Support Vector Machines (SVM) et méthodes à Noyaux

Nicolas Verzelen, Joseph Salmon (Pierre Pudlo)

INRA / Université de Montpellier

Plan

SVM

SVM linéaire pour des données séparables

SVM linéaire pour des données non séparables

SVM non linéaire : astuce du noyau

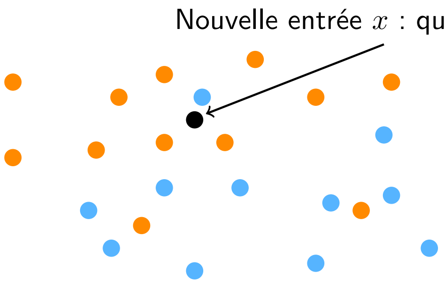
Lien avec la minimisation du risque empirique convexifié

Conclusion / questions ouvertes

Support Vector Machine, SVM

Machine à vecteurs supports (🇬🇧 : *Support Vector Machine*) : famille d'algorithmes d'apprentissage supervisé :classification (régression moins facile à voir) **TO DO: a la fin donner la forme Hinge**

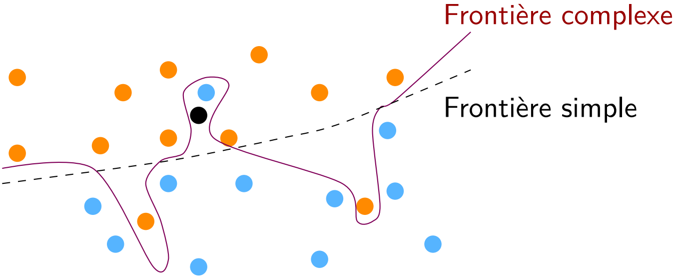
<u>Exemple</u> : classification binaire dimension 2 (orange : 1, bleu : -1)



Complexité de la frontière

- Fondements mathématiques solides ⇒ bonnes **propriétés de généralisation** Vapnik (1998)

Exemple : d'une règle de discrimination n'ayant pas de bonnes propriétés de généralisation



⚠ **sur-apprentissage (: *overfitting*)** phénomène fréquent en grande dimension

Plan

Données linéairement séparables

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

Définition

Les données observées $\mathcal{D}^n = (x_1, y_1), \ldots, (x_n, y_n)$ sont dites **linéairement séparables** s'il existe (w, w_0) tel que pour tout i,

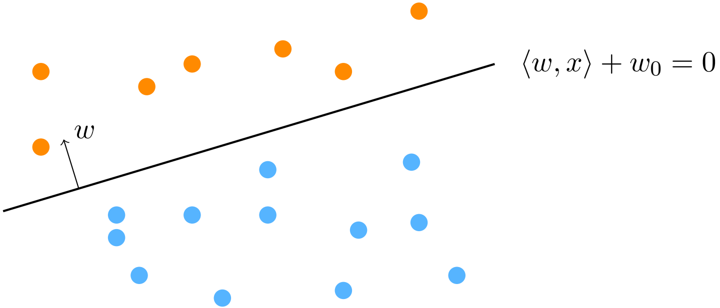
- $y_i = 1$ si $\langle w, x_i \rangle + w_0 > 0$,
- $y_i = -1$ si $\langle w, x_i \rangle + w_0 < 0$,

$$\iff \forall i = 1, \ldots, n \quad \boxed{y_i \cdot (\langle w, x_i \rangle + w_0) > 0}$$

Rem: w_0 : ordonnée à l'origine (: *intercept*)

Visualisation

- Équation $\langle w, x\rangle + w_0 = 0$: définit hyperplan séparateur $H_{w,b}$ de vecteur orthogonal w

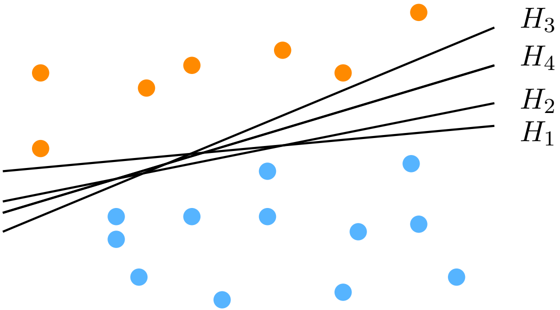


- Fonction $\phi_{w,w_0}(x) = \mathbb{1}_{\{\langle w,x\rangle + w_0 \geq 0\}} - \mathbb{1}_{\{\langle w,x\rangle + w_0 < 0\}}$: règle de discrimination linéaire

Remarque : pour tout $\kappa \neq 0$, $(\kappa w, \kappa w_0)$ et (w, w_0) définissent le même hyperplan

Dilemme de la complexité

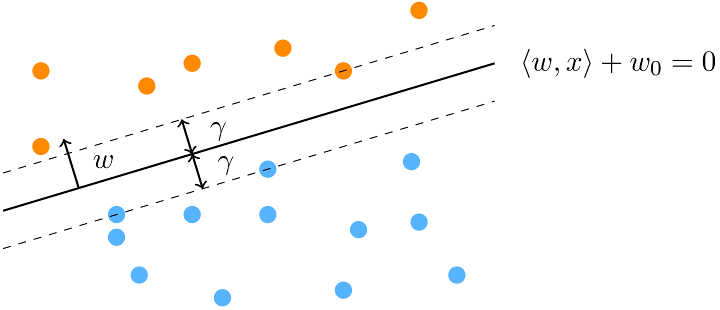
- **Problème** : une infinité d'hyperplans séparateurs $\Rightarrow$ une infinité de règles de discrimination linéaires potentielles !



- Lequel choisir ?

La marge

- Critère en sélection : hyperplan séparateur de **marge maximale** γ



La marge maximale

Soit x_{1^*} (resp. x_{-1^*}) de sortie 1 (resp. -1), se situant sur les frontières définissant la marge.

$$\text{La marge } \gamma \text{ satisfait :} \quad \gamma = \frac{\langle w, x_{1^*} \rangle}{\|w\|} = -\frac{\langle w, x_{-1^*} \rangle}{\|w\|}$$

<u>Forme canonique</u> pour $x_1, \ldots, x_n$: (à rescaling près) hyperplan $\langle w, x \rangle + w_0 = 0$ tel que

$$\min_{i=1,\ldots,n} |\langle w, x_i \rangle + w_0| = 1$$

Ainsi $\begin{cases} \langle w, x_{1^*} \rangle + w_0 = 1 \\ \langle w, x_{-1^*} \rangle + w_0 = -1 \end{cases}$ et donc $\langle w, x_{1^*} - x_{-1^*} \rangle = 2$,

d'où

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

Problème d'optimisation "primal"

Trouver l'hyperplan séparateur de marge maximale revient à trouver le couple (w, w_0) tel que

$$\min_{w \in \mathbb{R}^p, w_0 \in \mathbb{R}} \|w\|^2 \quad (\text{ou } \frac{1}{2}\|w\|^2)$$

$$\text{t.q.} \quad \forall i \in [\![1, n]\!],\, y_i \left(\langle w, x_i \rangle + w_0\right) \geq 1$$

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

Détours : multiplicateurs de Lagrange

Problème primal :

Minimiser $\forall u \in \mathbb{R}^d, h(u)$ sous contraintes $\forall i \in [\![1, n]\!], g_i(u) \geq 0$

Définition

Le **Lagrangien** est défini sur $\mathbb{R}^d \times \mathbb{R}^n$ par

$$\mathcal{L}(u, \alpha) = h(u) - \sum_{i=1}^{n} \alpha_i g_i(u)$$

Les variables α_i sont appelées les **variables duales**

Soit pour tout $\alpha \in \mathbb{R}_+^n$,

- $u_\alpha = \underset{u \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}(u, \alpha)$,
- $\theta(\alpha) = \mathcal{L}(u_\alpha, \alpha) = \min_{u \in \mathbb{R}^d} \mathcal{L}(u, \alpha)$: **fonction duale**.

Formulation duale

Problème dual :

$$\alpha^* = \arg\max_{\alpha \in \mathbb{R}^n} \theta(\alpha)$$

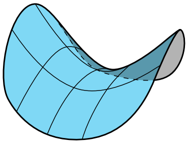
$$\text{t.q.} \quad \forall i \in [\![1, n]\!], \quad \alpha_i \geq 0$$

Solution du problème dual : α^* donne la solution du problème primal avec la relation

$$u^* = u_{\alpha^*}$$

Multiplicateurs de Lagrange : conditions de Karush-Kuhn-Tucker (KKT)

- $\alpha_i^* \geq 0$ pour tout $i = 1, \ldots, n$.
- $g_i(u_{\alpha^*}) \geq 0$ pour tout $i = 1, \ldots, n$.



Point scelle

Retour sur le problème dual

Minimiser $\mathcal{L}(u, \alpha) = h(u) - \sum_{i=1}^{n} \alpha_i g_i(u)$ par rapport à u
Maximiser $\mathcal{L}(u_\alpha, \alpha)$ associé par rapport aux variables duales α_i

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

$\Rightarrow$ Si $g_i(u_{\alpha^*}) > 0$, alors nécessairement $\alpha_i^* = 0$

Multiplicateurs de Lagrange : cas SVM

Lagrangien : $\mathcal{L}(w, w_0, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, x_i\rangle + w_0\right) - 1\right)$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_0}(w, w_0, \alpha) = -\sum_{i=1}^n \alpha_i y_i = 0 \;\Leftrightarrow\; \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial w}(w, w_0, \alpha) = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \;\Leftrightarrow\; w = \sum_{i=1}^n \alpha_i y_i x_i \end{cases}$$

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

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

- $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i,$
- w_0^* : résoudre (en w_0) pour un i t.q. $y_i(\langle w^*, x_i\rangle + w_0) = 1$

$$\text{où } \alpha^* = \arg\max_{\alpha} \theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$
$$\text{s. c. } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0, \forall i \in [\![1, n]\!]$$

Multiplicateurs de Lagrange : cas SVM

Lagrangien : $\mathcal{L}(w, w_0, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, x_i \rangle + \cancel{w_0}\right) - 1\right)$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_0}(w, w_0, \alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial w}(w, w_0, \alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \;\Leftrightarrow\; w = \sum_{i=1}^{n} \alpha_i y_i x_i \end{cases}$$

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

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

- $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$,
- w_0^* : résoudre (en w_0) pour un i t.q. $y_i(\langle w^*, x_i \rangle + w_0) = 1$

$$\text{où } \alpha^* = \arg\max_{\alpha} \theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{s. c. } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0, \forall i \in [\![1, n]\!]$$

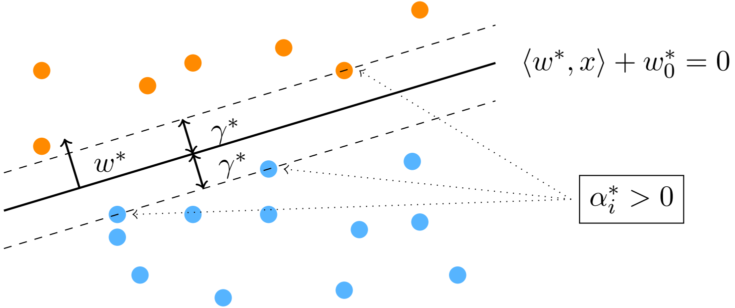
Conditions de Karush-Kuhn-Tucker (KKT)

- $\alpha_i^* \geq 0, \forall i \in [\![1, n]\!]$ (positivité)
- $y_i \left(\langle w^*, x_i \rangle + w_0^* \right) \geq 1, \quad \forall i \in [\![1, n]\!]$ (séparation)
- $\alpha_i^* \left(y_i \left(\langle w^*, x_i \rangle + w_0^* \right) - 1 \right) = 0, \forall i \in [\![1, n]\!]$ (complémentarité)

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

vecteurs supports : x_i tels que $\alpha_i^* > 0$; situés sur les frontières définissant la marge maximale *i.e.*, $y_i \left(\langle w^*, x_i \rangle + w_0^* \right) = 1$ (*cf.* condition complémentaire de KKT)

Représentation des vecteurs supports



Synthèse

Pour conclure, l'algorithme est défini par :

$$\phi_{\mathcal{D}_n}(x) = \mathbb{1}_{\{\langle w^*, x\rangle + w_0^* \geq 0\}} - \mathbb{1}_{\{\langle w^*, x\rangle + w_0^* < 0\}}$$

avec

- $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$
- w_0^* : résoudre (en w_0) pour un i t.q. $y_i(\langle w^*, x_i\rangle + w_0) = 1$

ou encore :

$$\phi_{\mathcal{D}_n}(x) = \begin{cases} 1, & \text{si } \sum_{x_i V.S.} \alpha_i^* y_i \langle x_i, x\rangle + w_0^* \geq 0 \\ -1, & \text{si } \sum_{x_i : V.S.} \alpha_i^* y_i \langle x_i, x\rangle + w_0^* < 0 \end{cases}$$

<u>Rem</u>: la marge maximale vaut $\gamma^* = \frac{1}{\|w^*\|}$

Plan

SVM

SVM linéaire pour des données séparables

SVM linéaire pour des données non séparables

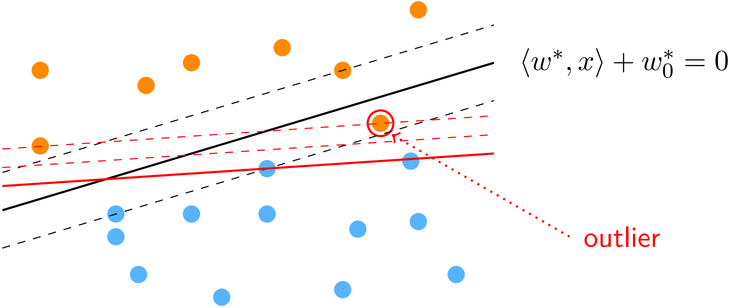
SVM non linéaire : astuce du noyau

Lien avec la minimisation du risque empirique convexifié

Conclusion / questions ouvertes

SVM linéaire pour des données non séparables

- Méthode non applicable pour données non linéairement séparables
- Méthode sensible aux "outliers"



SVM (sans séparation linéaire)

Nouvelle proposition : autoriser quelques vecteurs à être bien classés mais dans la région définie par la marge (voire mal classés)

<u>Contrainte associée</u> :
$y_i(\langle w, x_i \rangle + w_0) \geq 1 \implies y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i$, avec $\xi_i \geq 0$

$\xi_i \in [0, 1] \iff$ bien classé, mais région définie par la marge
$\xi_i > 1 \iff$ mal classé

<u>Vocabulaire</u> : **marge souple** (🇬🇧 : *soft margin*) et les ξ_i sont appelées les variables **ressorts** (🇬🇧 : *slacks*)

⚠ les contraintes relaxées ne peuvent pas être utilisées sans contrepartie sous peine d'obtenir une marge maximale infinie (prendre les ξ_i grands)

$\implies$ pénaliser les grandes valeurs de ξ_i

SVM (sans séparation linéaire, suite)

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

s. c. $\begin{cases} y_i\left(\langle w, x_i\rangle + w_0\right) \geq 1 - \xi_i, \forall i \in [\![1, n]\!] \\ \xi_i \geq 0 \end{cases}$

- $C > 0$ paramètre, **constante de tolérance** à ajuster

Solution du problème du primal :

- $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i,$
- w^* tel que $y_i\left(\langle w^*, x_i\rangle + w_0^*\right) = 1 - \xi_i, \forall x_i,\ 0 < \alpha_i^* < C,$

Solution duale : $\alpha^* = \arg\max_{\alpha} \theta(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$

s. c. $\sum_{i=1}^{n}\alpha_i y_i = 0$ et $0 \leq \alpha_i \leq C$, $\forall i \in [\![1, n]\!]$

Conditions de Karush-Kuhn-Tucker

- $0 \leq \alpha_i^* \leq C, \forall i \in [\![1, n]\!]$
- $y_i \left(\langle w^*, x_i \rangle + w_0^*\right) \geq 1 - \xi_i^* \, \forall i \in [\![1, n]\!]$
- $\alpha_i^* \left(y_i \left(\langle w^*, x_i \rangle + w_0^*\right) + \xi_i^* - 1\right) = 0, \forall i \in [\![1, n]\!]$
- $\xi_i^* (\alpha_i^* - C) = 0, \forall i \in [\![1, n]\!]$

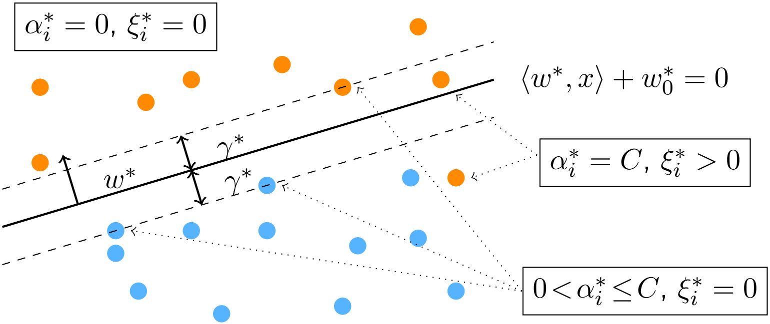
<u>Vocabulaire</u> : x_i tels que $\alpha_i^* > 0$, **vecteurs supports**

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 $\alpha_i^* = C$.

Vecteurs non supports : vérifient $\alpha_i^* = 0$ et $\xi_i^* = 0$

Représentation des vecteurs supports



Synthèse

Règle de classification du SVM linéaire :

$$\phi_{\mathcal{D}^n}(x) = \mathbb{1}_{\langle w^*, x\rangle + w_0^* \geq 0} - \mathbb{1}_{\langle w^*, x\rangle + w_0^* < 0}$$

avec

- $w^* = \sum_{i=1}^n \alpha_i^* x_i y_i$,
- w_0^* tel que $y_i \left(\langle w^*, x_i\rangle + w_0^*\right) = 1 - \xi_i, \forall x_i,\ 0 < \alpha_i^* < C$,

ou encore :

$$\phi_{\mathcal{D}_n}(x) = \begin{cases} 1, & \text{si } \sum\limits_{x_i : V.S.} \alpha_i^* y_i \langle x_i, x\rangle + w_0^* \geq 0 \\ -1, & \text{si } \sum\limits_{x_i : V.S.} \alpha_i^* y_i \langle x_i, x\rangle + w_0^* < 0 \end{cases}$$

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

Plan

SVM

SVM linéaire pour des données séparables

SVM linéaire pour des données non séparables

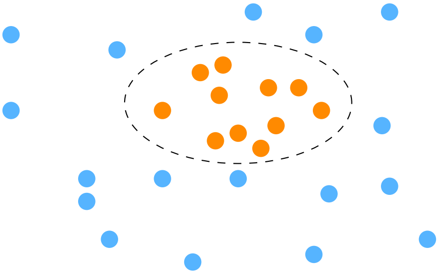
SVM non linéaire : astuce du noyau

Lien avec la minimisation du risque empirique convexifié

Conclusion / questions ouvertes

SVM non linéaire : astuce du noyau

Exemple de données difficiles à discriminer linéairement :



- SVM linéaire : mauvaise discrimination avec un nombre de vecteurs supports très élevé $\Rightarrow$ SVM non linéaire ?

Kernelisation

Boser,Guyon et Vapnik (1992)

Envoyer les entrées $x_1, \ldots, x_n$ dans un espace de Hilbert $\mathcal{H}$ (produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$) (dimension infinie), via une fonction φ, et appliquer un SVM linéaire à $\{(\varphi(x_i), y_i), i = 1, \ldots, n\}$.
Sortie attribuée à x : celle attribuée à son image $\varphi(x)$

vocabulaire :
φ : **fonction de représentation** (: *feature function*)
$\mathcal{H}$: **espace de représentation** (: *feature space*)

Exemple précédent : $\varphi(x) = (x_1^2, x_2^2, x_1, x_2)$; linéairement séparables dans $\mathbb{R}^4$

Choisir $\mathcal{H}$ et φ

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

$$\phi_{\mathcal{D}_n}(x) = \mathbb{1}_{\sum y_i \alpha_i^* \langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} + w_0^* \geq 0} - \mathbb{1}_{\sum y_i \alpha_i^* \langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} + w_0^* < 0},$$

α^* : solution du problème dual dans l'espace de représentation $\mathcal{H}$:

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

Solution duale :

$$\alpha^* = \arg\max_{\alpha} \theta(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$$

$$\text{s. c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C, \forall i \in [\![1, n]\!]$$

Remarque fondamentale

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*) : seule la connaissance de la fonction k définie par $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ est requise, sans déterminer explicitement $\mathcal{H}$ et φ

Définition

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 $\varphi : \mathcal{X} \to \mathcal{H}$ donnée est appelée un **noyau**

Rem: Un noyau est souvent plus facile à calculer que la fonction φ

Exemple : pour $x = (x_1, x_2) \in \mathbb{R}^2$, $\varphi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, prendre $k(x, x') = \langle x, x' \rangle^2$

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

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

Agrégation de noyaux

Soit k_1 et k_2 des noyaux, f une fonction : $\mathbb{R}^d \to \mathbb{R}$, $\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}$, B une matrice définie positive, P un polynôme à coefficients positifs, $\lambda \geq 0$.

La fonction définie par $k(x, x') = k_1(x, x') + k_2(x, x')$, $\lambda k_1(x, x')$, $k_1(x, x')k_2(x, x')$, $f(x)f(x')$, $k_1(\varphi(x), \varphi(x'))$, $x^T B x'$, $P(k_1(x, x'))$, ou $e^{k_1(x,x')}$ est encore un noyau.

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

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

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

Plan

SVM

SVM linéaire pour des données séparables

SVM linéaire pour des données non séparables

SVM non linéaire : astuce du noyau

Lien avec la minimisation du risque empirique convexifié

Conclusion / questions ouvertes

Éléments de théorie

Minimiser $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i$ s. c. $\begin{cases} y_i\left(\langle w, x_i\rangle + w_0\right) \geq 1 - \xi_i, \ \forall\ i \\ \xi_i \geq 0 \end{cases}$

est équivalent à minimiser

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \left(1 - y_i\left(\langle w, x_i\rangle + w_0\right)\right)_+ ,$$

ou encore

$$\frac{1}{n}\sum_i \left(1 - y_i\left(\langle w, x_i\rangle + w_0\right)\right)_+ + \frac{1}{2Cn}\|w\|^2.$$

$\gamma(w, w_0, x_i, y_i) = \left(1 - y_i\left(\langle w, x_i\rangle + w_0\right)\right)_+$ est un majorant convexe de l'erreur empirique $\mathbb{1}_{y_i(\langle w, x_i\rangle + w_0) \leq 0}$

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

Plan

Conclusion / questions ouvertes

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

Librairies / code sources

- Liblinear Fan *et al.* (2008)
- Libsvm Chang et Lin (2011)
- `sklearn` (charge cette implémentation)

Bibliographie I

- Boser, B. E., I. M. Guyon et V. N. Vapnik. "A training algorithm for optimal margin classifiers". In : *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, p. 144-152.
- Fan, R.-E. et al. "LIBLINEAR : A library for large linear classification". In : *J. Mach. Learn. Res.* 9 (2008), p. 1871-1874.
- Vapnik, V. N. *Statistical learning theory*. Wiley, 1998.