Machines à vecteurs supports

masedki.github.io

mer. 27 nov. 2019

Introduction

Les Support Vector Machines souvent traduit par l'appellation de **Séparateur à Vaste Marge (SVM)** sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire. Ils ont été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible.

Un peu d’histoire

Cette approche découle directement des travaux de Vapnik en théorie de l’apprentissage à partir de 1995. Elle s’est focalisée sur les propriétés de généralisation (ou prévision) d’un modèle en contrôlant sa complexité appelée **dimension de Vapnik Chernovenkis** qui est un indicateur du pouvoir séparateur d’une famille de fonctions associé à un modèle et qui en contrôle la qualité de prévision. Le principe fondateur des SVM est justement d’intégrer à l’estimation le contrôle de la complexité c’est-à-dire le nombre de paramètres qui est associé dans ce cas au nombre de vecteurs supports. L’autre idée directrice de Vapnik dans ce développement, est d’éviter de substituer à l’objectif initial : la discrimination, un ou des problèmes qui s’avèrent finalement plus complexes à résoudre comme par exemple l’estimation non-paramétrique de la densité d’une loi multidimensionnelle en analyse discriminante.

Principe

Le principe de base des SVM consiste de ramener le problème de la discrimination à celui, linéaire, de la recherche d'un hyperplan optimal. Deux idées ou astuces permettent d'atteindre cet objectif :

- La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle.
- Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (kernel) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire (feature space) de plus grande dimension. D'où l'appellation couramment rencontrée de machine à noyau ou kernel machine.

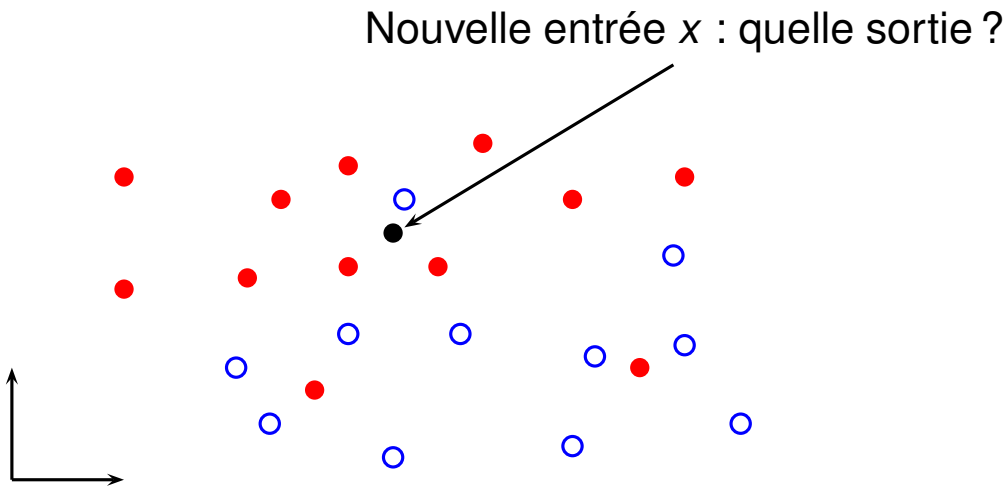
Utilisation

Cet outil est largement utilisé dans de nombreux types d'applications et s'avère un concurrent sérieux des algorithmes les plus performants (agrégation de modèles). L'introduction de noyaux, spécifiquement adaptés à une problématique donnée, lui confère une grande flexibilité pour s'adapter à des situations très diverses (reconnaissance de formes, de séquences génomiques, de caractères, détection de spams, diagnostics…). À noter que, sur le plan algorithmique, ces algorithmes sont plus pénalisés par le nombre d'observations, c'est-à-dire le nombre de vecteurs supports potentiels, que par le nombre de variables. Néanmoins, des versions performantes des algorithmes permettent de prendre en compte des bases de données volumineuses dans des temps de calcul acceptables.

Support Vector Machine : quoi et pourquoi ?

*Une **SVM (Support Vector Machine) ou Machine à Vecteurs Supports** est une famille d'algorithmes d'apprentissage supervisé pour des problèmes de discrimination (ou de régression).*

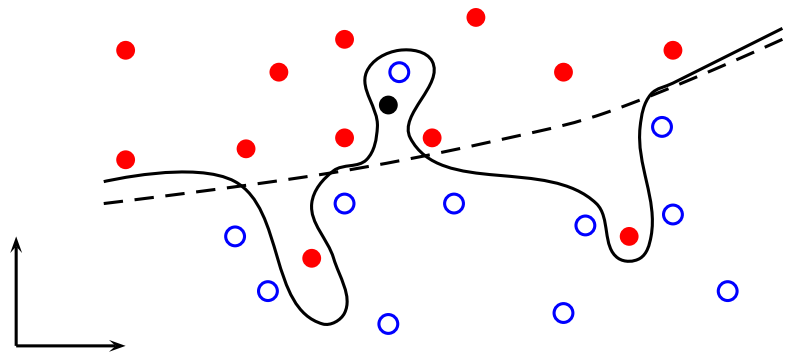
Exemple jouet : classification binaire en dimension 2



Support Vector Machine : quoi et pourquoi ?

Fondements mathématiques solides $\Rightarrow$ bonnes **propriétés de généralisation** *i.e.* bon compromis la classification des données et la prédiction de la sortie associée à une nouvelle entrée x.

Exemple de règle n'ayant pas de bonnes propriétés de généralisation



!!! Phénomène de sur-apprentissage (ou overfitting) !!!

particulièrement présent en grande dimension ou pour des règles de discrimination non linéaires complexes.

SVM linéaire pour des données séparables

Données linéairement séparables : On considère des données à valeurs dans $\mathbb{R}^p$, muni du produit scalaire usuel $\langle \cdot, \cdot \rangle$. Notons que $\langle x, y \rangle = x^\top y$.
Les données observées $(x_1, y_1), \ldots, (x_n, y_n)$ sont dites **linéairement séparables** s'il existe (w, b) tel que pour tout i :

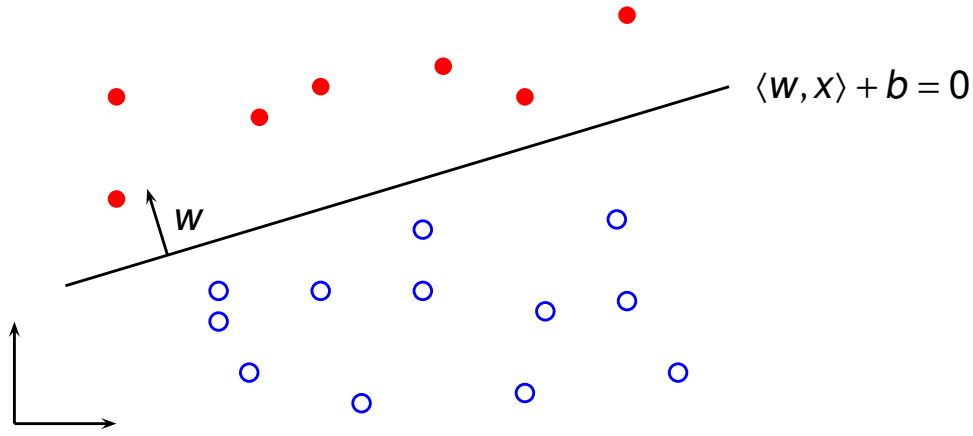
$$y_i = 1 \qquad \text{si} \quad w^\top x_i + b > 0,$$
$$y_i = -1 \qquad \text{si} \quad w^\top x_i + b < 0,$$

Autrement dit

$$\forall i = 1, \ldots, n \quad y_i \left(w^\top x_i + b \right) > 0.$$

Hyperplan séparateur

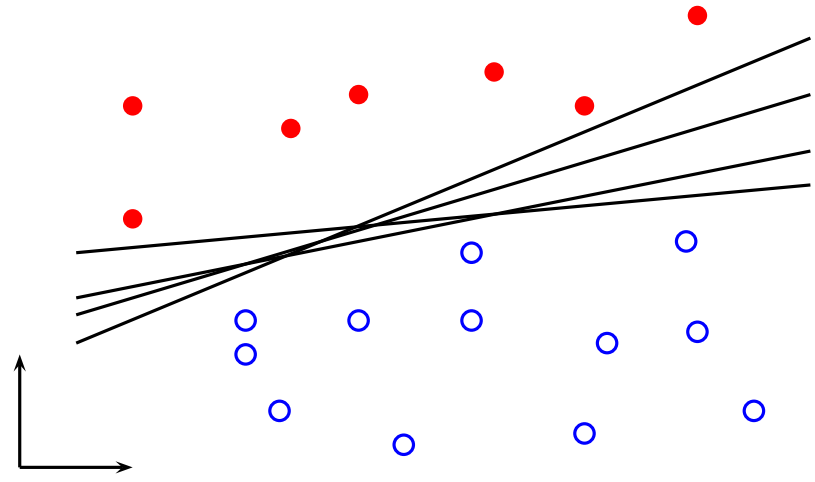
L'équation $w^\top x + b = 0$ définit un hyperplan séparateur de vecteur orthogonal w.



La règle de classification $g_{w,b}(x) = \mathbf{1}_{w^\top x + b \geq 0} - \mathbf{1}_{w^\top x + b < 0}$ est une règle de classification linéaire potentielle.

Choix de la règle de décision ?

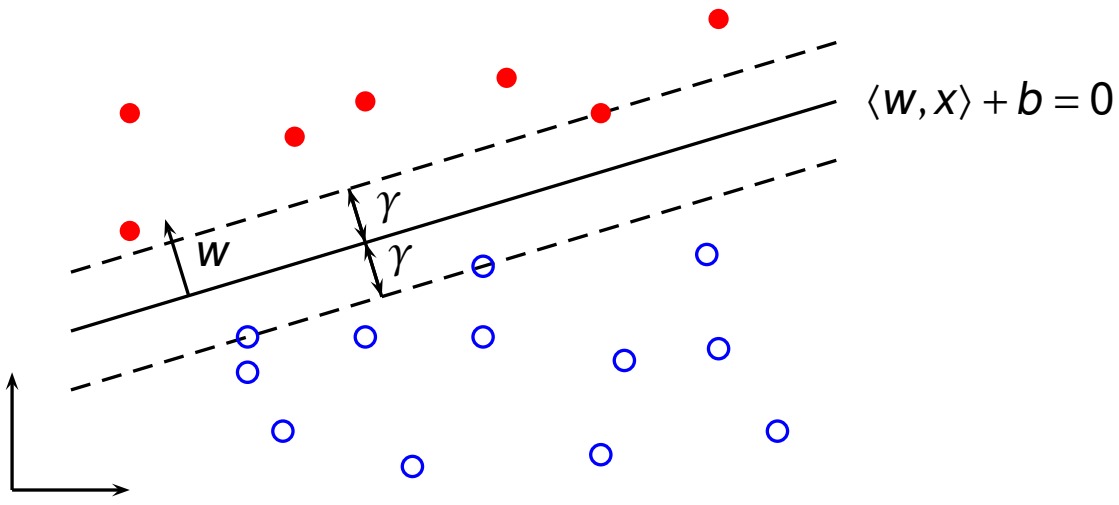
- **Problème:** une infinité d'hyperplans séparateurs $\Rightarrow$ une infinité de règles de discrimination linéaires potentielles !



- Laquelle choisir ?

Hyperplan séparateur de marge maximale

La réponse (Vapnik) : La règle de discrimination linéaire ayant les meilleures propriétés de généralisation correspond à l'hyperplan séparateur de **marge maximale** γ.



La marge maximale

- Soit deux entrées de l'ensemble d'apprentissage (re)notées x_1 et x_{-1} de sorties respectives 1 et -1, se situant sur les frontières définissant la marge. L'hyperplan séparateur correspondant se situe à mi-distance entre x_1 et x_{-1}. La marge s'exprime donc comme suit :

$$\gamma = \frac{1}{2} \frac{w^\top (x_1 - x_{-1})}{\|w\|}.$$

- **Remarque:** pour tout $\kappa \neq 0$, les couples $(\kappa w, \kappa b)$ et (w, b) définissent le même hyperplan.
- L'hyperplan $\langle w, x \rangle + b = 0$ est dit en forme canonique relativement à un ensemble de vecteurs $x_1, \ldots, x_m$ si $\min_{i=1,\ldots,m} |w^\top x + b| = 1$.
- L'hyperplan séparateur est en forme canonique relativement aux vecteurs $\{x_1, x_{-1}\}$ s'il est défini par (w, b) avec $w^\top x_1 + b = 1$ et $w^\top x_{-1} + b = -1$. On a alors $w^\top (x_1 - x_{-1}) = 2$, d'où

$$\gamma = \frac{1}{\|w\|}.$$

Problème d'optimisation *primal*

Trouver l'hyperplan séparateur de marge maximale revient à trouver le couple (w, b) tel que

$$\begin{cases} \underset{w,b}{\text{Minimiser}} & \|w\|^2 \quad \text{ou} \quad \frac{1}{2}\|w\|^2 \\ \text{sous la contrainte} & y_i\left(w^\top x_i + b\right) \geq 1 \quad \text{pour tout} \quad i. \end{cases}$$

- Problème d'optimisation convexe sous contraintes linéaires
- Existence d'un optimum global, obtenu par résolution du problème "dual" (méthode des multiplicateurs de Lagrange).

Programmation quadratique

Le problème de minimisation sous contraintes précédent s'écrit sous la forme d'un programme quadratique

$$\begin{cases} \min\limits_{z} & \frac{1}{2} z^\top A z - d^\top z \\ \text{avec} & Bz \leq e \end{cases}$$

où $z = (w, b)^\top \in \mathbb{R}^{p+1}$ et $d = (0, \ldots, 0)^\top \in \mathbb{R}^{p+1}$. La matrice A est donnée par

$$Z = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

où I est la matrice identité de $\mathbb{R}^p$, et $B = -[\text{diag}(y)X, y]$,
$e = -(1, \ldots, 1)^\top \in \mathbb{R}^n$, $y \in \mathbb{R}^n$ est la matrice des signes des observations et X la matrice $n \times p$ des données (la ligne i est le vecteur $x_i^\top$).
Ce problème est convexe car A est semidéfinie positive. Il admet donc une solution unique (qui existe puisque le problème est linéairement séparable par hypothèse) et les conditions nécessaires d'optimalité du premier ordre sont aussi suffisantes. *La solution est obtenue par la résolution du problème "dual".*

Multiplicateurs de Lagrange : problème primal

- Minimiser pour $u \in \mathbb{R}^2$ $h(u)$ sous les contraintes $g_i(u) \geq 0$ pour $i = 1 \ldots n$, h fonction quadratique, g_i fonction affines.
- Le **Lagrangien** est défini sur $\mathbb{R}^2 \times \mathbb{R}^n$ par $L(u, \alpha) = h(u) - \sum_{i=1}^{n} \alpha_i g_i(u)$. Les variables α_i sont appelées les **variables duales**. Soit pour tout $\alpha \in \mathbb{R}^n$,
  - $u_\alpha = \underset{u \in \mathbb{R}^2}{\operatorname{argmin}} L(u, \alpha)$,
  - $\theta(\alpha) = L(u_\alpha, \alpha) = \min_{u \in \mathbb{R}^2} L(u, \alpha)$ (**fonction duale**).
- **Multiplicateurs de Lagrange : problème dual**
  - Maximiser $\theta(\alpha)$ sous les contraintes $\alpha_i \geq 0$ pour $i = 1 \ldots n$.
  - La solution du problème dual α^* donne la solution du problème primal : $u^* = u_{\alpha^*}$.

Conditions de Karush-Kuhn-Tucker

- $\alpha^* \geq 0$ pour tout $i = 1 \ldots n$.
- $g_i\left(u_{\alpha^*}\right) \geq 0$ pour tout $i = 1 \ldots n$.
- **Retour sur le problème dual :**

On doit minimiser $L\left(u, \alpha\right) = h(u) - \sum_{i=1}^{n} \alpha_i g_i(u)$ par rapport à u et maximiser $L\left(u_\alpha, \alpha\right)$ correspondant par rapport aux variables duales α_i.
$\Rightarrow$ Si $g_i\left(u_{\alpha^*}\right) > 0$, alors nécessairement $\alpha_i^* = 0$.

- **Condition complémentaire de Karush-Kuhn-Tucker** qui s'exprime sous la forme $\alpha_i^* g_i\left(u_{\alpha^*}\right) = 0$.

Lagrangien pour SVM en classification binaire

Lagrangien : $L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \Big(y_i (w^\top x_i + b) - 1 \Big)$

Fonction duale :

$$\begin{cases} \frac{\partial L(w,b,\alpha)}{\partial w} & = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \frac{\partial L(w,b,\alpha)}{\partial b} & = -\sum_{i=1}^{n} \alpha_i y_i = 0 \iff \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

$$\theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

La solution du problème d'optimisation primal est donnée par :

- $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$,
- $b^* = -\frac{1}{2} \left\{ \min_{y_i=1} w^{*\top} x_i + \min_{y_i=-1} w^{*\top} x_i \right\}$, où $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ est la solution du problème d'optimisation dual :

$$\begin{gathered} \text{Maximiser } \theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{S.C. } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0 \quad \forall i = 1 \ldots n. \end{gathered}$$

La solution α^* du problème dual est indépendante de la dimension p : SVM linéaire ne souffre pas du "fléau de la dimension".

Conditions de Karush-Kuhn-Tucker

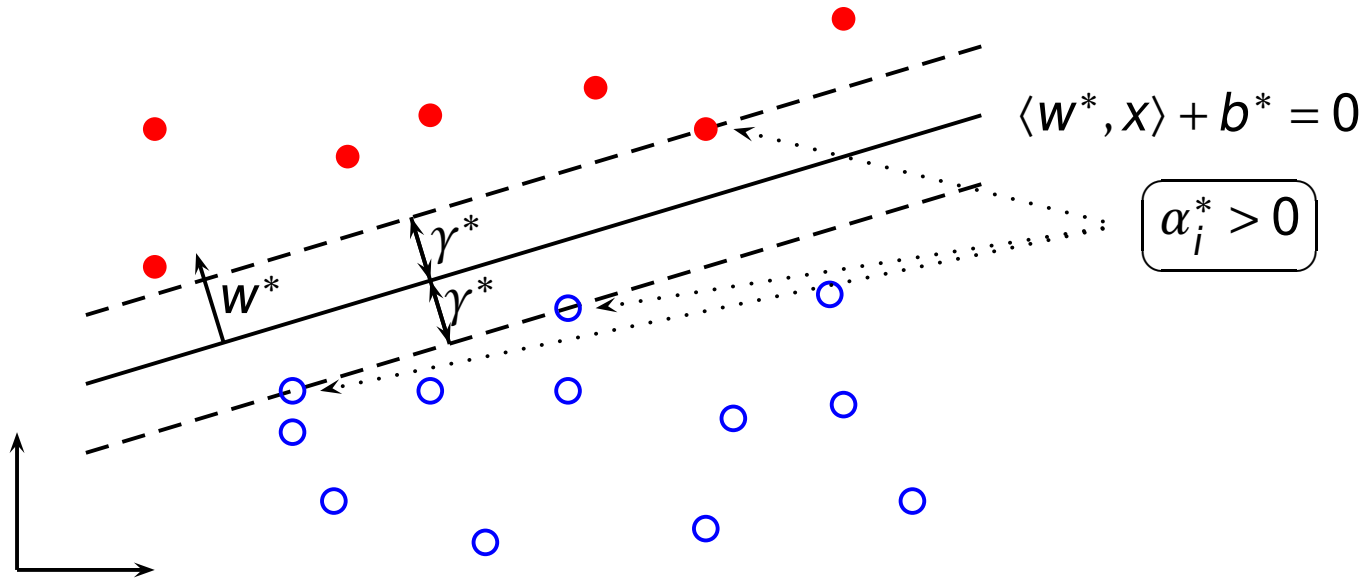
- $\alpha_i^* \geq 0 \quad \forall i = 1 \ldots n.$
- $y_i\left(w^{*\top} x_i + b^*\right) \geq 1 \quad \forall i = 1 \ldots n.$
- $\alpha_i^* \left(y_i\left(w^{*\top} x_i + b^*\right) - 1\right) = 0 \quad \forall i = 1 \ldots n.$ (condition complémentaire)

1. Le nombre de $\alpha_i^* > 0$ peut être petit : on dit que la solution du problème dual est **parcimonieuse (sparse)**.
2. Efficacité algorithmique.

Les x_i tels que $\alpha_i^* > 0$ sont appelés les **vecteurs supports**. Ils sont situés sur les frontières définissant la marge maximale i.e.

$$y_i\left(w^{*\top} x_i + b^*\right) = 1 \quad \text{(c.f. condition complémentaire de KKT).}$$

Représentation des vecteurs supports



En conclusion

La règle de classification obtenue

$$\hat{g}_n(x) = \mathbf{1}_{w^{*\top} x + b^* \geq 0} - \mathbf{1}_{w^{*\top} x + b^* < 0},$$

où

- $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$,
- $b^* = -\frac{1}{2}\left\{ \min_{y_i=1} w^{*\top} x_i + \min_{y_i=-1} w^{*\top} x_i \right\}$,

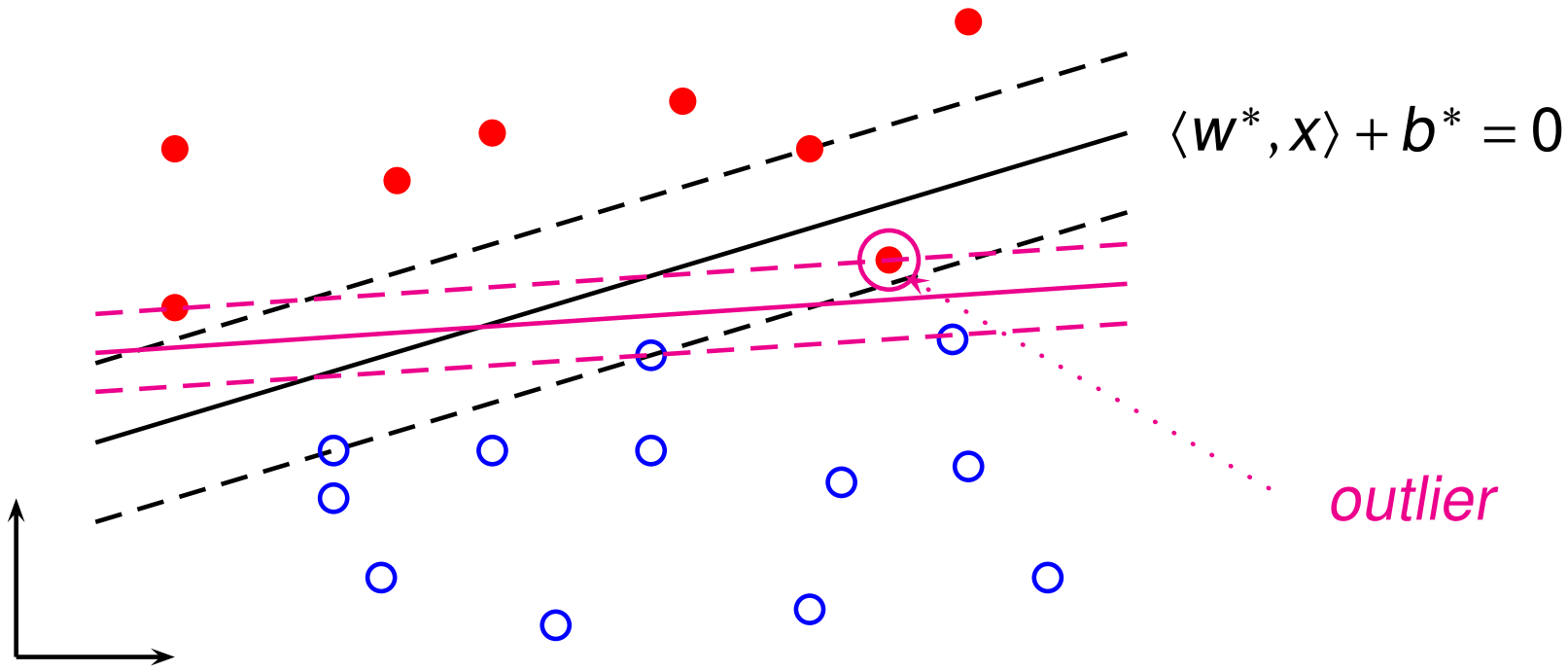
ou encore :

$$\hat{g}_n(x) = \mathbf{1}_{\sum_{x_i V.S} \alpha_i^* y_i x_i^\top x + b^* \geq 0} - \mathbf{1}_{\sum_{x_i V.S} \alpha_i^* y_i x_i^\top x + b^* < 0}.$$

La marge maximale vaut $\gamma^* = \frac{1}{\|w^*\|} = \left(\sum_{i=1}^n \left(\alpha_i^*\right)^2 \right)^{-\frac{1}{2}}$.

SVM pour des données non séparables

- La méthode précédente ne peut être appliquée si les données ne sont pas linéairement séparables
- Sensibilité aux *outliers*



La solution

- Autoriser quelques vecteurs à être bien classés mais dans la région définie par la marge, voire mal classés.
- La contrainte $y_i\left(w^\top x_i + b\right) \geq 1$ devient $y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i$, avec $\xi_i \geq 0$.
  - $\xi_i \in [0, 1] \leftrightarrow$ bien classé, mais région définie par la marge.
  - $\xi_i > 1 \leftrightarrow$ mal classé.
- On parle de **marge souple** ou marge relaxée.
- Les variables ξ_i sont appelées les **variables ressort** (slacks).
- Un souci : les contraintes relaxées ne peuvent pas être utilisées sans contrepartie sous peine d'obtenir une marge maximale infinie (en prenant des valeurs de ξ_i suffisamment grandes).

Pénaliser les grandes valeurs des ξ_i

Problème d'optimisation primal

$$\text{Minimiser en } (w, b, \xi) \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.c.} \begin{cases} y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i \quad \forall i \\ \xi_i \geq 0 \end{cases}$$

$\hookrightarrow C > 0$ paramètre (**constante de tolérance**) à ajuster.

La solution du problème d'optimisation primal est donnée par :

- $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$,
- b^* tel que $y_i\left(w^{*\top} x_i + b^*\right) = 1 \quad \forall x_i, 0 < \alpha_i^* < C$,

où $\alpha^* = \left(\alpha_1^*, \ldots, \alpha_n^*\right)$ est la solution du **problème d'optimisation dual** :

$$\text{Maximiser } \theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$
$$\text{s.c. } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \; \forall i.$$

Conditions de Karush-Kuhn-Tucker

- $0 \leq \alpha_i^* \leq C \quad \forall i = 1 \ldots n.$
- $y_i \left(w^{*\top} x_i + b^* \right) \geq 1 - \xi_i^* \quad \forall i = 1 \ldots n.$
- $\alpha_i^* \left(y_i \left(w^{*\top} x_i + b^* \right) + \xi_i^* - 1 \right) = 0 \quad \forall i = 1 \ldots n.$
- $\xi_i^* \left(\alpha_i^* - C \right) = 0.$

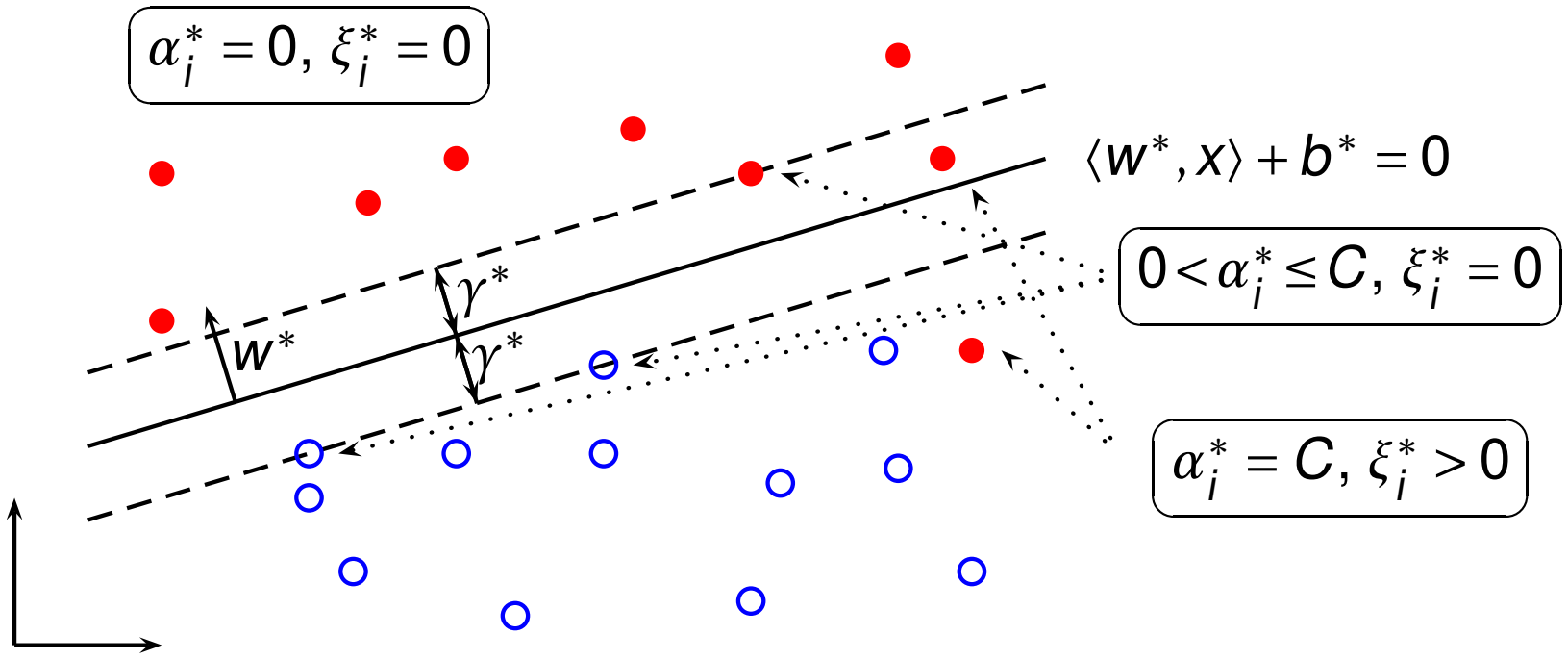
Les x_i tels que $\alpha_i^* > 0$ sont les **vecteurs supports**.

Deux types de vecteurs supports :

- Les vecteurs correspondant à des variables ressort nulles. Ils sont situés sur les frontières de la région définissant la marge.
- Les vecteurs correspondant à des variables ressort non nulles : $\xi_i^* > 0$ et dans ce cas $\alpha_i^* = C$.

Les vecteurs qui ne sont pas supports vérifient $\alpha_i^* = 0$ et $\xi_i^* = 0$

Représentation des vecteurs supports



En conclusion

La règle de classification obtenue

$$\hat{g}_n(x) = \mathbf{1}_{w^{*\top} x + b^* \geq 0} - \mathbf{1}_{w^{*\top}, x + b^* < 0},$$

où

- $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$,
- b^* tel que $y_i \left(w^{*\top}, x_i + b^* \right) = 1 \quad \forall x_i, 0 < \alpha_i^* < C$,

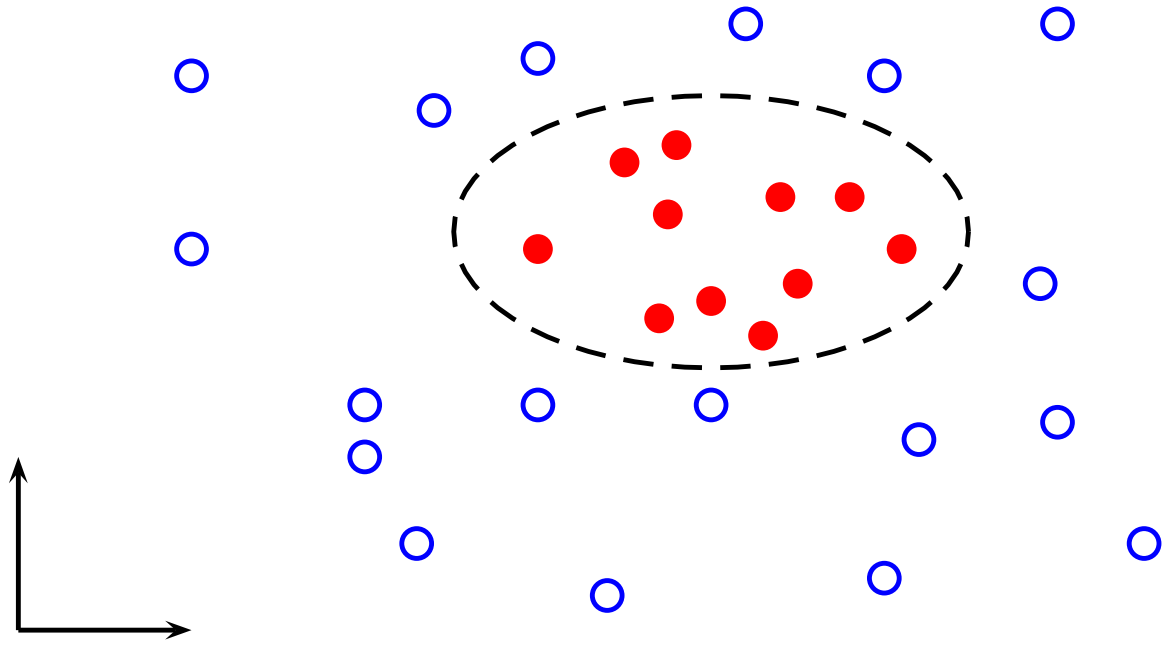
ou encore :

$$\hat{g}_n(x) = \mathbf{1}_{\sum_{x_i V.S} \alpha_i^* y_i x_i^\top x + b^* \geq 0} - \mathbf{1}_{\sum_{x_i V.S} \alpha_i^* y_i x_i^\top x + b^* < 0}.$$

La marge maximale vaut $\gamma^* = \frac{1}{\|w^*\|} = \left(\sum_{i=1}^n \left(\alpha_i^* \right)^2 \right)^{-\frac{1}{2}}$.

SVM non linéaire : astuce du noyau

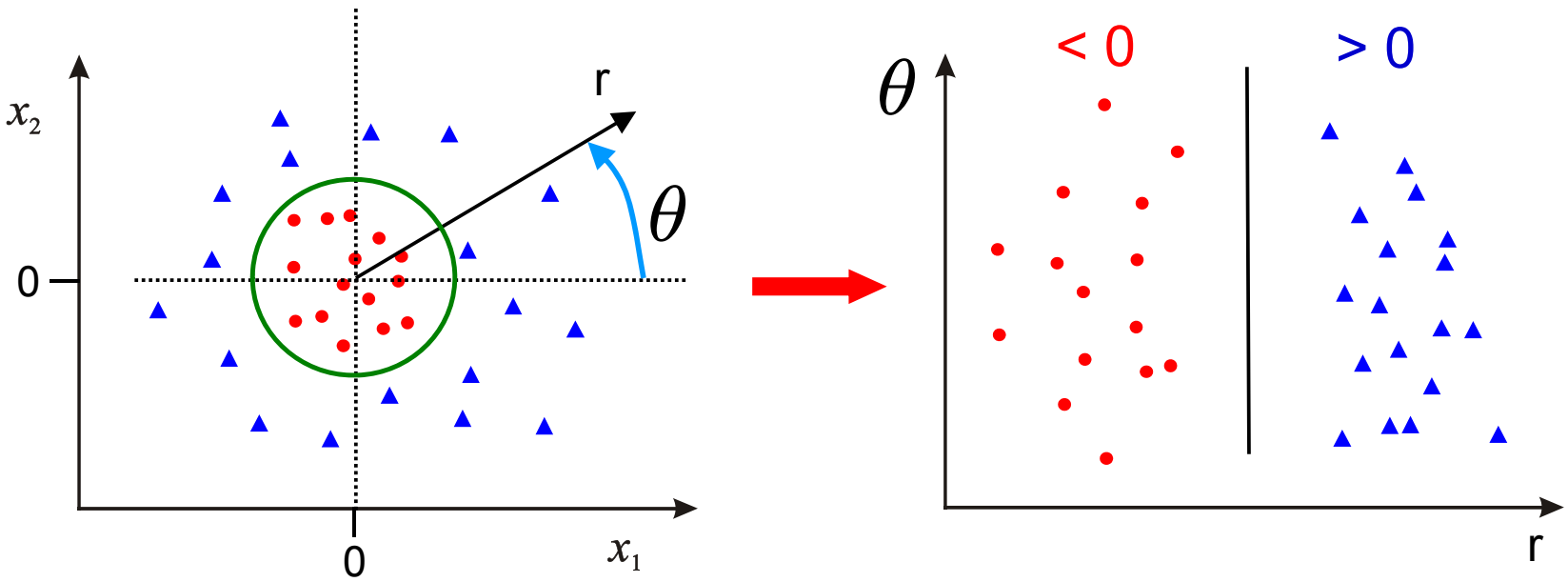
Exemple de données difficiles à discriminer linéairement :



Une règle SVM linéaire donnera une très mauvaise classification avec un nombre de vecteurs supports très élevé $\Rightarrow$ SVM non linéaire ?

Solution 1

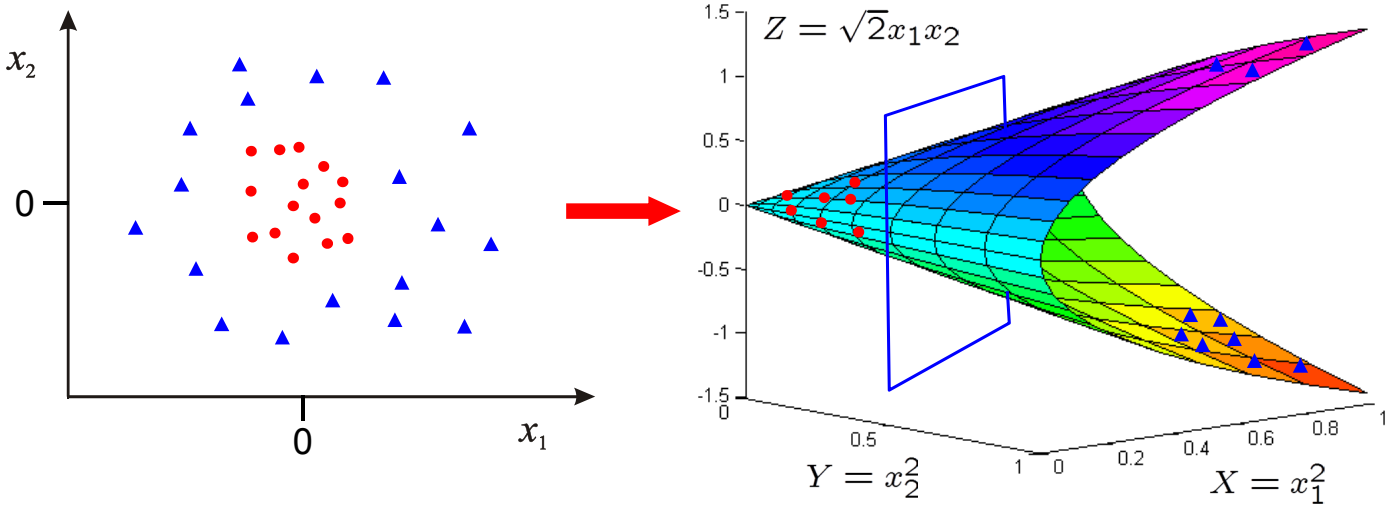
Passer aux coordonnées polaires



Solution 2 (qu'on va retenir !)

Augmenter la dimension des données à l'aide d'une transformation

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \qquad \mathbb{R}^2 \rightarrow \mathbb{R}^3$$



- En dimension 3, le problème devient linéairement séparable !

L'idée (Boser, Guyon, Vapnik, 1992)

Envoyer les entrées $\{x_i, i = 1 \dots n\}$ dans un espace de Hilbert $\mathcal{H}$ (muni d'un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$), de grande dimension, voire de dimension infinie, via une fonction φ, et appliquer une SVM linéaire aux nouvelles données $\left\{ \left(\varphi(x_i), y_i\right), i = 1 \dots n \right\}$.
La sortie attribuée à l'entrée x est celle attribuée à son image $\varphi(x)$.

- La fontion φ est appelée la **fonction de représentation (feature function)**.
- L'espace $\mathcal{H}$ est appelé l'**espace de représentation (feature space)**.

Dans l'exemple précédent, pour $\varphi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$, les données deviennent linéairement séparables dans $\mathbb{R}^3$.

Comment choisir $\mathcal{H}$ et φ ?

La règle de discrimination de la SVM non linéaire est définie par :

$$\hat{g}_n(x) = \mathbf{1}_{\sum y_i \alpha_i^* \langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} + b^* \geq 0} - \mathbf{1}_{\sum y_i \alpha_i^* \langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} + b^* < 0},$$

où

$$\begin{aligned} \text{Maximiser } \theta(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} \\ \text{s.c. } \sum_{i=1}^n \alpha_i y_i &= 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i. \end{aligned}$$

- Remarque fondamentale :

La règle de discrimination de la SVM non linéaire ne dépend de φ qu'à travers des produits scalaires de la forme $\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}}$ ou $\langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$.

- **Astuce du noyau (kernel trick)** : La connaissance de la seule fonction k définie par $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ permet de lancer la SVM dans $\mathcal{H}$, sans déterminer explicitement $\mathcal{H}$ et φ.

Le noyau

- Une fonction $k : \mathcal{X}, \mathcal{X} \to \mathbb{R}$ telle que $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ pour une fonction $\varphi : \mathcal{X} \to \mathcal{X}$ donnée est appelée un **noyau**.
- Un noyau est souvent plus facile à calculer que la fonction φ. Par exemple, si pour $x = (x_1, x_2) \in \mathbb{R}^2$, $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, alors que $k(x, x') = \langle x, x' \rangle^2$.
- **Quelques noyaux classiques pour $\mathcal{X} = \mathbb{R}^p$**
  - Noyau **polynomial** : $k(x, x') = \left(\langle x, x' \rangle + c\right)^p$

  $\hookrightarrow \quad \varphi(x) = \left(\varphi_1(x), \ldots, \varphi_k(x)\right)$ avec $\varphi_i(x) =$ monôme de degré inférieur à p de certains composantes de x.

  - Noyau **gaussien** ou **radial** (RBF) : $k(x, x') = e^{-\frac{||x-x'||^2}{2\sigma^2}}$

  $\hookrightarrow$ à valeurs dans un espace de dimension infinie.

  - Noyau **laplacien** : $k(x, x') = e^{-\frac{||x-x'||}{\sigma}}$.

Agrégation de noyaux

- Soit k_1 et k_2 des noyaux, f une fonction $\mathbb{R}^p \to \mathbb{R}$, $\psi : \mathbb{R}^p \to \mathbb{R}^{p'}$, B une matrice définie positive, P une polynôme à coefficients positifs, $\lambda \geq 0$.
- La fonction définie $k(x, x') = k_1(x, x') + k_2(x, x'), \lambda k_1(x, x'), k_1(x, x')k_2(x, x'), f(x)f(x'), k(\psi(x), \psi(x')), x^T B x', P(k_1(x, x'))$ ou $e^{k_1(x,x')}$ est encore un noyau.
- **Noyaux pour $\mathcal{X} \neq \mathbb{R}^p$** : quelques noyaux ont été proposés pour d'autres types d'objets comme des
  - ensembles,
  - arbres,
  - graphes,
  - chaînes de symboles,
  - documents textuels…

Éléments de théorie

$$\text{Minimiser} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.c.} \quad \begin{cases} y_i\big(\langle w, x_i\rangle + b\big) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 \end{cases}$$

est équivalent à minimiser

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\Big(1 - y_i\big(\langle w, x_i\rangle + b\big)\Big)_+,$$

ou encore

$$\frac{1}{n}\sum_{i=1}^{n}\Big(1 - y_i\big(\langle w, x_i\rangle + b\big)\Big)_+ + \frac{1}{2Cn}\|w\|^2.$$

$\gamma\big(w, b, x_i, y_i\big) = \Big(1 - y_i\big(\langle w, x_i\rangle + b\big)\Big)_+$ est une majorant convexe de l'erreur empirique $\mathbf{1}_{y_i\left(\langle w, x_i\rangle + b\right) \leq 0}$

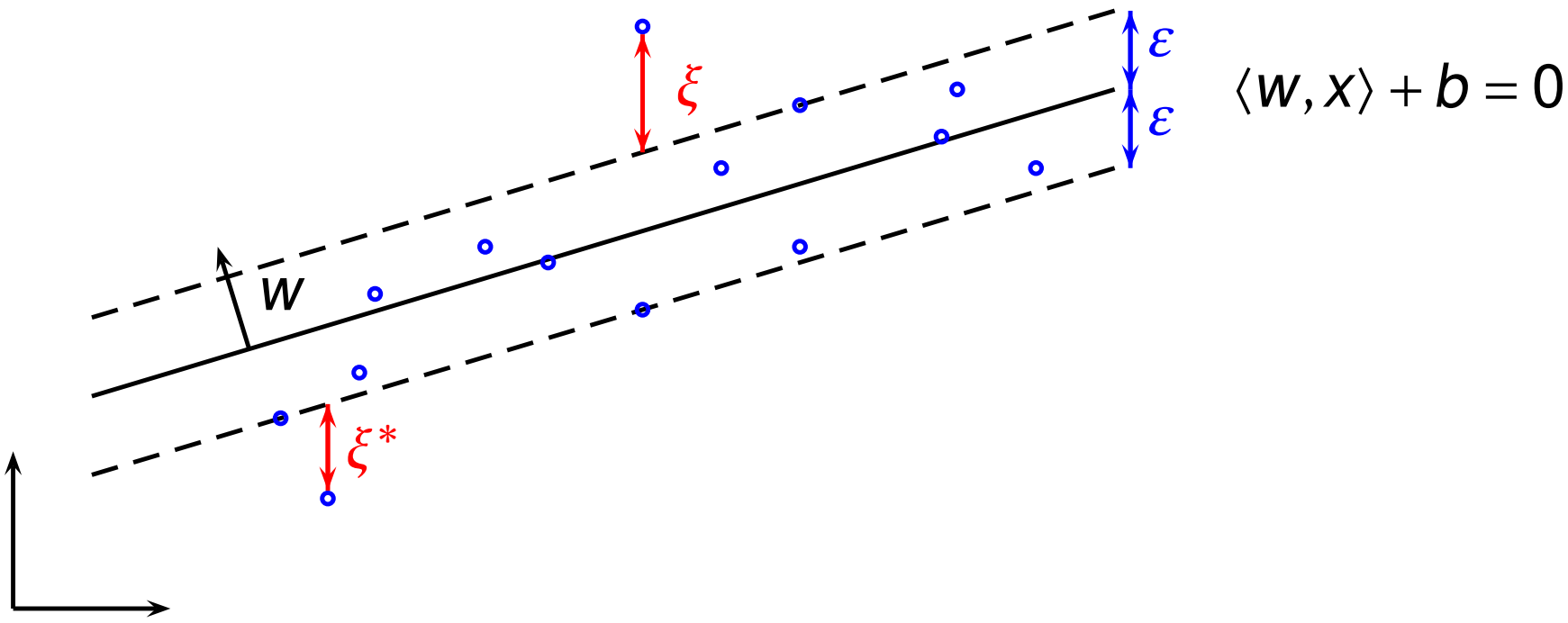
$\hookrightarrow$ SVM = Minimisation de risque empirique convexifié régularisé

Conclusion et questions ouvertes

- Le rééquilibrage des données
- La renormalisation des données
- Les règlages à effectuer :
    - Le noyau et ses paramètres (validation croisée)
    - La constante de tolérance C (validation croisée)
- La sélection de variables
- Généralisation à la discrimination multiclasses : one-versus-all, one-versus-one

SVM pour la régression (SVR) : un mot

La dite ε—SVR linéaire en dimension 2



Le principe

Trouver la règle de régression linéaire la plus *plate* possible, sous certaines contraintes.

Problème d'optimisation prima pour la $\varepsilon-$SVR linéaire

$$\text{Minimiser en } (w, b, \xi, \xi^*) \qquad \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)$$

$$\text{s.c.} \quad \begin{cases} y_i - w^\top x_i - b \leq \varepsilon + \xi_i & \forall i \\ w^\top x_i + b - y_i \leq \varepsilon + \xi_i^* & \forall i \\ \xi_i, \xi_i^* \geq 0 & \forall i \end{cases}$$

$\hookrightarrow$ $C > 0$ paramètre à ajuster.

$\hookrightarrow$ SVR non linéaire : astuce du noyau !

Application : le package e1071

- La fonction svm du package e1071
- La fonction ksvm du package kernlab
- Le package caret pour le choix des paramètres

Table of Contents