# Apprentissage Artificiel

## Séparateurs à Vastes Marges
## Support Vector Machines

Ewa Kijak

ESIR/Univ. Rennes





## Sommaire







## Apprentissage de surfaces séparatrices linéaires

- Dans $\mathbb{R}^d$, une surface linéaire est un hyperplan $H$, défini par :

$$w_0 + \boldsymbol{w}^T \boldsymbol{x} = 0$$

  avec $\boldsymbol{w}$ vecteur de dimension $d$ et $w_0$ scalaire.
- Un hyperplan divise l'espace en deux régions, celle où :

$$\boldsymbol{x} \in \omega_1 \Rightarrow w_0 + \boldsymbol{w}^T \boldsymbol{x} \geq 0$$

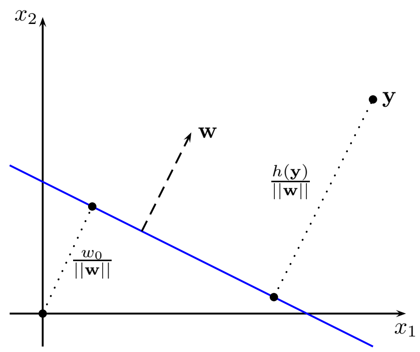
  et l'autre :

$$\boldsymbol{x} \in \omega_2 \Rightarrow w_0 + \boldsymbol{w}^T \boldsymbol{x} \leq 0$$





## La géométrie d'un hyperplan




$$h(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 = 0$$





## Sommaire







## Séparation linéaire dans $\mathcal{X}$

- Soit $\mathcal{X}$, l'espace de représentation des données.
- Si les points d'apprentissage sont séparables, il existe une infinité d'hyperplans séparateurs.
- Comment trouver le meilleur hyperplan séparateur dans $\mathcal{X}$ ?
- Et comment faire si les données ne sont pas séparables ?

## Passage dans un autre espace $\Phi(\mathcal{X})$

- Deux classes peuvent être linéairement séparées dans un espace de grande dimension $\Phi(\mathcal{X})$.
- Comment trouver le meilleur hyperplan séparateur dans $\Phi(\mathcal{X})$ avec des calculs peu coûteux ?





## Séparateurs linéaires et marge

Pour le moment, on se place dans $\mathcal{X} = \mathbb{R}^d$.
$\mathcal{S} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), ..., (\boldsymbol{x}^{(m)}, y^{(m)})\}$, avec $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{-1, 1\}$.

On suppose qu'il existe une séparatrice linéaire permettant de distinguer les exemples positifs (supervisés par $y^{(i)} = +1$) des exemples négatifs (supervisés par $y^{(i)} = -1$).

Il existe une donc hypothèse d'apprentissage $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0$ telle que :

$$h(\boldsymbol{x}^{(i)}) = \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + w_0 \begin{cases} > 0 \\ < 0 \end{cases} \implies \hat{y}^{(i)} = \begin{cases} +1 \\ -1 \end{cases}$$





## Séparateurs linéaires et marge

Ce qui peut se réécrire :

$$\forall\, 1 \leq i \leq m \;:\; y^{(i)}\, h(\boldsymbol{x}^{(i)}) \; > 0$$

ou

$$\forall\, 1 \leq i \leq m \;:\; y^{(i)}\, (\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + w_0) \; > 0$$

$h(\boldsymbol{x})$ est l'équation d'un hyperplan dans $\mathcal{X}$ de vecteur normal $\boldsymbol{w}$.
La distance d'un point $\boldsymbol{x}^{(i)}$ à l'hyperplan d'équation $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0 = 0$ est égale à : $|h(\boldsymbol{x}^{(i)})|/\|\, \boldsymbol{w} \,\|$.

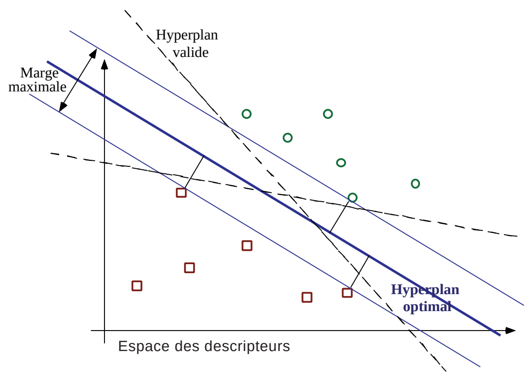
L'hyperplan est dit sous *forme canonique* lorsque $\boldsymbol{w}$ et $w_0$ sont normalisés de façon à satisfaire :

$$\forall\, 1 \leq i \leq m \;:\; y^{(i)}\, (\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + w_0) \; \geq 1$$





## Séparateurs linéaires et marge








## Le meilleur séparateur linéaire

Lorsqu'il existe *une* séparatrice linéaire entre les points d'apprentissage, il en existe une infinité. On peut alors chercher parmi ces séparatrices celle qui est la meilleure : la plus écartée des deux nuages de points exemples et contre-exemples.

Cet hyperplan optimal est défini par :

$$\underset{\boldsymbol{w}, w_0}{\text{argmax}} \; \min_i \; \{\|\, \boldsymbol{x} - \boldsymbol{x}^{(i)} \,\| \;:\; \boldsymbol{x} \in \mathbb{R}^d \,,\, (\boldsymbol{w}^\top \boldsymbol{x} + w_0) = 0 \,,\, i = 1, \ldots, m\}$$

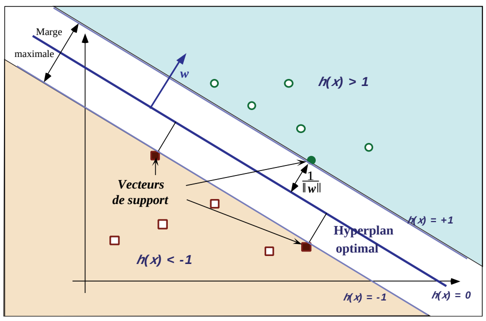
Il maximise la distance minimale aux exemples d'apprentissage.

La *marge* vaut : $2/\|\, \boldsymbol{w} \,\|$.





## Le meilleur séparateur linéaire




L'hyperplan optimal est perpendiculaire au segment de droite le plus court joignant un exemple d'apprentissage à l'hyperplan. Ce segment a pour longueur $\frac{1}{\|\boldsymbol{w}\|}$ lorsqu'on normalise $\boldsymbol{w}$ et $w_0$.

## Sommaire







## L'expression primale du problème des SVM

La recherche de l'hyperplan optimal revient donc à minimiser $\| \mathbf{w} \|$, soit à résoudre le problème d'optimisation suivant qui porte sur les paramètres $\mathbf{w}$ et $w_0$ :

$$\begin{cases} \text{Minimiser} & \frac{1}{2}\| \mathbf{w} \|^2 \\ \text{sous les contraintes} & y^{(i)}\,(\mathbf{w}^\top \mathbf{x}^{(i)} \; + \; w_0) \; \geq 1 \quad i = 1, \ldots, m \end{cases} \tag{1}$$

Cette écriture du problème, appelée *formulation primale*, implique le réglage de $d+1$ paramètres, $d$ étant la dimension de l'espace des entrées $\mathcal{X}$.

- Possible quand $d$ est petit avec méthodes d'optimisation quadratique
- inenvisageable pour des valeurs de $d$ très élevées ($> 10^3$), en particulier quand on travaillera dans $\Phi(\mathcal{X})$.





## L'expression duale du problème des SVM

Le *lagrangien* est la somme de la fonction objectif et d'une combinaison linéaire des contraintes. Les coefficients $\alpha_i \geq 0$ sont appelés *multiplicateurs de Lagrange* ou *variables duales*.

$$L(\mathbf{w}, w_0, \boldsymbol{\alpha}) \;=\; \frac{1}{2}\| \mathbf{w} \|^2 \;-\; \sum_{i=1}^{m} \alpha_i \, \left( y^{(i)}(\mathbf{x}^{(i)}.\mathbf{w} \;+\; w_0) \;-\; 1 \right) \tag{2}$$

Le problème primal et sa formulation duale ont la même solution qui correspond à un *point-selle* du lagrangien (il faut le minimiser par rapport aux variables primaires $\mathbf{w}$ et $w_0$ et le maximiser par rapport aux variables duales $\alpha_i$).





## L'expression duale (suite)

Au point-selle, la dérivée du Lagrangien par rapport aux variables primaires doit s'annuler. Ceci s'écrit :

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) \;=\; 0, \quad \frac{\partial}{\partial w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) \;=\; 0 \tag{3}$$

et conduit à :

$$\mathbf{w} \;=\; \sum_{i=1}^{m} \alpha_i y^{(i)} \mathbf{x}^{(i)} \tag{4}$$

et à :

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \;=\; 0 \tag{5}$$





## Solution

En substituant (5) et (4) dans (2), on élimine les variables primaires et l'on obtient la *forme duale* du problème d'optimisation. C'est un problème *quadratique*.

Trouver les multiplicateurs de Lagrange $\alpha_i \geq 0$ tels que :

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{m} \alpha_i \;-\; \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (\mathbf{x}^{(i)} \,.\, \mathbf{x}^{(j)}) \right\} \\ \alpha_i \geq 0, \quad i = 1, \ldots, m \\ \sum_{i=1}^{m} \alpha_i y^{(i)} \;=\; 0 \end{cases} \tag{6}$$





## Intuitif et remarquable

- En pratique, seuls les points qui sont sur les hyperplans frontière $(\mathbf{x}^{(i)} \,.\, \mathbf{w}) \;+\; w_0 \;=\; \pm 1$ jouent un rôle, car les multiplicateurs de Lagrange sont non nuls pour ces seuls points.
- Ils sont appelés **vecteurs de support** ou *exemples critiques*.
- Le vecteur solution $\mathbf{w}^\star$ a donc une expression en termes d'un sous-ensemble des exemples d'apprentissage : les exemples critiques $(\mathbf{x}^{(c)}, y(c))$.

$$\mathbf{w}^\star \;=\; \sum_{i=1}^{m_c} \alpha_c^\star y^{(c)} \mathbf{x}^{(c)}$$

- C'est en même temps intuitivement satisfaisant puisque l'on voit bien que l'hyperplan solution est entièrement déterminé par ces exemples.

## Solution

L'hyperplan solution correspondant peut alors être écrit :

$$h(\boldsymbol{x}) \;=\; (\boldsymbol{w}^\star \,.\, \boldsymbol{x}) \;+\; w_0^\star \;=\; \sum_{i=1}^{m} \alpha_i^\star \, y^{(i)} \, (\boldsymbol{x} \,.\, \boldsymbol{x}^{(i)}) \;+\; w_0^\star \qquad (7)$$

où les $\alpha_i^\star$ sont solution de (6) et $w_0$ est obtenue en utilisant n'importe quel exemple critique $(\boldsymbol{x}^{(c)}, y^{(c)})$ dans l'équation :

$$y^{(c)} \, ((\boldsymbol{x}^{(c)} \,.\, \boldsymbol{w}^\star) \;+\; w_0) \;-\; 1 \;=\; 0 \qquad (8)$$





## Remarques sur la solution

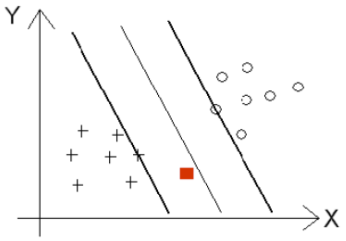
- L'hyperplan solution ne requiert que le calcul des produits scalaires $(\boldsymbol{x} \,.\, \boldsymbol{x}^{(i)})$ entre des vecteurs de l'espace d'entrée $\mathcal{X}$.
- La solution ne dépend plus de la dimension $d$ de l'espace d'entrée, mais de la taille $m$ de l'échantillon de données et même du nombre $m_c$ d'exemples critiques qui est généralement bien inférieur à $m$.
  - nécessite de déterminer $m$ variables duales $(\alpha_i)$ et non $d+1$ paramètres $(\boldsymbol{w}, w_0)$.
- Les méthodes d'optimisation quadratique standard suffisent pour la plupart des cas (env. $10^5$ exemples).
- Les fonctions de coût et les contraintes sont strictement convexes (Th. de Kuhn-Tucker)





## Classification d'une nouvelle donnée

La classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal.




Dans le schéma ci-dessus, le nouvel élément sera classé dans la catégorie des « + ».





## Classification d'une nouvelle donnée

La classe d'une nouvelle donnée $\boldsymbol{x}$ est fournie par le signe de $h(\boldsymbol{x}) = \boldsymbol{w}^{\star\top} \boldsymbol{x} \;+\; w_0^\star$ :

$$h(\boldsymbol{x}) \begin{cases} > 0 \\ < 0 \end{cases} \Longrightarrow \; \hat{y} \;=\; \begin{cases} +1 \\ -1 \end{cases}$$

On calcule :

$$h(\boldsymbol{x}) \;=\; \sum_{c=1}^{m_c} \alpha_c^\star \, y^{(c)} \, (\boldsymbol{x} \,.\, \boldsymbol{x}^{(c)}) \;+\; w_0^\star \qquad (9)$$

où $m_c$ est le nombre de vecteurs support $(\boldsymbol{x}^{(c)}, y^{(c)})$.





## Cas d'un échantillon non linéairement séparable 1

Si les exemples ne peuvent pas être linéairement séparés, on peut employer la technique dite des *variables ressort* (*slack variables*). On modifie les contraintes en les relâchant grâce à des variables ressort $\xi_i \;\geq\; 0$ :

$$y^{(i)} \, \left( (\boldsymbol{w} \,.\, \boldsymbol{x}^{(i)}) \;+\; w_0 \right) \;\geq\; 1 - \xi_i \qquad (10)$$

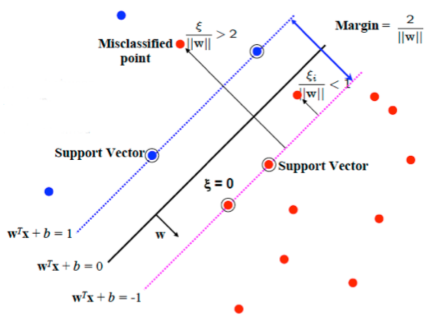
Remarque :

- si $\xi_i = 0$, l'exemple $x^{(i)}$ est bien classé et sur la marge ou à l'extérieur de la marge
- si $0 < \xi_i < 1$, l'exemple $x^{(i)}$ est bien classé mais à *l'intérieur* de la marge
- si $\xi_i = 1$, l'exemple $x^{(i)}$ est sur l'hyperplan optimal (séparatrice)
- si $\xi_i > 1$, l'exemple $x^{(i)}$ est mal classé





## Cas d'un échantillon non linéairement séparable 2




$\xi_i$ est d'autant plus grand que l'exemple $x^{(i)}$ est loin de la séparatrice.

## Echantillon non linéairement séparable 3

Le problème d'optimisation sous contraintes s'écrit alors :

$$\begin{cases} \text{Minimiser} & \frac{1}{2}\| \boldsymbol{w} \|^2 + C \sum_{i=1}^m \xi_i \\ \text{sous les contraintes} & y^{(i)}\,(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + w_0) \geq 1 - \xi_i \quad \text{et} \quad \xi_i \geq 0 \\ & \forall i = 1, \dots, m \end{cases} \qquad (11)$$

- Le coefficient $C$ (appelé constante de régularisation - *trade-off parameter*) règle le compromis entre la marge possible entre les exemples et les erreurs admissibles
  - Remarque : ce n'est pas le nombre de mauvaises classifications qui est minimisé (problème NP-complet), mais la somme des distances aux hyperplans marges





## Echantillon non linéairement séparable 4

On obtient alors, comme dans le cas séparable, une formulation duale, avec des contraintes légèrement différentes.

$$\begin{cases} \max_{\alpha} \left\{ \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (\boldsymbol{x}^{(i)} . \boldsymbol{x}^{(j)}) \right\} \\ \forall i, \quad 0 \leq \alpha_i \leq C \\ \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{cases} \qquad (12)$$

Les vecteurs support sont toujours tels que $\alpha_i > 0$ :

- $0 < \alpha_i < C \quad \Rightarrow \quad \xi_i = 0 \quad \Leftrightarrow \quad x^{(i)}$ est sur un hyperplan marge
- $\alpha_i = C \quad \Leftrightarrow \quad \xi_i > 0 \quad \Leftrightarrow \quad x^{(i)}$ est à l'intérieur de la marge (bien ou mal classé)





## Echantillon non linéairement séparable 5

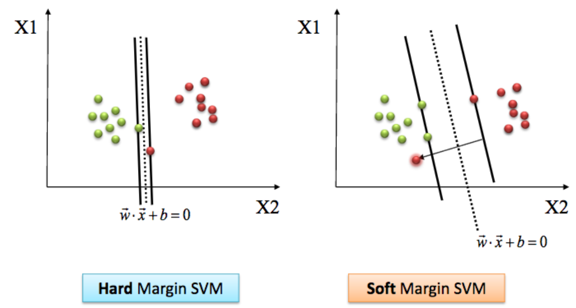
- On parle de SVM "à marge molle" (*Soft margin SVM*) ou "marge souple"
  - Plus $C$ est grand, moins on admet d'exemples à l'intérieur de la marge (moins on régularise)
  - $C$ très grande $\Leftrightarrow$ SVM à marge dure
- Les SVM à marges molles ont toujours une solution
- Ils sont plus robustes aux *outliers*
- Les SVM à marges dures n'ont pas d'hyper-paramètre (pas de constante $C$ à déterminer)





## Echantillon non linéairement séparable 6

Les vecteurs supports ne sont pas nécessairement sur les marges ($\neq$ SVM à marges dures)




http://chem-eng.utoronto.ca/~datamining/





## Sommaire







## Rappel : Les idées de base

- Certains calculs ne sont pas plus coûteux en transformant l'espace de représentation $\mathcal{X}$ en un espace de grande dimension $\Phi(\mathcal{X})$, par exemple les produits scalaires.
- Deux classes peuvent être linéairement séparées dans un espace de grande dimension.
- Comment trouver le meilleur hyperplan séparateur dans $\Phi(\mathcal{X})$ avec des calculs peu coûteux ?

## Le passage par un espace de redescription

Plus la dimension de l'espace de description est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples et les contre-exemples est élevée.

En transformant l'espace d'entrée en un espace de redescription de très grande dimension, éventuellement infinie, il devient donc possible d'envisager d'utiliser la méthode des SVM.

Notons $\Phi$ une transformation non linéaire de l'espace d'entrée $\mathcal{X}$ en un *espace de redescription* $\Phi(\mathcal{X})$ :

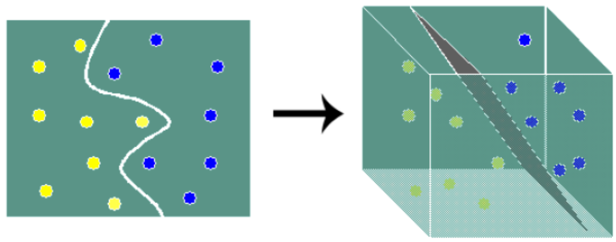
$$\boldsymbol{x} = (x_1, \ldots, x_d)^\top \overset{\Phi}{\longmapsto} \Phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_d(\boldsymbol{x}), \ldots))^\top \qquad (13)$$





## Le passage par un espace de redescription

$$\begin{aligned} \Phi : \mathbb{R}^d &\longmapsto \mathbb{R}^{d'} \quad (d' > d) \\ \boldsymbol{x} &\longmapsto \Phi(\boldsymbol{x}) \end{aligned}$$







## Le passage par un espace de redescription

Le problème d'optimisation se transcrit dans ce cas par :

$$\begin{cases} \max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle \Phi(\boldsymbol{x}^{(i)}), \Phi(\boldsymbol{x}^{(j)}) \rangle \right\} \\ \alpha_i \geq 0, \quad i = 1, \ldots, m \\ \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{cases} \qquad (14)$$

où $\langle ., . \rangle$ dénote le produit scalaire dans le nouvel espace.
L'équation de l'hyperplan séparateur dans le nouvel espace devient :

$$h(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i^\star y^{(i)} \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}^{(i)}) \rangle + w_0^\star \qquad (15)$$

où les coefficients $\alpha_i^\star$ et $w_0^\star$ sont obtenus comme précédemment par résolution de (14).





## Le passage par un espace de redescription : les fonctions noyau

$\langle \Phi(\boldsymbol{x}^{(i)}), \Phi(\boldsymbol{x}^{(j)}) \rangle$ devient rapidement impossible à calculer quand la dimension de $\Phi(\mathcal{X})$ augmente (sans parler du cas de la dimension infinie), ceci d'autant plus que l'on utilisera des transformations non linéaires des descripteurs d'entrée.

On peut dans certains cas s'arranger pour court-circuiter le passage par les calculs dans l'espace de redescription.

Il existe des fonctions bilinéaires symétriques positives $K(x, y)$, appelées *fonctions noyau*, faciles à calculer et dont on peut montrer qu'elles correspondent à un produit scalaire $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle$ dans un espace de grande dimension.





## Les fonctions noyau : exemple

$$\mathcal{X} \overset{\Phi}{\longmapsto} \Phi(\mathcal{X})$$

On connaît une *fonction noyau* $K(\boldsymbol{x}, \boldsymbol{y})$, avec $\boldsymbol{x}$ et $\boldsymbol{y}$ dans $\mathcal{X}$ telle que

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle$$

Dans ce cas, le produit scalaire dans $\Phi(\mathcal{X})$ est un calcul dans $\mathcal{X}$.

Par exemple : soient $\mathcal{X} = \mathbb{R}^2$ et $\Phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2) \in \mathbb{R}^3$

Montrer que : $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2$





## Les fonctions noyau

Lorsqu'une telle correspondance est exploitable, le problème d'optimisation (14) est équivalent au problème suivant :

$$\begin{cases} \max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \right\} \\ \alpha_i \geq 0, \quad i = 1, \ldots, m \\ \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{cases} \qquad (16)$$

dont la solution est l'hyperplan séparateur d'équation :

$$h(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i^\star y^{(i)} K(\boldsymbol{x}, \boldsymbol{x}^{(i)}) + w_0^\star \qquad (17)$$

où les coefficients $\alpha_i^\star$ et $w_0^\star$ sont obtenus comme précédemment par résolution du problème d'optimisation quadratique.

## Exemple de fonction noyau : la fonction noyau polynomiale

Il peut être montré que la *fonction noyau polynomiale* :

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \langle \boldsymbol{x}\,,\,\boldsymbol{y}\rangle^n \tag{18}$$

réalise implicitement un produit scalaire dans l'espace des descripteurs correspondant à tous les produits d'exactement $n$ dimensions.

Ainsi pour $n = 2$ et $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^2$, on a :

$$\langle \boldsymbol{x}\,,\,\boldsymbol{y}\rangle^2 \;=\; \langle \Phi(\boldsymbol{x})\,,\,\Phi(\boldsymbol{y})\rangle$$

cf. exemple ci-dessus

qui correspond au changement de description par la fonction :

$$\Phi(\boldsymbol{x}) \;=\; (x_1^2\,,\,x_2^2\,,\,\sqrt{2}x_1x_2)^\top \in \mathbb{R}^3$$





## La fonction noyau polynomiale

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \langle \boldsymbol{x}\,,\,\boldsymbol{y}\rangle^n$$

$$\Phi : \mathbb{R}^d \longmapsto \mathbb{R}^{d'} \quad {\scriptstyle (d'>d)}$$

Pour $n = 2$ et $d = 2$, on peut aussi transformer autrement en dimension $d' = 3$, ou $d' = 4$ :

$$\Phi(\boldsymbol{x}) \;= (x_1^2\,,\,\sqrt{2}x_1x_2\,,\,x_2^2)^\top$$

$$\Phi(\boldsymbol{x}) \;= \left(\,\frac{1}{\sqrt{2}}(x_1^2 - x_2^2)\,,\,\sqrt{2}x_1x_2\,,\,\frac{1}{\sqrt{2}}(x_1^2 + x_2^2)\,\right)^\top$$

$$\Phi(\boldsymbol{x}) \;= (\,x_1^2\,,\,x_1x_2\,,\,x_1x_2\,,\,x_2^2\,)^\top$$





## Autre exemple : un polynôme de degré 3

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \langle \boldsymbol{x}\,,\,\boldsymbol{y}\rangle^n$$

$$\Phi : \mathbb{R}^d \longmapsto \mathbb{R}^{d'} \quad {\scriptstyle (d'>d)}$$

Pour $n = 3$, $d = 2$ et $d' = 4$, avec :

$$\Phi(\boldsymbol{x}) \;=\; (x_1^3\,,\,\sqrt{3}\,x_1^2x_2\,,\,\sqrt{3}\,x_1x_2^2\,,\,x_2^3)^\top$$

on peut également vérifier que :

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle^3 \;= \langle \Phi(\boldsymbol{x})\,,\,\Phi(\boldsymbol{y})\rangle$$





## Les conditions de Mercer

Une fonction $K$ symétrique est un noyau **ssi** $(K(x^{(i)}, x^{(j)}))_{i,j}$ est une matrice définie positive.
Dans ce cas, il existe un espace $\mathcal{F}$ et une fonction $\Phi$ tels que

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \langle \Phi(\boldsymbol{x})\,,\,\Phi(\boldsymbol{y})\rangle$$

Si cette condition est vérifiée, on peut appliquer les SVMs

Problèmes :

- Cette condition est très difficile à vérifier
- Elle donne pas d'indication pour la construction de noyaux $K$
- Elle ne permet pas de savoir comment est $\Phi$





## D'autres fonctions noyau

- Noyaux polynomiaux (non homogènes)

$$K(\boldsymbol{x},\boldsymbol{x}') \;=\; (\boldsymbol{x}\,.\,\boldsymbol{x}' + c)^n$$

- Fonctions à Base Radiale (RBF)

$$K(\boldsymbol{x},\boldsymbol{x}') \;=\; e^{-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2}$$

- Noyaux gaussiens

$$K(\boldsymbol{x},\boldsymbol{x}') \;=\; e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}}$$

- Fonctions sigmoïdes

$$K(\boldsymbol{x},\boldsymbol{x}') \;=\; \tanh(a\,(\boldsymbol{x}\,.\,\boldsymbol{x}')\;-\;b)$$



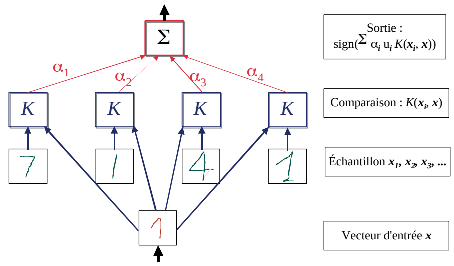


## D'autres fonctions noyau

En pratique, on combine des noyaux simples pour en obtenir de plus complexes.

- Construction de nouvelles fonctions noyau par combinaison linéaire de fonctions noyau.

Remarque : Les hyperparamètres (constante de régularisation $C$, écart-type des gaussiennes $\sigma$ si on utilise des noyaux gaussiens, degré $d$ si on utilise des noyaux polynomiaux, etc.) doivent être déterminés par validation croisée.

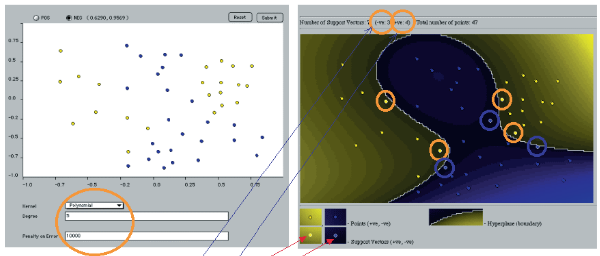



Lors de l'apprentissage, les exemples critiques sont retenus pour définir la fonction de décision. Lorsqu'une nouvelle entrée est présentée au système, elle est comparée aux exemples critiques à l'aide des fonctions noyau qui réalisent un produit scalaire dans l'espace de redescription $\Phi(\mathcal{X})$. La sortie est calculée en faisant une combinaison linéaire de ces comparaisons.





Exemple

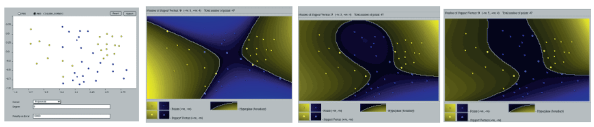


- http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml
- 47 exemples (22 +, 25 -)
- *Exemples critiques* : 4 + et 3 -
- Ici *fonction polynomiale* de *degré 5* et *C = 10000*





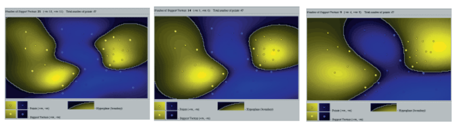
Exemple : Effets de différents choix de la fonction noyau



(5-, 4+) (3-, 4+) (5-, 4+)

- 47 exemples (22 +, 25 -)
- *Exemples critiques* : 4 + et 3 -

Ici *fonction polynomiale* de *degré 2, 5, 8* et *C = 10000*

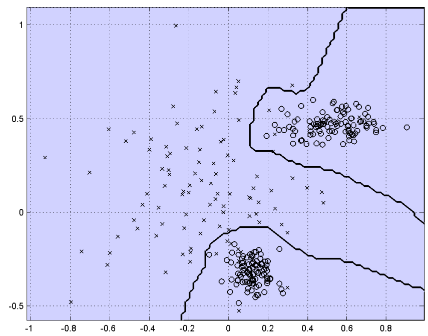


(10-, 11+) (8-, 6+) (4-, 5+)

Ici *fonction Gaussienne* de $\sigma = 2, 5, 10$ et $C = 10000$







Séparation de deux classes par SVM avec $K$ fonction à base radiale (RBF).





## Sommaire







## Implémentation des SVMs

→Minimisation de fonctions différentiables convexes à plusieurs variables

- pas d'optima locaux
- mais problèmes de stockage de la matrice noyau (long si milliers d'exemples)

⇒ mise au point de méthodes spécifiques

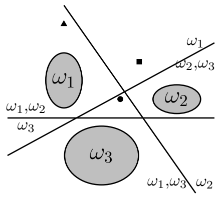
- Méthodes itératives, optimisation par morceaux...

Plusieurs librairies publiques disponibles :

- SVMLight, SVMTorch
- libSVM, SMO
- ...

## Et pour plus de 2 classes ?

One-versus-all




- chaque classe est séparée de toutes les autres : il y a $C$ hyperplans
- exemple : le triangle est assigné à la classe $\omega_1$, le carré est ambigu entre $\omega_1$ et $\omega_2$, le point central est ambigu entre les 3 classes





## Et pour plus de 2 classes ?

One-versus-all

- Technique :
  - pour chaque classe $\{\omega_1, \omega_2, ...\omega_C\}$, apprendre un classifieur binaire $h_{\omega_i}$
  - pour une observation $\boldsymbol{x}$ :

$$\omega^\star = \underset{i}{\text{argmax}}\ h_{\omega_i}(\boldsymbol{x})$$

$\rightarrow$le classifieur avec la valeur de plus grande confiance (marge) gagne

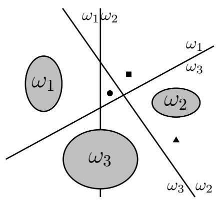
- Problème :
  - calibration : les scores des classifieurs ne sont pas forcément comparables
  - données non équilibrées : plus d'exemples négatifs que positifs
  $\rightarrow$néanmoins : simple et fréquemment utilisé en pratique





## Et pour plus de 2 classes ?

One-versus-one




- les classes sont séparées les unes des autres : il y a $\frac{C(C-1)}{2}$ hyperplans.
- exemple : le triangle et le carré sont assignés à la classe $\omega_2$, le point central est ambigu entre les 3 classes





## Et pour plus de 2 classes ?

One-versus-one

- Technique :
  - pour chaque paire de classes $(\omega_i, \omega_j)$, apprendre un classifieur binaire :

$$\text{sign } h_{ij}$$

  - combiner les classifieurs binaires par un mécanisme de vote majoritaire. Pour une observation $\boldsymbol{x}$ :

$$\omega^\star = \underset{j \in \mathcal{C}}{\text{argmax}}\ |\{i\ :\ h_{ij}(\boldsymbol{x}) = 1\}|$$

$\rightarrow$la classe assignée le plus grand nombre de fois gagne

- Problème :
  - computationnel : entraîner $\frac{C(C-1)}{2}$ classifieurs binaires
  - surapprentissage : la taille de l'ensemble d'apprentissage peut devenir trop petite pour une paire de classes donnée
  - il peut rester une ambiguité s'il n'y a pas de classe majoritaire (égalité)





## Conclusion

- Méthode d'apprentissage complètement issue de considérations théoriques (bien fondée mathématiquement).
- SVM faciles à mettre en oeuvre et donnent souvent de bons résultats en apprentissage (bonne capacité de généralisation : $R$ est proche de $R_{emp}$)
- ne permettent pas l'extraction d'un modèle compréhensible
- inadaptés aux très grands volumes de données (calculs lourds, max actuel = 10 000 exemples)
- pas de solution pour le choix du noyau





## Exercice : apprentissage du XOR

- On souhaite construire un SVM permettant de classer les points $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$ selon le résultat de l'opérateur XOR.

Q1 Représenter les points et leur classe dans $\mathbb{R}^2$. Ce problème est-il linéairement séparable ?

- On note $\Phi$ une transformation non linéaire de l'espace d'entrée $\mathcal{X}$ en un espace de redescription $\Phi(\mathcal{X})$ :
$\Phi : \boldsymbol{x} = (x_1, ..., x_d) \rightarrow \Phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), ..., \phi_d(\boldsymbol{x}), ...)$ et on considère la fonction noyau polynomiale de degré 2 :

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 1)^2$$

Cette fonction réalise implicitement un produit scalaire dans un espace des descripteurs de plus grande dimension : $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle$.

## Exercice : apprentissage du XOR

- Q2 Déduire du développement de $K(\boldsymbol{x}, \boldsymbol{y})$ la transformation $\Phi$ et la dimension de l'espace de redescription.
- Q3 Rappeler le problème d'optimisation à résoudre faisant intervenir un Lagrangien avec $\boldsymbol{\alpha}$ le vecteur des multiplicateurs de Lagrange.
- Q4 Donner dans un tableau les valeurs $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ pour tous les couples $(i, j)$.
- Q5 Ecrire le problème d'optimisation sous la forme d'un système d'équations, et le résoudre.
- Q6 Quel est, dans l'espace $\Phi(\mathcal{X})$, le vecteur poids optimal $\boldsymbol{w}^\star$ ? En déduire l'équation de l'hyperplan optimal.
- Q7 Tracer les séparatrices résultantes dans l'espace d'entrée $\mathbb{R}^2$.