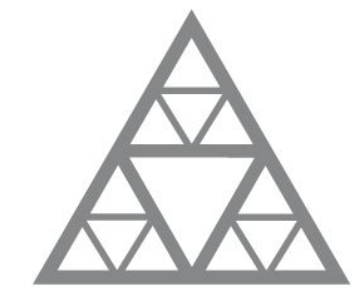
Pratique de l'apprentissage statistique

8. SVM

V. Lefieux



École des Ponts
ParisTech

Plan

Plan

Généralités I

- Les SVM (Support Vector Machine) (en français : *séparateurs à vaste marge* ou *machines à vecteurs supports*) sont issus de la théorie de Vapnik-Tchervonenkis (dénommée théorie VC) : (Cortes et Vapnik, 1995), (Vapnik, 1995).
- L'objectif historique des SVM est de classifier une variable binaire via un hyperplan de marge maximale, les SVM constituent une généralisation des classifieurs linéaires.
- Les SVM intègrent le contrôle de la complexité, ce qu'on peut appréhender via la dimension de Vapnik-Tchervonenkis qui est un indicateur du pouvoir séparateur d'une famille de fonctions.
- C'est une méthode souvent utilisée en pratique au vu des bons résultats obtenus.

Généralités II

- On parle de marge (*hard margin*) lorsque les données sont linéairement séparables et de marge souple (*soft margin*) lorsque les données ne le sont pas.
- Dans le cas où les données ne sont pas linéairement séparables, on utilise ce qu'on appelle l'astuce du noyau (*kernel trick*).
- Il existe également les SVR dans le cadre de la régression.

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}} \ .$$

  On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

- On considère dans la suite que :
  - $X \in \mathbb{R}^p$:
    Toutes les covariables sont considérés quantitatives.
    Mais il est également possible de considérer des covariables qualitatives.
  - $Y \in \{-1, 1\}$:
    On se place dans le cadre d'une classification supervisée binaire

Généralités sur les hyperplans I

- Dans $\mathbb{R}^p$, un hyperplan $\mathcal{H}$ admet comme équation :

$$w_0 + w_1 x_1 + \ldots + w_p x_p = 0 \ ,$$

ce qu'on peut noter également :

$$w_0 + \langle w, x \rangle = 0$$

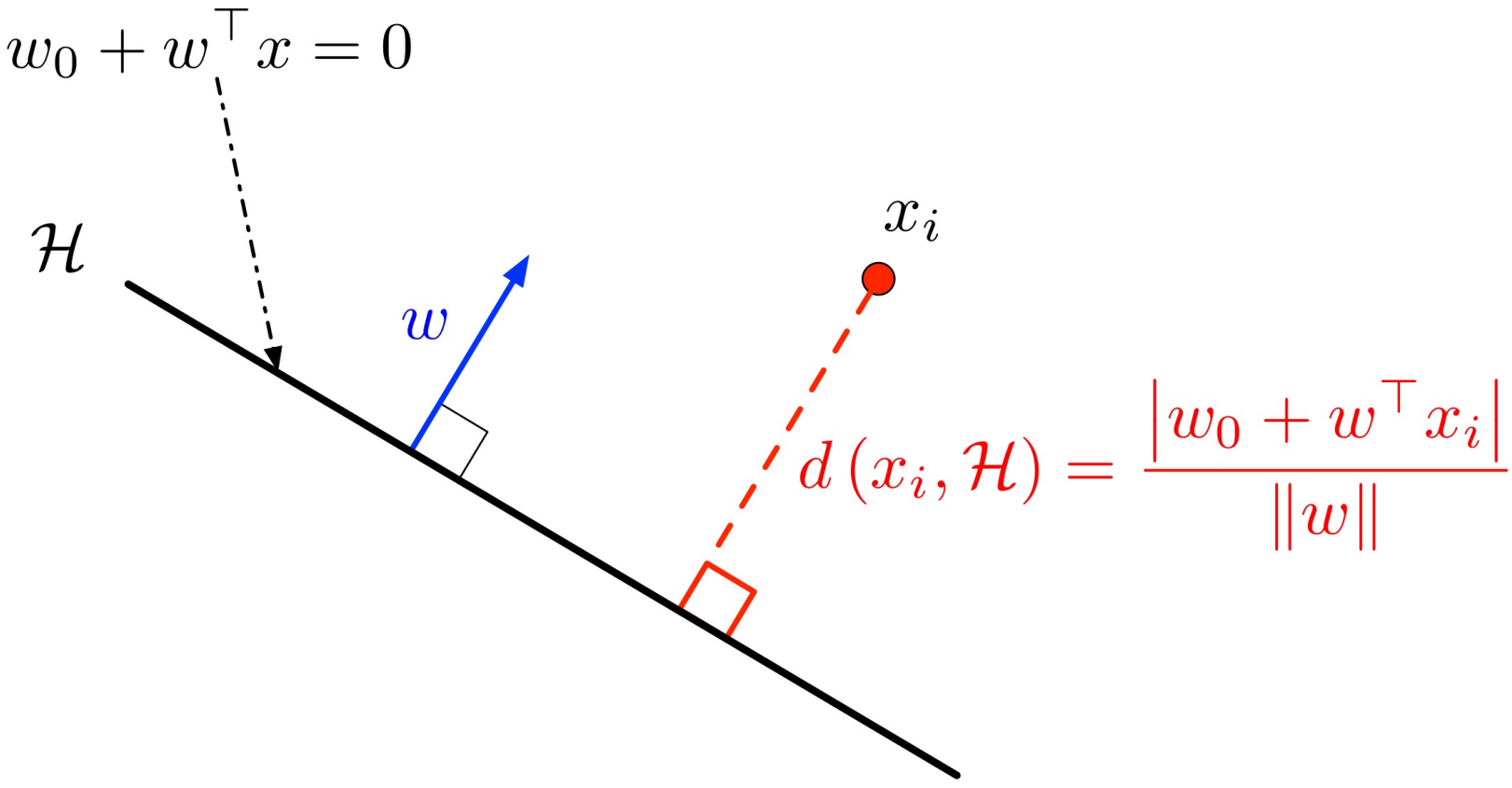
ou encore :

$$w_0 + w^\top x = 0$$

où $w = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p$ et $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$.

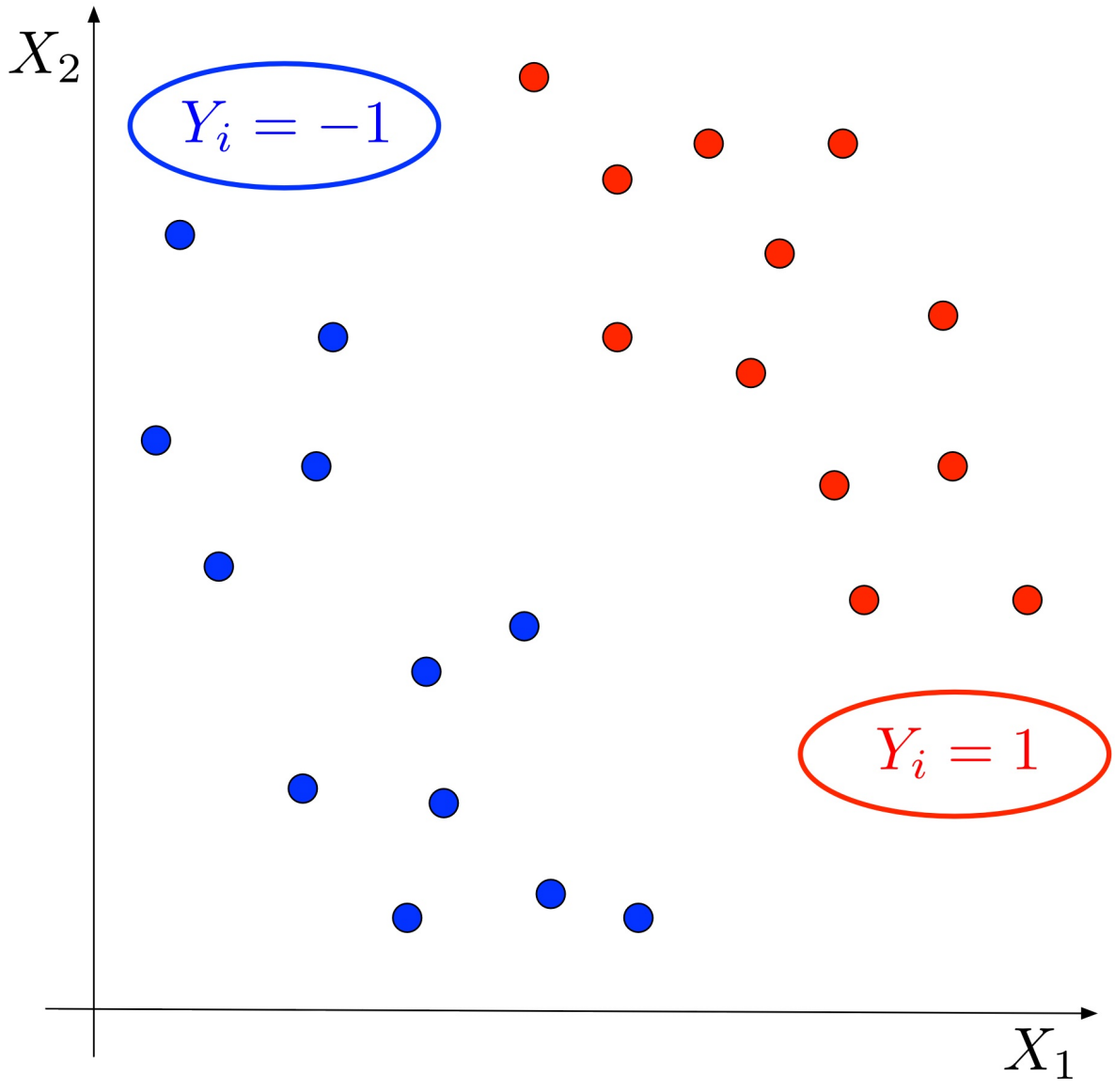
- w est le vecteur normal de l'hyperplan $\mathcal{H}$.
- Par exemple : un hyperplan dans $\mathbb{R}^2$ est une droite, un hyperplan dans $\mathbb{R}^3$ est un plan.

Généralités sur les hyperplans II



Plan

Données linéairement séparables I



Données linéairement séparables II

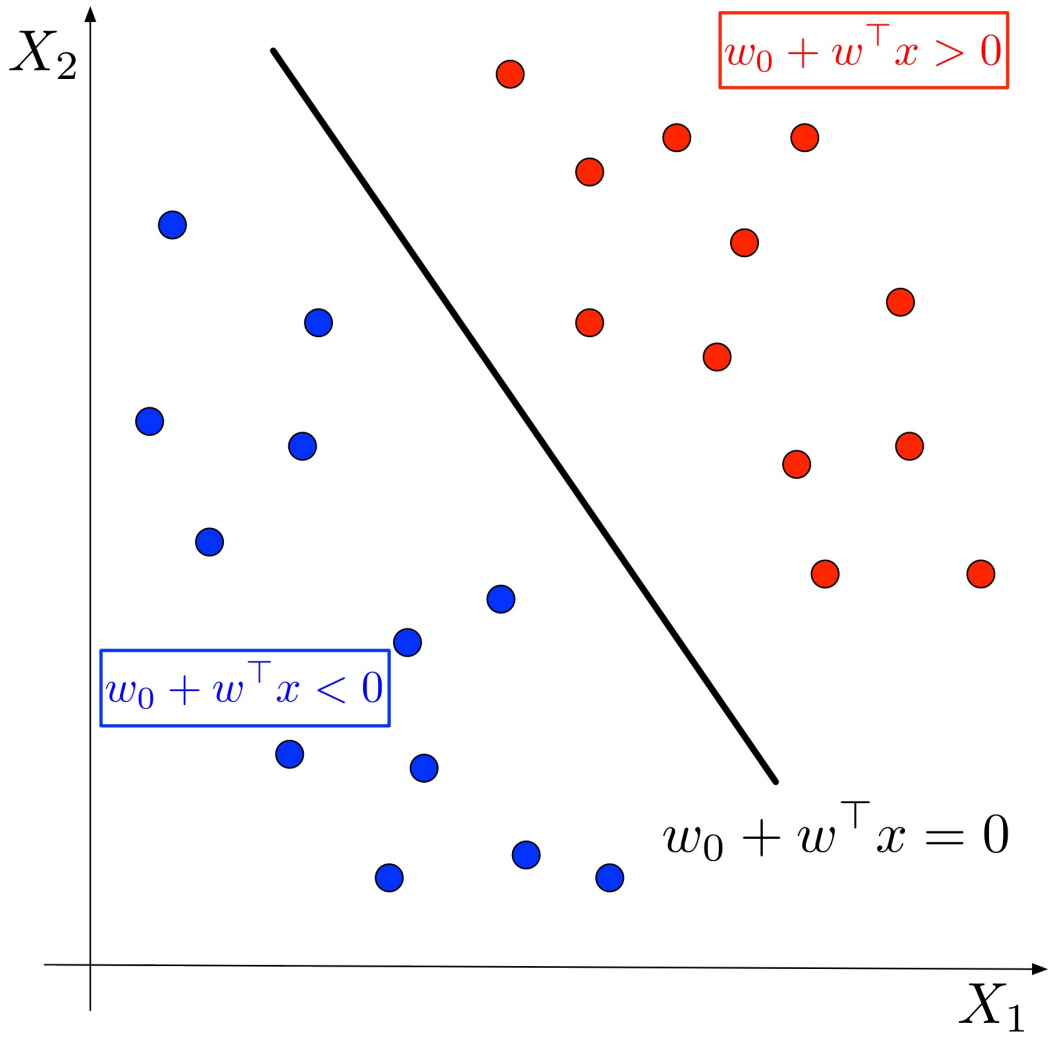
- On dit que $(x_1, y_1), \ldots, (x_n, y_n)$ sont linéairement séparables s'il existe $(w_0, w) \in \mathbb{R} \times \mathbb{R}^d$ tels que :

$$\forall i \in \{1, \ldots, n\} : y_i = \begin{cases} 1 & \text{si } w_0 + w^\top x_i > 0 \\ -1 & \text{si } w_0 + w^\top x_i < 0 \end{cases} .$$

- Cette propriété est équivalente à :

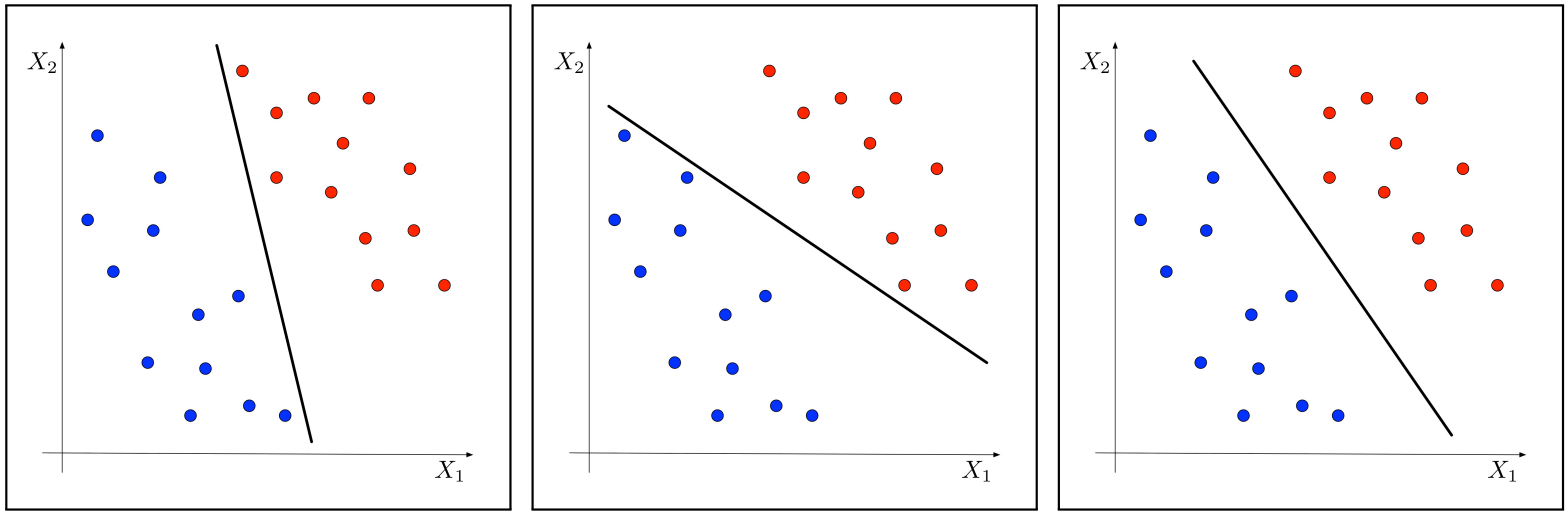
$$y_i \left(w_0 + w^\top x \right) > 0 .$$

Données linéairement séparables III



Le choix de l'hyperplan séparateur

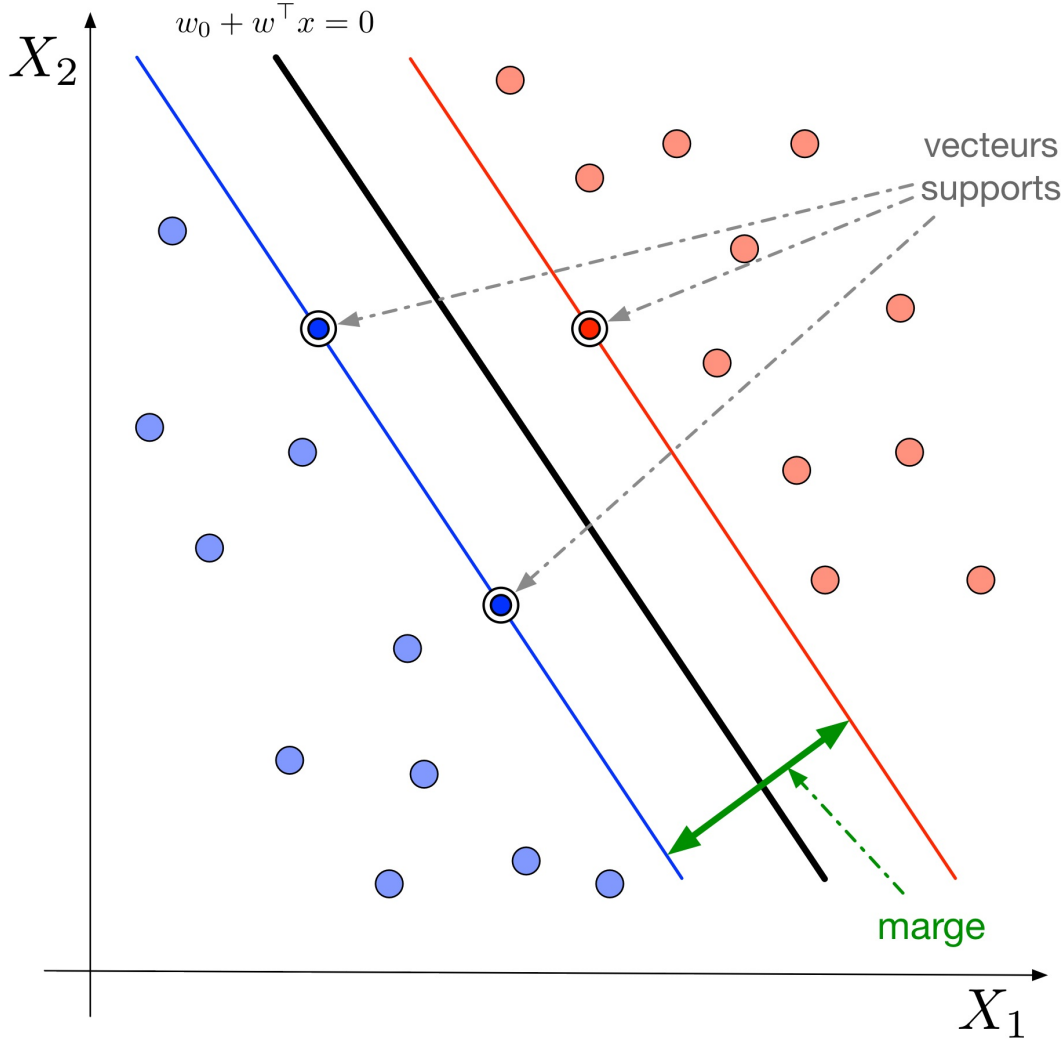
- Il existe une infinité d'hyperplans séparateurs possibles :



- Vapnik a proposé de maximiser la marge, soit la distance minimale entre les 2 classes déterminées par l'hyperplan séparateur.

- Les points pour lesquels la distance minimale est observée sont appelés vecteurs supports.

Marge et vecteurs supports



Formalisation du problème I

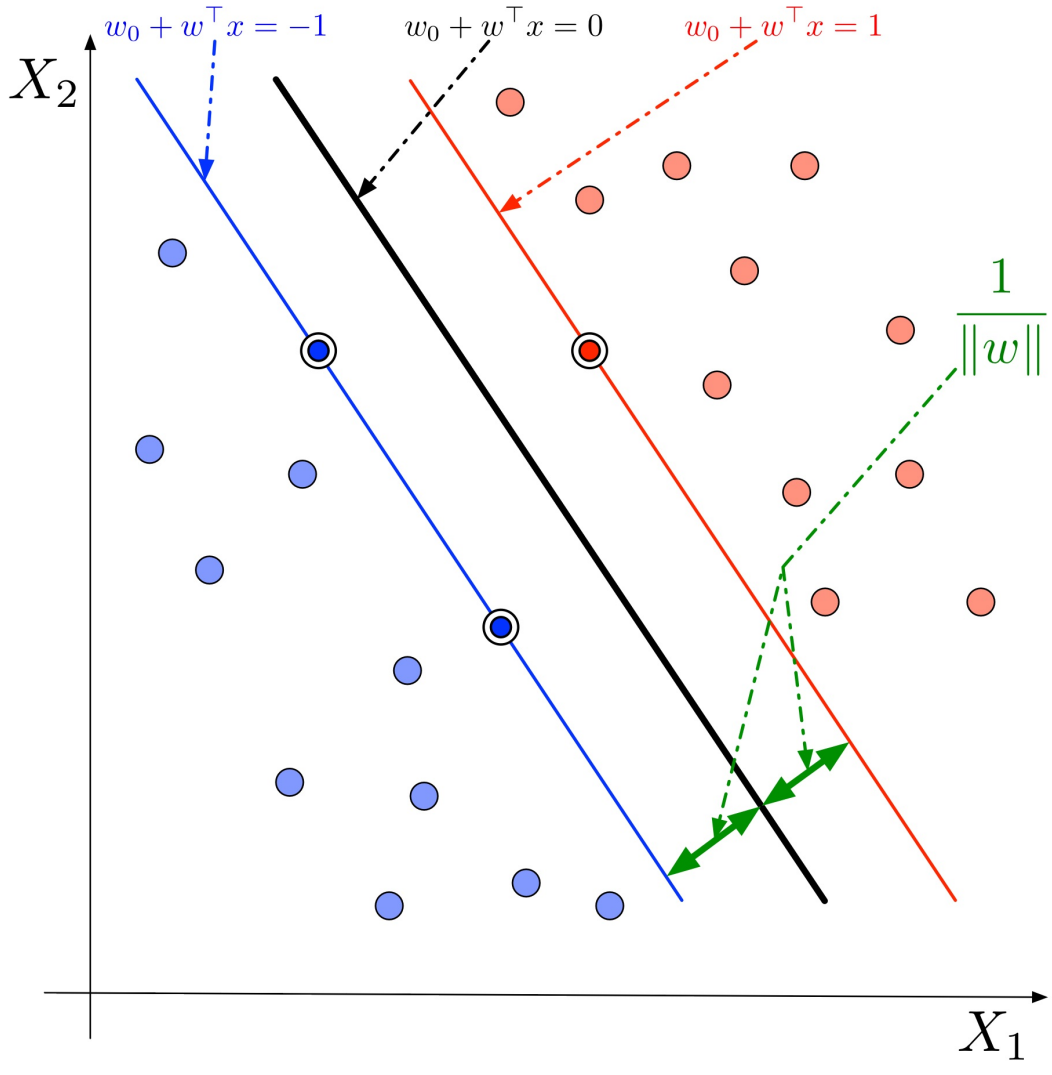
- On pose comme contrainte que les vecteurs supports sont situés sur les hyperplans canoniques d'équations :

$$\begin{cases} w_0 + w^\top x = -1 \\ w_0 + w^\top x = 1 \end{cases} .$$

- La marge vaut dans ce cas :

$$\frac{2}{\|w\|} .$$

Formalisation du problème II



Formalisation du problème III

- On obtient donc le problème suivant :

$$\max_{w_0, w} \frac{2}{\|w\|}$$
$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : y_i \left(w_0 + w^\top x_i \right) \geq 1 \ .$$

- Dans la suite, on considère le problème primal équivalent :

$$\min_{w_0, w} \frac{1}{2} \|w\|^2$$
$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : y_i \left(w_0 + w^\top x_i \right) \geq 1 \ .$$

- Le carré et la division par 2 ont comme seul objectif d'améliorer la lisibilité des résultats obtenus.
- Il s'agit d'un programme d'optimisation quadratique classique.

Résolution du problème I

- On considère le lagrangien du problème primal :

$$\mathcal{L}(w_0, w; \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(w_0 + w^\top x_i \right) - 1 \right]$$

où $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$.

- Le lagrangien doit être minimisé par rapport à w_0 et w, et maximisé par rapport à α.

Résolution du problème II

- Les conditions de Karush-Kuhn-Tucker (conditions KKT) sont les conditions que doivent vérifier un problème d'optimisation afin que la solution soit optimale.
- Les conditions KKT du problème primal sont :

$$\frac{\partial \mathcal{L}(w_0, w, \alpha)}{\partial w_0} = 0 \ ,$$

$$\forall j \in \{1, \ldots, p\} : \frac{\partial \mathcal{L}(w_0, w, \alpha)}{\partial w_j} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}(w_0, w, \alpha)}{\partial \alpha_i} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ .$$

Résolution du problème III

- On obtient :

$$w = \sum_{i=1}^{n} \alpha_i \, x_i \, y_i \ , \tag{1}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ , \tag{2}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \left[y_i \left(w_0 + w^\top x_i \right) - 1 \right] = 0 \ , \tag{3}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ . \tag{4}$$

- L'équation (3) implique que $\alpha_i = 0$ pour tous les points qui ne sont pas des vecteurs supports, donc tels que :

$$y_i \left(w_0 + w^\top x_i \right) - 1 > 0 \ .$$

- Donc, seuls les vecteurs supports participent à la définition de l'hyperplan optimal, ce qui diminue la complexité du problème.

Résolution du problème IV

- En substituant les équations (1) et (2) dans le lagrangien du problème primal, on obtient le lagrangien du problème dual :

$$\mathcal{L}_{dual}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i^\top \, x_j \ .$$

- On obtient au final le problème dual suivant :

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

$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$

$$\sum_{i=1}^{n} \alpha_i \, y_i = 0 \ .$$

Résolution du problème V

- Les conditions KKT du problème dual sont :

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}_{dual}(\alpha_i)}{\partial \alpha} = 0 \ ,$$
$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$
$$\sum_{i=1}^{n} \alpha_i \, y_i = 0 \ .$$

- Soit encore :

$$\forall i \in \{1, \ldots, n\} : \alpha_i^\star \left[y_i \left(w_0^\star + w^{\star\top} x_i \right) - 1 \right] = 0 \ , \quad (5)$$
$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ , \quad (6)$$
$$\sum_{i=1}^{n} \alpha_i \, y_i = 0 \ . \quad (7)$$

Résolution du problème VI

- Il existe de nombreux algorithmes de résolution pour ce problème d'optimisation quadratique classique, parmi lesquels SMO, SimpleSVM et LASVM.
- La résolution s'effectue itérativement :
  1. On obtient la solution $(\alpha_1^\star, \ldots, \alpha_n^\star)$ du problème dual.
  2. On en déduit $w^\star = \sum_{i=1}^{n} \alpha_i^\star x_i y_i$.
  3. On détermine $w_0^\star$ en résolvant l'équation (5) (une des conditions KKT du problème dual) :

$$\alpha_i^\star \left[y_i \left(w_0^\star + w^{\star\top} x_i \right) - 1 \right] = 0$$

  pour un $\alpha_i^\star$ non nul.

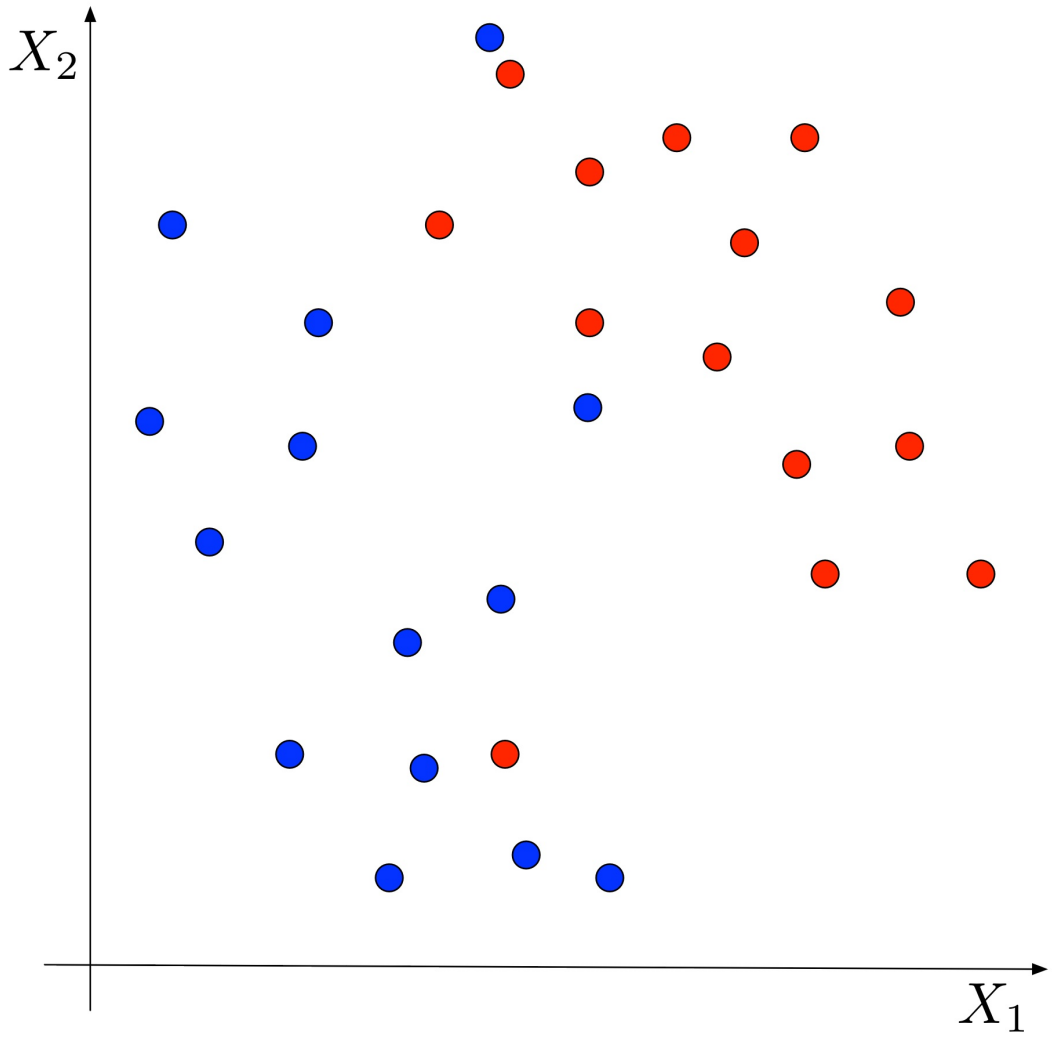
Règle de classification

La règle de classification obtenue est :

$$\forall x \in \mathbb{R}^p : g(x) = \begin{cases} 1 & \text{si } w_0^\star + w^{\star\top} x > 0 \\ -1 & \text{si } w_0^\star + w^{\star\top} x < 0 \end{cases}.$$

Plan

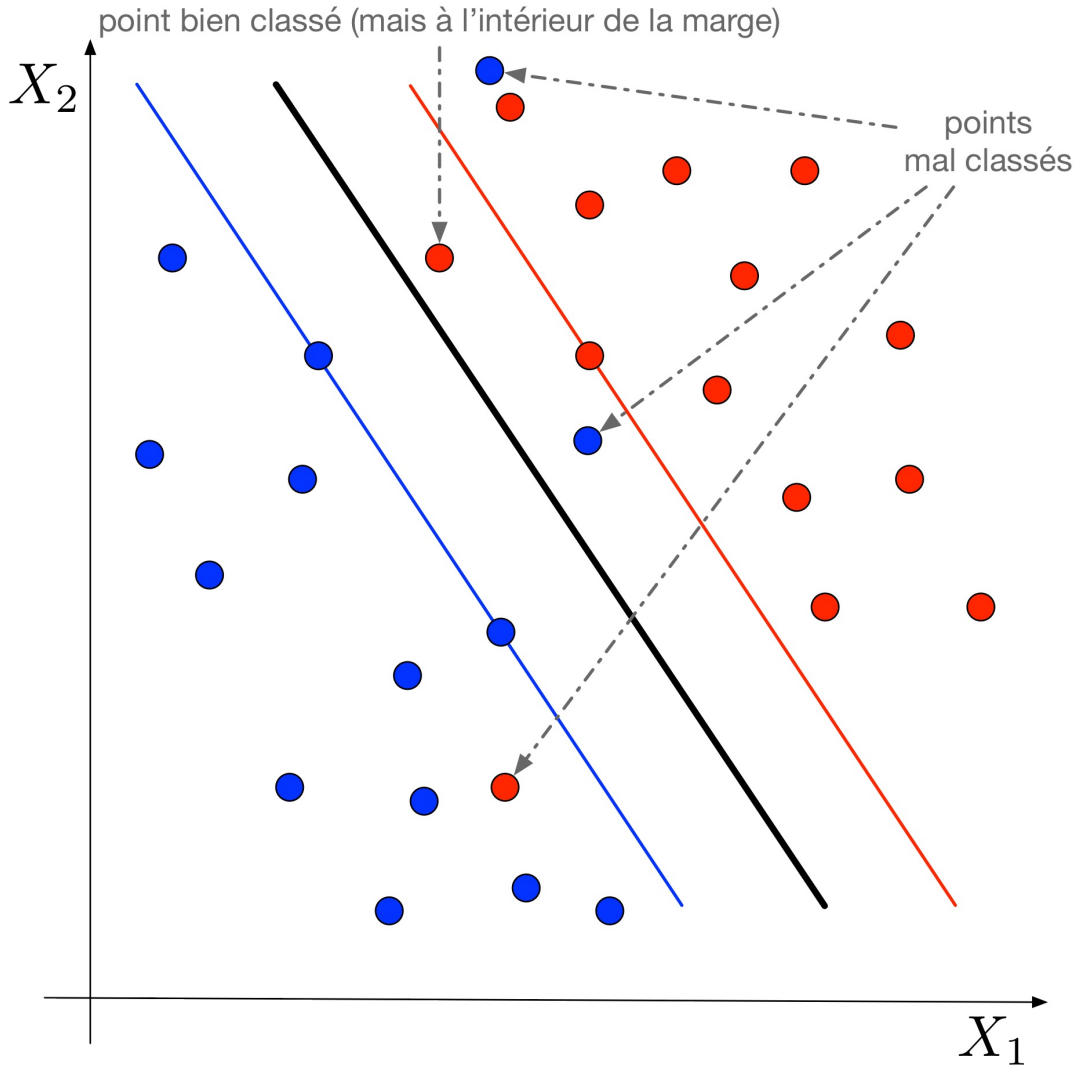
Exemple non-séparable



Lever les contraintes I

- Il est rare d'être confronté à un problème linéairement séparable.
- On lève la contrainte en tolérant que :
  - certains points soient bien classés mais à l'intérieur de la zone définie par la marge,
  - certains points soient mal classés.

Lever les contraintes II



Un outil : les variables ressorts I

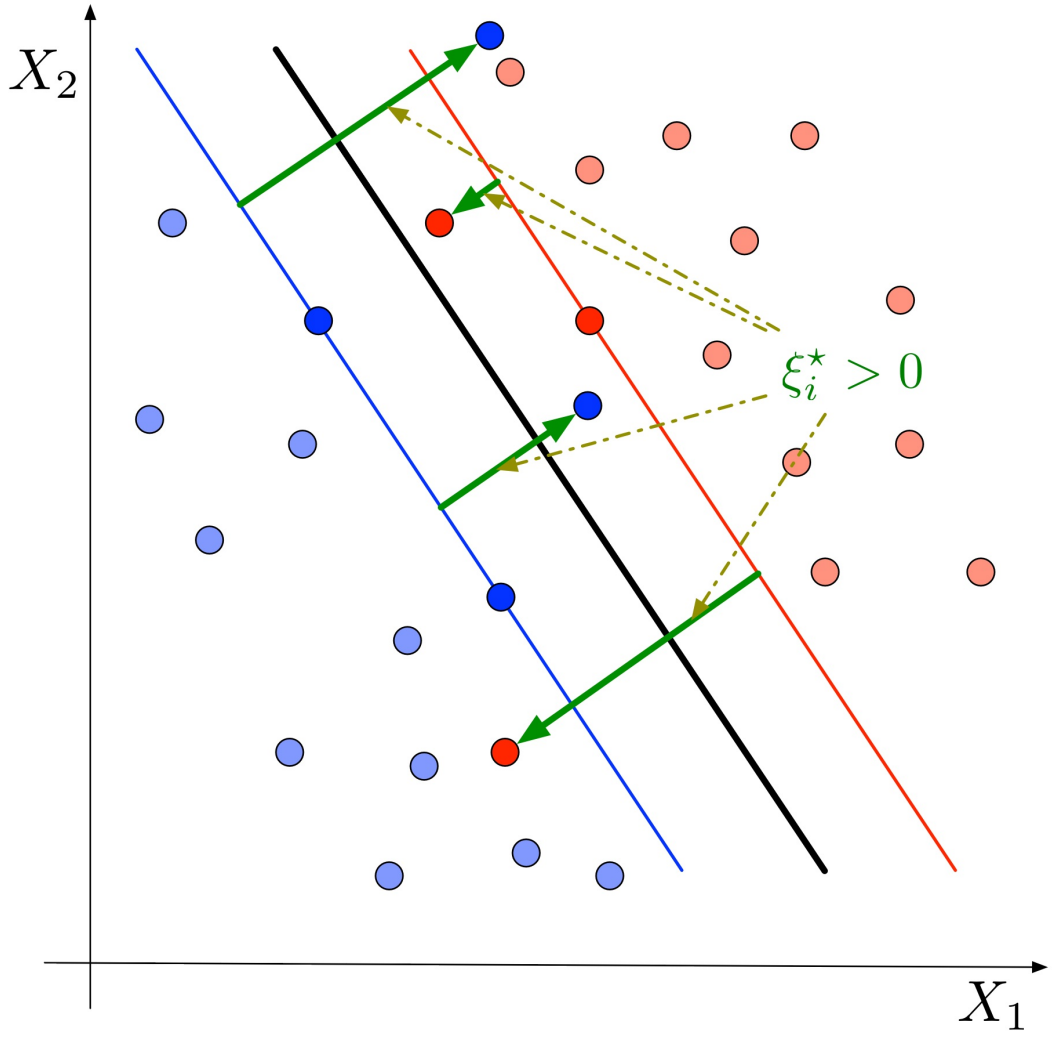
- On créé des variables ressorts (*slack variables*) $(\xi_1, \ldots, \xi_n)$ telles que :

$$y_i \left(w_0 + w^\top x_i \right) \geq 1 - \xi_i \ .$$

- On peut distinguer les cas suivants :
  - $\xi_i \in]0, 1]$: les points sont bien classés mais à l'intérieur (strictement) de la zone définie par la marge.
  - $\xi_i > 1$: les points sont mal classés.
  - $\xi_i = 0$: les points sont bien classés et à l'extérieur de la zone définie par la marge.

- L'enjeu est de ne pas pas avoir trop de variables ressorts non nulles (et lorsqu'elles le sont, qu'elles soient les plus faibles possibles).

Un outil : les variables ressorts II



Un nouveau problème I

On considérait le problème suivant :

$$\begin{aligned} & \min_{w_0, w} \ \frac{1}{2} \|w\|^2 \\ \text{sc} \quad & \forall i \in \{1, \ldots, n\} : y_i \left(w_0 + w^\top x \right) \geq 1 \ . \end{aligned}$$

Un nouveau problème II

On considère maintenant le problème suivant, avec $\xi = (\xi_1, \ldots, \xi_n)^\top$:

$$\min_{w_0, w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : y_i \left(w_0 + w^\top x \right) \geq 1 - \xi_i \ ,$$

$$\forall i \in \{1, \ldots, n\} : \xi_i \geq 0 \ .$$

Choix de l'hyper-paramètre C

- L'hyper-paramètre C contrôle de le compromis entre le nombre d'erreurs de classification et le niveau de la marge.
- Le cas linéairement séparable correspond à une valeur C infinie.
- On choisit l'hyper-paramètre C par validation croisée.

Résolution du problème I

On considérait le lagrangien du problème primal :

$$\mathcal{L}(w_0, w; \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(w_0 + w^\top x_i \right) - 1 \right]$$

où $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$.

Résolution du problème I

On considère maintenant le lagrangien du problème primal :

$$\begin{aligned}\mathcal{L}(w_0, w; \alpha, \beta) = & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ & - \sum_{i=1}^{n}\alpha_i\left[y_i\left(w_0 + w^\top x_i\right) - 1 + \xi_i\right] \\ & - \sum_{i=1}^{n}\beta_i\xi_i\end{aligned}$$

où $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ et $\beta = (\beta_1, \ldots, \beta_n)^\top$.

Résolution du problème II

- On obtient (conditions KKT) :

$$w = \sum_{i=1}^{n} \alpha_i x_i y_i \ , \tag{1}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ , \tag{2}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \left[y_i \left(w_0 + w^\top x_i \right) - 1 + \xi_i \right] = 0 \ , \tag{3}$$

$$\forall i \in \{1, \ldots, n\} : C - \alpha_i - \beta_i = 0 \ , \tag{4}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ , \tag{5}$$

$$\forall i \in \{1, \ldots, n\} : \beta_i \geq 0 \ . \tag{6}$$

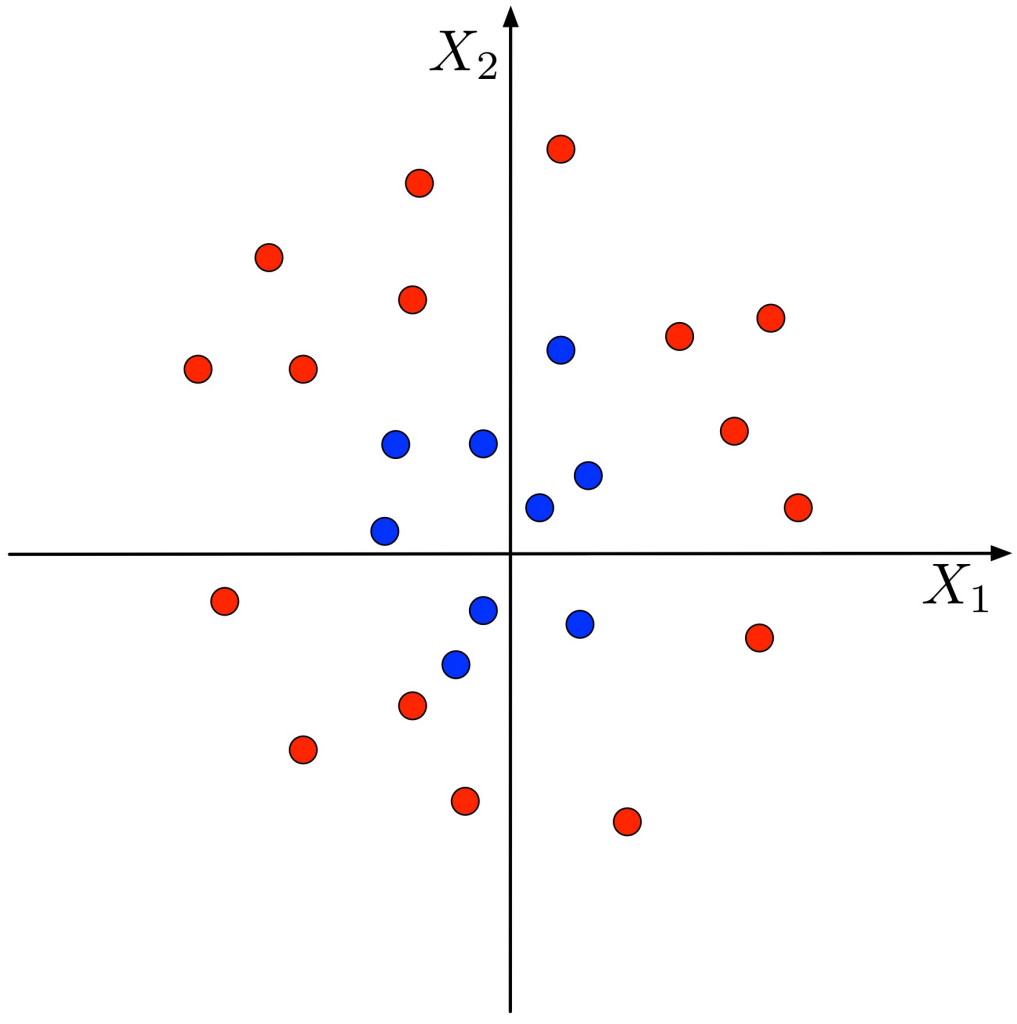
- Les équations (3) et (6) impliquent $\forall i \in \{1, \ldots, n\} : \alpha_i \leq C$.

Résolution du problème III

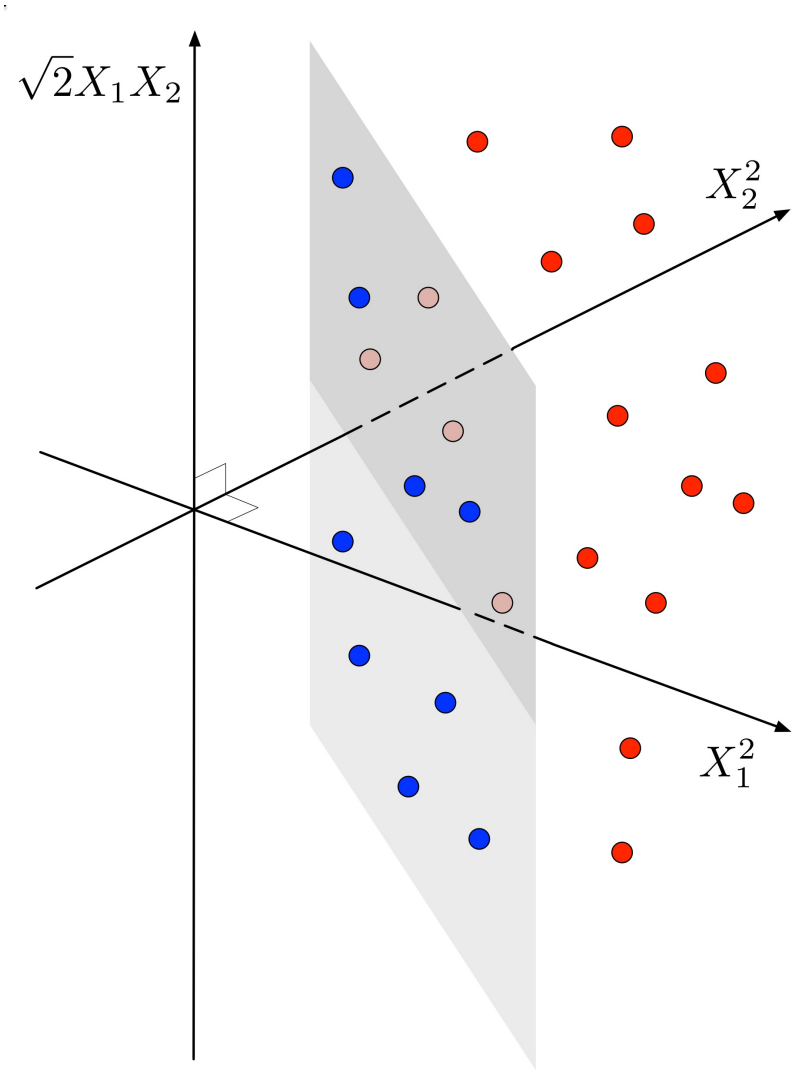
- La résolution s'effectue de manière analogue au cas linéairement séparable.
- On obtient des points :
  - vecteurs supports :
    - sur la frontière : $\xi_i^\star = 0$ et $\alpha_i^\star > 0$,
    - en dehors de la frontière : $\xi_i^\star > 0$ et $\alpha_i^\star = C$,
  - non vecteurs supports : $\xi_i^\star = 0$ et $\alpha_i^\star = 0$.

Plan

Changer la dimension I



Changer la dimension II



Astuce du noyau

- Déterminer un classifieur linéaire dans l'espace des observations n'est pas toujours opportun. On espère que la séparation linéaire sera plus simple dans un nouvel espace.
- On « envoie » les observations (dans l'espace $\mathcal{X}$) dans un nouvel espace $\mathcal{X}'$: l'espace de représentation (*feature space*).
- On considère pour cela une fonction Φ définie sur $\mathcal{X}$ et à valeurs dans $\mathcal{X}'$.
- Dans le problème d'optimisation des SVM, on retrouve les produits $x_i^\top x_j$ dans l'espace des observations, donc des produits $\Phi(x_i)^\top \Phi(x_j)$ dans l'espace de représentation.
- Il n'est pas nécessaire de déterminer Φ, on utilisera des noyaux K tels que :

$$K(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$$

pour i et j dans $\{1, \ldots, n\}$.

Retour au problème d'optimisation I

Dans le cas linéairement séparable, on devait résoudre le problème dual suivant dans l'espace des observations :

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

$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ .$$

Retour au problème d'optimisation II

Dans le cas linéairement séparable, on doit maintenant résoudre le problème dual suivant dans l'espace de représentation :

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

$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ .$$

Retour au problème d'optimisation III

Dans le cas linéairement séparable, on doit résoudre le problème dual suivant dans l'espace de représentation :

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

$$\text{sc} \quad \forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ .$$

Noyau

Une fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau si et seulement si :

- K est une fonction symétrique :

$$\forall \left(x, x'\right) \in \mathcal{X} \times \mathcal{X} : K\left(x, x'\right) = K\left(x', x\right) \ .$$

- K est une fonction semi-définie positive :

$$\forall n \in \mathbb{N}^{\star}, \forall \left(x_1, \ldots, x_n\right) \in \mathcal{X}^n, \forall \left(a_1, \ldots, a_n\right) \in \mathbb{R}^n :$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K\left(x_i, x_j\right) \geq 0 \ .$$

Un exemple de noyau

- Pour une observation $x_i = (x_{i1}, x_{i2})^\top$, on considère la fonction suivante :

$$\begin{array}{rccc} \Phi : & \mathbb{R}^2 & \to & \mathbb{R}^3 \\ & (x_{i1}\,,\, x_{i2})^\top & \mapsto & \left(x_{i1}^2\,,\, \sqrt{2}\, x_{i1}\, x_{i2}\,,\, x_{i2}^2\right)^\top \end{array}$$

- On peut montrer que pour 2 observations x_i et x_j :

$$\begin{aligned} K(x_i, x_j) &= \Phi(x_i)^\top \Phi(x_j) \\ &= (x_{i1}x_{j1})^2 + 2(x_{i1}x_{j1})(x_{i2}x_{j2}) + (x_{i2}x_{j2})^2 \\ &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 \\ &= \left(x_i^\top x_j\right)^2 \ . \end{aligned}$$

Quelques noyaux (parmi bien d'autres)

- Noyau affine :

$$K(x_i, x_j) = x_i^\top x_j + c \ .$$

- Noyau polynomial :

$$K(x_i, x_j) = \left(x_i^\top x_j + c\right)^d \ .$$

- Noyau laplacien :

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\sigma}\right) \ .$$

- Noyau gaussien (ou RBF : Radial Basis Function) :

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \ .$$

En pratique

On choisit :

- l'hyper-paramètre C,
- le noyau K,

par validation croisée.

Compléments

- Dans le cas où on dispose de $K > 2$ classes, on peut par exemple considérer K discriminations binaires « classe k » contre « classe autre que k » pour $k \in \{1, \ldots, K\}$.
- Il est également possible d'utiliser ces méthodes pour la régression : on parle alors de SVR : (Drucker et collab., 1997), (Vapnik et collab., 1997).

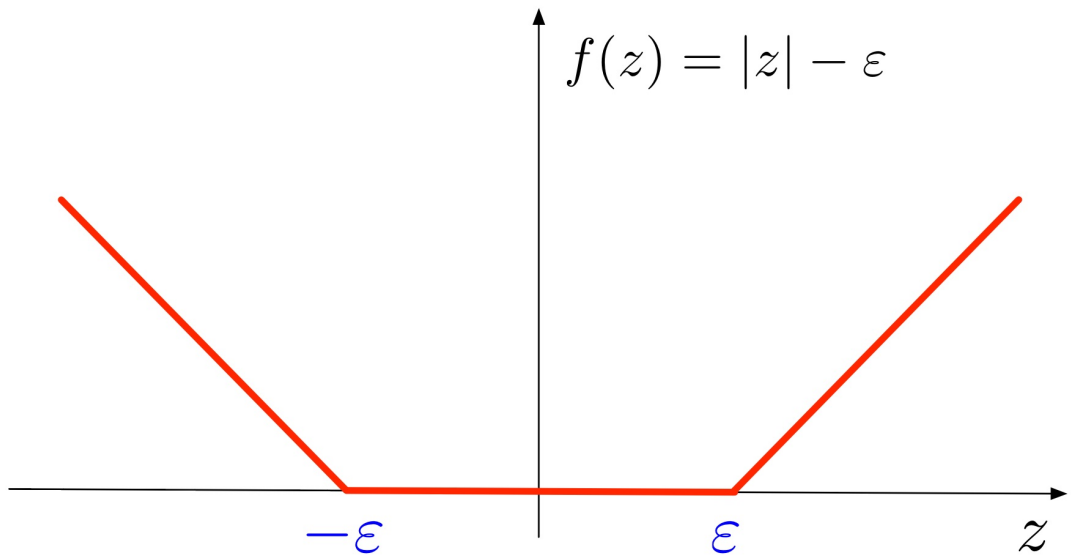
Plan

Fonction de perte

Vapnik a introduit la fonction de perte suivante (ε-*insensitive loss function*) pour mesurer la qualité de l'ajustement de la fonction de régression m :

$$\ell\left(m(x), y\right) = \begin{cases} |m(x) - y| - \varepsilon & \text{si } |m(x) - y| > \varepsilon \\ 0 & \text{sinon} \end{cases}$$

où $\varepsilon > 0$.



Risque empirique

Le risque empirique vaut :

$$R_n(m) = \sum_{i=1}^{n} \left(|m(x_i) - y_i| - \varepsilon \right) = \sum_{i=1}^{n} \left(\xi_i + \xi_i^\star \right)$$

où :

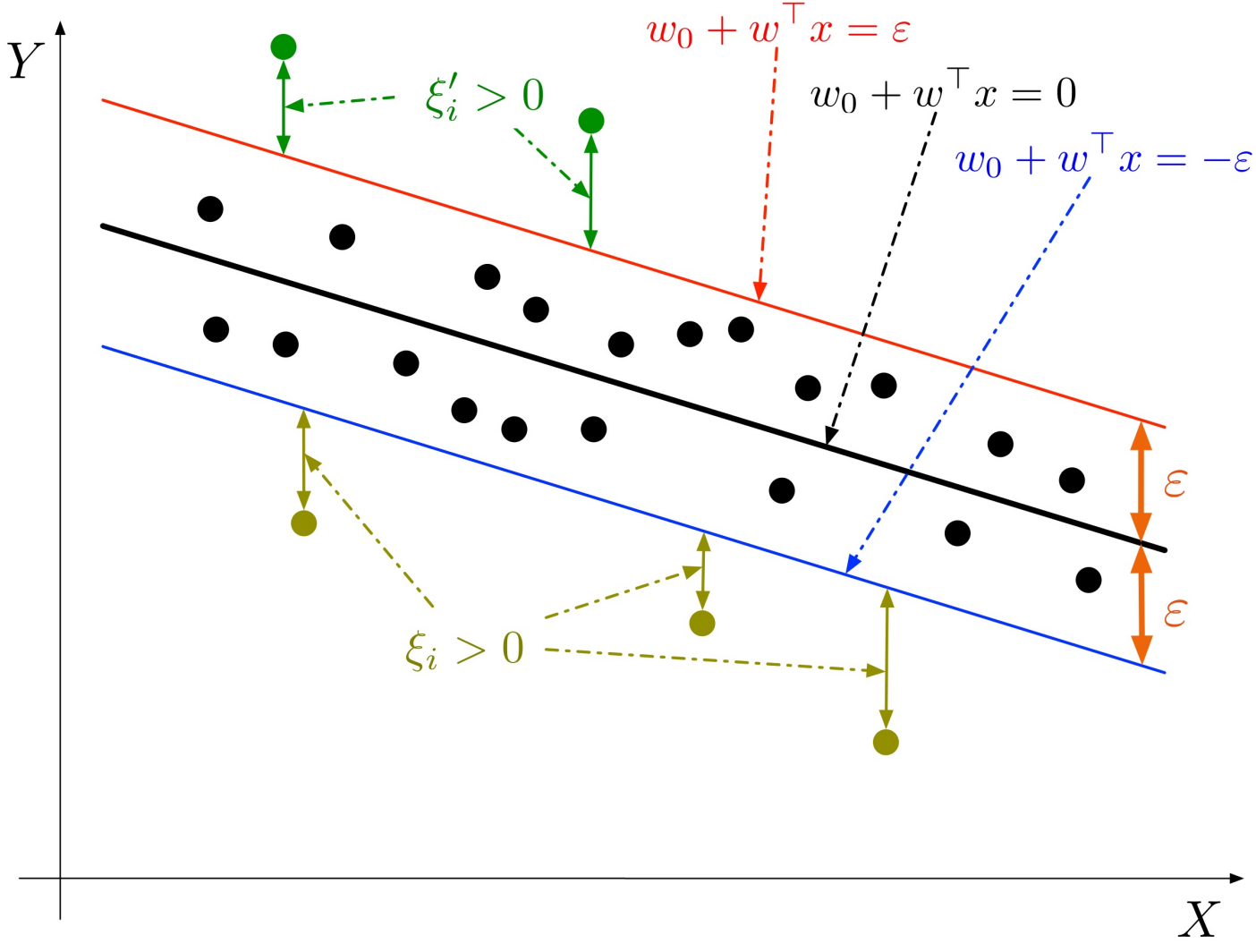
$$\begin{cases} \xi_i = m(x_i) - \varepsilon - y_i & \text{si } y_i < m(x_i) - \varepsilon \\ 0 & \text{sinon} \end{cases}$$

et :

$$\begin{cases} \xi_i^\star = y_i - m(x_i) - \varepsilon & \text{si } y_i > m(x_i) + \varepsilon \\ 0 & \text{sinon} \end{cases}$$

.

Cas linéaire I



Cas linéaire II

- On considère la fonction de régression :

$$\forall x \in \mathbb{R}^p : m_{w_0,w}(x) = w_0 + w^\top x \ .$$

- On cherche w_0 et w de manière à minimiser la somme de la perte qui traduit l'ajustement et d'un terme de régularisation (assurant la parcimonie) $\|w\|^2$.

- On considère le problème suivant :

$$\begin{aligned}
\min_{w_0,w} \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^\star) \\
\text{sc} \quad & \forall i \in \{1,\dots,n\} : m_{w_0,w}(x_i) - y_i \leq \varepsilon + \xi_i \ , \\
& \forall i \in \{1,\dots,n\} : y_i - m_{w_0,w}(x_i) \leq \varepsilon + \xi_i^\star \ , \\
& \forall i \in \{1,\dots,n\} : \xi_i \geq 0 \ , \\
& \forall i \in \{1,\dots,n\} : \xi_i^\star \geq 0 \ .
\end{aligned}$$

Résolution du problème I

On considère le lagrangien du problème primal :

$$\begin{aligned}\mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right) = & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i'\right) \\ & - \sum_{i=1}^{n}\alpha_i\left(\varepsilon + \xi_i - w_0 - w^\top x_i + y_i\right) \\ & - \sum_{i=1}^{n}\alpha_i'\left(\varepsilon + \xi_i' - y_i + w_0 + w^\top x_i\right) \\ & - \sum_{i=1}^{n}\beta_i\,\xi_i \\ & - \sum_{i=1}^{n}\beta_i'\,\xi_i' \ .\end{aligned}$$

Résolution du problème II

Les conditions KKT sont :

$$\frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial w_0} = 0 \ ,$$

$$\forall j \in \{1, \ldots, p\} : \frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial w_j} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial \alpha_i} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial \alpha'_i} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial \beta_i} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \frac{\partial \mathcal{L}\left(w_0, w; \alpha, \alpha', \beta, \beta'\right)}{\partial \beta'_i} = 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \alpha'_i \geq 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \beta_i \geq 0 \ ,$$

$$\forall i \in \{1, \ldots, n\} : \beta'_i \geq 0 \ .$$

Résolution du problème III

On obtient :

$$\sum_{i=1}^{n} \left(\alpha_i' - \alpha_i\right) = 0 \ , \tag{1}$$

$$\sum_{i=1}^{n} \left(\alpha_i' - \alpha_i\right) x_i = w \ , \tag{2}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i = C - \beta_i \ , \tag{3}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \prime = C - \beta_i' \ , \tag{4}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0 \ , \tag{5}$$

$$\forall i \in \{1, \ldots, n\} : \alpha_i' \geq 0 \ , \tag{6}$$

$$\forall i \in \{1, \ldots, n\} : \beta_i \geq 0 \ , \tag{7}$$

$$\forall i \in \{1, \ldots, n\} : \beta_i' \geq 0 \ . \tag{8}$$

Résolution du problème IV

- La résolution s'effectue de manière analogue au cas linéairement séparable :
  - On résout les conditions KKT du problème primal.
  - On en déduit le problème dual.
  - On résout le problème dual (via les conditions KKT).
- La résolution du problème dual conduit à la solution :

$$w^\star = \sum_{i=1}^{n} \left(\alpha_i'^\star - \alpha_i^\star\right) x_i \ .$$

- On peut là-encore utiliser travailler dans un espace de représentation via un noyau.

Choix des hyper-paramètres ε et C

- L'hyper-paramètre ε contrôle la largeur du « tube » : plus ε est important, moins on a de vecteurs support et plus lisse est l'estimation.
- L'hyper-paramètre C contrôle de le compromis entre l'erreur d'ajustement et le niveau de la marge. On le choisit par validation croisée.

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

- Le package `e1071` :
  - La fonction de base est `svm`.
  - L'option `kernel` permet de choisir le noyau.
  - L'option `cost` correspond à l'hyper-paramètre C.
- Le package `kernlab` qui offre un choix plus large de noyaux :
  - La fonction de base est `ksvm`.
  - L'option `kernel` permet de choisir le noyau.
  - L'option `C` correspond à l'hyper-paramètre C.
- Le package `caret` :
  - La fonction de base est `train`.
  - On utilise une des options correspondant aux SVM, par exemple :
    - `method="svmLinear"`,
    - `method="svmPoly",degree=2`.

Références I

Boyd, S. et L. Vandenberghe. 2003, *Convex optimization*, Cambridge University Press.

Cortes, C. et V. N. Vapnik. 1995, «Support-vector networks», *Machine Learning*, vol. 20, n° 3, p. 273–297.

Drucker, H., C. J. Burges, L. Kaufman, A. Smola et V. N. Vapnik. 1997, *Advances in Neural Information Processing Systems*, vol. 9, chap. Support vector regression machines, MIT Press, p. 155–161.

Schölkopf, B. et A. J. Smola. 2001, *Learning with Kernels. Support vector machines, regularization, optimization, and beyond*, MIT Press.

Vapnik, V. N. 1995, *The nature of statistical learning theory*, Springer.

Références II

Vapnik, V. N., S. E. Golowich et A. Smola. 1997, *Advances in Neural Information Processing Systems*, vol. 9, chap. Support vector method for function approximation, regression estimation, and signal processing, MIT Press, p. 281–287.