Support Vector Machines (SVM) et méthodes à Noyaux

Nicolas Verzelen, Alexis Joly

INRA,Inria

M2 MIASH

Sommaire

SVM linéaire pour des données séparables

Données linéairement séparables

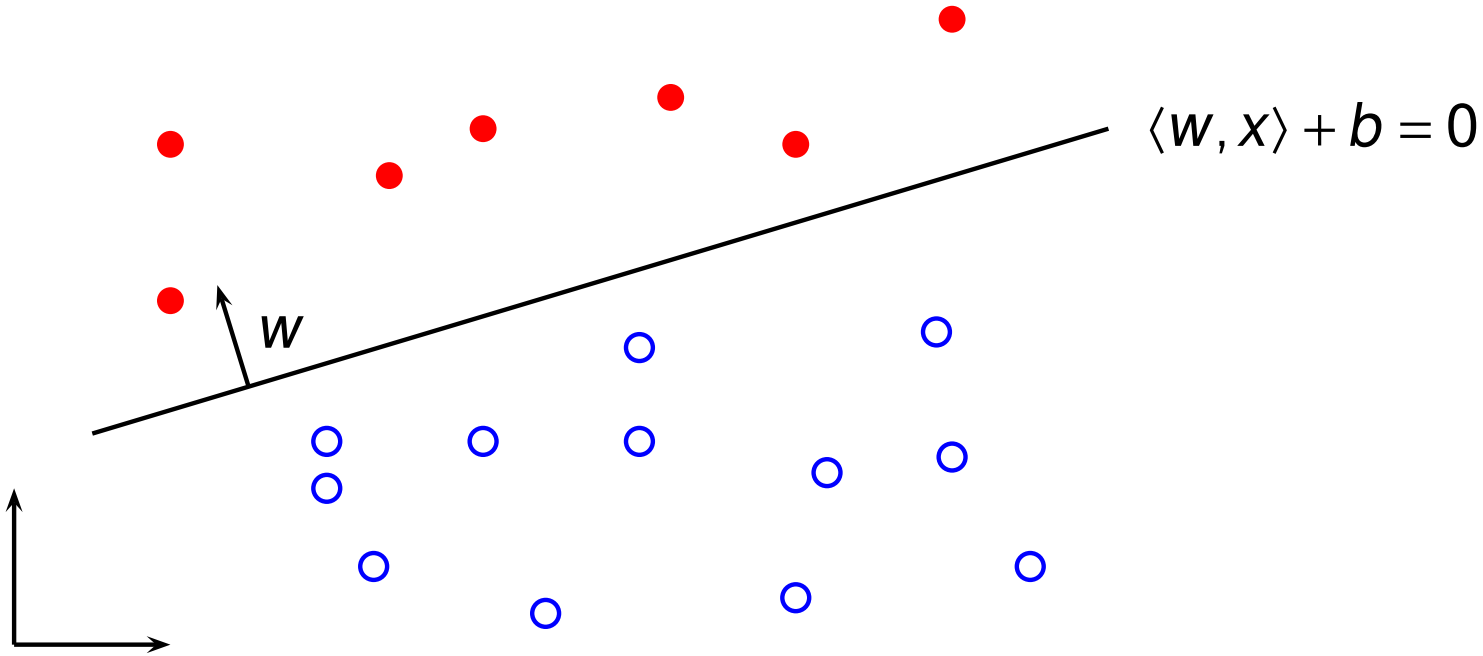
On considère $\mathscr{X} = \mathbb{R}^d$, muni du produit scalaire usuel $\langle ., . \rangle$.

Les données observées $d_1^n = (x_1, y_1), \ldots, (x_n, y_n)$ sont dites **linéairement séparables** s'il existe (w, b) tel que pour tout i,
- $y_i = 1$ si $\langle w, x_i \rangle + b > 0$,
- $y_i = -1$ si $\langle w, x_i \rangle + b < 0$,

c'est-à-dire

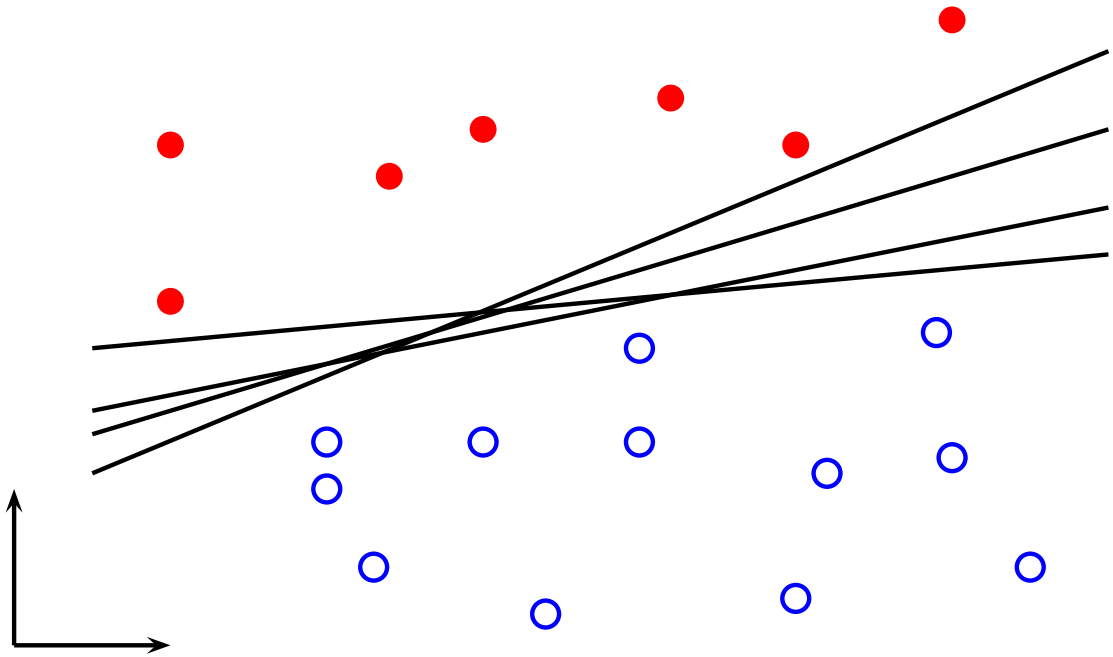
$$\forall i = 1, \ldots, n \quad y_i (\langle w, x_i \rangle + b) > 0$$

L'équation $\langle w, x \rangle + b = 0$ définit un hyperplan séparateur de vecteur orthogonal w.



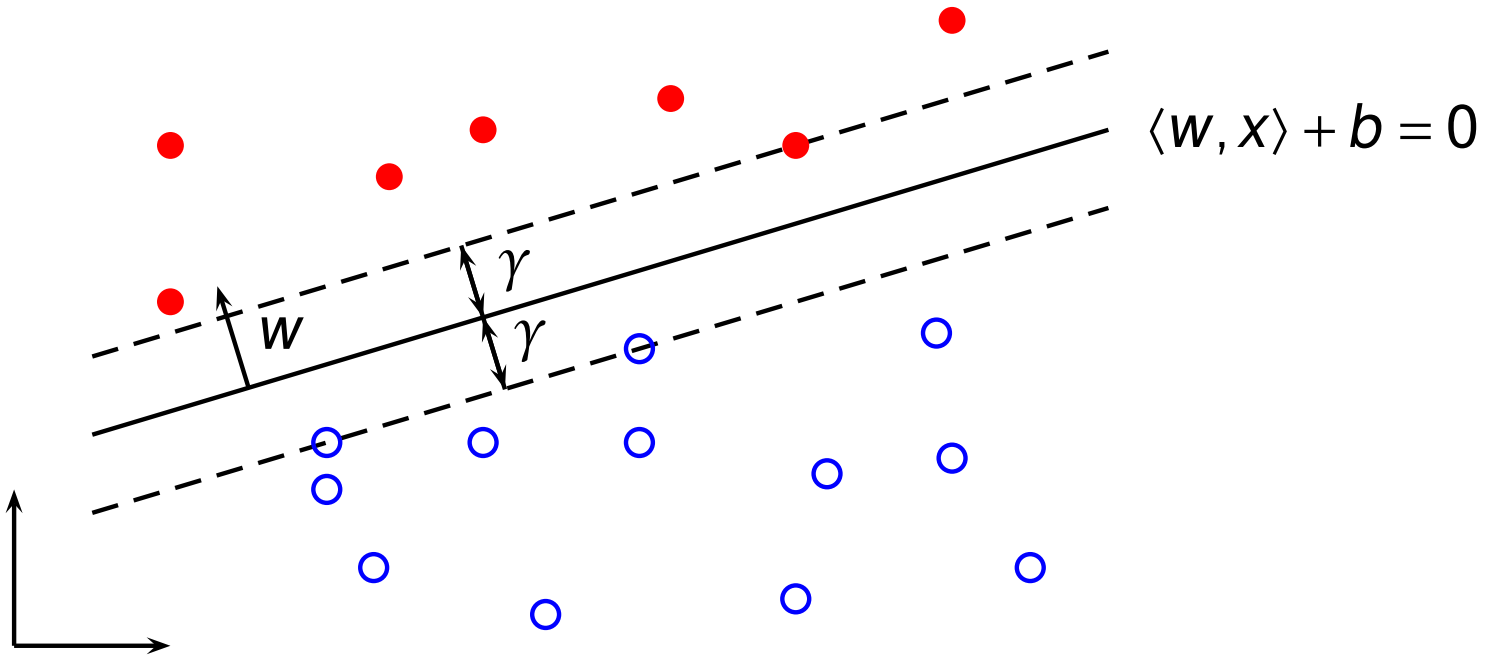
La fonction $\phi_{w,b}(x) = 1_{\langle w,x \rangle + b \geq 0} - 1_{\langle w,x \rangle + b < 0}$ est une règle de discrimination linéaire potentielle.

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



- Laquelle choisir ?

- **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 :

$$\gamma = \frac{1}{2} \frac{\langle w, x_1 - x_{-1} \rangle}{\|w\|}.$$

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 $x_1, \ldots, x_k$ si $\min_{i=1\ldots k} |\langle w, x_i \rangle + b| = 1$.

L'hyperplan séparateur est en forme canonique relativement aux vecteurs $\{x_1, x_{-1}\}$ s'il est défini par (w, b) avec $\langle w, x_1 \rangle + b = 1$ et $\langle w, x_{-1} \rangle + b = -1$. On a alors $\langle w, x_1 - x_{-1} \rangle = 2$, 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

$$\|w\|^2 \text{ soit minimal}$$
$$\text{sous la contrainte}$$
$$y_i(\langle w, x_i \rangle + b) \geq 1 \text{ pour tout } i.$$

- Problème d'optimisation convexe sous contraintes linéaires

- Existence d'un optimum global, obtenu par résolution du problème "dual" (méthode des multiplicateurs de Lagrange).

La solution du problème est donnée par :

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

Avec :

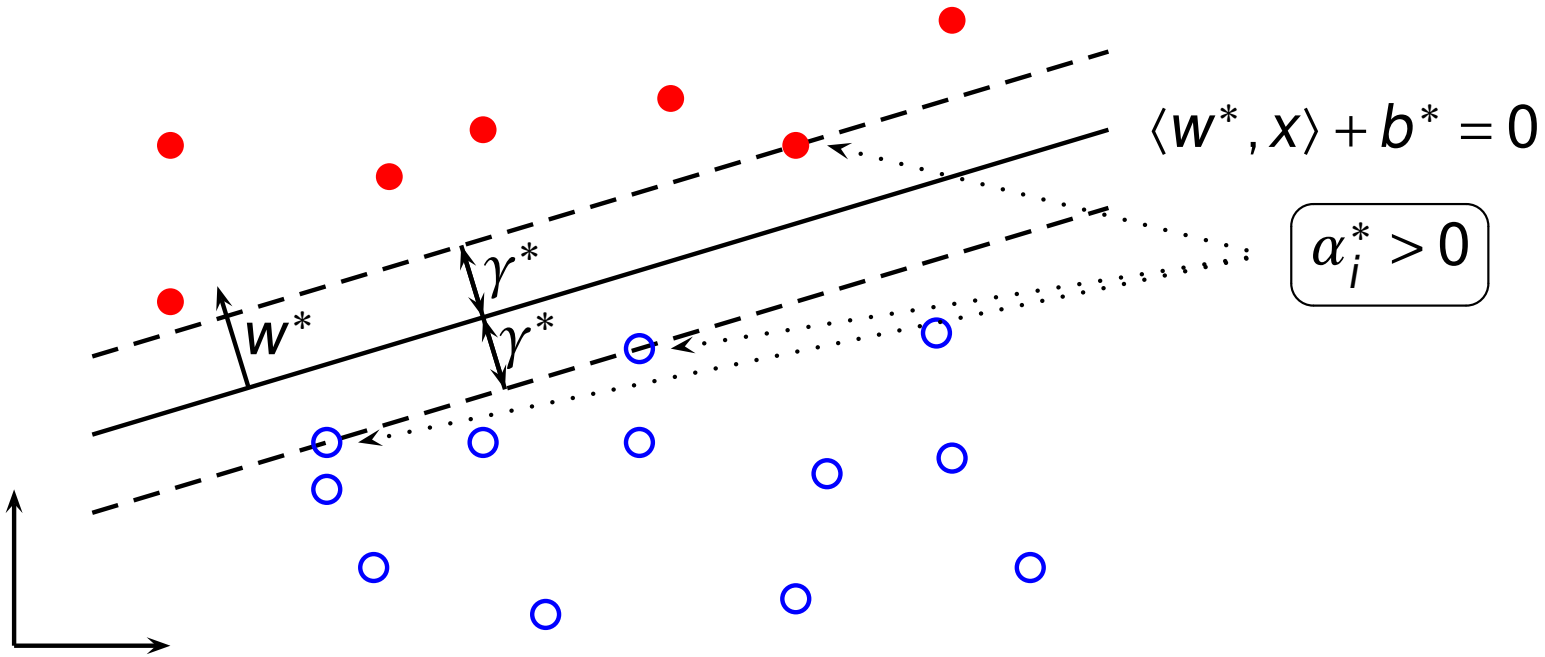
- $\alpha_i^* \geq 0 \ \forall i = 1 \ldots n$.
- $y_i \left(\langle w^*, x_i \rangle + b^*\right) \geq 1 \ \forall i = 1 \ldots n$.
- $\alpha_i^* \left(y_i \left(\langle w^*, x_i \rangle + b^*\right) - 1\right) = 0 \ \forall \ i = 1 \ldots n$.

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

• 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. $y_i \left(\langle w^*, x_i \rangle + b^*\right) = 1$.

Représentation des vecteurs supports



Pour conclure, l'algorithme est défini par :

avec

$$\phi_{d_1^n}(x) = 1_{\langle w^*, x\rangle + b^* \geq 0} - 1_{\langle w^*, x\rangle + b^* < 0},$$

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

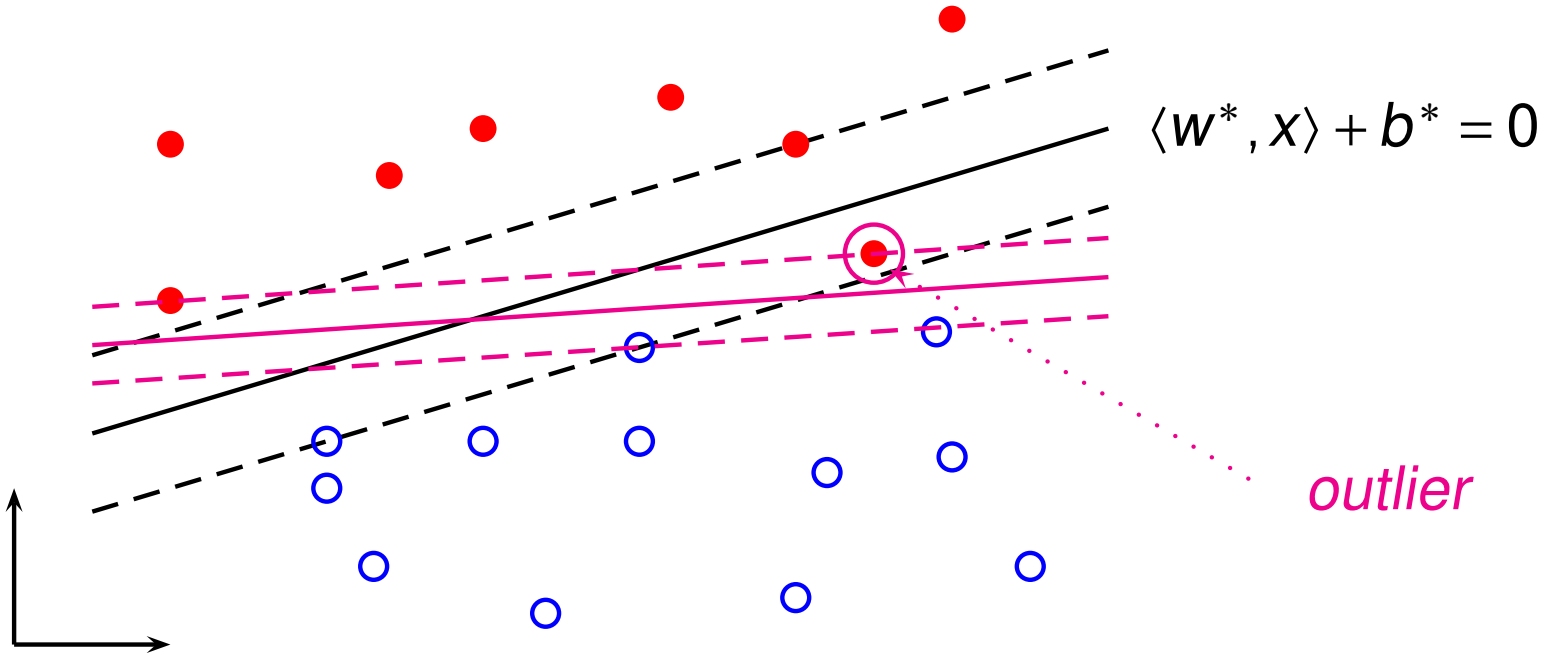
ou encore :

$$\phi_{d_1^n}(x) = 1_{\sum_{x_i V.S.} \alpha_i^* y_i \langle x_i, x\rangle + b^* \geq 0} - 1_{\sum_{x_i V.S.} \alpha_i^* y_i \langle x_i, x\rangle + b^* < 0}.$$

La marge maximale vaut $\gamma^* = \frac{1}{\|w^*\|}$.

SVM linéaire pour des données non séparables

- Méthode précédente non applicable si les données ne sont pas linéairement séparables
- Méthode très sensible aux "outliers"



- La solution : autoriser quelques vecteurs à être bien classés mais dans la région définie par la marge, voire mal classés.

La contrainte $y_i(\langle w, x_i \rangle + b) \geq 1$ devient $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$, 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**).

Problème d'optimisation primal

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

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

La solution du problème est donnée par :

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

Avec les conditions suivantes :

- $0 \leq \alpha_i^* \leq C \ \forall i = 1 \ldots n.$
- $y_i \left(\langle w^*, x_i \rangle + b^* \right) \geq 1 - \xi_i^* \ \forall i = 1 \ldots n.$
- $\alpha_i^* \left(y_i \left(\langle w^*, x_i \rangle + b^* \right) + \xi_i^* - 1 \right) = 0 \ \forall \ i = 1 \ldots n.$
- $\xi_i^* (\alpha_i^* - C) = 0.$

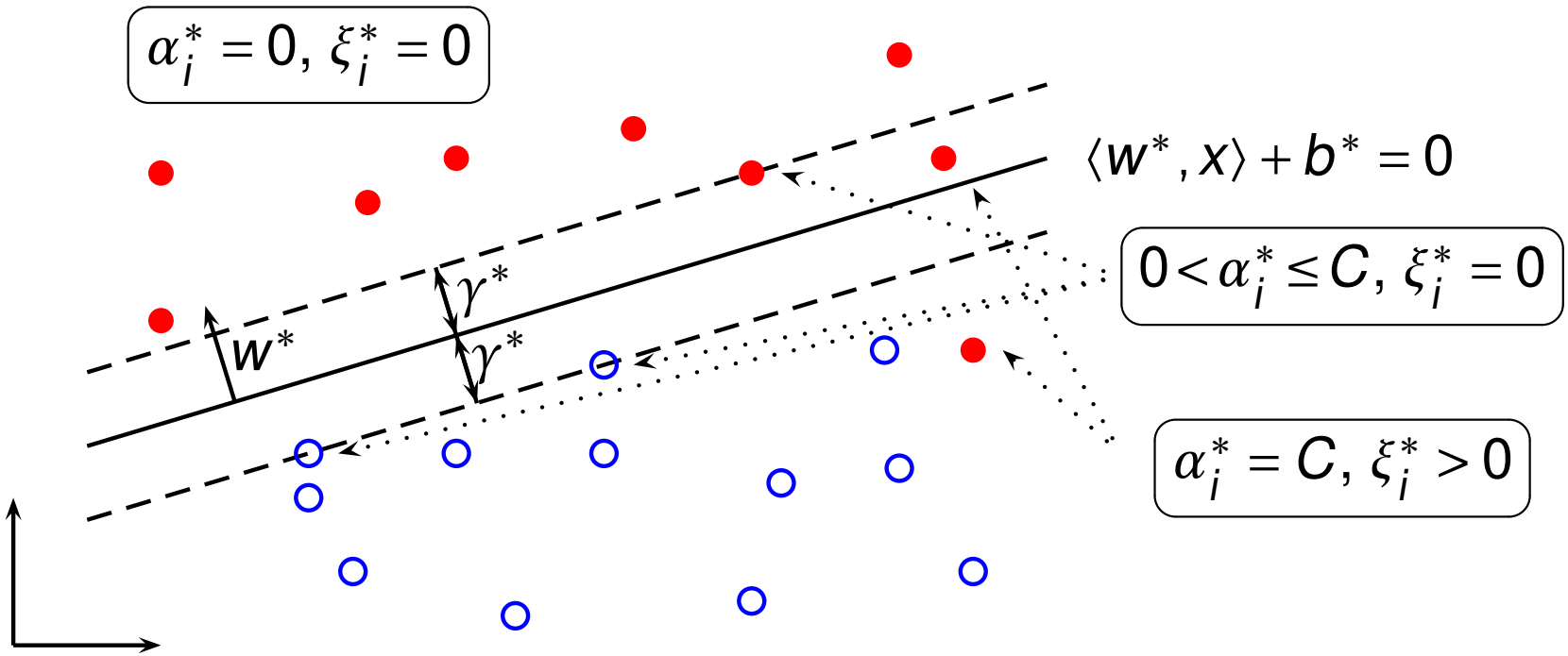
Les x_i tels que $\alpha_i^* > 0$ sont les **vecteurs supports**.

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 $\alpha_i^* = C$.

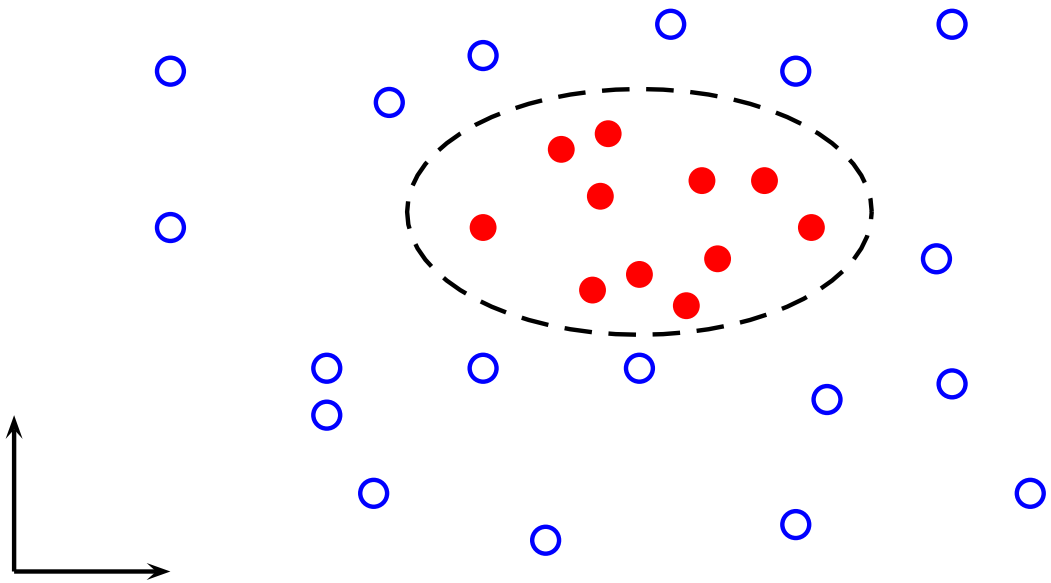
Les vecteurs qui ne sont pas supports vérifient $\alpha_i^* = 0$ et $\xi_i^* = 0$.

Représentation des vecteurs supports



SVM non linéaire : astuce du noyau

Exemple de données difficiles à discriminer linéairement :



- Une SVM linéaire donnera une très mauvaise discrimination avec un nombre de vecteurs supports très élevé ⇒ SVM non linéaire ?

Remarque fondamentale :

La règle de discrimination de la SVM linéaire ne dépend que de produits scalaires avec les x_i :

$$\langle w^*, x\rangle + b^* = \langle \sum_{i=1}^{n} \alpha_i^* y_i x_i, x\rangle + b^* = \sum_{i=1}^{n} \alpha_i^* y_i \langle x_i, x\rangle + b^*$$

- **Astuce du noyau (kernel trick)** : on remplace le produit scalaire $\langle x, x'\rangle$ par une fonction k définie par $k(x,x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}}$ permettant de lancer la SVM sans déterminer explicitement φ.

Une fonction k $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que $k(x,x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}}$ pour une fonction φ $\mathcal{X} \to \mathcal{H}$ donnée est appelée un **noyau**.

Quelques noyaux classiques pour $\mathscr{X} = \mathbb{R}^d$

- Noyau **polynomial** : $k(x, x') = (\langle x, x' \rangle + c)^p$
  $\hookrightarrow \varphi(x) = (\varphi_1(x), \ldots, \varphi_k(x))$ avec $\varphi_i(x)$=monôme de degré p de certaines composantes de x.
- Noyau **gaussien** ou **radial** (RBF) : $k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$
  $\hookrightarrow \varphi$ à valeurs dans un espace de dimension infinie.
- Noyau **laplacien** : $k(x, x') = e^{-\frac{\|x-x'\|}{\sigma}}$.

Fonction de décision dans le cas des noyaux :

$$sign(\langle \varphi(w^*), \varphi(x) \rangle_{\mathscr{H}} + b^*) = sign(\sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x) \rangle + b^*)$$

Noyaux pour $\mathcal{X} \neq \mathbb{R}^d$

Quelques noyaux ont été proposés pour d'autres types d'objets comme des

- ensembles,
- arbres,
- graphes,
- chaînes de symboles,
- documents textuels...

Conclusion / questions ouvertes

- Les règlages à effectuer :
  - Le noyau et ses paramètres (validation croisée, bootstrap ?)
  - La constante de tolérance C (validation croisée, bootstrap ?)
- Généralisation à la discrimination multiclasses : one-versus-all, one-versus-one ?