Apprentissage et noyaux séparateur à vaste marge (SVM)

Pour quoi faire ?

9 Novembre 2006

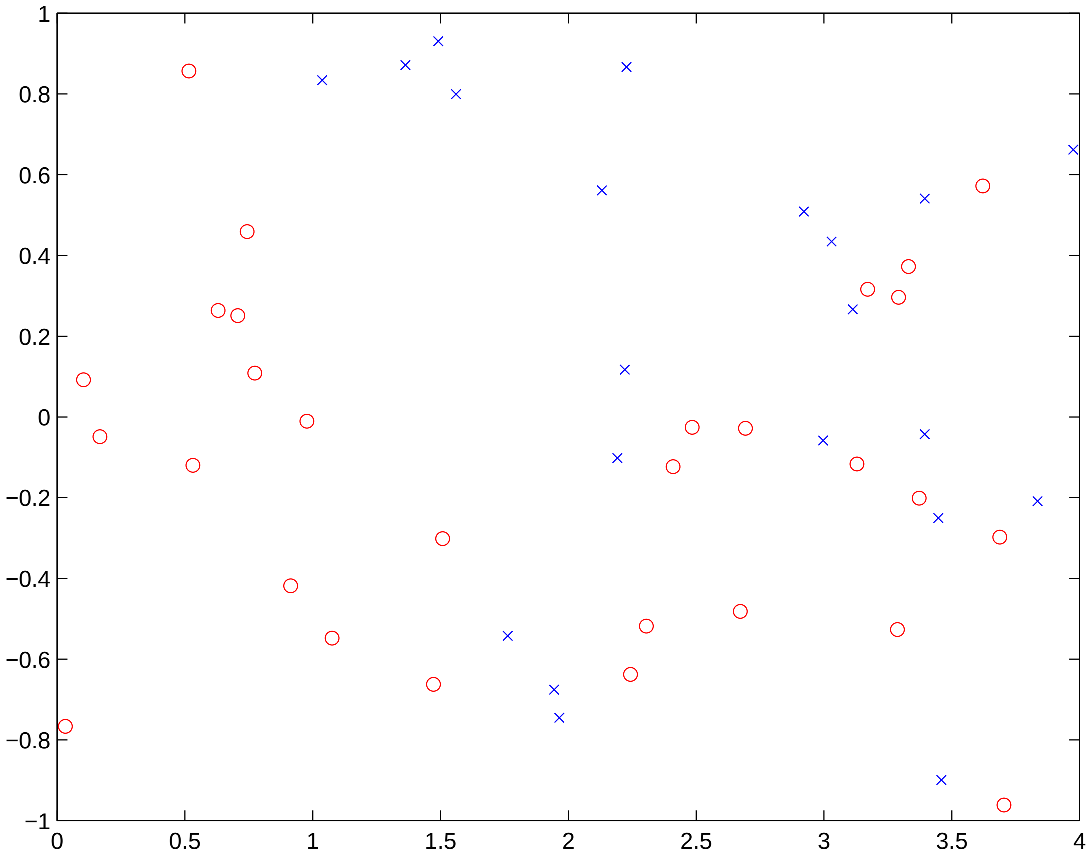
Stéphane Canu
`stephane.canu@insa-rouen.fr`

`asi.insa-rouen.fr/~scanu`

INSA Rouen - Département ASI

Laboratoire LITIS

A la recherche d'une règle de décision universelle



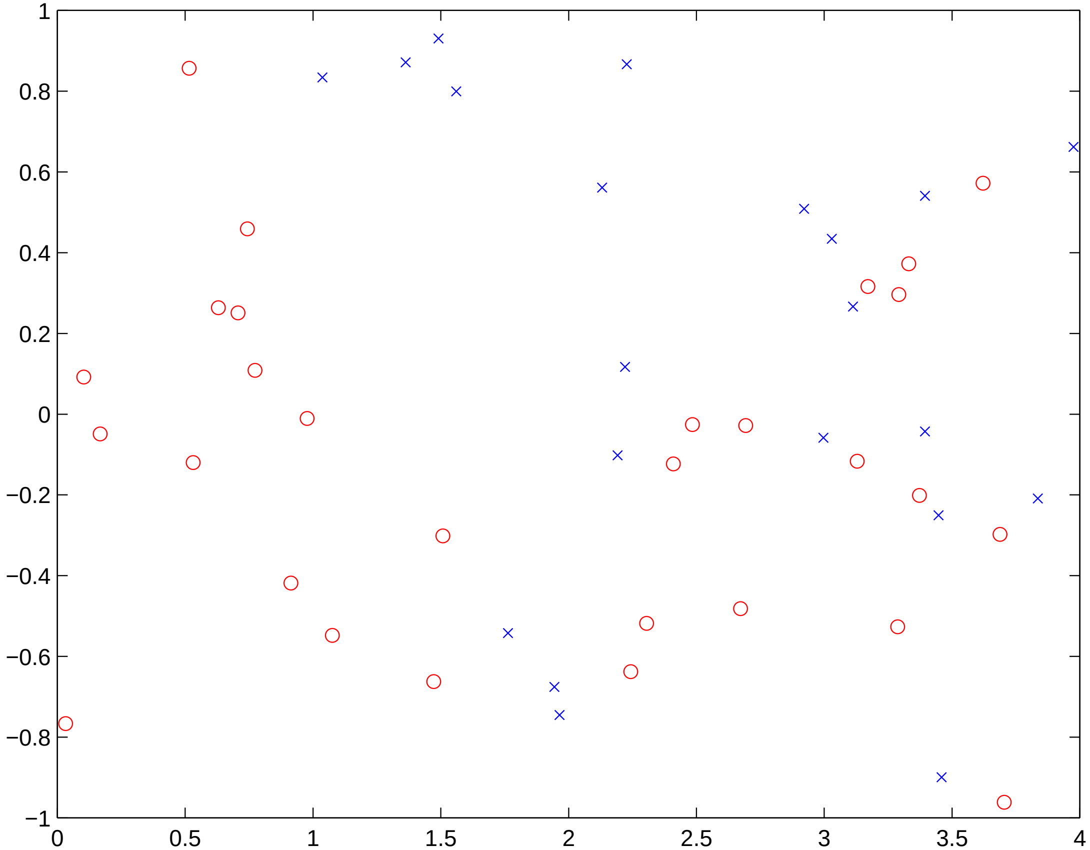
on cherche un algorithme $\mathcal{A}$
capable de résoudre
tous les problèmes

l'échantillon $(x_i, y_i)_{i=1,n}$

$$\underbrace{\mathbb{P}(err(f, x_i, y_i))}_{\text{erreur de } f} \xrightarrow[n\to\infty]{} \underbrace{\mathbb{P}_\mathsf{b}(err)}_{\text{erreur de bayes}}$$

Tracez la frontière de décision entre ces deux classes ?

A la recherche d'une règle de décision universelle



on cherche un algorithme $\mathcal{A}$
capable de résoudre
tous les problèmes

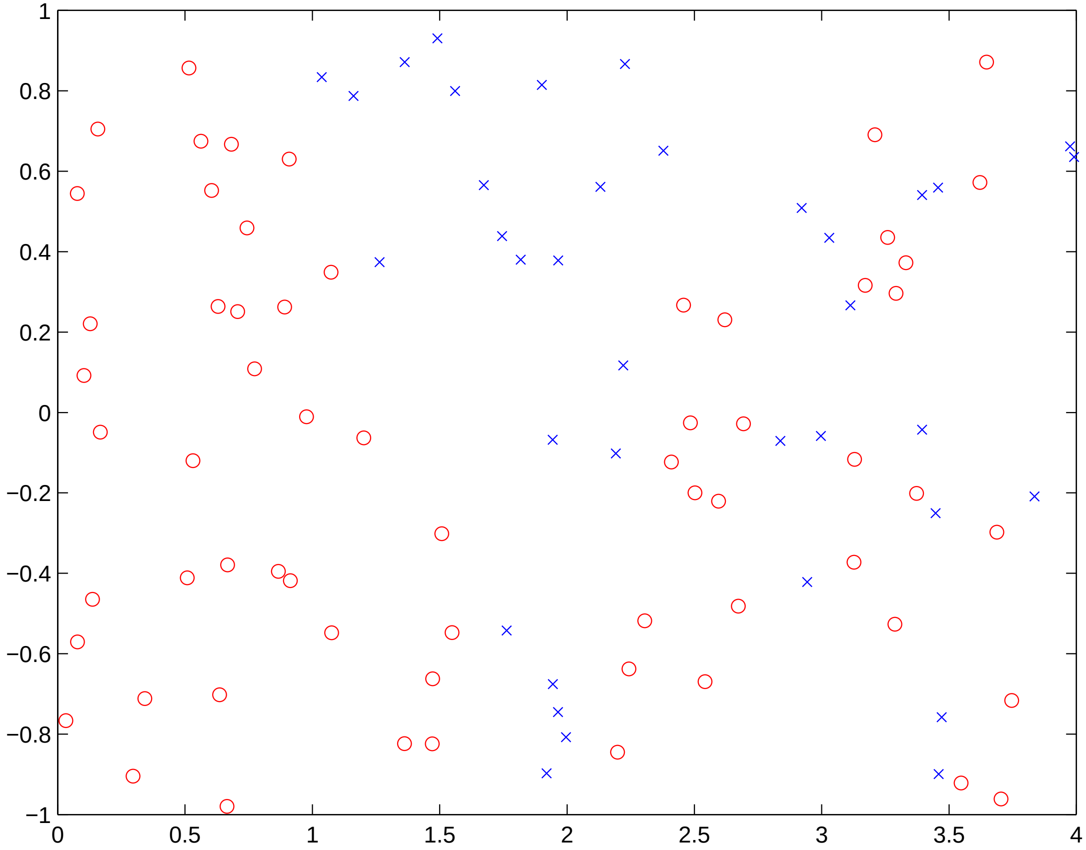
l'échantillon $(x_i, y_i)_{i=1,n}$

$$\underbrace{\mathbb{P}(err(f, x_i, y_i))}_{\text{erreur de } f} \xrightarrow[n \to \infty]{} \underbrace{\mathbb{P}_{\text{b}}(err)}_{\text{erreur de bayes}}$$

Tracez la frontière de décision entre ces deux classes ?

Universelle : pour tous les problèmes

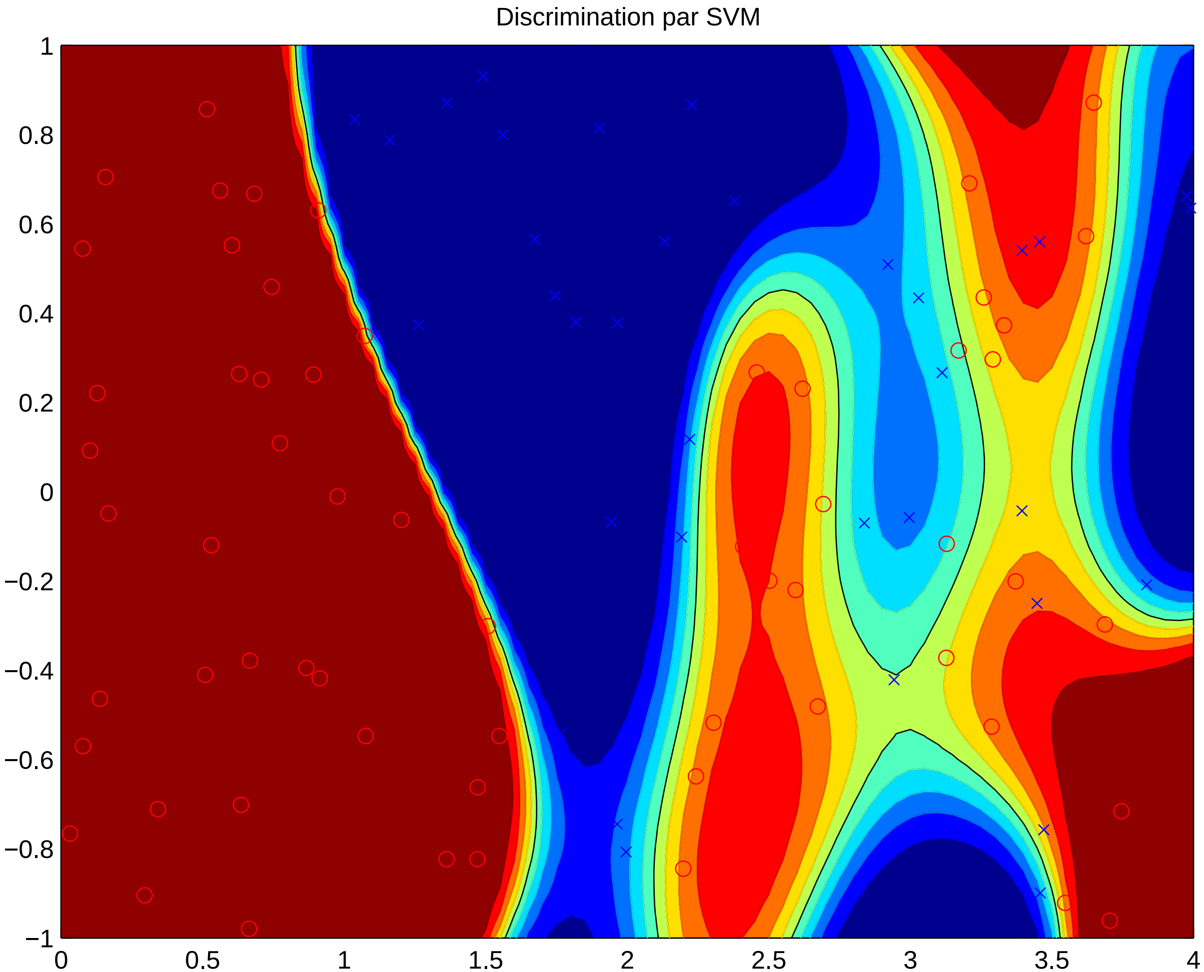
Introduction



c'est plus facile...
...avec un peu plus de points

Tracez la frontière de décision entre ces deux classes ?

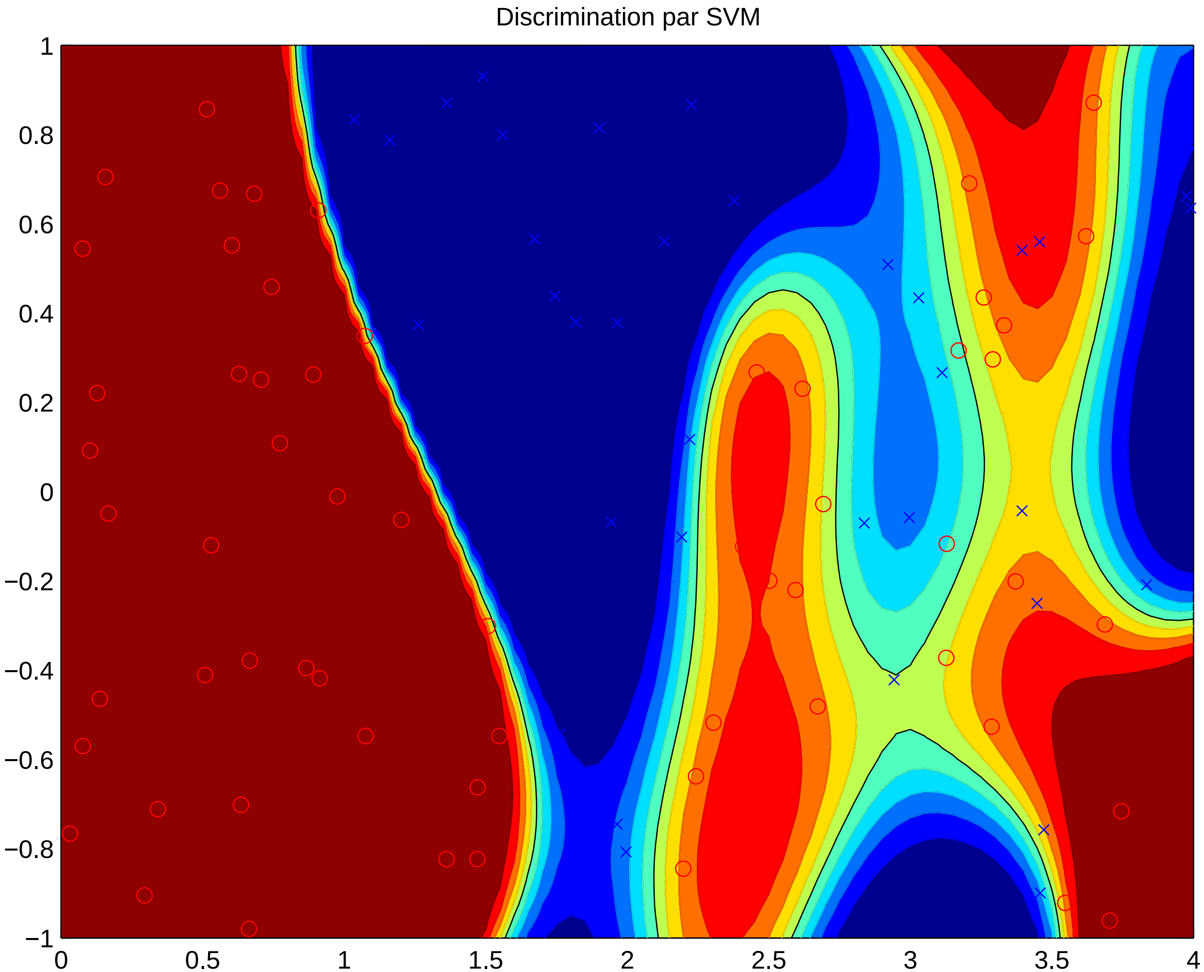
Introduction : le plan



Une solution
⇒ Quels critères ?

- (1) Fidélité

Introduction : le plan

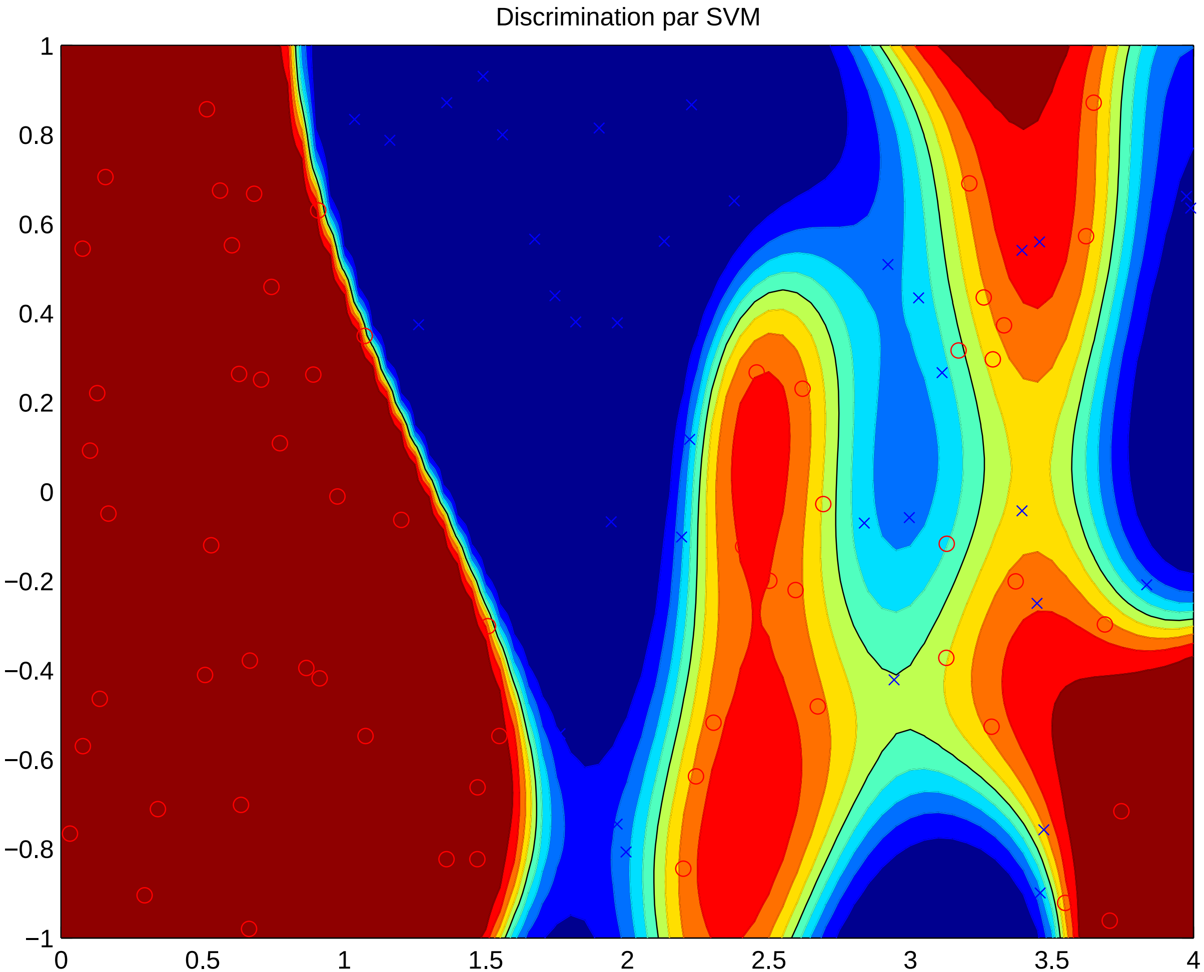


Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité

<u>Théorème</u> [Vapnik, 1979] : $\underbrace{\mathbb{P}(\text{err}[f])}_{\text{erreur}} \leq \underbrace{R_{emp}[f]}_{\text{fidélité}} + \underbrace{\sqrt{\frac{h(log(2n/h)+1) - log(\eta/4)}{n}}}_{\text{régularité}}$

Introduction : le plan

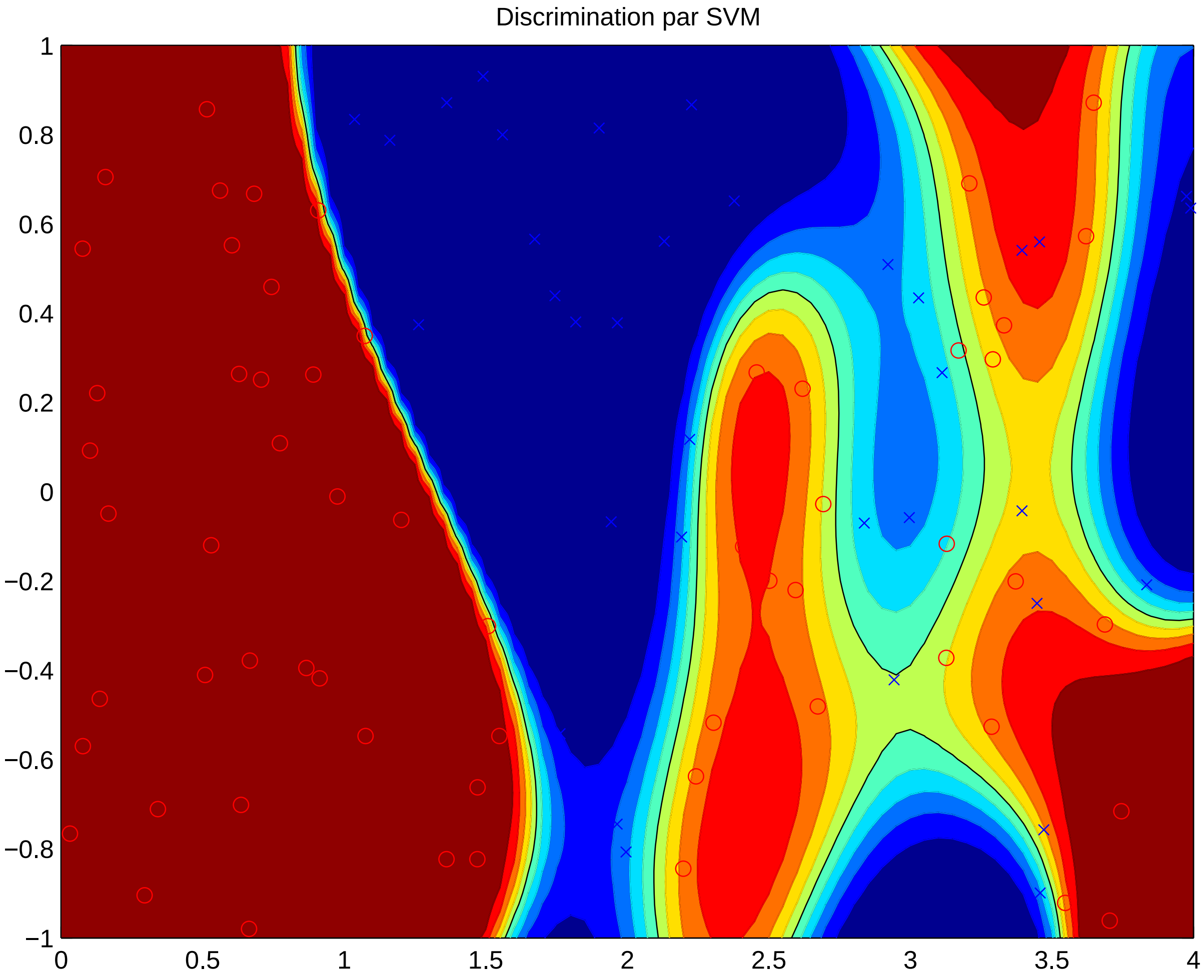


Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité
- (3) Décision locale

<u>Théorème</u> [Vapnik, 1979] : $\underbrace{\mathbb{P}(\text{err}[f])}_{\text{erreur}} \leq \underbrace{R_{emp}[f]}_{\text{fidélité}} + \underbrace{\sqrt{\frac{h(log(2n/h)+1) - log(\eta/4)}{n}}}_{\text{régularité}}$

Introduction : le plan

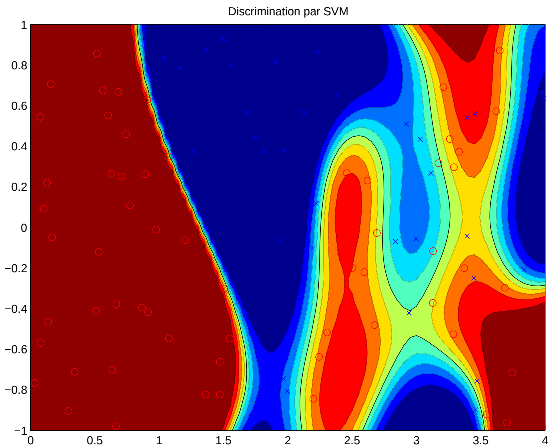


Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité
- (3) Décision locale
- (4) Points frontière

<u>Théorème</u> [Vapnik, 1979] : $\underbrace{\mathbb{P}(\text{err}[f])}_{\text{erreur}} \leq \underbrace{R_{emp}[f]}_{\text{fidélité}} + \underbrace{\sqrt{\frac{h(log(2n/h)+1) - log(\eta/4)}{n}}}_{\text{régularité}}$

Fidélité



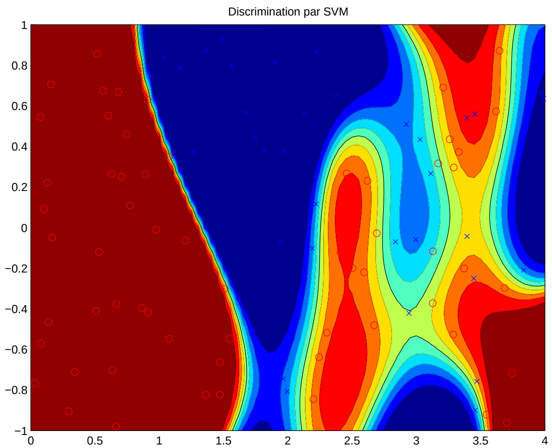
l'échantillon $(x_i, y_i)_{i=1,n}$
$y_i \in \{-1, 1\}$ (codage -1/1)
la fonction de décision : signe$(f(x_i))$
(f fonction de discrimination)
$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière de décision.

Bien classer tout le monde :

$$\text{signe}(f(x_i)) = y_i \qquad i = 1, n$$

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité



l'échantillon $(x_i, y_i)_{i=1,n}$
$y_i \in \{-1, 1\}$ (codage -1/1)
la fonction de décision : signe$(f(x_i))$
(f fonction de discrimination)
$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière de décision.

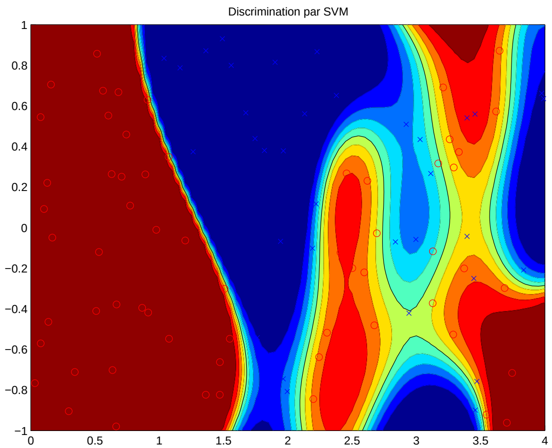
Bien classer tout le monde :

$$\text{signe}(f(x_i)) = y_i \qquad i = 1, n$$ critère non dérivable

$$f(x_i)y_i \geq 0 \qquad i = 1, n$$

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité



l'échantillon $(x_i, y_i)_{i=1,n}$
$y_i \in \{-1, 1\}$ (codage -1/1)
la fonction de décision : signe$(f(x_i))$
(f fonction de discrimination)
$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière de décision.

Bien classer tout le monde :

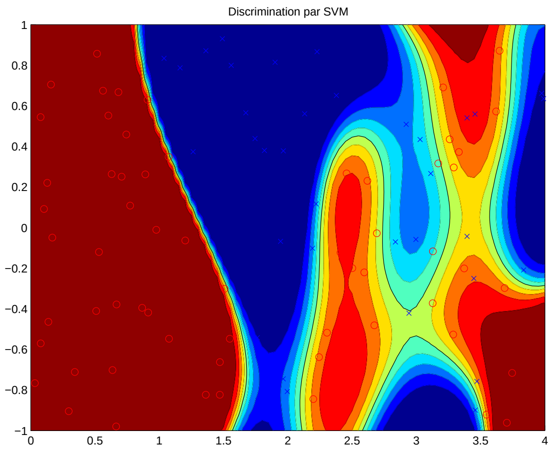
signe$(f(x_i)) = y_i \qquad i = 1, n$ critère non dérivable

$f(x_i) y_i \geq 0 \qquad i = 1, n$ solution triviale $f = 0$

$$\boxed{f(x_i) y_i \geq k \qquad k > 0,\ i = 1, n}$$

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité



l'échantillon $(x_i, y_i)_{i=1,n}$
$y_i \in \{-1, 1\}$ (codage -1/1)
la fonction de décision : signe$(f(x_i))$
(f fonction de discrimination)
$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière de décision.

Bien classer tout le monde :

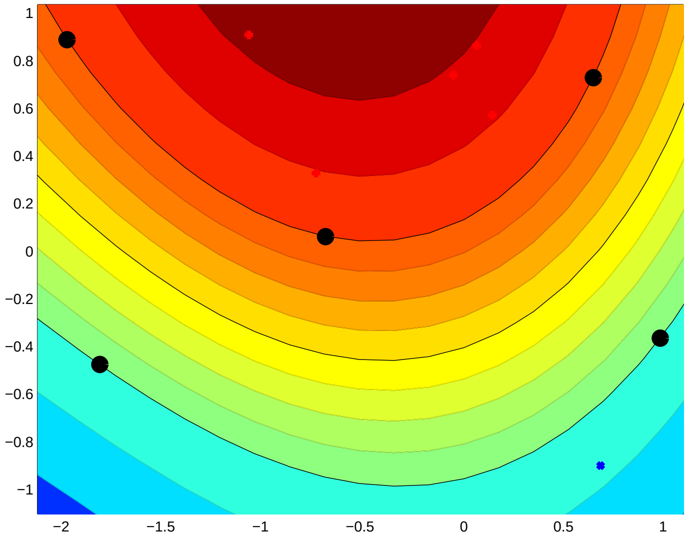
signe$(f(x_i)) = y_i$	$i = 1, n$	critère non dérivable
$f(x_i)y_i \geq 0$	$i = 1, n$	solution triviale $f = 0$

$$\boxed{f(x_i)y_i \geq k \qquad k > 0,\ i = 1, n}$$

Marge

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité et marge



$f(x_i)y_i > k \qquad k > 0,\ i = 1, n$

$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière

Marge :

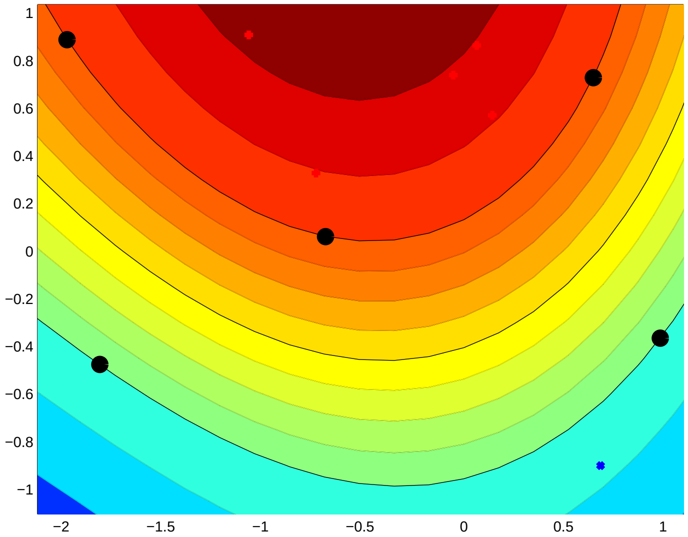
$$\min_{i=1,n} d(\mathcal{H}, x_i) = \min_{i=1,n} max(1 - f(x_i)y_i, 0)$$

Bien classer tout le monde $(k = 1)$:

$$f(x_i)y_i > 1 \qquad i = 1, n$$

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité et marge



$$f(x_i)y_i > k \qquad k > 0,\ i = 1, n$$

$\mathcal{H} = \{x \mid f(x) = 0\}$: frontière

Marge :

$$\min_{i=1,n} d(\mathcal{H}, x_i) = \min_{i=1,n} max(1 - f(x_i)y_i, 0)$$

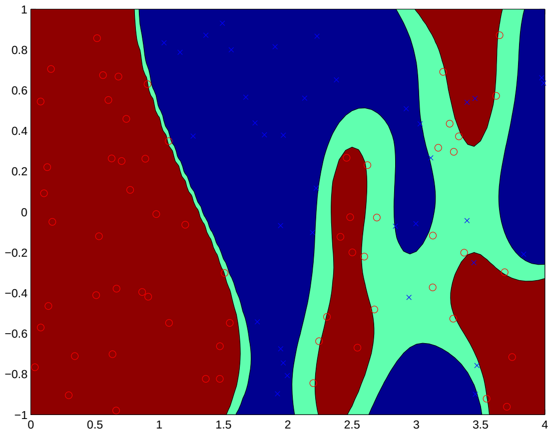
Bien classer tout le monde $(k = 1)$:

$$f(x_i)y_i > 1 \qquad i = 1, n$$

1 est la marge minimale

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité et droit à l'erreur : minimiser l'erreur



$$f(x_i)y_i > 1 \qquad i = 1, n$$

Introduisons une variable d'écart ξ_i

Bien classer **a peu près** tout le monde :

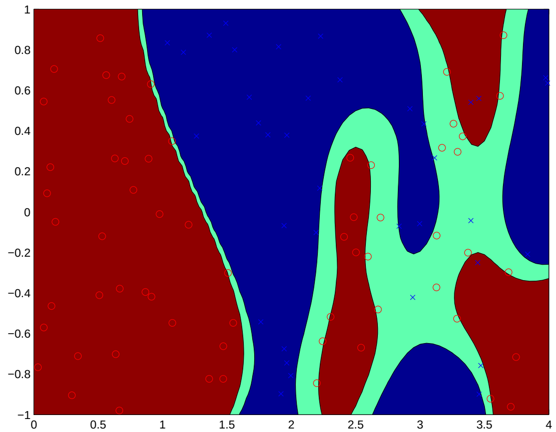
$$f(x_i)y_i > 1 \; - \xi_i \qquad \xi_i > 0 \; , \; i = 1, n$$

où ξ_i est une variable d'écart

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Fidélité et droit à l'erreur : minimiser l'erreur



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

Introduisons une variable d'écart ξ_i

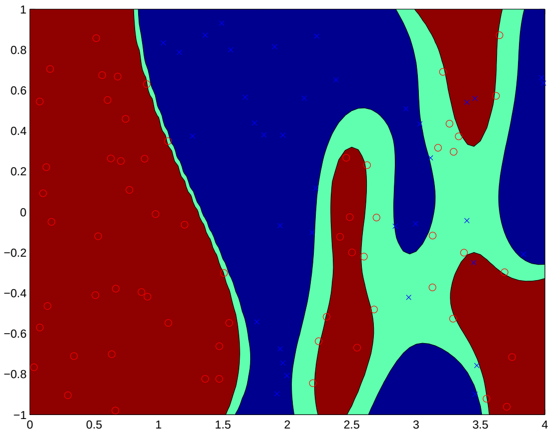
Bien classer **a peu près** tout le monde :

$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0\ ,\ i = 1, n$$

où ξ_i est une variable d'écart

(1) Fidélité - (3) Décision « locale »
(2) Régularité - (4) Points « frontière »

Fidélité et droit à l'erreur : minimiser l'erreur



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

Introduisons une variable d'écart ξ_i

$$\boxed{\min_{\xi_i} \sum_{i=1}^{n} \xi_i}$$

Bien classer **a peu près** tout le monde :

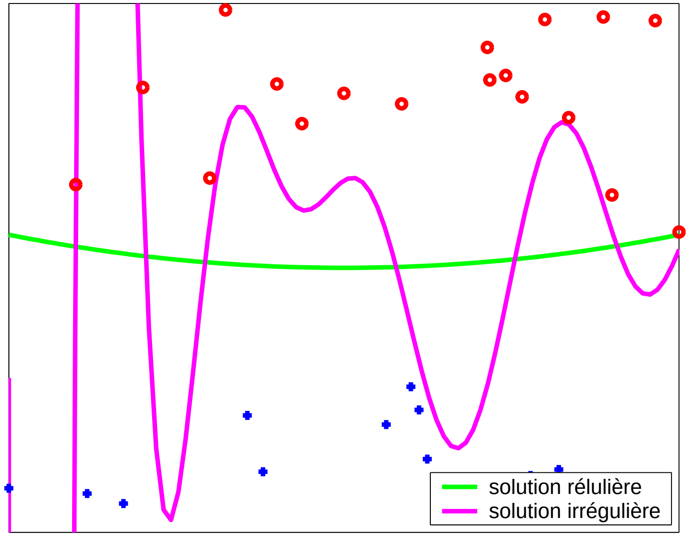
$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0\ ,\ i = 1, n$$

où ξ_i est une variable d'écart $\quad \xi_i = 0; \qquad \underbrace{\xi_i \geq 1}_{\text{mal classé}} \quad ; \quad 0 < \xi_i < 1$

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Régularité



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

les deux solutions vérifient $\xi_i = 0, \quad i = 1, n$

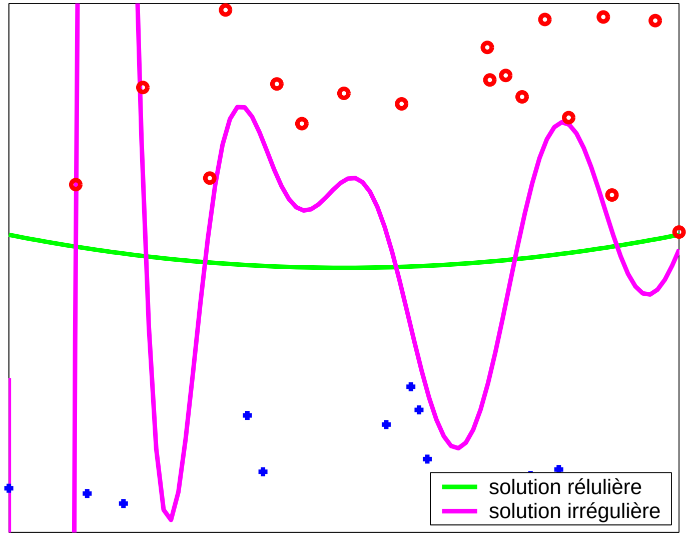
par exemple

- « l'énergie » de f : la norme de sa dérivée (*cf* les splines)

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Régularité



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

les deux solutions vérifient $\xi_i = 0, \quad i = 1, n$

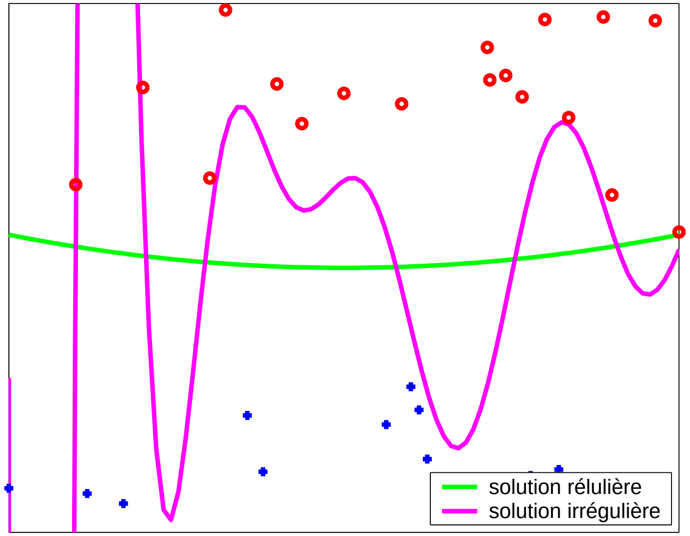
par exemple

- « l'énergie » de f : la norme de sa dérivée (*cf* les splines)
- la longueur de f - la taille du code calculant f

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Régularité



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

les deux solutions vérifient $\xi_i = 0, \quad i = 1, n$

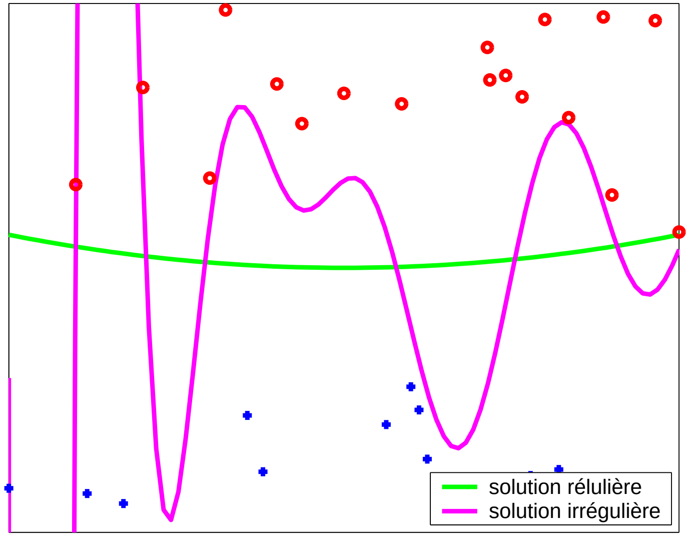
par exemple

- « l'énergie » de f : la norme de sa dérivée (*cf* les splines)
- la longueur de f - la taille du code calculant f
- une norme de f au sens de $\mathcal{H}$ (défini a priori) : $\|f\|_{\mathcal{H}}$

(1) Fidélité - (3) Décision « locale »

$\boxed{\textbf{(2) Régularité}}$ - (4) Points « frontière »

Régularité



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

$$\min_{f} \ \|f\|_{\mathcal{H}} \qquad \left(f(x) = \sum_{j \in J} w_i \phi_i(x) + b\right)$$

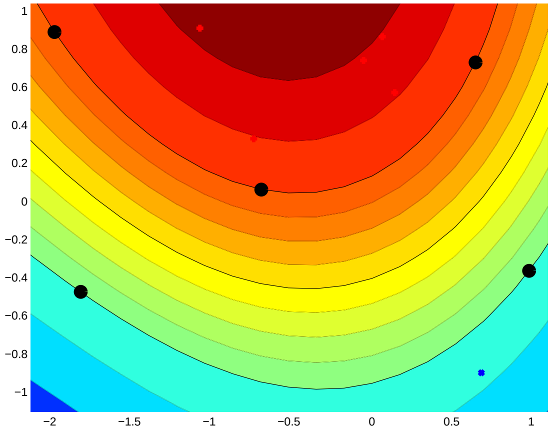
par exemple

- « l'énergie » de f : la norme de sa dérivée (*cf* les splines)
- la longueur de f - la taille du code calculant f
- une norme de f au sens de $\mathcal{H}$ (défini a priori) : $\|f\|_{\mathcal{H}}$
- un terme de régularisation : une fonctionelle positive assurant l'unicité de la solution

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Régularité et marge



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

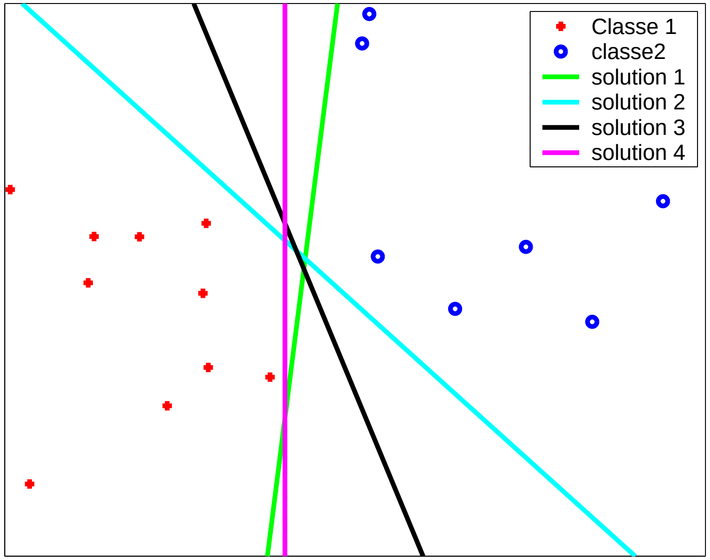
- $\mathbb{P}(err) \leq \underbrace{\sum_{i=1}^{n} 1\!\!1_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{Fidélité}} + \varphi\left(\frac{1}{\text{marge}}\right)$

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Régularité et marge

Cas linéaire : quelle solution choisir ?



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

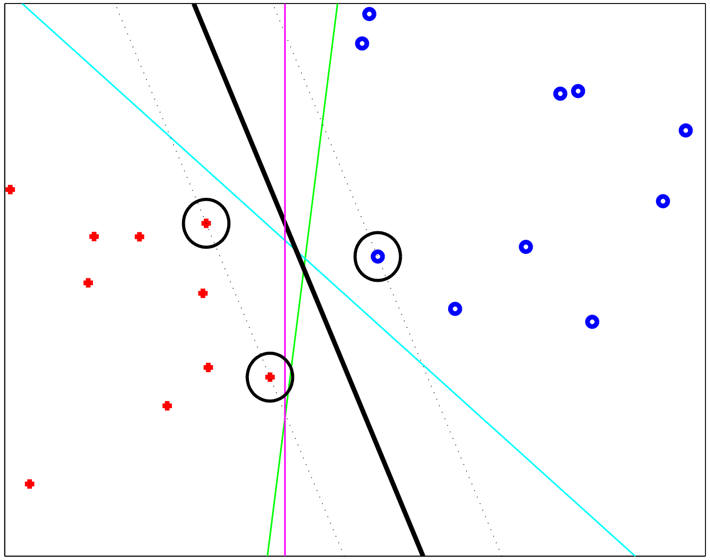
$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

- $\mathbb{P}(err) \leq \underbrace{\sum_{i=1}^{n} \mathbb{1}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{Fidélité}} + \varphi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err) \Leftrightarrow$ **maximiser** la marge

(1) Fidélité - (3) Décision « locale »

$\boxed{\textbf{(2) Régularité}}$ - (4) Points « frontière »

Régularité et marge



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

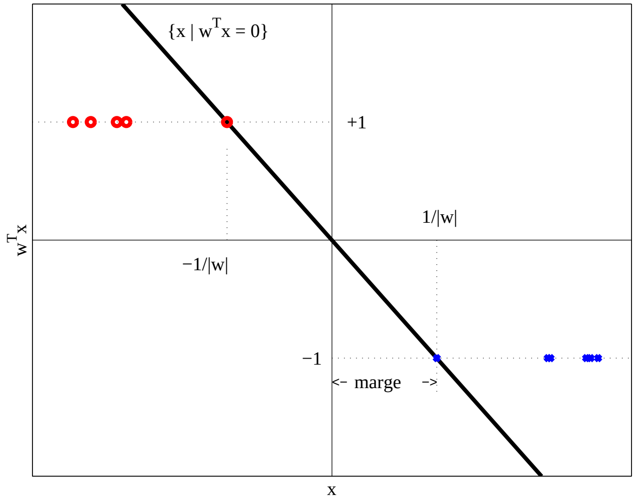
- $\mathbb{P}(err) \leq \underbrace{\sum_{i=1}^{n} \mathbb{1}_{\{\text{signe}_{(f(x_i))\neq y_i}\}}}_{\text{Fidélité}} + \varphi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err) \Leftrightarrow$ **maximiser** la marge
- maximiser la robustesse $\Leftrightarrow$ maximiser la marge

(1) Fidélité - (3) Décision « locale »

$\boxed{\textbf{(2) Régularité}}$ - (4) Points « frontière »

Régularité et marge

Valeur de la marge dans le cas monodimensionnel



$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n$$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

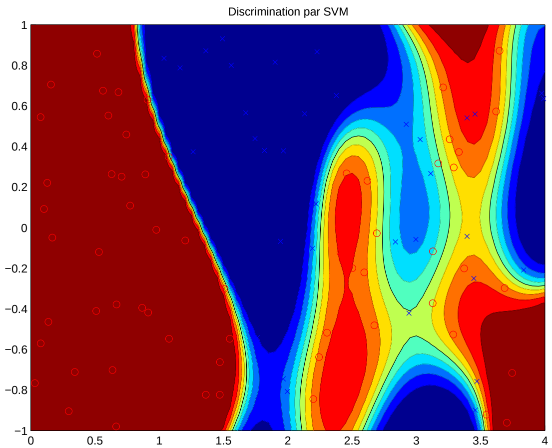
$$\min_{f} \ \|f\|^2_{\mathcal{H}} \quad \Leftrightarrow \quad \min_{\mathbf{w}} \sum_{j=1}^{\infty} w_j^2$$

- $\mathbb{P}(err) \leq \underbrace{\sum_{i=1}^{n} \mathbb{1}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{Fidélité}} + \varphi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err) \Leftrightarrow$ **maximiser** la marge
- maximiser la robustesse $\Leftrightarrow$ maximiser la marge
- maximiser la marge $\Leftrightarrow$ minimiser $\|f\|^2_{\mathcal{H}}$

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Apprendre : choisir une hypothèse



- se donner un ensemble $\mathcal{H}$ assez grand
- trouver $f \in \mathcal{H}$: $f(x_i)y_i > 1 - \xi_i \quad \xi_i > 0,\ i = 1, n$

$$\min_{\xi_i} \sum_{i=1}^{n} \xi_i$$

$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

soit $(\phi_j)_{j \in J}$ une base orthonormé de fonctions (polynômes, fourier, ondelettes...)

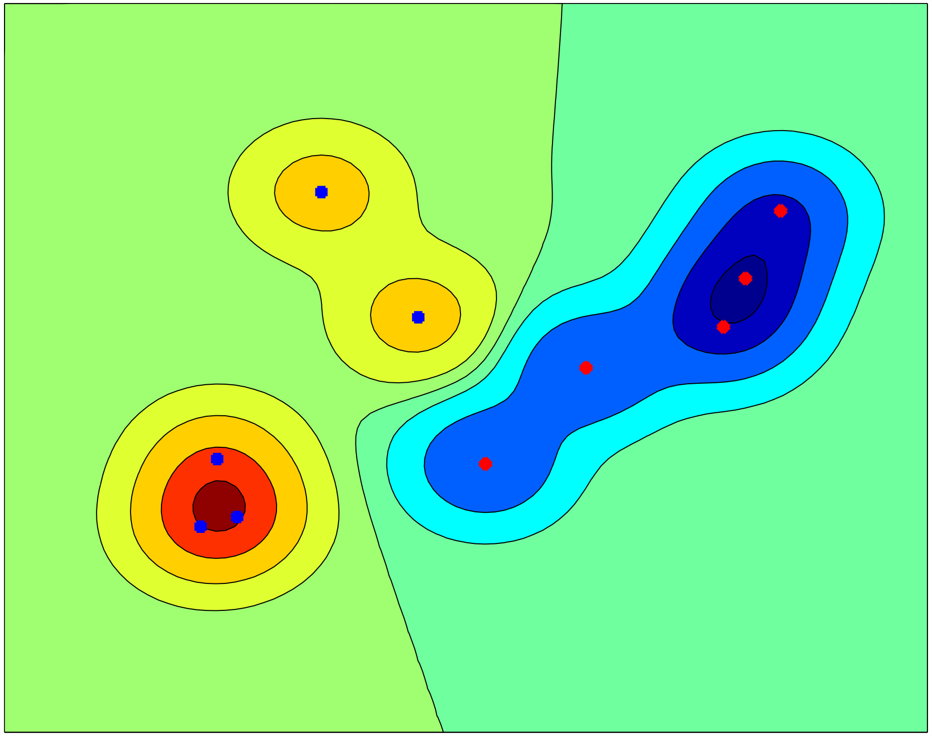
$$f(x) = \sum_{j=1}^{\infty} w_j \phi_j(x) + b$$

l'ensemble des hypothèses est alors de la forme

$$\mathcal{H} = \left\{ f \;\middle|\; f(x) = \sum_{j=1}^{\infty} w_j \phi_j(x) + b \right\}$$

f est « linéaire » en ϕ et non linéaire en x

Principe : « qui se ressemble s'assemble »



Principes :

– Mesure de similarité

(pas nécéssairement symètrique)

– Zone d'influence

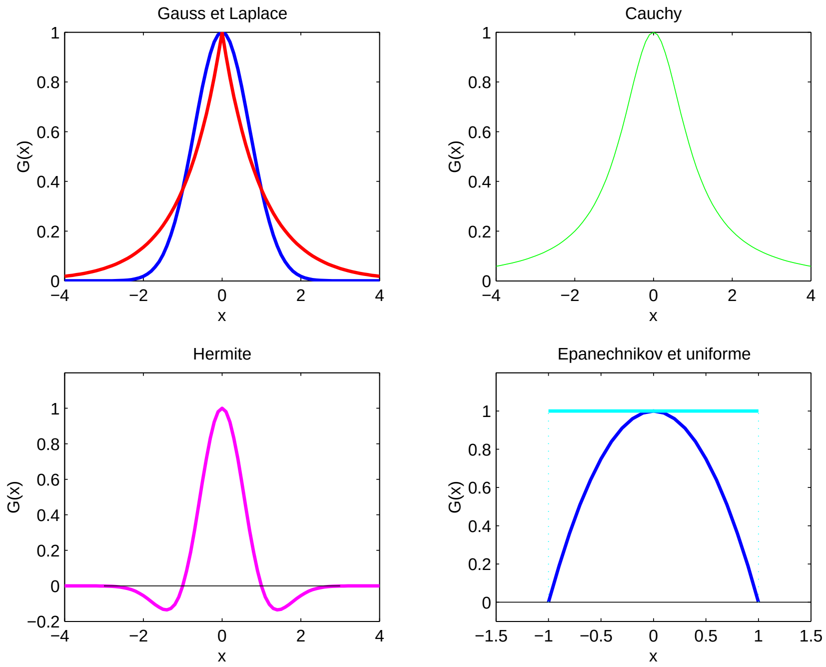
Les noyaux :

– fonction de deux variables

$$K(x, y)$$

Estimateur à base de **noyaux**

Malédiction de la dimensionnaleté



noyaux défini positifs

noyau multidimensionnel produit :

$$K_b(\mathbf{u}, \mathbf{v}) = \prod_{\ell=1}^{L} K_b(u_\ell, v_\ell)$$

noyaux « radiaux » $\rho = \|\mathbf{u} - \mathbf{v}\|^2$ (distance entre les deux variables)

noyaux « projectifs » $\mathbf{u}^\top \mathbf{v} = \sum_{\ell=1}^{L} u_\ell v_\ell$

formule de « passage »

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}^\top \mathbf{v}$$

Quelques exemples de noyaux

le noyau gaussien

$$K_b(u,v) = \frac{1}{Z}\exp^{-\frac{\rho}{b}}$$

le noyau de Cauchy

$$K_b(u,v) = \frac{1}{Z}\frac{1}{1+\frac{\rho}{b}}$$

le noyau uniforme

$$K_b(u,v) = \frac{1}{Z}\mathbb{1}_{\{\rho \le b\}}$$

le noyau de Fourier régularisé

$$K_b(u,v) = \frac{1}{Z}\cosh\left(\pi - \frac{|u-v|}{b}\right)$$

le noyau scalaire

$$K_b(\mathbf{u},\mathbf{v}) = (\mathbf{u}^\top\mathbf{v}+1)^b$$

le noyau de Laplace

$$K_b(u,v) = \frac{1}{Z}\exp^{-\frac{|u-v|}{b}}$$

le noyau d'Hermite

$$K_b(u,v) = \frac{1}{Z}(b-\rho)\ \exp^{-\frac{\rho}{b}}$$

le noyau d'Epanechnikov

$$K_b(u,v) = \frac{1}{Z}(b-\rho)\ \mathbb{1}_{\{\rho \le b\}}$$

le noyau sigmoïde

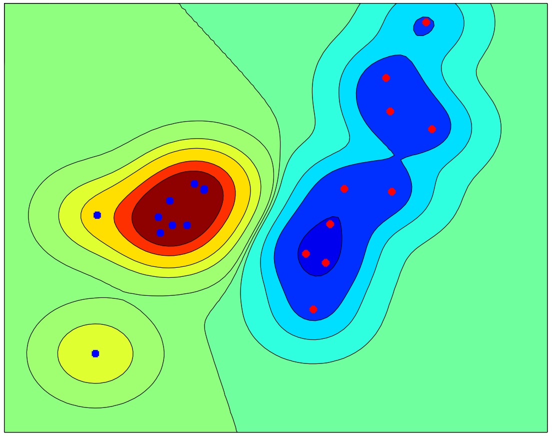
$$K_b(\mathbf{u},\mathbf{v}) = \frac{1}{Z}\tanh\left(b(\mathbf{u}^\top\mathbf{v}) + b_0\right)$$

le noyau de Hardy

$$K_b(u,v) = \frac{1}{(\mathbf{u}^\top\mathbf{v}+1)^b}$$

noyaux de chaines de caractères, de graphes, d'automates...

Comment choisir $\mathcal{H}$?



$$f(x) = \sum_{j=1}^{\infty} w_j \phi_j(x)$$

Influence locale $\Rightarrow$ Noyaux. Influence fixe (Parzen + MAP)

$$f(x) = \sum_{i=1}^{n} y_i K_b(x, x_i)$$

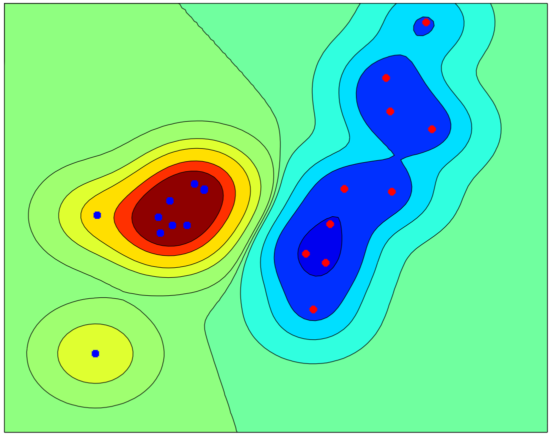
Influence ajustée

$$f(x) = \sum_{i=1}^{n} a_i K_b(x, x_i)$$

(1) Fidélité - **(3) Décision « locale »**

(2) Régularité - (4) Points « frontière »

Comment choisir $\mathcal{H}$?



$$f(x) = \sum_{j=1}^{\infty} w_j \phi_j(x)$$

Influence locale $\Rightarrow$ Noyaux. Influence fixe (Parzen + MAP)

$$f(x) = \sum_{i=1}^{n} y_i K_b(x, x_i)$$

Influence ajustée

$$f(x) = \sum_{i=1}^{n} a_i K_b(x, x_i)$$

on construit $\mathcal{H}$ à partir du noyau K

(1) Fidélité - **(3) Décision « locale »**

(2) Régularité - (4) Points « frontière »

Comment choisir construire $\mathcal{H}$ (les hypothèses ?)

- au commencement était le noyau...
  - noyau : $k(x,y) \qquad \forall x,y \in \Omega$
  - ...positif $\sum_{i=1}^n \sum_{j=1}^m \alpha_i \alpha_j k(x_i,x_j) > 0$
  - $\mathcal{H}_0 = \{f \in \mathbb{R}^\Omega | f(x) = \displaystyle\sum_{i=1}^n \alpha_i k(x,x_i), n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \Omega\}$
  - et un produit scalaire on $\mathcal{H}_0$

$$\langle f(.), g(.)\rangle_{\mathcal{H}_0} = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i,x_j)$$

- propriétés du produit scalaire sur $\mathcal{H}_0$
  - $\langle f(.), k(x,.)\rangle_{\mathcal{H}_0} = f(x)$ Evaluation
  - $\langle k(x,.), k(y,.)\rangle_{\mathcal{H}_0} = k(x,y)$ (reproduction)

le noyau représente une grande partie de la connaissance a priori

L'astuce du noyau

Théorème de Mercer : Si K est un noyau défini positif, il existe une famille $(\phi_j)_{j\in\mathcal{J}}$ orthonormée telle que :

$$K_b(\mathbf{x},\mathbf{y}) = \sum_{j\in J} \phi_j(\mathbf{x})\,\phi_j(\mathbf{y}) \tag{1}$$

toute fonction $f \in \mathcal{H}$ s'écrit alors :

$$f(\mathbf{x}) = \sum_{j\in J} w_j \phi_j(\mathbf{x}) = \sum_{i=1}^{n} a_i K_b(\mathbf{x},\mathbf{x}_i)$$

$$\|f\|^2_{\mathcal{H}} = \mathbf{w}^\top \mathbf{w} = \mathbf{a}^\top K \mathbf{a}$$

où K est la matrice d'influence.
$K_{ij} = K_b(\mathbf{x}_i,\mathbf{x}_j)$

(1) Fidélité - $\boxed{\textbf{(3) Décision « locale »}}$
(2) Régularité - (4) Points « frontière »

L'astuce du noyau

Théorème de Mercer : Si K est un noyau défini positif, il existe une famille $(\phi_j)_{j \in \mathcal{J}}$ orthonormée telle que :

$$K_b(\mathbf{x}, \mathbf{y}) = \sum_{j \in J} \phi_j(\mathbf{x}) \phi_j(\mathbf{y}) \qquad (2)$$

toute fonction $f \in \mathcal{H}$ s'écrit alors :

$$f(\mathbf{x}) = \sum_{j \in J} w_j \phi_j(\mathbf{x}) = \sum_{i=1}^{n} a_i K_b(\mathbf{x}, \mathbf{x}_i)$$

$$\|f\|_{\mathcal{H}}^2 = \mathbf{w}^\top \mathbf{w} = \mathbf{a}^\top K \mathbf{a}$$

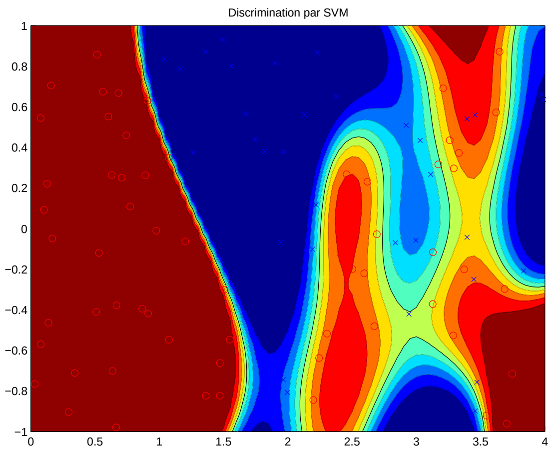
où K est la matrice d'influence. dim ∞ dim n

$K_{ij} = K_b(\mathbf{x}_i, \mathbf{x}_j)$

(1) Fidélité - **(3) Décision « locale »**

(2) Régularité - (4) Points « frontière »

Quel critère minimiser ?



$$\left.\begin{array}{l} f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0,\ i = 1, n \\ \displaystyle\min_{\xi_i} \sum_{i=1}^{n} \xi_i \end{array}\right\} \text{Fidélité}$$

$$\min_{\mathbf{w}} \sum_{j=1}^{\infty} w_j^2 \quad \ldots\ldots \qquad \text{Régularité}$$

Comment choisir la fonction de discrimination ?

$$f(x) = \sum_{j=1}^{\infty} w_j \phi_j(x)$$

(1) Fidélité - (3) Décision « locale »

(2) Régularité - (4) Points « frontière »

Ensemble d'hypothèses + Critère = Le problème SVM

$$\mathcal{H} = \left\{ f : \mathbb{R}^L \to \mathbb{R} \,\middle|\, \exists \mathbf{a}, \mathbf{c} \;;\; f(\mathbf{x}) = \sum_{j=1}^{m} c_j \varphi_j(\mathbf{x}) + \sum_{\ell=1}^{n_{\text{sup}}} a_\ell K_b(\mathbf{x}, \mathbf{x}_\ell) \right\}$$

où n_{sup} est le nombre de vecteurs supports

problème de minimisation sous contraintes :

$$\left\{ \begin{array}{lll} \min\limits_{\mathbf{w}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum\limits_{i=1}^{n} \xi_i & \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 - \xi_i & i = 1, n \\ \text{et} & \xi_i > 0 & i = 1, n \end{array} \right. \tag{3}$$

où : $f(\mathbf{x}) = \sum_{k=1}^{\infty} w_k \phi_k(\mathbf{x}) + \sum_{j=1}^{m} c_j \varphi_j(\mathbf{x})$

(1) Fidélité - **(3) Décision « locale »**

(2) Régularité - (4) Points « frontière »

Ensemble d'hypothèses + Critère = Le problème SVM

$$\mathcal{H} = \left\{ f : \mathbb{R}^L \to \mathbb{R} \,\middle|\, \exists \mathbf{a}, \mathbf{c} \; ; \; f(\mathbf{x}) = \sum_{j=1}^{m} c_j \varphi_j(\mathbf{x}) + \sum_{\ell=1}^{n_{\text{sup}}} a_\ell K_b(\mathbf{x}, \mathbf{x}_\ell) \right\}$$

où n_{sup} est le nombre de vecteurs supports
problème de minimisation sous contraintes :

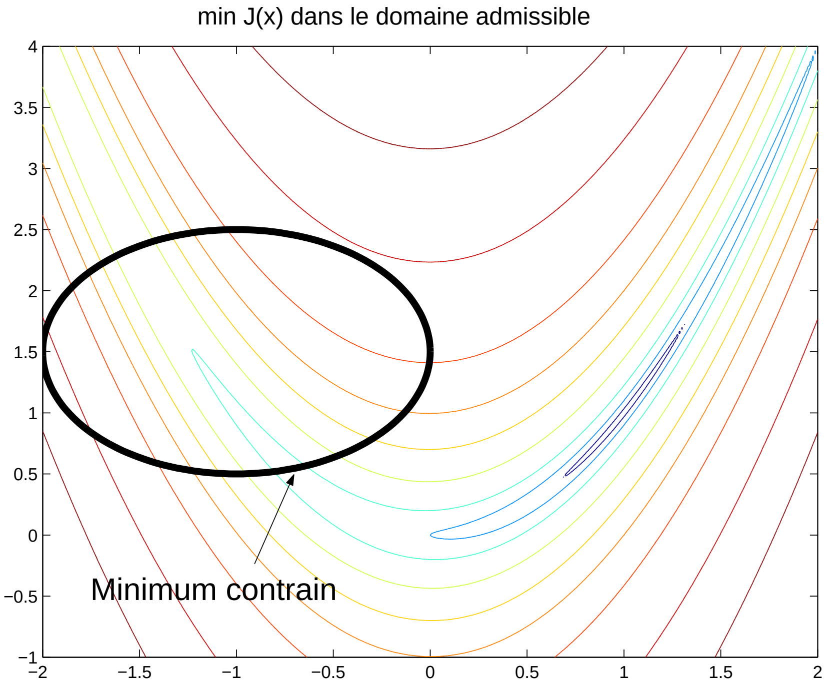
$$\begin{cases} \min\limits_{\mathbf{w}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^{n} \xi_i & \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 - \xi_i & i = 1, n \\ \text{et} & \xi_i > 0 & i = 1, n \end{cases} \tag{4}$$

où : $f(\mathbf{x}) = \sum_{k=1}^{\infty} w_k \phi_k(\mathbf{x}) + \sum_{j=1}^{m} c_j \varphi_j(\mathbf{x})$

(1) Fidélité - **(3) Décision « locale »**

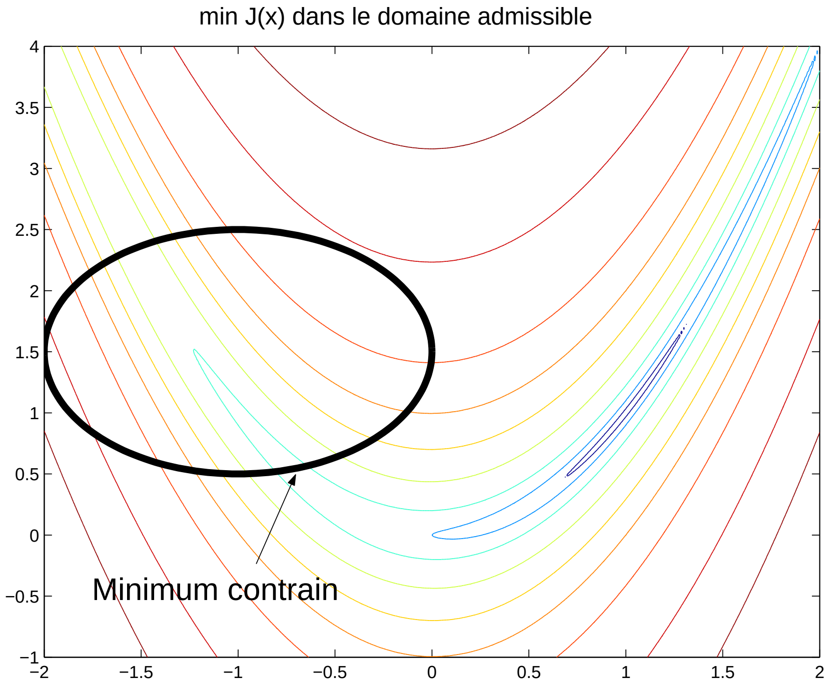
(2) Régularité - (4) Points « frontière »

Minimisation sous contraintes (cas séparable)



$$
\begin{cases}
\min\limits_{\mathbf{w}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\
\text{avec} \quad & y_i f(\mathbf{x}_i) > 1 \qquad i = 1, n
\end{cases}
$$

Minimisation sous contraintes (cas séparable)



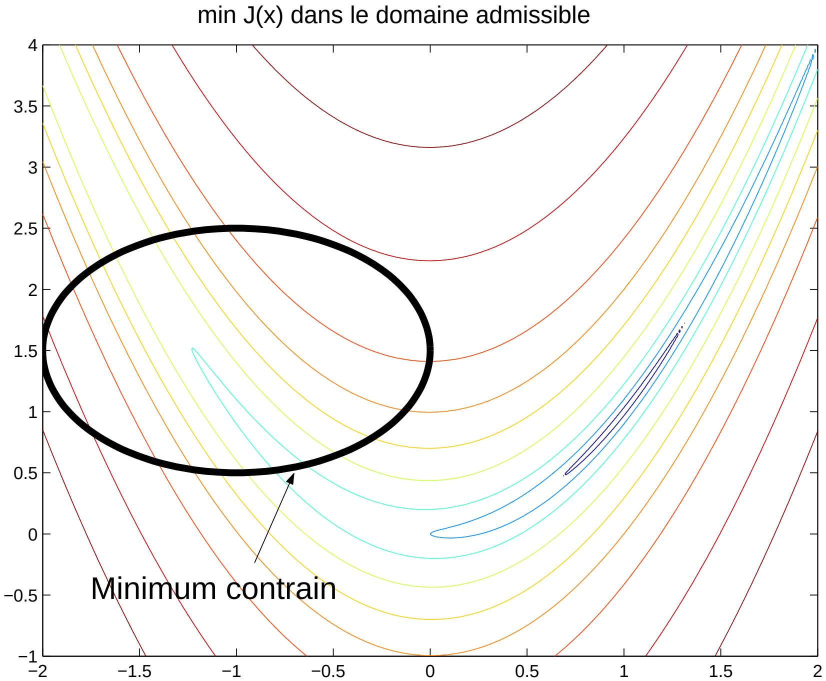
$$\begin{cases} \displaystyle\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} & \\ \text{avec} \quad y_i f(\mathbf{x}_i) > 1 & i = 1, n \end{cases}$$

$$\Leftrightarrow$$

$$\min_{\mathbf{w},\mathbf{c}} \max_{\lambda} \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda)$$ Lagrangien

$$\mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda) = \frac{1}{2}\|\mathbf{w}\|^2 - \underbrace{\sum_{i=1}^{n}}_{\text{les exemples}} \lambda_i \; \big(y_i f(\mathbf{x}_i) - 1\big)$$

Minimisation sous contraintes (cas séparable)



$$\begin{cases} \min\limits_{\mathbf{w}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 \qquad i = 1, n \end{cases}$$

$$\Leftrightarrow$$

$$\min_{\mathbf{w},\mathbf{c}} \max_{\lambda} \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda) \qquad \text{Lagrangien}$$

$$\mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda) = \frac{1}{2}\|\mathbf{w}\|^2 - \underbrace{\sum_{i=1}^{n}}_{\text{les exemples}} \lambda_i \big(y_i f(\mathbf{x}_i) - 1\big)$$

Multiplicateur de Lagrange λ_i = influence de l'exemple i dans la solution

interprètation : $\lambda_i = 0 \rightarrow$ pas d'influence $\qquad \lambda_i > 0 \rightarrow$ exemple *support*

Reformulation dans l'espace des exemples

$$\mathcal{L}(\mathbf{w}, \mathbf{a}, \lambda) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \lambda_i \big(y_i f(\mathbf{x}_i) - 1\big)$$

dont on tire les conditions de Kuhn et Tucker :

$$\left\{ \begin{array}{ccc} \dfrac{\partial \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda)}{\partial \mathbf{w}} & = & 0 \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda)}{\partial \mathbf{c}} & = & 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ccc} \mathbf{w} - \displaystyle\sum_{i=1}^{n} \lambda_i y_i \phi(\mathbf{x}_i) & = & 0 \\ \displaystyle\sum_{i=1}^{n} \lambda_i y_i \varphi(\mathbf{x}_i) & = & 0 \end{array} \right.$$

conséquence pour f :

$$\begin{aligned} f(\mathbf{x}) = \sum_{k=1}^{\infty} w_k \phi_k(\mathbf{x}) \quad & = \sum_{k=1}^{\infty} \left(\sum_{i=1}^{N} \lambda_i y_i \phi(\mathbf{x}_i) \right) \phi_k(\mathbf{x}) \\ & = \sum_{i=1}^{N} \underbrace{\lambda_i y_i}_{a_i} \underbrace{\sum_{k=1}^{\infty} \phi_k(\mathbf{x}) \phi(\mathbf{x}_i)}_{K_b(\mathbf{x}, \mathbf{x}_i)} \end{aligned}$$

Reformulation dans l'espace des exemples

$$\mathcal{L}(\mathbf{w}, \mathbf{a}, \lambda) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \lambda_i \big(y_i f(\mathbf{x}_i) - 1\big)$$

dont on tire les conditions de Kuhn et Tucker :

$$\left\{ \begin{array}{lcl} \dfrac{\partial \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda)}{\partial \mathbf{w}} & = & 0 \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, \mathbf{c}, \lambda)}{\partial \mathbf{c}} & = & 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{lcl} \mathbf{w} - \displaystyle\sum_{i=1}^{n} \lambda_i y_i \phi(\mathbf{x}_i) & = & 0 \\ \displaystyle\sum_{i=1}^{n} \lambda_i y_i \varphi(\mathbf{x}_i) & = & 0 \end{array} \right.$$

conséquence pour f :

$$\begin{aligned} f(\mathbf{x}) = \sum_{k=1}^{\infty} w_k \phi_k(\mathbf{x}) \quad &= \sum_{k=1}^{\infty} \left(\sum_{i=1}^{N} \lambda_i y_i \phi(\mathbf{x}_i) \right) \phi_k(\mathbf{x}) \\ &= \sum_{i=1}^{N} \underbrace{\lambda_i y_i}_{a_i} \underbrace{\sum_{k=1}^{\infty} \phi_k(\mathbf{x}) \phi(\mathbf{x}_i)}_{K_b(\mathbf{x}, \mathbf{x}_i)} \end{aligned}$$

$\boxed{\text{Stratégie}}$

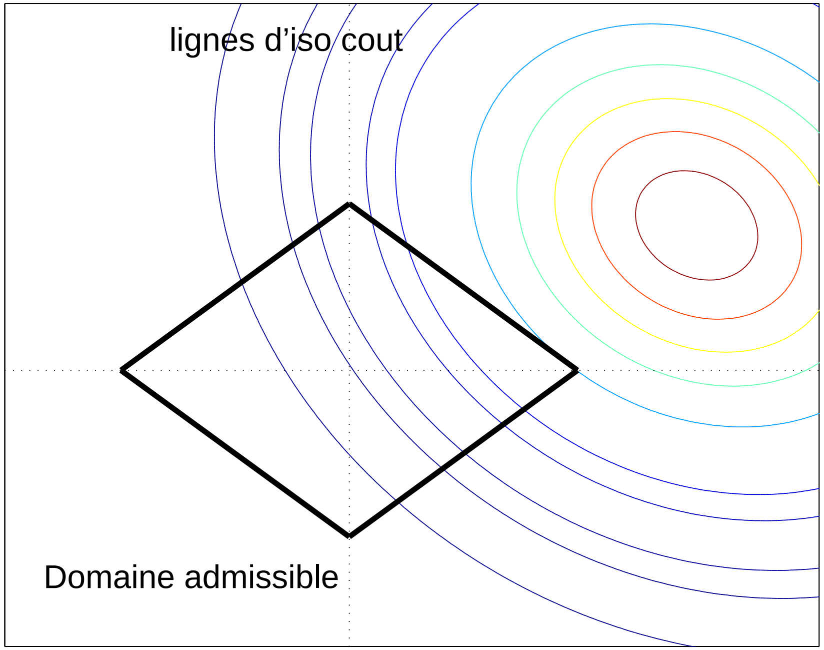
calcul de K

calcul des λ

calcul des $\mathbf{a}$

calcul de f

calcul des λ : problème Dual (2)



Reformulation de Girosi (97)

$$\left\{\begin{array}{lll} & \min_{\lambda} & \frac{1}{2}\lambda^\top H\lambda + \mathbf{c}^\top\lambda \\ \text{avec} & & \sum_{i=1}^{N}\lambda_i y_i \varphi_j(x_i) = 0 \quad j = 1, m \\ \text{et} & & 0 \leq \lambda_i \leq C \qquad i = 1, n \end{array}\right.$$

où H est la matrice de terme général $H_{ij} = y_i y_j K_b(\mathbf{x}_i, \mathbf{x}_j)$
et $\mathbf{c}$ un vecteur de 1.

$$\min_{\mathbf{a}} \|f(\mathbf{x}_i) - y_i\|^2_{\mathcal{H}} + \mu \sum_{i=1}^{n} |a_i|$$

(1) Fidélité - (3) Décision « locale »

(2) Régularité - $\boxed{\textbf{(4) Points « frontière »}}$

Solution pratique : Problème d'optimisation

- dans le pire des cas...Simplex
- Solution « hors lignes » :
  - Contraintes actives ($\mathcal{O}(n^{1.5})$)
    - asi.insa-rouen.fr/~gloosli
    - asi.insa-rouen.fr/~arakotom
  - Points intérieurs : lent
  - stochastiques : SMO
    - libsvm rapide et complet
    - coresvm (pas très fiable)
- Solution « en ligne » : LaSVM
  - La SVM : très très rapides - $8\ 10^6$ exemples...

si n est grand : plus efficace que les autres méthodes

Conclusion

- Les méthodes à noyaux
  - approximateur universel
  - minimum global unique
  - Parcimonieux (des coef. = 0)

  $\Rightarrow$ très rapide (en général)

- Classification : les SVM
  - SVM vs Réseaux de neurones (PMC) : optimisation
  - SVM vs Parzen : parcimonie et vitesse (L^2 vs L^1)
  - SVM vs régression logistique : parcimonie et vitesse
  - SVM : des résultats

- Régression : le kLAR
  - kLAR vs Réseaux de neurones (PMC) : optimisation
  - kLAR vs Noyaux (FBR) : optimisation
  - kLAR vs splines : parcimonie et vitesse (L^2 vs L^1)

- autres applications : one class SVM, kACP, kPLS...

Références

- SVM
  - V. Vapnik : The Nature of Statistical Learning Theory. Springer, 1995.
  - N. Cristianini and J. Shawe-Taylor :An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000
  - B. Scholkopf et A. Smola : Kernel Machines, 2002
- Reconnaissance des formes statistiques
  - R. O. Duda, P. E. Hart and D. G. Stork : Pattern Classification (2nd ed.), John Wiley and Sons, 2001.
  - T. Hastie, R. Tibshirani, and J. Fridman : The Elements of Statistical Learning :Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- et sur le réseau
  - http ://kernel-machines.org
  - http ://www.ph.tn.tudelft.nl/PRInfo/
  - http ://citeseer.nj.nec.com/
  - http ://asi.insa-rouen.fr/~scanu