Courte introduction aux SVMs

C. Hudelot

7 septembre 2022

CentraleSupélec

Modèles discriminants : SVMs

Séparateurs à Vaste Marge linéaires

But

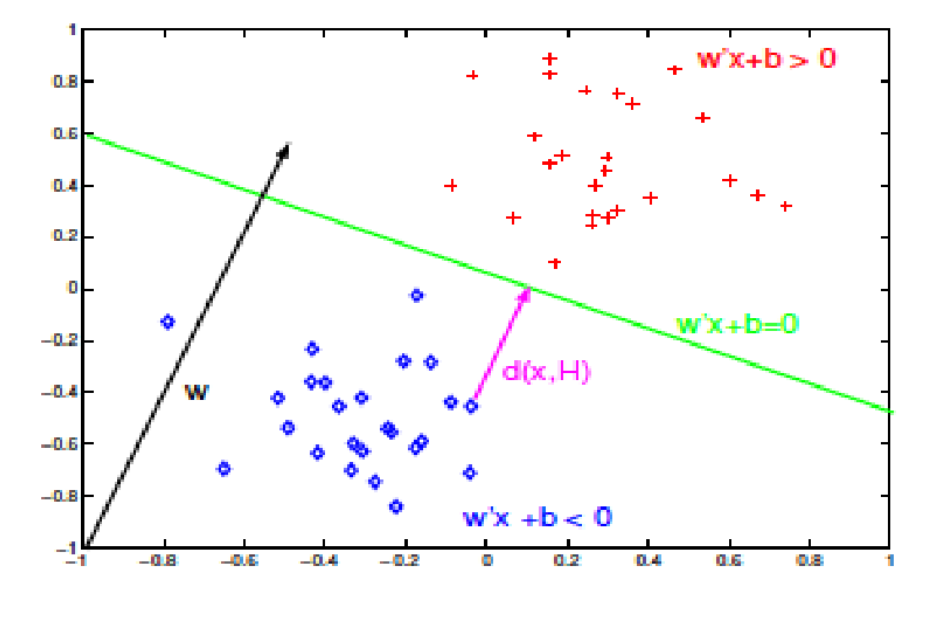
- Données d'apprentissage $D_n = \{(x_i, y_i)\}_{1 \leq i \leq n}, x_i \in \mathcal{X}, y_i \in \{-1, +1\}$, ensemble de points étiquetés.
- On cherche à construire à partir de D_n une fonction de décision $f : \mathcal{X} \to \{-1, 1\}$ ou $f : \mathcal{X} \to \mathbb{R}$ qui permet de prédire la classe -1 ou 1 d'un point $x \in \mathcal{X}$.

Fonction de décision

- $\mathcal{X} = \mathbb{R}^d$ et $x = (x^{(1)}, x^{(2)}, ..., x^{(d)})^T$
- Fonction de décision $f : \mathbb{R}^d \to \mathbb{R}$ telle que x soit affecté à la classe -1 si $f(x) < 0$ et à la classe $+1$ sinon.
- Fonction de décision linéaire :

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x^{(j)} + b = \mathbf{w}^T \mathbf{x} + b, \ \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$$

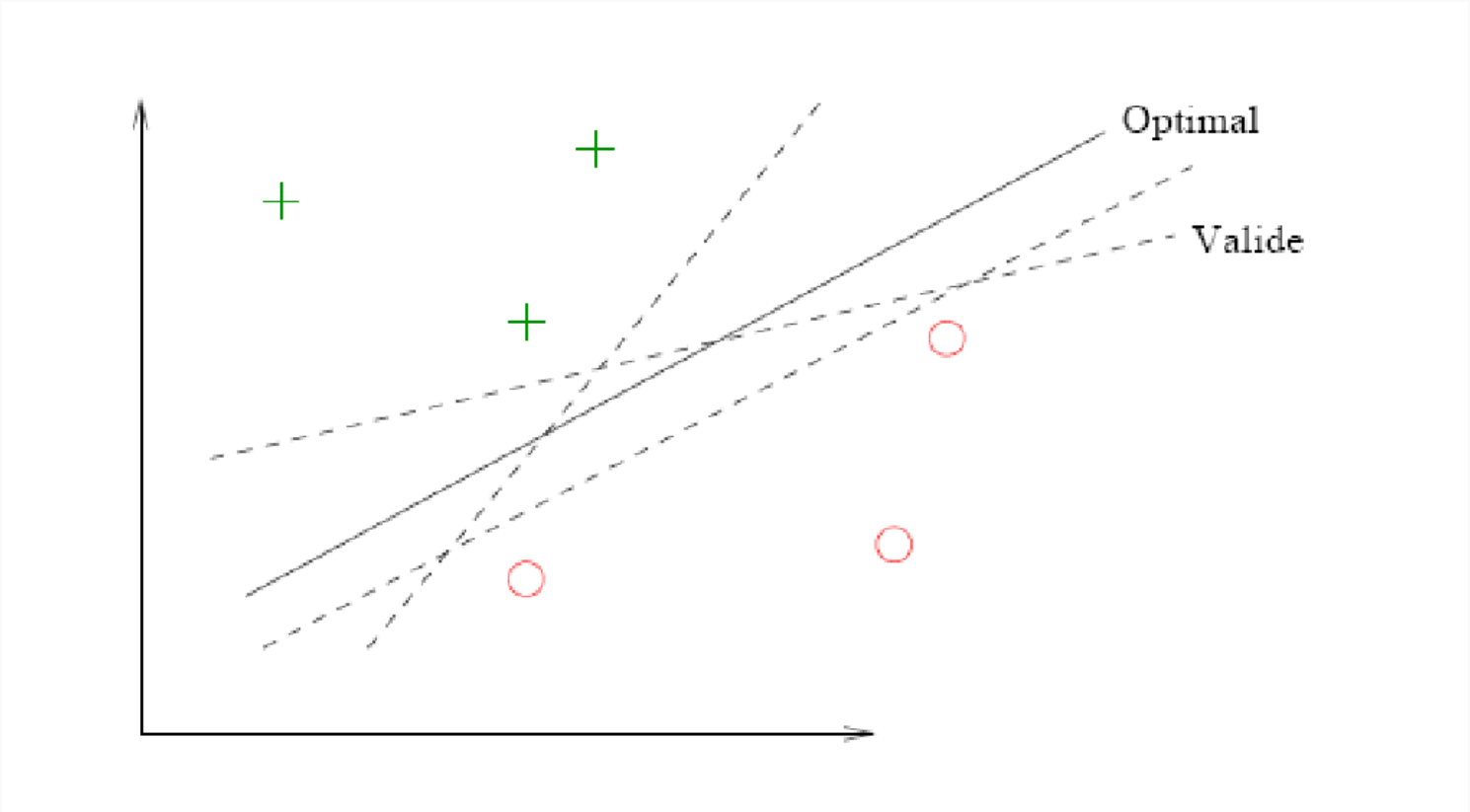
Séparateur linéaire : exemple dans $\mathbb{R}^2$



- Le plan est séparé en 2 par un hyperplan
- $f(\mathbf{x}) = \texttt{sign}(\mathbf{w}^T\mathbf{x} + b)$.
- Distance d'un point $\mathbf{x}$ à l'hyperplan de séparation est : $d(x, H) = \frac{|\mathbf{w}^T\mathbf{x}+b|}{||w||}$
- Distance de l'hyperplan à l'origine : $\frac{|b|}{||\mathbf{w}||}$

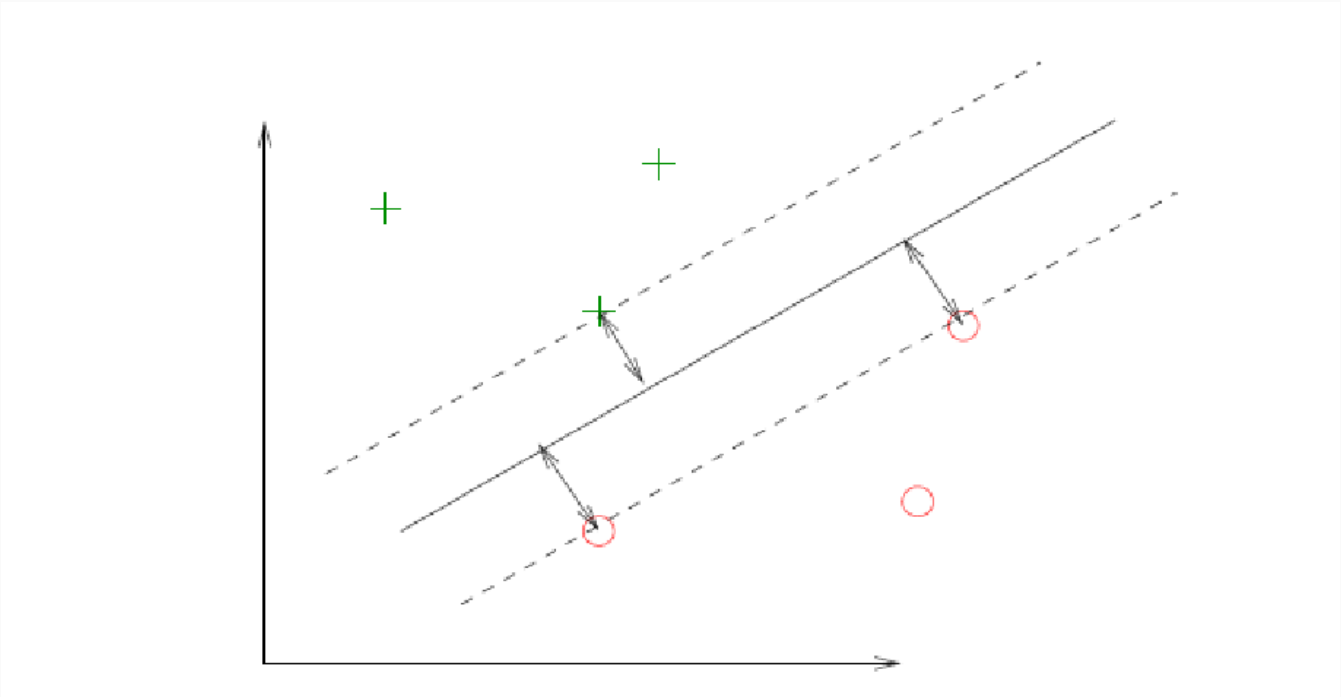
Séparateur linéaire : exemple dans $\mathbb{R}^2$

Plusieurs séparateurs peuvent être possibles



Séparateur linéaire : exemple dans $\mathbb{R}^2$

- Hyperplan qui classifie correctement les données et qui se trouve *le plus loin possible de tous les exemples.*
- Hyperplan de marge maximale ($\frac{1}{2}$ marge = distance minimale entre un exemple et la surface de séparation)



Séparateur linéaire : Définitions

Linéairement séparables
Les points $\{(x_i, y_i)\}$ sont linéairement séparables si il existe un hyperplan qui permet de discriminer correctement l'ensemble des données. Dans le cas contraire, on parle d'exemples non séparables.

Séparateur linéaire : quantification de la marge

Pour limiter l'espace des possibles, on considère que les points les plus proches sont situés sur les hyperplans canoniques donnés par :

$$\mathbf{w}^T\mathbf{x} + b = \pm 1$$

Dans ce cas, la marge est définie par

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

Les conditions d'une bonne classification sont :

$$\forall i,\ y_i f(\mathbf{x}_i) > 1$$

(chaque point est bien classé)

Formulation du problème de maximisation de la marge

Séparateur à vaste marge : formulation

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1..n}$: ensemble de points linéairement séparables.
- Objectif : trouver un hyperplan qui maximise la marge et discrimine correctement les points de $\mathcal{D}$

$$min_{\mathbf{w},b} \ \frac{1}{2}||\mathbf{w}||^2 \ \texttt{maximisation de la marge}$$

$$s.c. \ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \ \forall i = 1, \ldots, n \ \texttt{tous les points bien classés}$$

Problème de minimisation sous contraintes qui peut être résolu par des approches numériques comme la programmation quadratique (minimiser le carré de la norme) .

Résolution : passage au Lagrangien

- Un problème d'optimisation possède une forme duale si la fonction objectif et les contraintes sont strictement convexes. Alors la solution du problème dual est la solution du problème original.
- Optimisation sous contraintes : passage au Lagrangien

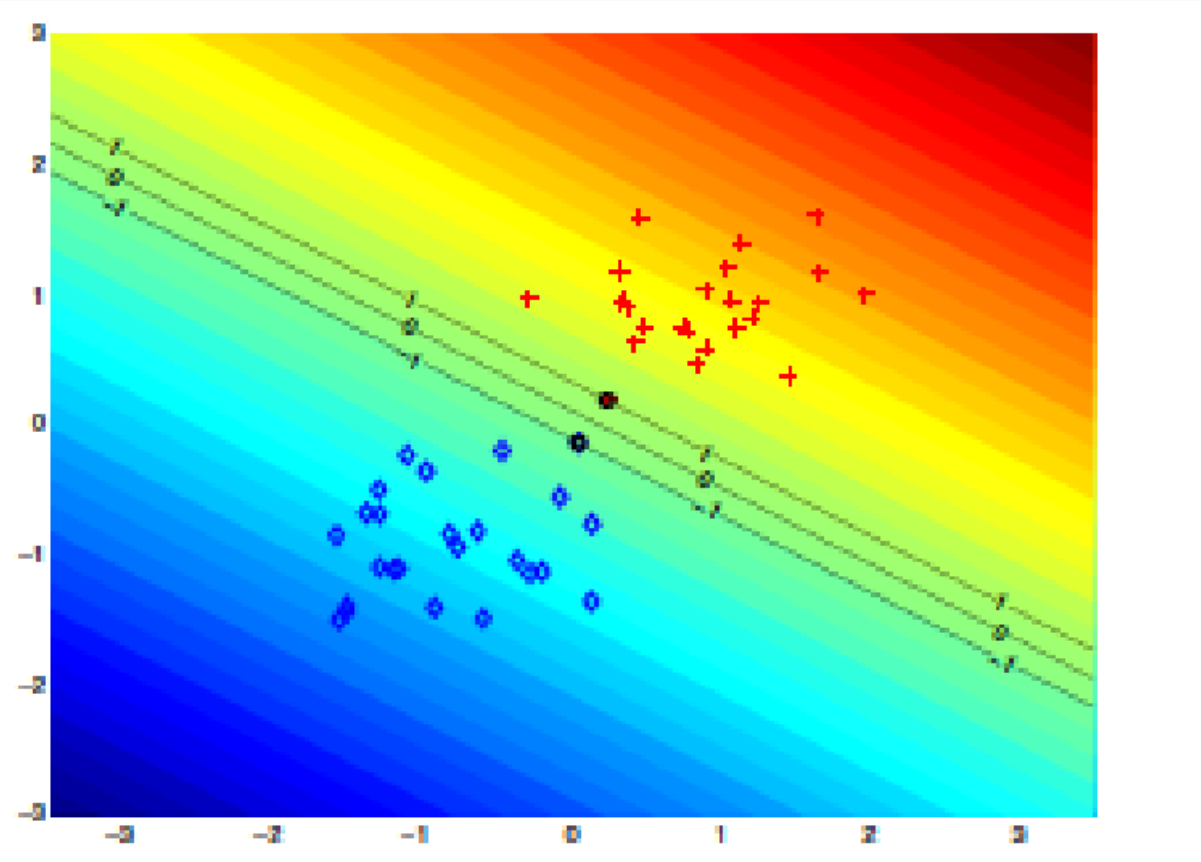
Lagrangien du problème SVM
Introduction des multiplicateurs de Lagrange $\alpha_i \geq 0$ associés aux contraintes d'inégalités, i.e. n paramètres α_i

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1)$$

Nouvelle formulation du problème où la contrainte est intégrée dans la fonction objectif

Vecteurs supports

- $\mathbf{w}$ est défini comme $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$.
- On sait que α_i est nul si $y_i(\mathbf{w}^T \mathbf{x}_i + b) > 1$ donc $\mathbf{w}$ n'est défini que par les points tels que $y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$. Ces points sont les vecteurs supports.



SVM en pratique

Calcul de w

- On utilise les données $\mathcal{D}$ pour résoudre le dual : on obtient les paramètres $\{\alpha_i, i = 1..n\}$.
- On en déduit la solution $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

Calcul de b

- Les $\alpha_i > 0$ correspondent aux points supports qui vérifient la relation

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$$

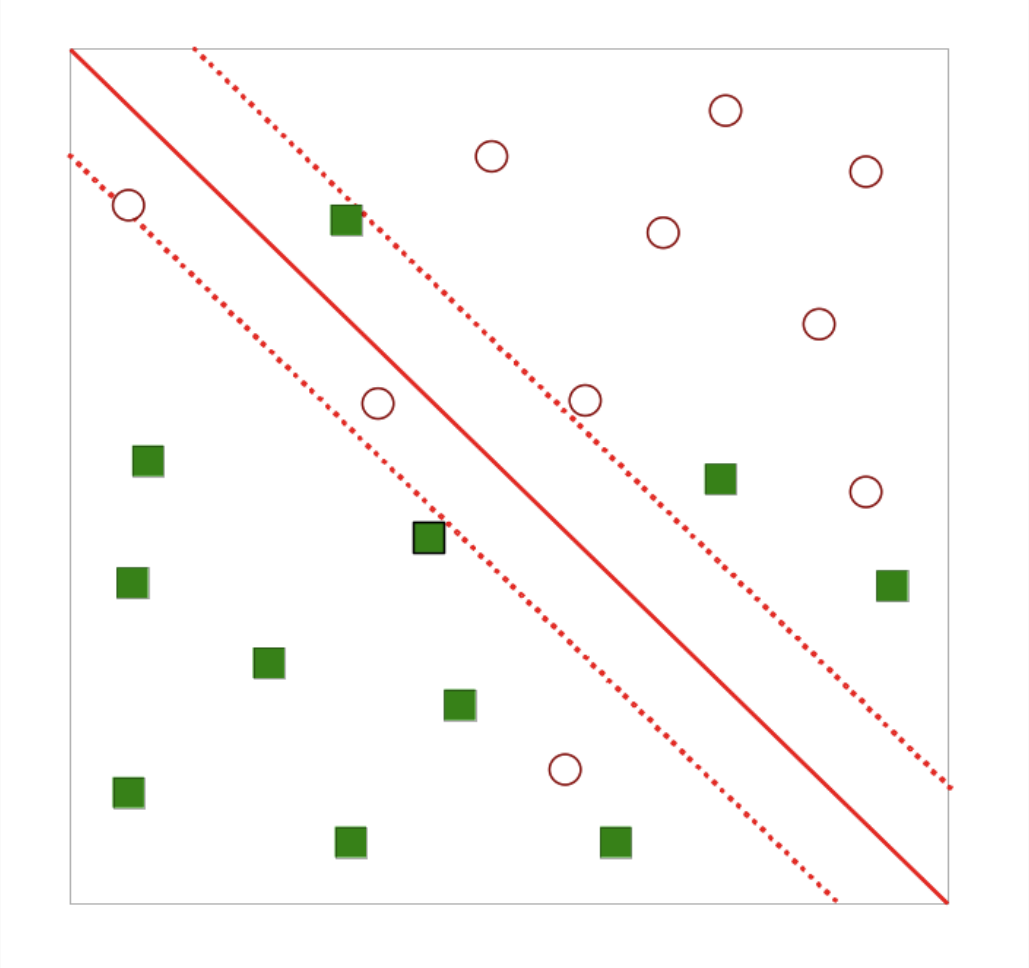
- On peut donc en déduire la valeur de b. En pratique, on fait la moyenne de ces termes pour l'ensemble SV des vecteurs supports pour obtenir une valeur numérique stable.

Fonction de décision

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

Cas non séparable

- Le problème de l'hyperplan optimal est défini dans le cas où les données sont linéairement séparables.
- Que se passe t-il quand cette hypothèse n'est pas vérifiée ?



Cas non séparable

Dans le cas non séparable, il faut :

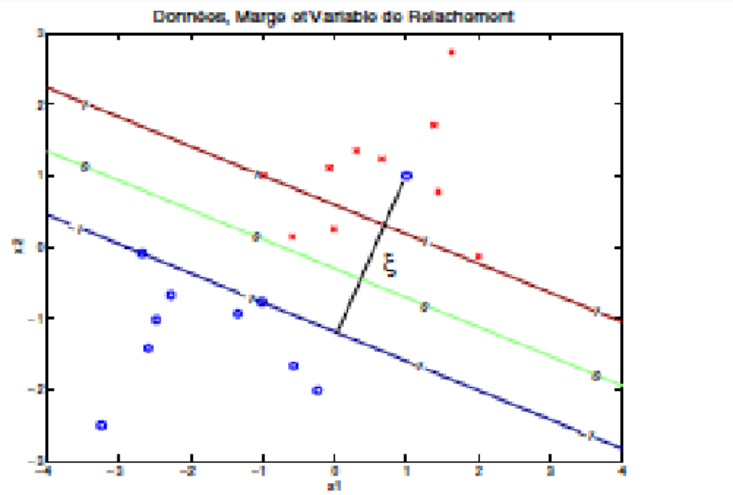
- Relacher les contraintes $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ et rajouter des variables de relachement ϵ_i dans ces contraintes.
- Pénaliser les relachements dans la fonction objectif.

Cas non séparable : formulation

SVM : cas non séparable

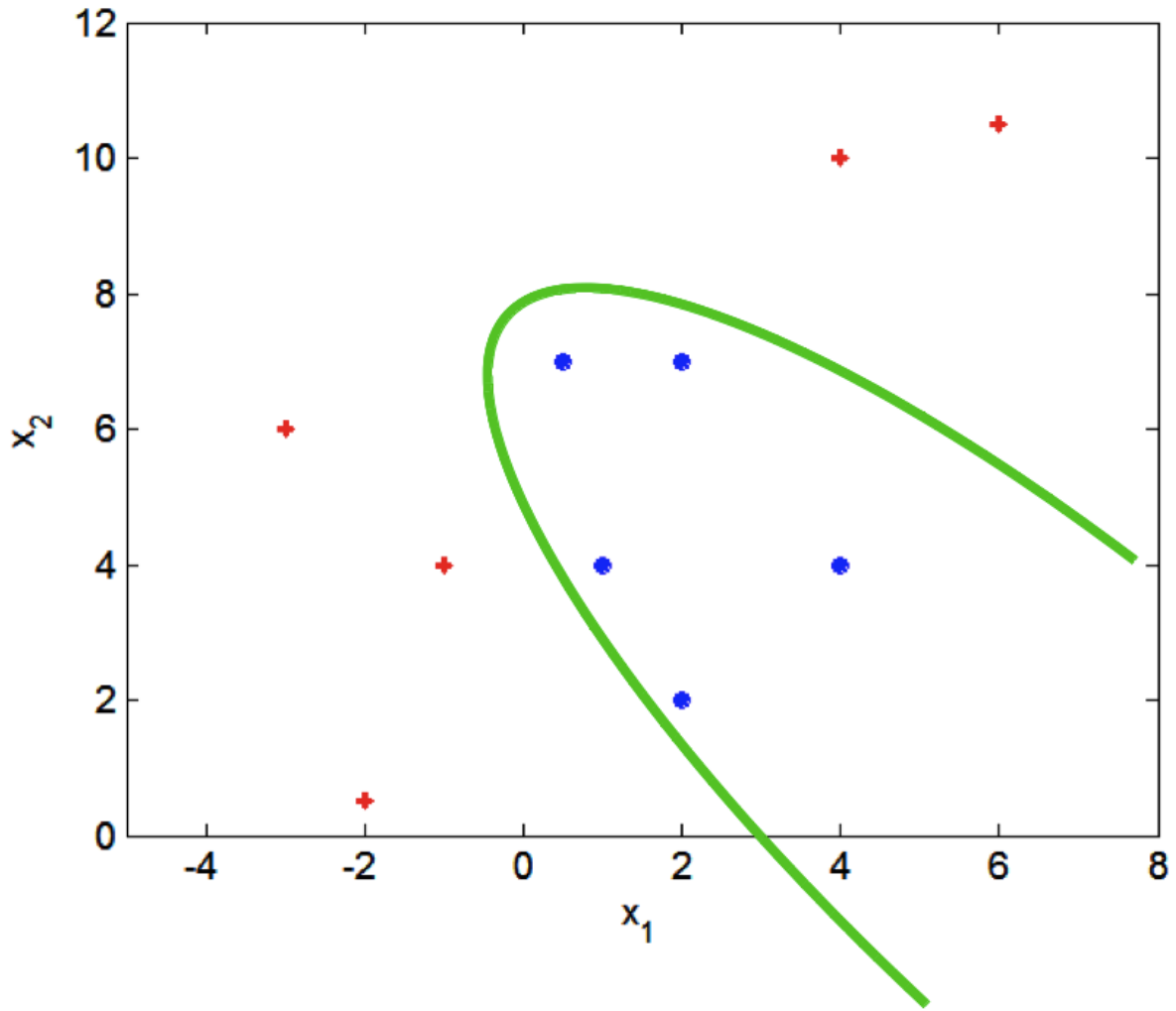
$$min_{\mathbf{w},b,\{\epsilon_i\}} \ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\epsilon_i$$

$$s.c. y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \epsilon_i, \forall i = 1..n \quad \epsilon_i \geq 0 \forall i = 1..n$$



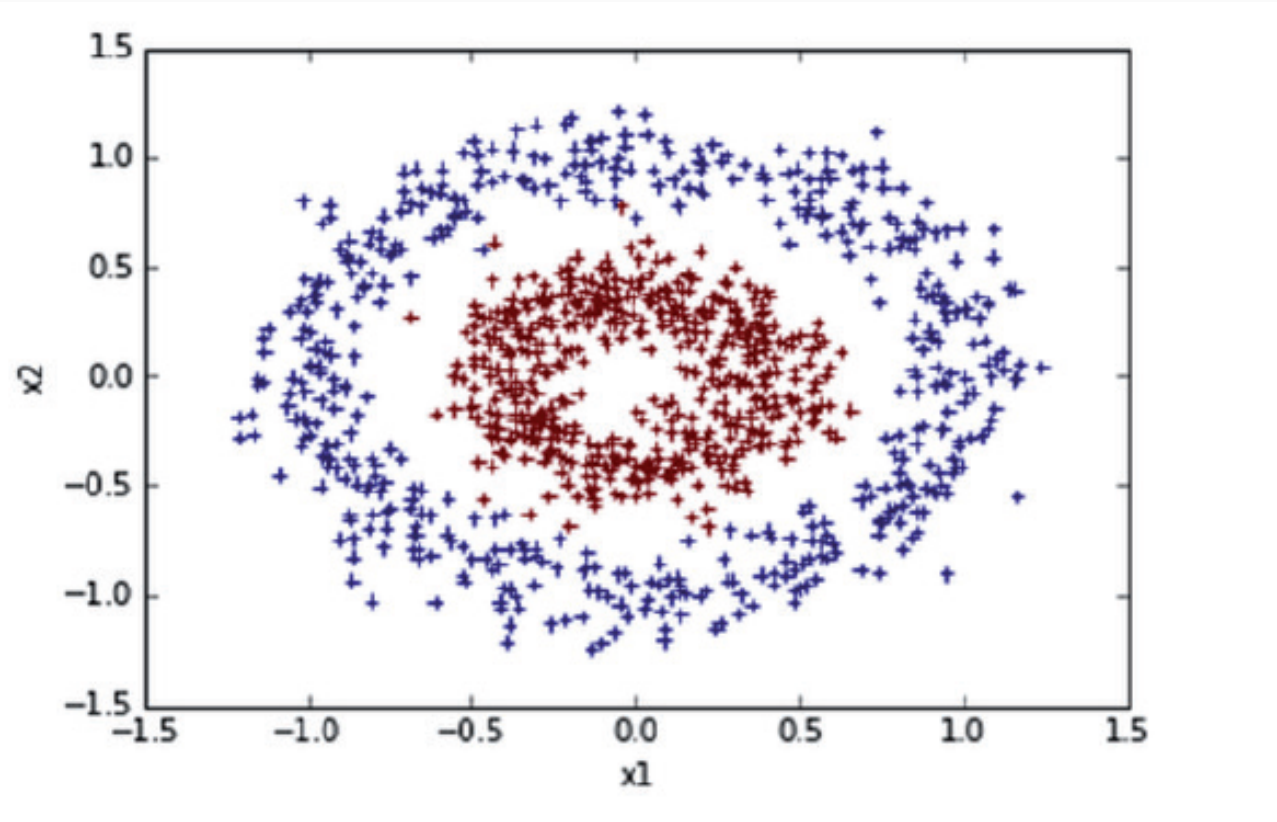
Séparation non linéaire

Que se passe t'il quand la séparation n'est pas linéaire ?



Séparation non linéaire

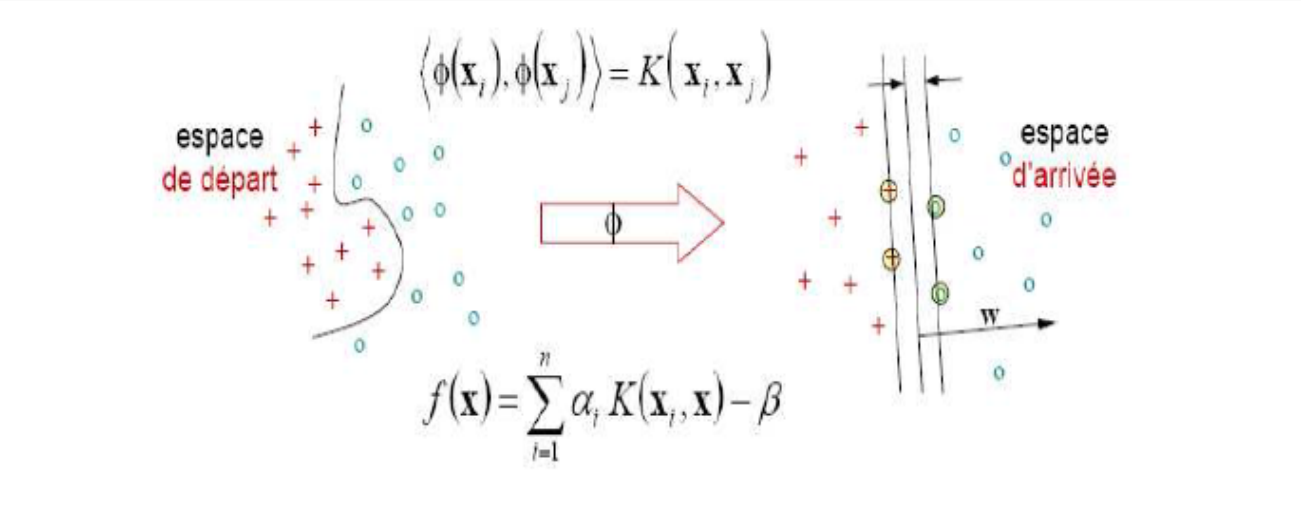
Exemple de séparation non linéaire



Séparation non linéaire

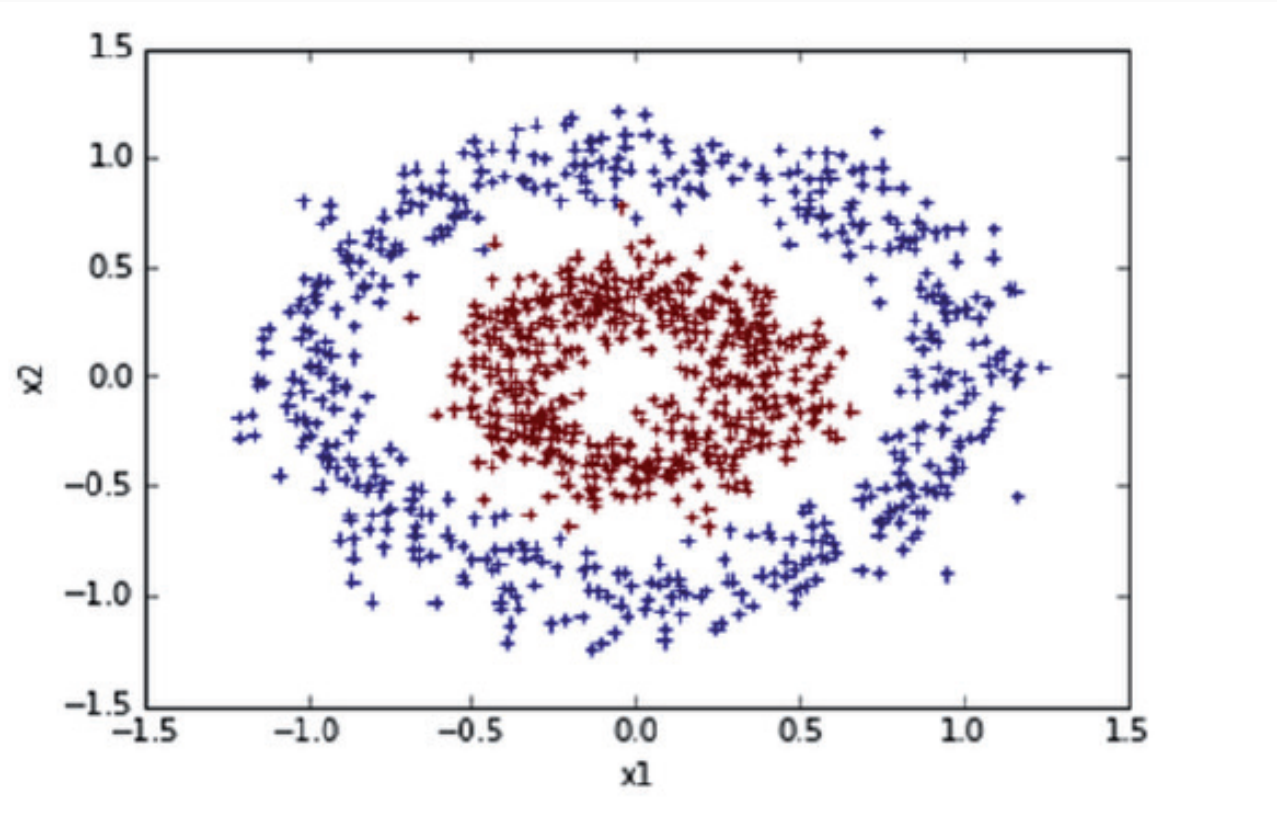
Astuces des noyaux

- Extensions à des séparateurs non linéaires
- Principe : on transpose les données dans un autre espace dans lequel elles sont linéairement séparables
- Transformation : $\phi : \mathbb{R}^d \rightarrow \mathcal{H}, \mathbf{x} \rightarrow \phi(x)$ (H : espace de Hilbert)



Séparation non linéaire

Retour à l'exemple



Séparation non linéaire

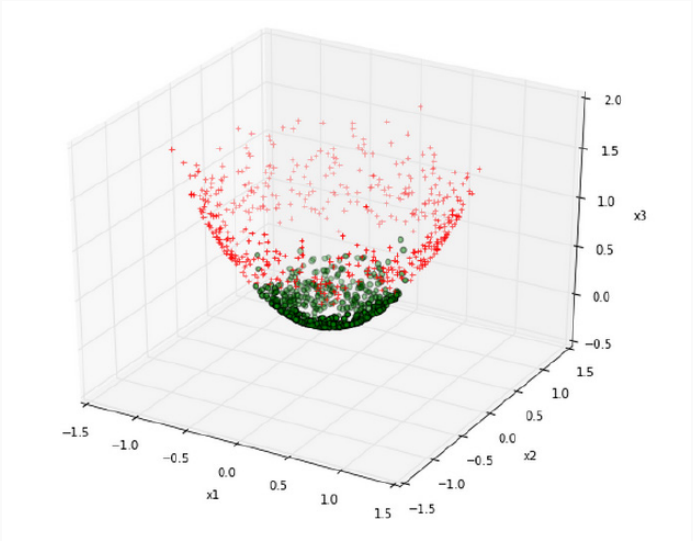
Retour à l'exemple

- Dans notre cas, nous avons un espace initial à deux dimensions dans lequel il n'est pas possible de séparer linéairement nos données.
- Il faut choisir une transformation ϕ qui doit permettre une séparation linéaire dans le nouvel espace $\mathcal{H}$

$$\phi : \mathbb{R}^d \to \mathcal{H}, \mathbf{x} \to \phi(x)$$

On prend

$$\phi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$$



Séparation non linéaire

Changement de représentation : comment faire ?

- On souhaite un changement de représentation :
  - permettant une séparation linéaire de deux classes.
  - respectant la vraie similarité entre les données.
- En général, cela veut dire :
  - Trouver un espace de redescription $\Theta(\mathcal{X})$ de grande dimension.
  - Comment ?
  - Comment garantir la réalisation des calculs ?

Astuce des fonctions noyaux : on va éviter de calculer explicitement la transformation ϕ.

Astuces des noyaux

- On ne calcule pas directement la transformation de changement de représentation, mais on définit une fonction noyau K telle que :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

  On cherche donc une fonction K qui correspond au produit scalaire dans l'espace $\mathcal{H}$.

- De telles fonctions existent :
  Théorème de Mercer : une fonction noyau K continue, symétrique et semi-définie positive peut s'exprimer comme un produit scalaire dans un espace de grande dimension.

- La fonction de décision dans l'espace d'origine est :

$$\sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Astuce des fonctions noyaux.

On appelle noyau toute fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ qui peut être interprétée comme un produit scalaire dans un plongement Φ.

$$\forall x, y \in \mathcal{X} : K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

On peut appliquer les algorithmes précédents de séparation optimale avec marges souples ou dures en remplaçant :

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \text{ par } K(\mathbf{x}_i, \mathbf{x}_j).$$

On obtient alors un classifieur :

$$f : \mathbf{x} \to \mathit{sign}(\sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_j))$$

linéaire dans l'espace de plongement.

Fonctions noyaux les plus populaires

- Noyau polynomial :

$$K(x_1, x_2) = (1 + x_1^T x_2)^q$$

- Noyau gaussien (RBF (Radial Basis Function))

$$K(x_1, x_2) = \exp^{-\gamma(x_1 - x_2)^2}$$

- Noyau sigmoid :

$$K(x_1, x_2) = tanh(kx_1 x_2 - \delta)$$

Le choix du noyau est important : il doit maximiser les chances d'être dans le bon espace

Cas multi-classe

On considère que l'on a C classes c_i. Comment gérer $C > 2$?

one-versus-all

- exemples positifs $(+1)$ sont c_i et négatifs (-1) tous les autres $c_{j \neq i}$
- apprentissage de C classifieurs binaires
- la classe de plus fort score est retenue

one-versus-one

- exemples positifs $(+1)$ sont c_{i1} et négatifs (-1) sont c_{i2}
- apprentissage de $C(C-1)/2$ classifieurs binaires
- vote de chaque classifieur : une classe *gagne* à chaque fois
- vainqueur = classe ayant le maximum de votes

SVM : bilan

- Un approche d'apprentissage relativement puissante capable de trouver des motifs non linéaires.
- Deux idées principales :
  - Maximisation de la marge entre la frontière de décision et les exemples les plus proches, les vecteurs de support.
  - Redescription des observations dans un nouvel espace où une séparation linéaire sera possible.
- Extension facile au cas multiclasse.