} . \boldsymbol{x_i} + b\right)$$

Ou encore

$$\boldsymbol{f(x)} = \boldsymbol{signe}\left(\sum_{\alpha_i > 0} \alpha_i \, y_i \boldsymbol{x_i}^T \boldsymbol{x_i} + \boldsymbol{b}\right)$$

Formulation duale

Reprenons l'exercice précédant :

$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix}$ et $y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$

Le solver renvoie $\alpha = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \\ 0 \end{bmatrix}$. Quels sont les vecteurs support ? Retrouver les paramètres $\boldsymbol{w}$ et $\boldsymbol{b}$ de l'hyperplan

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \text{ et } y = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

Reprenons l'exercice précédant. Le solver renvoie $\alpha = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \\ 0 \end{bmatrix}$. Quels sont les vecteurs support ? Donner l'équation de la fonction de décision

Les vecteurs supports correspondent aux positifs. Donc les points 1, 2 et 3.
L'hyperplan séparateur est obtenu avec :
$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \, y_i \boldsymbol{x_i}$
$\boldsymbol{b} = y_i - \boldsymbol{w}^T . \boldsymbol{x_i}$ pour un des vecteurs supports

Soit, $\boldsymbol{w} = -0.5 \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.5 \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
En prenant le premier point b=-1
On retrouve bien les mêmes résultats que précédemment

Problèmes non linéaires

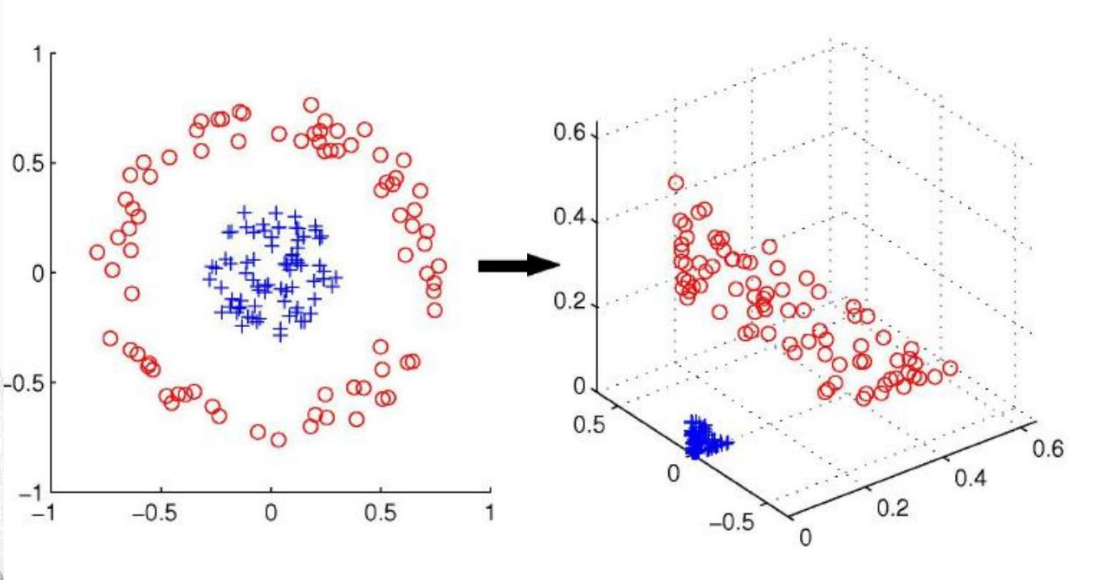
Ce problème est mis sous une forme quadratique. Mais alors, quel est l'intérêt ?

- **L'astuce du noyau permet de résoudre des problèmes non linéaires**

Problèmes non linéaires

Une solution est de transformer les données pour les rendre linéairement séparables,
Ceci demande souvent d'augmenter la dimension du problème

Figure issue de https://www.lri.fr/~antoine/Courses/Master-ISI/ISI-10/Tr-cours-SVM_2014_2x2.pdf



(x_1, x_2)

$(x_1^2, \sqrt{2}x_1x_2, x_2^2)$

Ainsi, les données $\boldsymbol{x}$ seront transformées en $\Phi(\boldsymbol{x})$.
En reprenant la formulation duale :

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j \, y_i y_j \boldsymbol{x_i^T x_j} \qquad sc \sum_{i=1}^{N} \alpha_i \alpha y_i = 0 \text{ et } \alpha_i \geq 0$$

Dans le nouvel espace,

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j \, y_i y_j \, \Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x_j}) \qquad sc \sum_{i=1}^{N} \alpha_i \alpha y_i = 0 \text{ et } \alpha_i \geq 0$$

$\Phi(\boldsymbol{x})$ peut être de grande dimension mais il n'a pas besoin d'être calculé dès lors que l'on peut calculer le produit scalaire $\Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x_j})$ de manière rapide avec une fonction noyau

Dans le cas précédant,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \; et \; \Phi(\boldsymbol{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}$$

Or,

$$\Phi(\boldsymbol{x})^T.\Phi(\boldsymbol{z}) = \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix}.\begin{bmatrix} z_1^2 \\ \sqrt{2}z_1z_2 \\ z_2^2 \end{bmatrix}$$

$$\Phi(\boldsymbol{x})^T.\Phi(\boldsymbol{z}) = x_1^2z_1^2 + 2x_1x_{2z_1z_2} + x_2^2z_2^2$$

$$\Phi(\boldsymbol{x})^T.\Phi(\boldsymbol{z}) = (x_1z_1 + x_2z_2)^2$$

$$\Phi(\boldsymbol{x})^T.\Phi(\boldsymbol{z}) = (\boldsymbol{x}^T\mathbf{z})^2$$

On a donc $\Phi(\boldsymbol{x})^T.\Phi(\boldsymbol{z}) = K(\boldsymbol{x},\boldsymbol{z}) = (\boldsymbol{x}^T\mathbf{z})^2$

➔ Le produit scalaire peut être calculé sans passer par la transformation des données

Problèmes non linéaires

Les noyaux fréquemment utilisés sont

- Les noyaux polynomiaux d'ordre p : $\boldsymbol{K(x,z) = (x^T z + 1)^p}$
- Les noyaux gaussiens : $K(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{e}^{-\frac{\|x-z\|^2}{2\sigma^2}}$

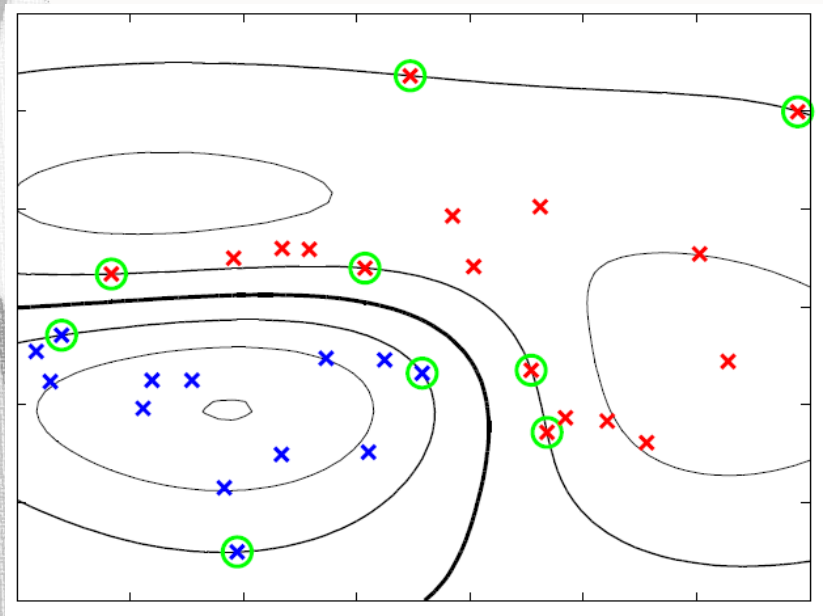


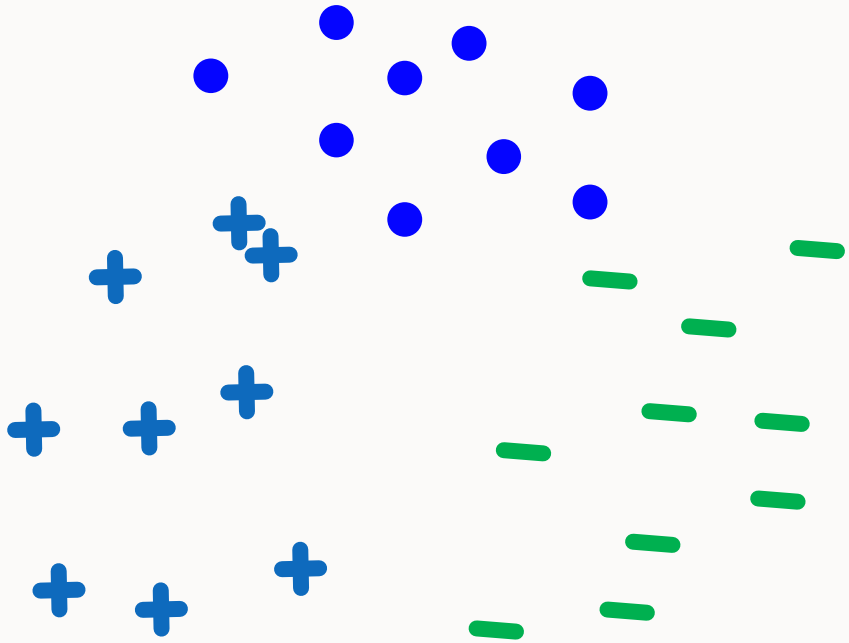
Figure issue de C. Bishop, Pattern Recognition and Machine Learning, 2006

Séparation obtenue avec un noyau gaussien
Les points verts représentent les vecteurs support

Représentation dans l'espace original

SVM multi-classes

Les SVM ont été conçus pour résoudre des problèmes binaires.
Comment faire pour les problèmes multi-classes ?

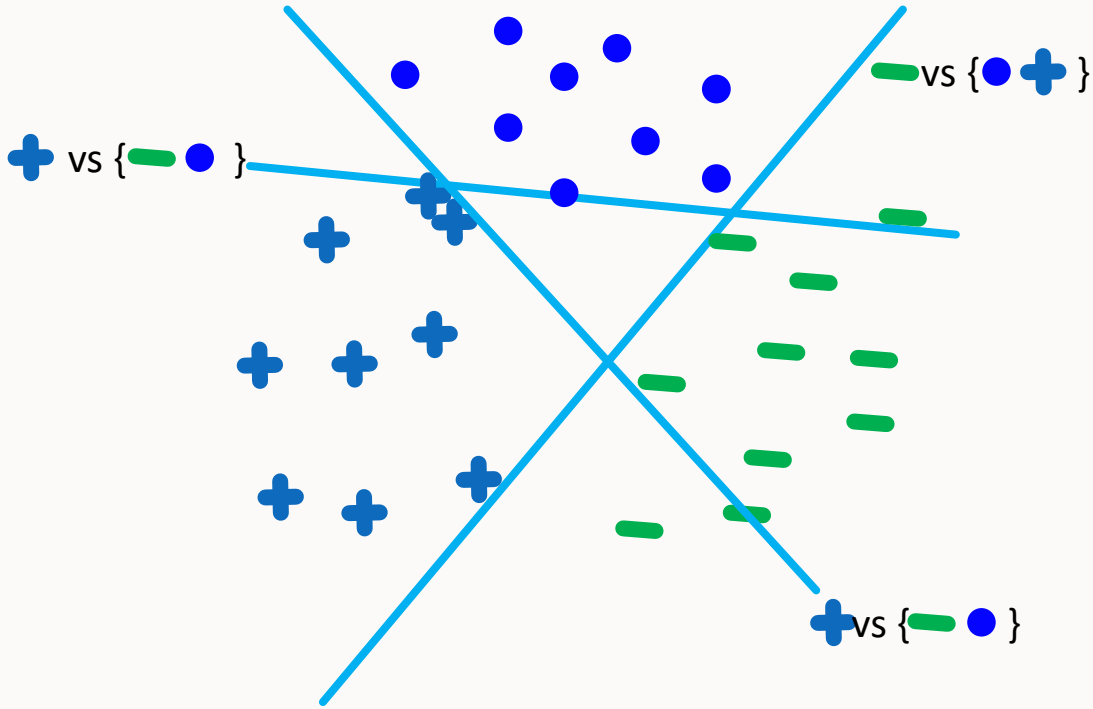


SVM multi-classes

L'approche 1 contre tous

Dans un problème à K classes, K SVM qui discriminent une classe contre les autres

Exemple



On entraine 3 SVM binaires

vs { }

vs { }

vs { }

En test, l'étiquette est donnée par le SVM avec la plus grande fonction de décision f(***x***)>0 (l'exemple classé >0 qui répond le mieux)

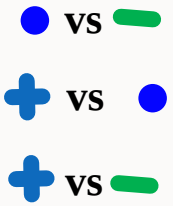
Inconvénients de l'approche 1 contre tous

- Pour chaque SVM, il y a beaucoup plus d'exemples négatifs

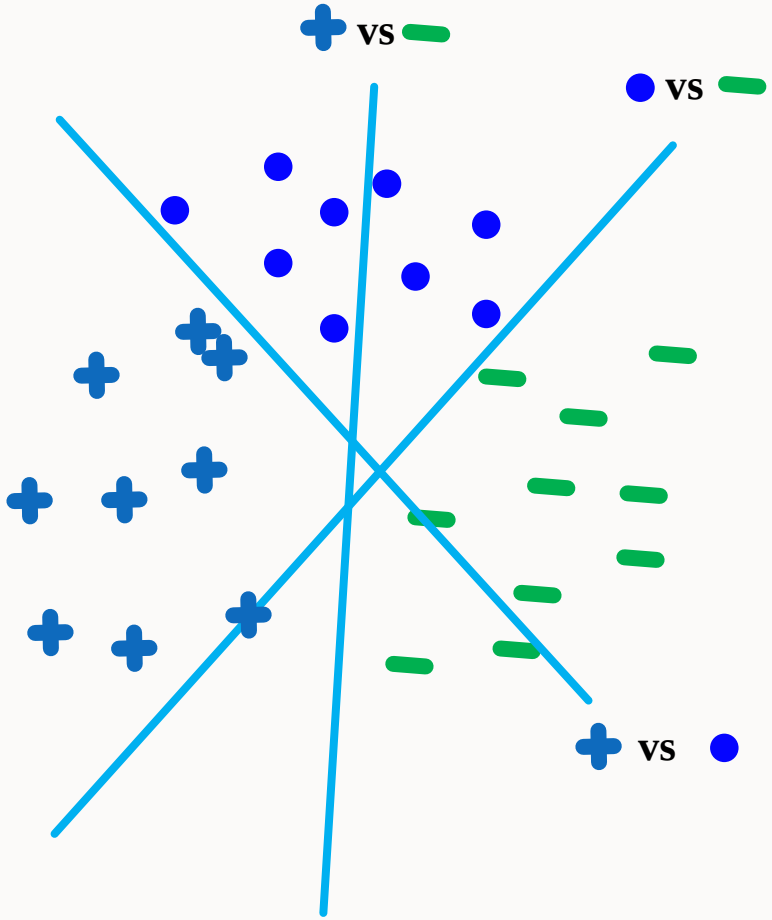
L'approche 1 contre 1

On construit un SVM pour chaque paire d'étiquettes ➔ K(K-1)/2 SVM

Exemple



La prédiction correspond au label qui a gagné le plus de « duels »



Inconvénients de l'approche 1 contre 1

- Très long : K(K-1)/2 SVM à apprendre

Avantage de l'approche 1 contre 1

- Moins de déséquilibre entre classes