Apprentissage, reproduction, régularisation et noyaux séparateur à vaste marge (*SVM*)

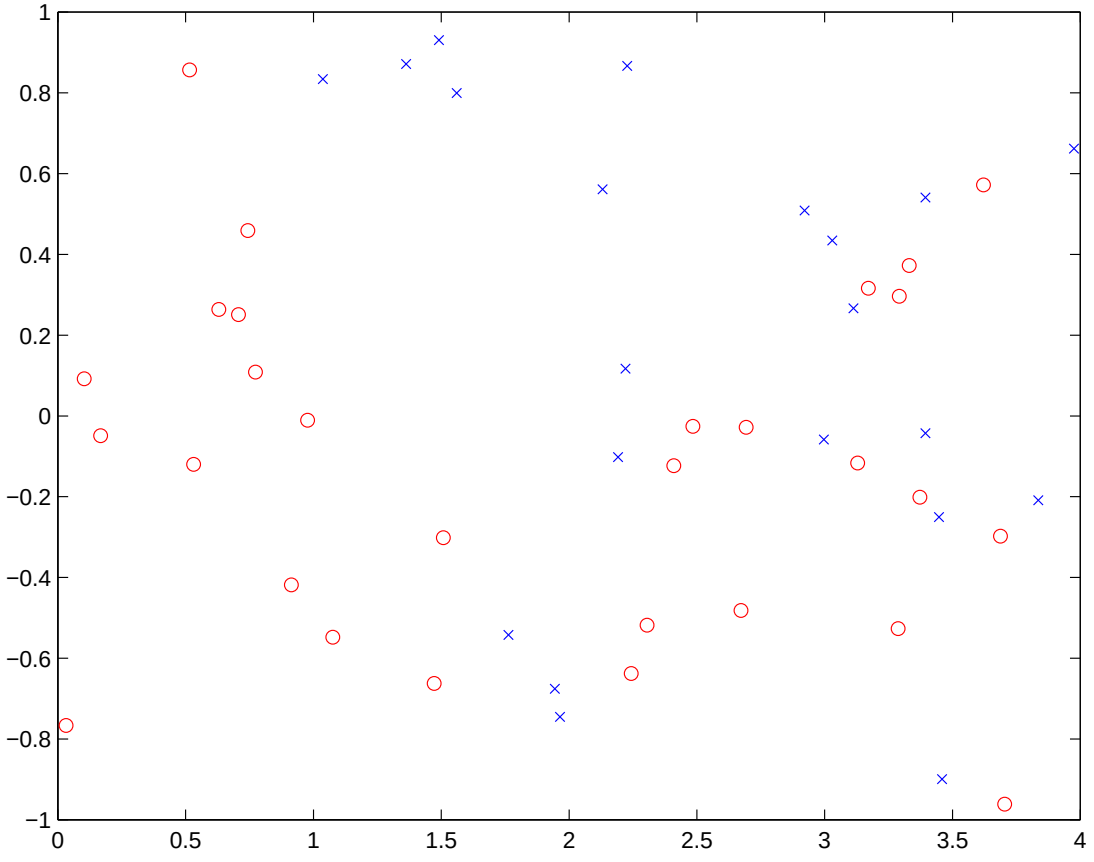
Stéphane Canu, Xavier Mary et Alain Rakotomamonjy
stephane.canu@insa-rouen.fr

asi.insa-rouen.fr/~scanu

INSA Rouen - Département ASI

Laboratoire PSI, FRE CNRS 2645

A la recherche d'une règle de décision universelle



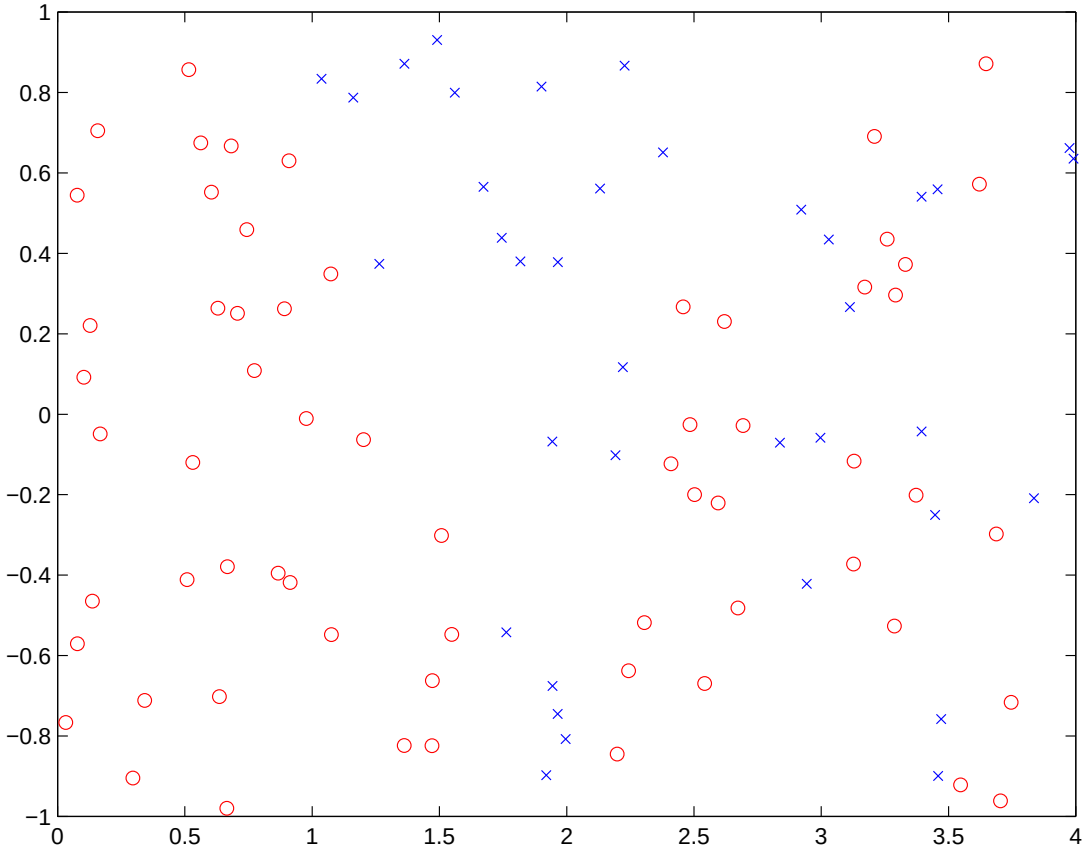
on cherche un algorithme $\mathcal{A}$
capable de résoudre
tous les problèmes

l'échantillon $(x_i, y_i)_{i=1,n}$

$$\underbrace{\mathbb{P}(err(\mathcal{A}, x_i, y_i))}_{\text{erreur de } \mathcal{A}} \underset{n\to\infty}{\longrightarrow} \underbrace{\mathbb{P}_{\mathsf{b}}(err)}_{\text{erreur de bayes}}$$

Tracez la frontière de décision entre ces deux classes ?

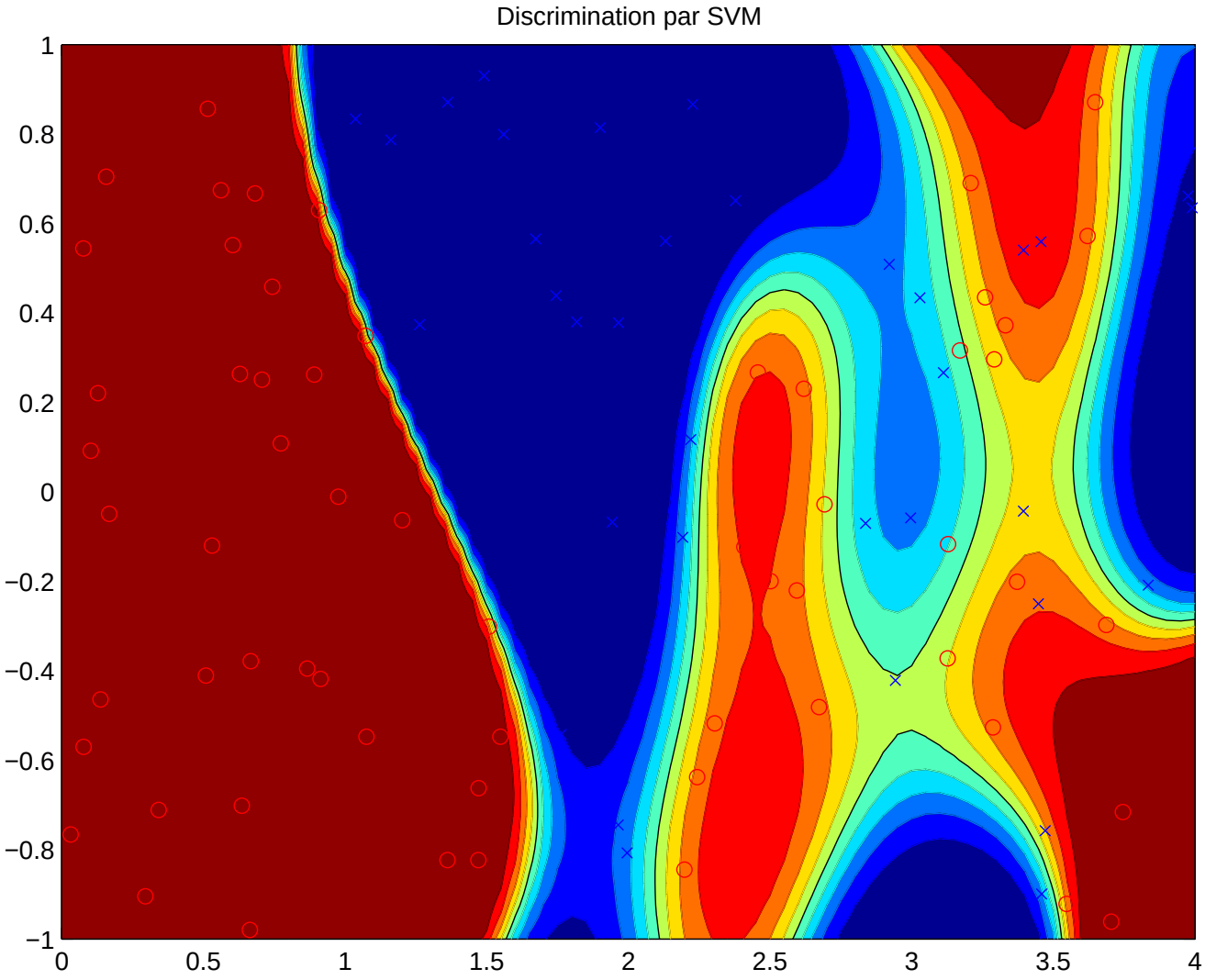
Introduction



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

Tracez la frontière de décision entre ces deux classes ?

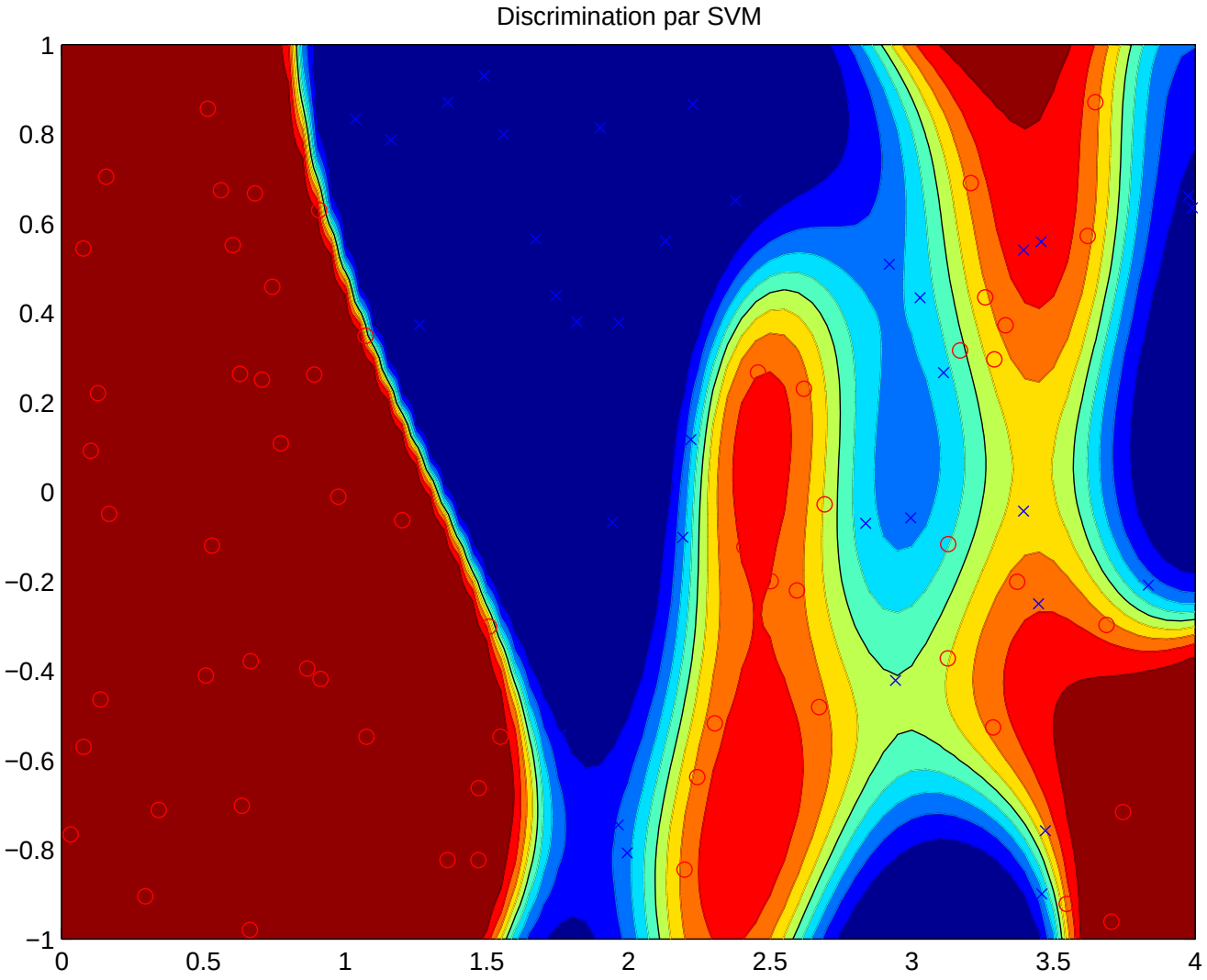
Introduction : le plan



Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité

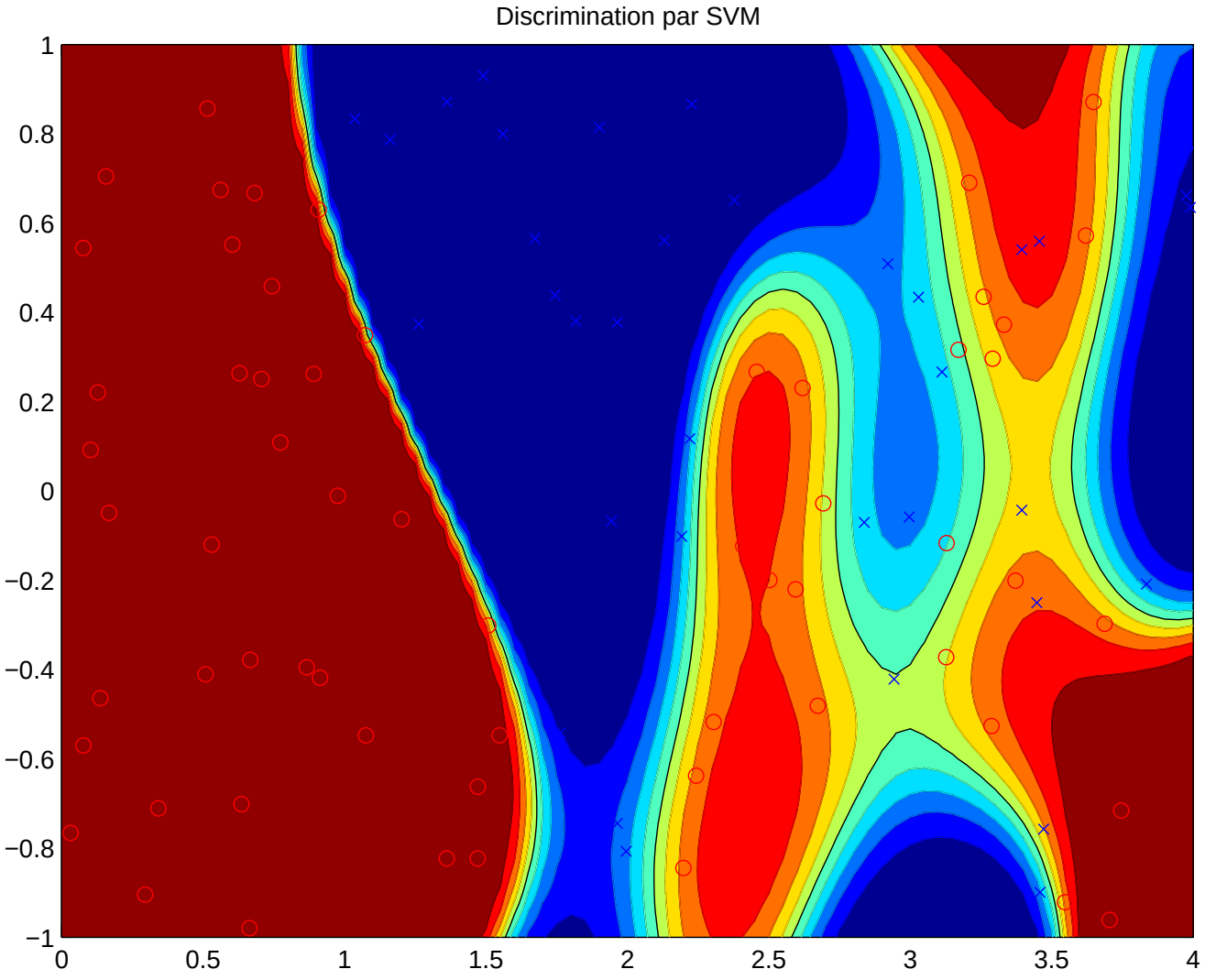
Introduction : le plan



Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité

Introduction : le plan

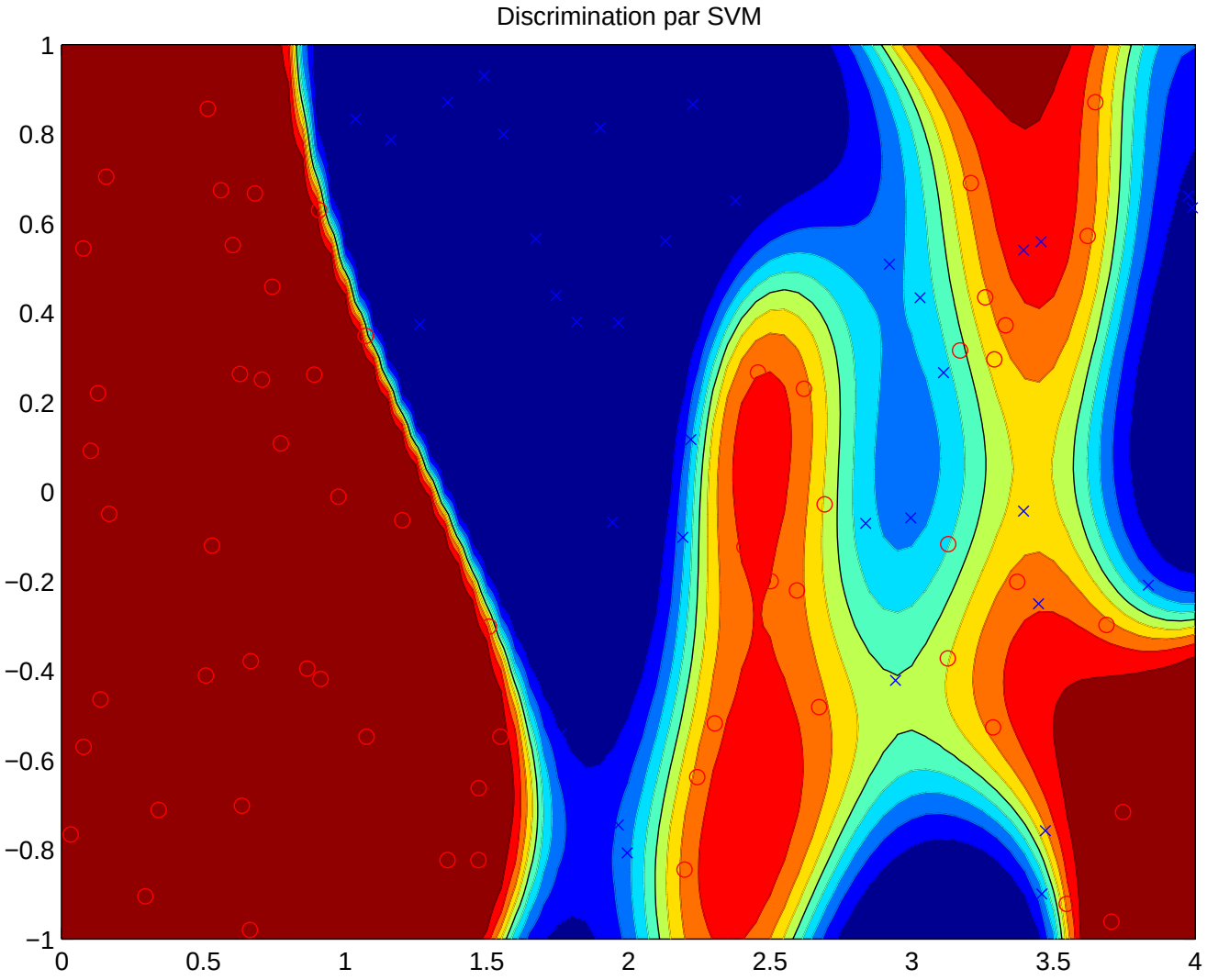


Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité

- (3) Décision locale

Introduction : le plan

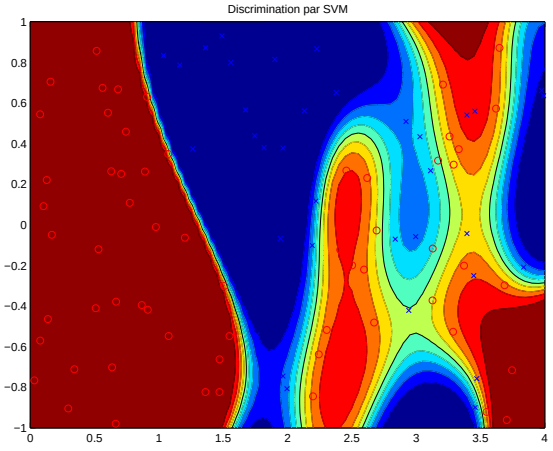


Une solution
$\Rightarrow$ Quels critères ?

- (1) Fidélité
- (2) Régularité

- (3) Décision locale
- (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)

$\mathcal{F} = \{x \mid f(x) = 0\}$: frontière de décision.

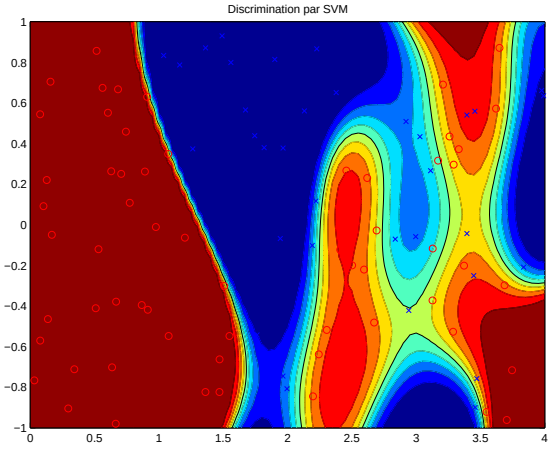
Bien classer tout le monde :

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

(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)

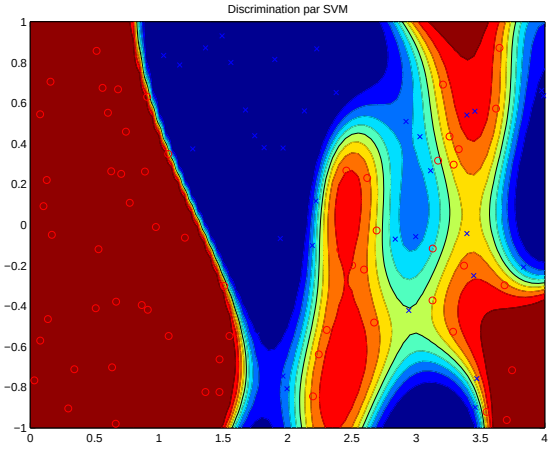
$\mathcal{F} = \{x \mid f(x) = 0\}$: frontière de décision.

Bien classer tout le monde :

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

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

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)

$\mathcal{F} = \{x \mid f(x) = 0\}$: frontière de décision.

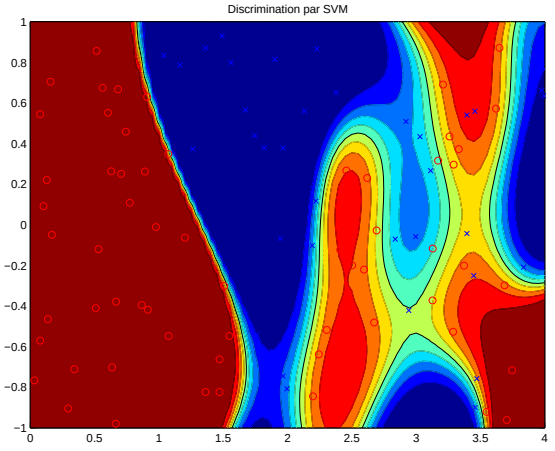
Bien classer tout le monde :

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

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

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

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)

$\mathcal{F} = \{x \mid f(x) = 0\}$: frontière de décision.

Bien classer tout le monde :

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

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

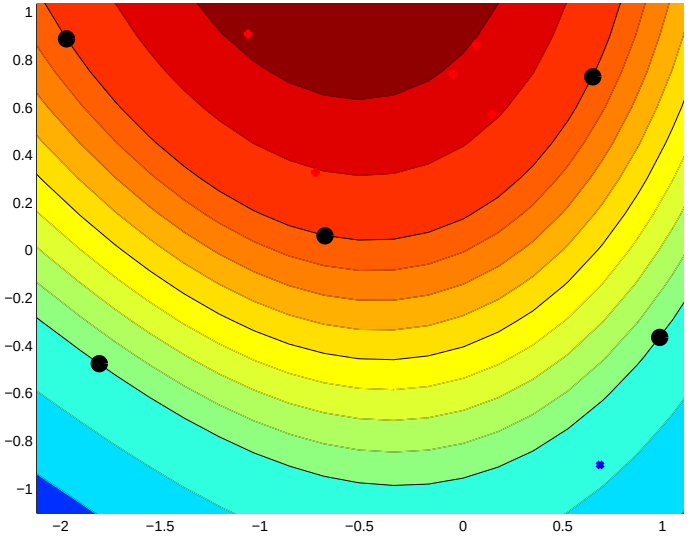
$$f(x_i)y_i \geq k \qquad k > 0,\ i = 1, n$$

Marge

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

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

Fidélité et marge



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

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

Marge :

$$\min_{i=1,n} d(\mathcal{F}, x_i) = \min_{i=1,n} f(x_i)y_i$$

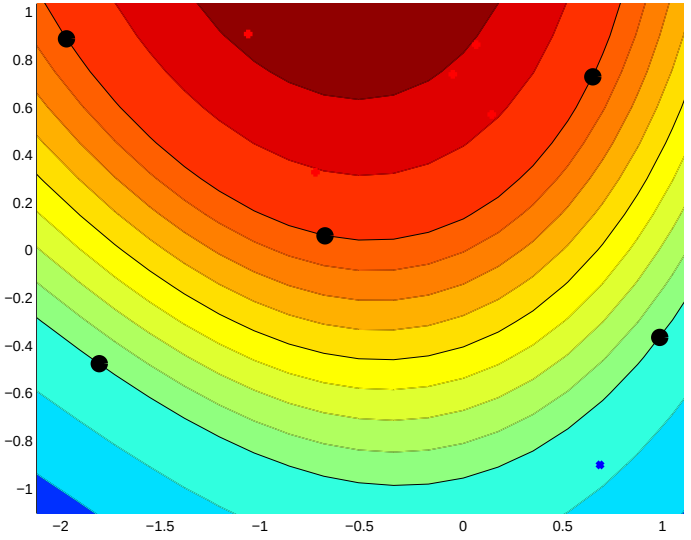
Bien classer tout le monde $(k = 1)$:

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

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

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

Fidélité et marge



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

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

Marge :

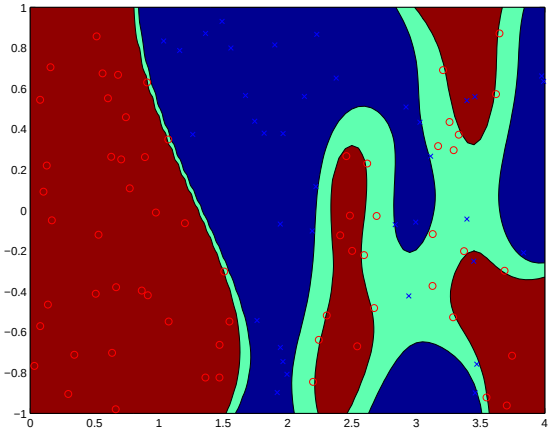
$$\min_{i=1,n} d(\mathcal{F}, x_i) = \min_{i=1,n} f(x_i)y_i$$

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

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

1 est la marge minimale

Fidélité et droit à l'erreur



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

Introduisons une variable d'écart ξ_i

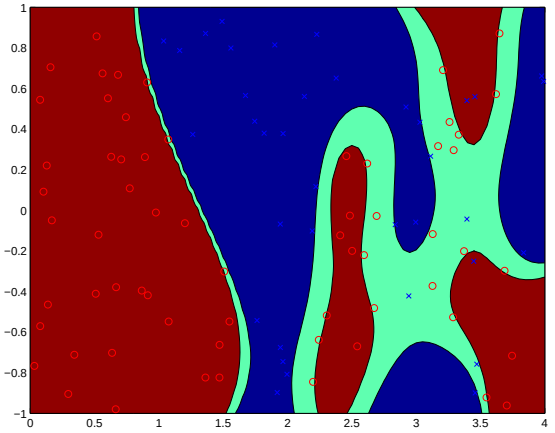
Bien classer **a peu près** tout le monde :

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

où ξ_i est une variable d'écart

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

Fidélité et droit à l'erreur



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

Introduisons une variable d'écart ξ_i

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

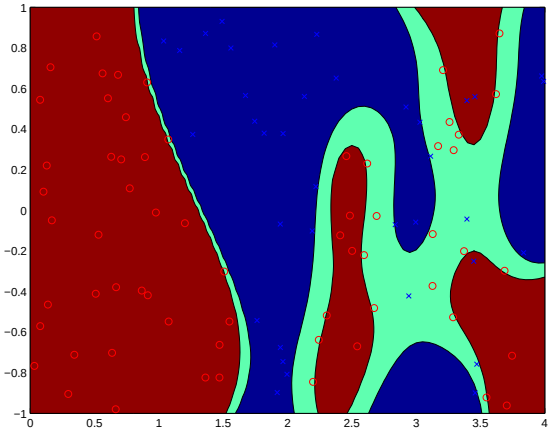
$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0\ ,\ i = 1, n$$

où ξ_i est une variable d'écart

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

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

Fidélité et droit à l'erreur



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

Introduisons une variable d'écart ξ_i

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

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

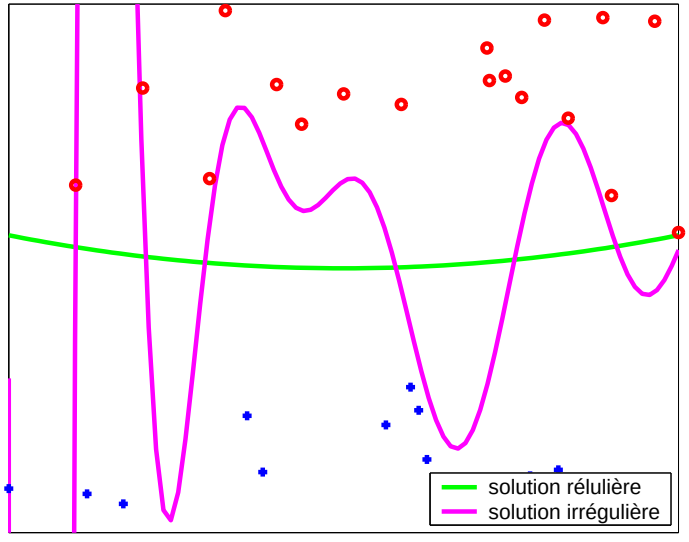
$$f(x_i)y_i > 1 - \xi_i \qquad \xi_i > 0\ ,\ i = 1, n$$

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

(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 $\xi_i = 0,\ i = 1, n$

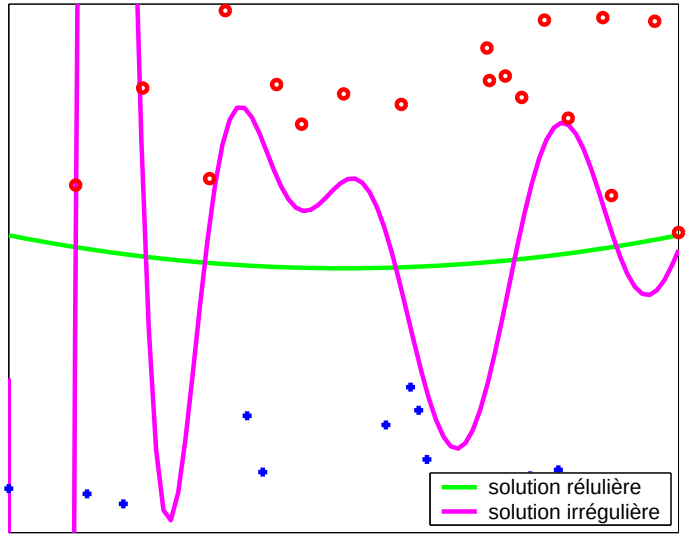
par exemple

- « 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 $\xi_i = 0,\ i = 1, n$

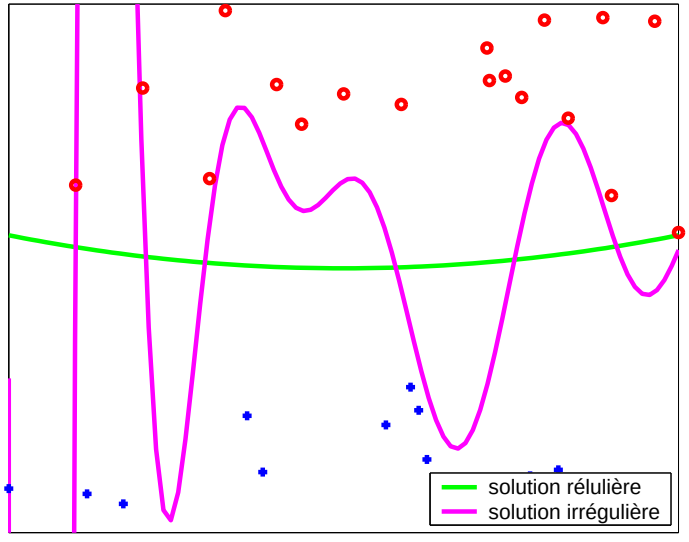
par exemple

- « 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 $\xi_i = 0,\ i = 1, n$

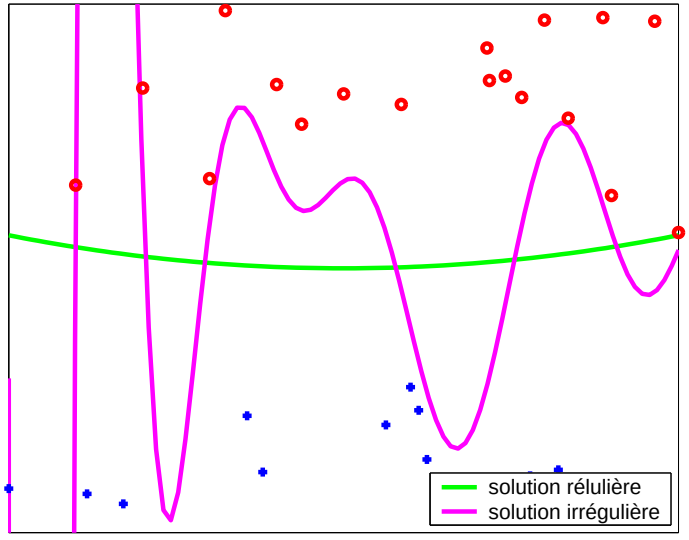
par exemple

- « 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 »

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

Régularité



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

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

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

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 »

Comment choisir $\mathcal{H}$ (l'ensemble des hypothèses ?)

1. $\mathcal{H}$ doit être un espace vectoriel

Comment choisir $\mathcal{H}$ (l'ensemble des hypothèses ?)

1. $\mathcal{H}$ doit être un espace vectoriel
2. $f(x)$ doit avoir un sens...

$$\begin{array}{rccc} \delta_x : & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & \delta_x f = f(x) \end{array}$$

L^2 ne convient pas !

Comment choisir $\mathcal{H}$ (l'ensemble des hypothèses ?)

1. $\mathcal{H}$ doit être un espace vectoriel
2. $f(x)$ doit avoir un sens...

$$\begin{array}{rccc} \delta_x : & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & \delta_x f = f(x) \end{array}$$

L^2 ne convient pas !

3. la fonctionnelle d'évaluation doit être continue ($\delta_x \in \mathcal{H}'$) :

$$f_n \xrightarrow{\mathcal{H}} t \Longrightarrow \forall x \in \mathcal{X},\ f_n(x) \xrightarrow{\mathbb{R}} t(x)$$

dans le cas général, ce n'est pas la convergence uniforme

Comment choisir $\mathcal{H}$ (l'ensemble des hypothèses ?)

1. $\mathcal{H}$ doit être un espace vectoriel
2. $f(x)$ doit avoir un sens...

$$\begin{array}{rccc} \delta_x : & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & \delta_x f = f(x) \end{array}$$

L^2 ne convient pas !

3. la fonctionnelle d'évaluation doit être continue ($\delta_x \in \mathcal{H}'$) :

$$f_n \xrightarrow{\mathcal{H}} t \Longrightarrow \forall x \in \mathcal{X},\ f_n(x) \xrightarrow{\mathbb{R}} t(x)$$

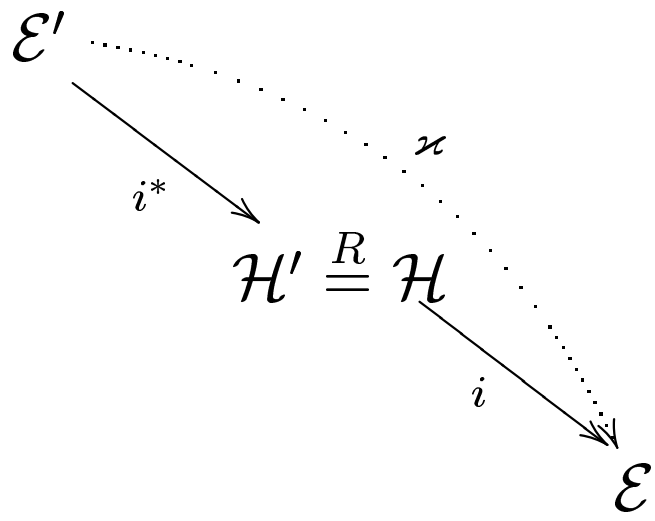
dans le cas général, ce n'est pas la convergence uniforme

(1) + (2) + (3) $\Leftrightarrow$ $\mathcal{H}$ est un espace à noyau reproduisant :
pas nécessairement de Hilbert

$\mathcal{H}$: un espace à noyau reproduisant

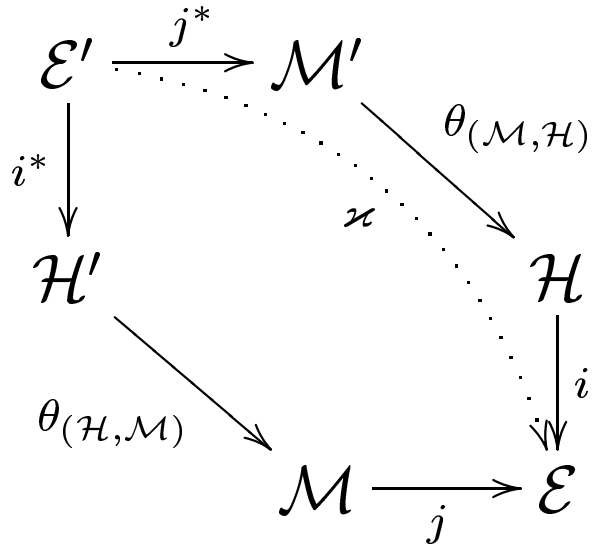
$\mathcal{H}$ espace vectoriel d'évaluation $\Rightarrow$ il existe un noyau $K(s,t)$

Cas Hilbertien



$$K(s,t) = \langle K(s,.), K(.,t) \rangle_{\mathcal{H}}$$

Cas général



$$K(s,t) = \left(\varkappa^*(\delta_s), \varkappa(\delta_t)\right)_{\mathcal{M},\mathcal{H}}$$

mesures fonctions

Généralisation des noyaux reproduisants et de Schwarz

Construction de $\mathcal{H}$ dans le cas Hilbertian à partir de L^2

- soit S un operateur linéaire de L^2 sur $\mathbb{R}^{\mathcal{X}}$
- soit $\Gamma(t,\tau)$ une fonction de deux variables, $\Gamma(t,.) = \Gamma_t \in L^2$

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(t) = Sf(t) = \int \Gamma(t,\tau) f(\tau)\, d\tau \end{array}$$

opérateur de Carleman

Construction de $\mathcal{H}$ dans le cas Hilbertian à partir de L^2

- soit S un operateur linéaire de L^2 sur $\mathbb{R}^{\mathcal{X}}$
- soit $\Gamma(t,\tau)$ une fonction de deux variables, $\Gamma(t,.) = \Gamma_t \in L^2$

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(t) = Sf(t) = \int \Gamma(t,\tau) f(\tau)\, d\tau \end{array}$$

opérateur de Carleman

- à partir de S nous définisions $\mathcal{H} = Im(S)$

Construction de $\mathcal{H}$ dans le cas Hilbertian à partir de L^2

- soit S un operateur linéaire de L^2 sur $\mathbb{R}^{\mathcal{X}}$
- soit $\Gamma(t,\tau)$ une fonction de deux variables, $\Gamma(t,.) = \Gamma_t \in L^2$

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(t) = Sf(t) = \int \Gamma(t,\tau) f(\tau)\, d\tau \end{array}$$

opérateur de Carleman

- à partir de S nous définisions $\mathcal{H} = Im(S)$

$$K(s,t) = \int \Gamma(t,\tau)\Gamma(s,\tau)\, d\tau = \langle \Gamma(t,\tau)\Gamma(s,\tau) \rangle_{L^2} = \langle S\Gamma(t,\tau) S\Gamma(s,\tau) \rangle_{\mathcal{H}}$$

$$\|g\|^2_{\mathcal{H}} = \|Sf\|^2_{\mathcal{H}} = \|f\|^2_{L^2} = \| \ S^{-1} \ g\|^2_{L^2}$$

Construction de $\mathcal{H}$ dans le cas Hilbertian à partir de L^2

- soit S un operateur linéaire de L^2 sur $\mathbb{R}^{\mathcal{X}}$
- soit $\Gamma(t,\tau)$ une fonction de deux variables, $\Gamma(t,.) = \Gamma_t \in L^2$

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(t) = Sf(t) = \int \Gamma(t,\tau) f(\tau)\, d\tau \end{array}$$

opérateur de Carleman

- à partir de S nous définisions $\mathcal{H} = Im(S)$

$$K(s,t) = \int \Gamma(t,\tau)\Gamma(s,\tau)\, d\tau = \langle \Gamma(t,\tau)\Gamma(s,\tau) \rangle_{L^2} = \langle S\Gamma(t,\tau) S\Gamma(s,\tau) \rangle_{\mathcal{H}}$$

$$\|g\|^2_{\mathcal{H}} = \|Sf\|^2_{\mathcal{H}} = \|f\|^2_{L^2} = \|S^{-1} g\|^2_{L^2} = \|Pf\|^2_{L^2}$$

$$\boxed{S = P^{-1}}$$

Construction de $\mathcal{H}$ dans le cas Hilbertian à partir de L^2

- soit S un operateur linéaire de L^2 sur $\mathbb{R}^{\mathcal{X}}$
- soit $\Gamma(t,\tau)$ une fonction de deux variables, $\Gamma(t,.) = \Gamma_t \in L^2$

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(t) = Sf(t) = \int \Gamma(t,\tau) f(\tau)\, d\tau \end{array}$$

opérateur de Carleman

- à partir de S nous définisions $\mathcal{H} = Im(S)$

$$K(s,t) = \int \Gamma(t,\tau)\Gamma(s,\tau)\, d\tau = \langle \Gamma(t,\tau)\Gamma(s,\tau) \rangle_{L^2} = \langle S\Gamma(t,\tau) S\Gamma(s,\tau) \rangle_{\mathcal{H}}$$

$$\|g\|^2_{\mathcal{H}} = \|Sf\|^2_{\mathcal{H}} = \|f\|^2_{L^2} = \|S^{-1} g\|^2_{L^2} = \|Pf\|^2_{L^2}$$

$$S = P^{-1}$$

L^2 étant separable $\mathcal{H}$ l'est aussi : il exite une base dénombrable telle que $K(x,y) = \sum \Phi(x)\Phi(y)$

frame - *structures obliques*

- Famille « interprètable de fonctions » $\left\{\varphi_k\right\}_{k\in\mathbb{N}}$
- base + dépendance + redondance
- représentation

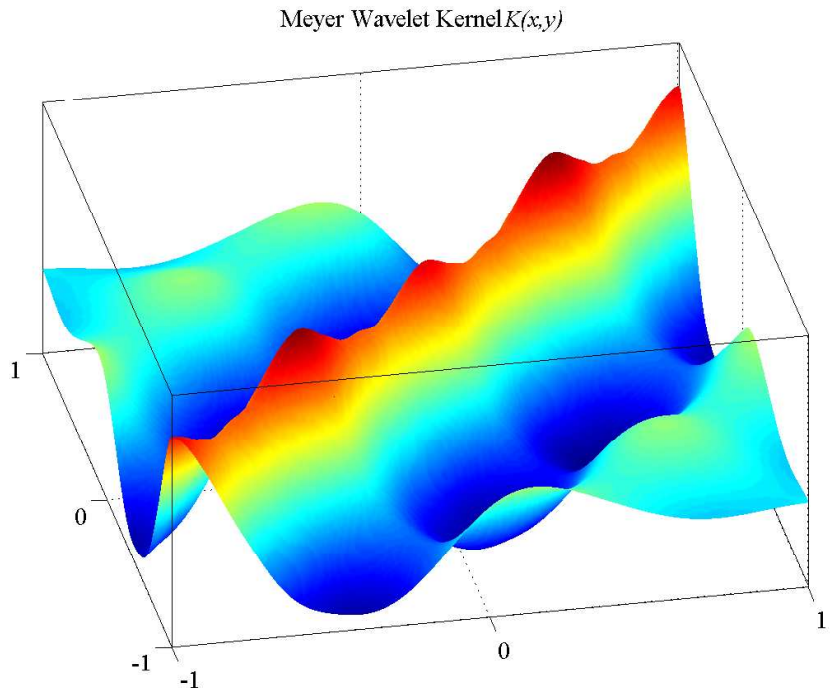
$$f = \sum_{k=1}^{\infty} \langle f, \varphi^* \rangle \varphi$$

φ^* : frame dual

- exemple : les ondelettes

bonnes propriétés statistiques - mais une infinité de terme !

construction de frame *apprenables (de noyau)*



- Opérateur de *Carleman*

$$S: \quad L^2 \longrightarrow \mathbb{R}^{\mathcal{X}}$$
$$f \longmapsto Sf = \int \Gamma(\tau, t) f(\tau)\, d\tau$$

  - $\mathcal{H} = Im(S)$
    $K(s,t) = \int \Gamma(s,\tau)\Gamma(\tau,t)\, d\tau$

- Noyaux d'ondelettes
  - ψ_i frame de L^2, $\quad \varphi_i = S\psi_i$
  - φ_i frame de $Im(S)$
  - $K(s,t) = \sum_i \varphi_i^*(s)\varphi_i(t)$

La fonction Γ est représentée par une matrice

L'astuce du noyau

Si K est un noyau défini positif d'un k.r.h.s. séparable, il existe une famille $(\phi_j)_{j\in\mathcal{J}}$ orthonormée telle que :

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

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

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

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

L'astuce du noyau

Si K est un noyau défini positif d'un k.r.h.s. séparable, 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 :

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

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

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

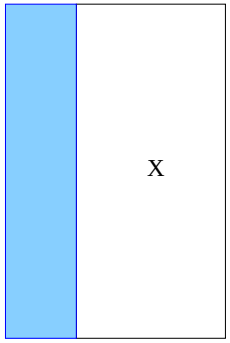
dim ∞ dim n

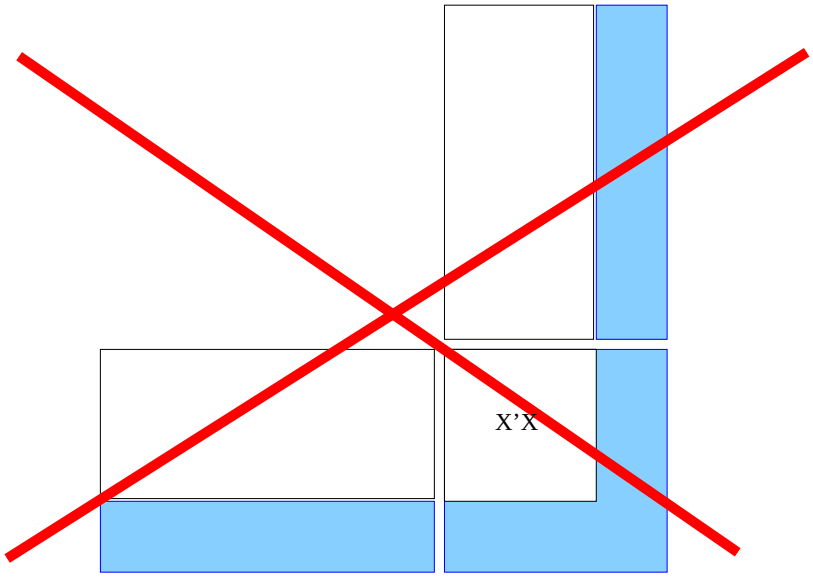
(1) Fidélité - **(3) Décision « locale »**

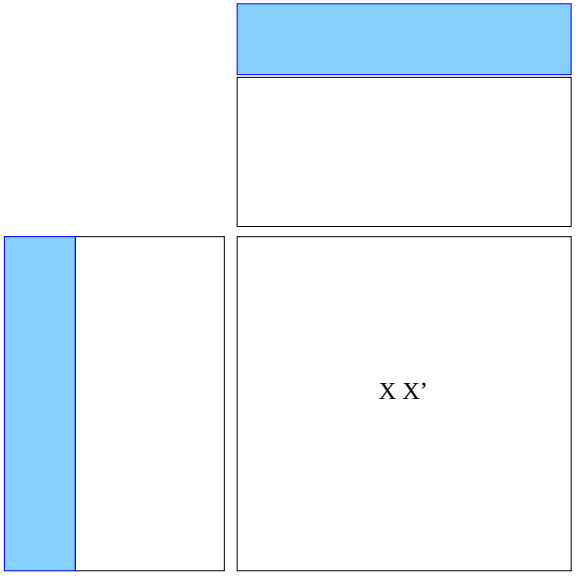
(2) Régularité - (4) Points « frontière »

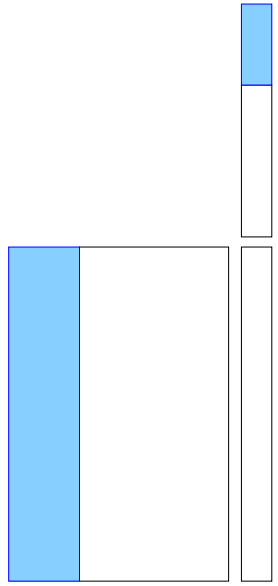
Illustration : le cas de l'ACP...

Nouvelles variables...
par exemple x^2 !







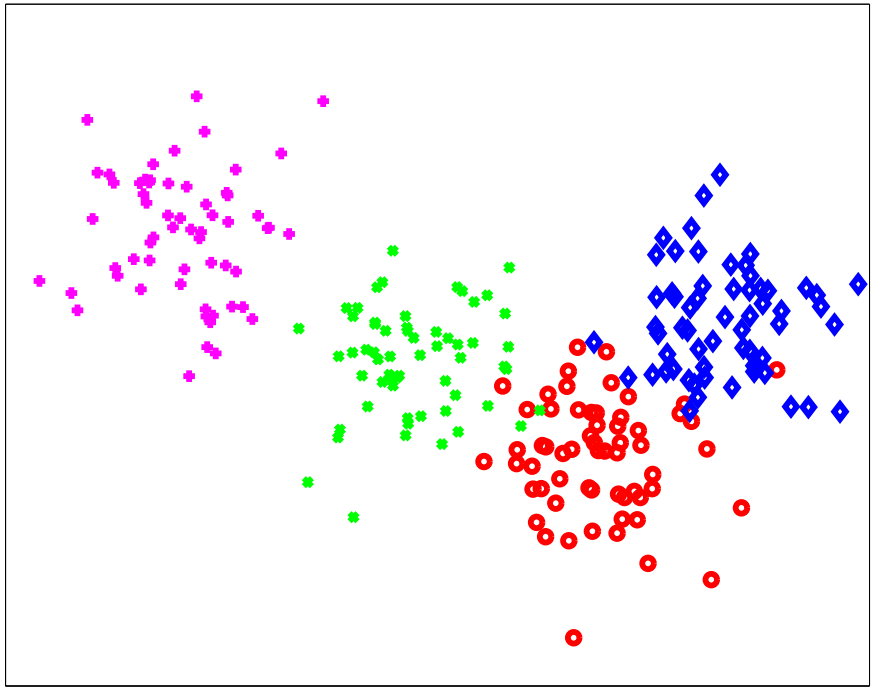


les variables...

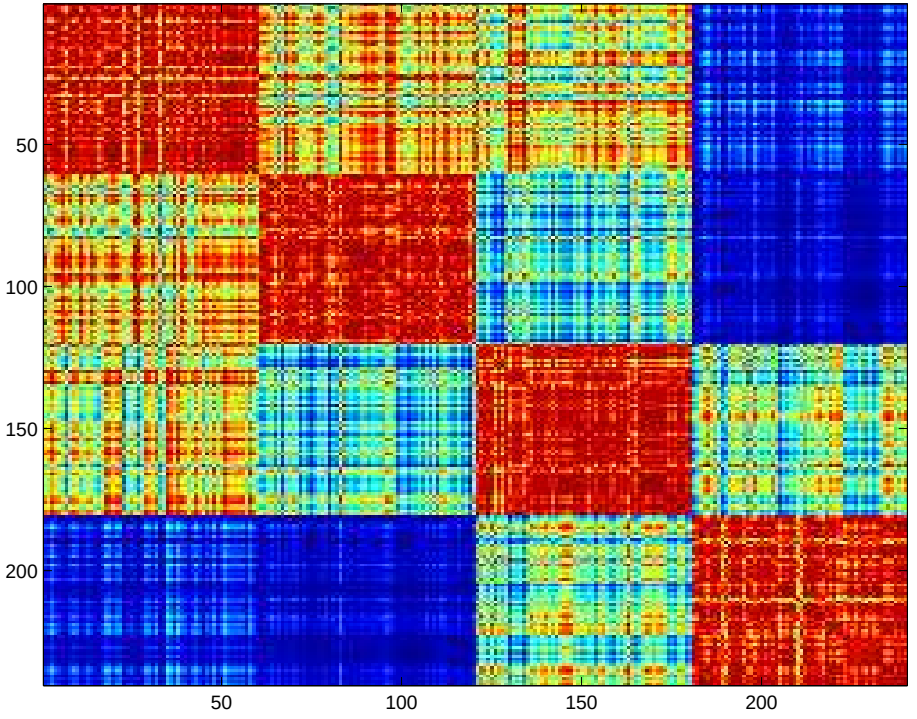
$X^\top X$ où $X X^\top$

ou les individus

La matrice de Gram : illustration



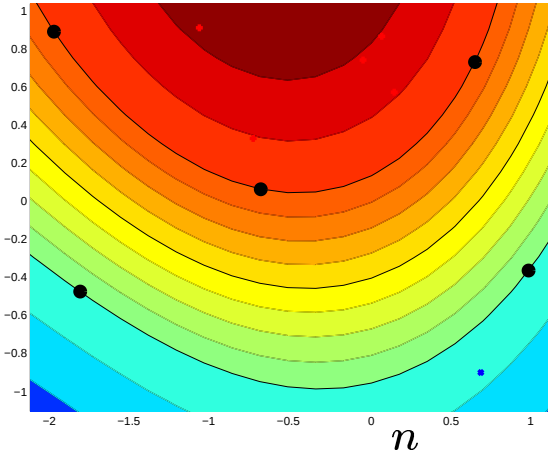
4 classes de 60 points chacune



K matrice de gram associée

$$K_{ij} = \exp - \frac{\|x_i - x_j\|^2}{b}$$

Régularité et marge



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

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

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

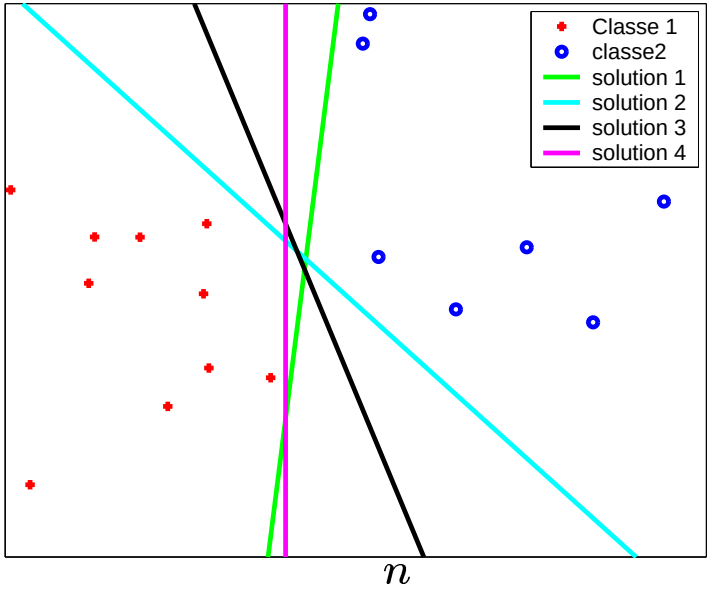
- $\mathbb{P}(err) < \underbrace{\sum_{i=1} \mathbb{I}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{erreur empirique}} + \phi\left(\frac{1}{\text{marge}}\right)$

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

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

Régularité et marge

Cas linéaire : quelle solution choisir ?



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

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

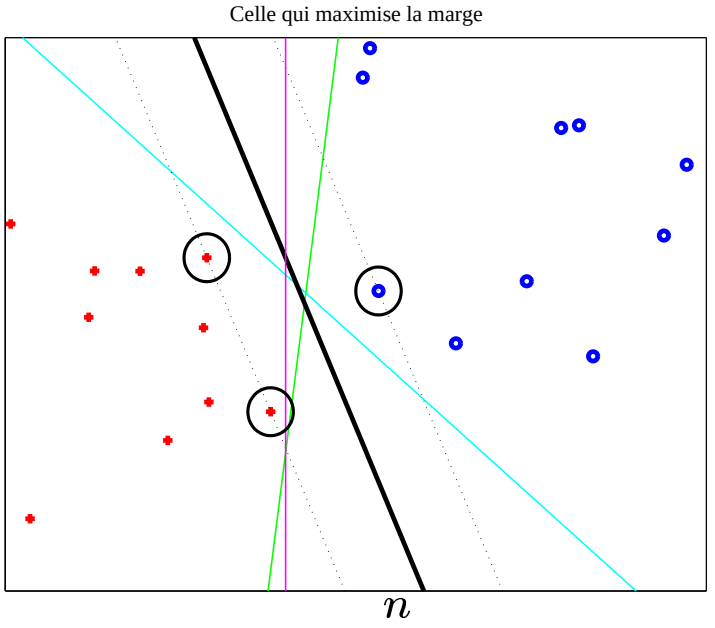
$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

- $\mathbb{P}(err) < \underbrace{\sum_{i=1} \mathbb{I}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{erreur empirique}} + \phi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err) \Leftrightarrow$ **maximiser** la marge

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

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

Régularité et marge



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

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

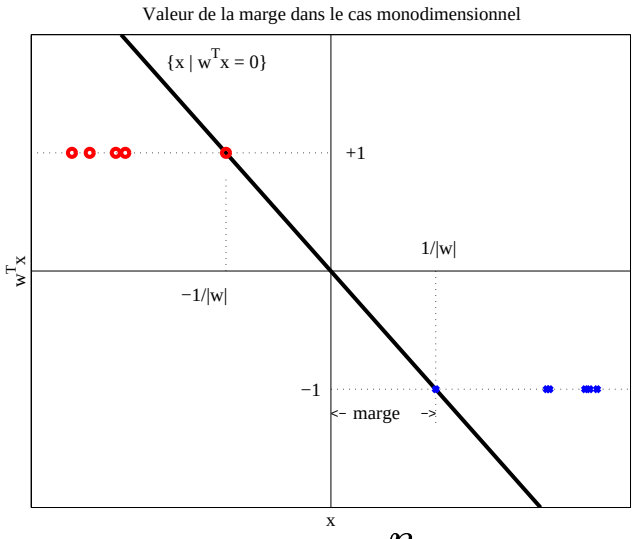
$$\min_{f} \ \|f\|^2_{\mathcal{H}}$$

- $\mathbb{P}(err) < \underbrace{\sum_{i=1} \mathbb{I}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{erreur empirique}} + \phi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err) \Leftrightarrow$ **maximiser** la marge
- maximiser la robustesse $\Leftrightarrow$ maximiser la marge

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

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

Régularité et marge



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

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

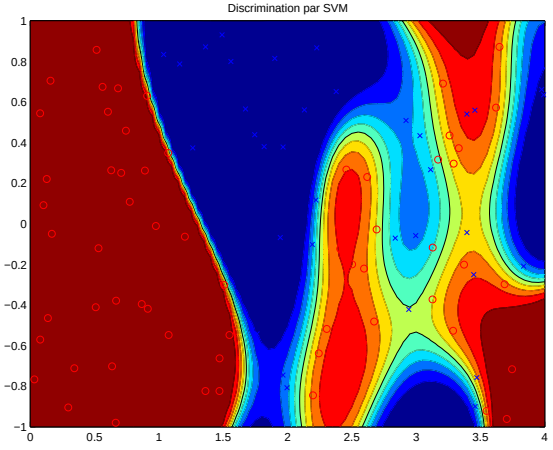
$$\min_{f} \ \|f\|_{\mathcal{H}}^2 \quad \Leftrightarrow \quad \min_{\mathbf{w}} \sum_{j=1}^{\infty} w_j^2$$

- $\mathbb{P}(err) < \underbrace{\sum_{i=1}^{n} \mathbb{1}_{\{\text{signe}(f(x_i)) \neq y_i\}}}_{\text{erreur empirique}} + \phi\left(\frac{1}{\text{marge}}\right)$
- **minimiser** $\mathbb{P}(err)$ $\Leftrightarrow$ **maximiser** la marge
- maximiser la robustesse $\Leftrightarrow$ maximiser la marge
- maximiser la marge $\Leftrightarrow$ minimiser $\|\mathbf{w}\|^2$

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

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

Normes...



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

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

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

l'ensemble des hypothèses est alors de la forme

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

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

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 :

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

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 :

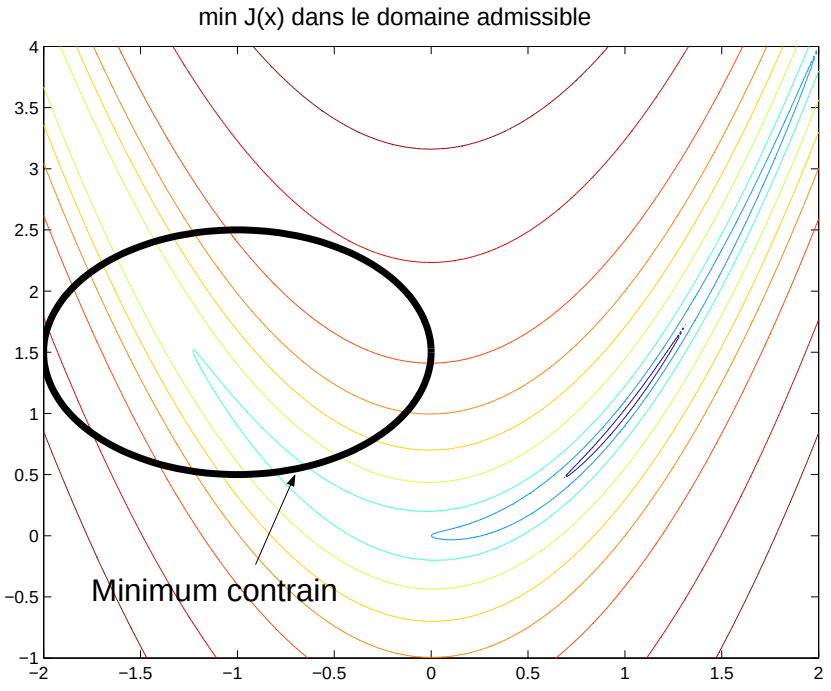
$$\begin{cases} \min\limits_{\mathbf{w}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum\limits_{i=1}^{n} \xi_i & \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 - \xi_i & i = 1, n \\ \text{et} & \xi_i > 0 & i = 1, n \end{cases} \qquad (4)$$

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 »**

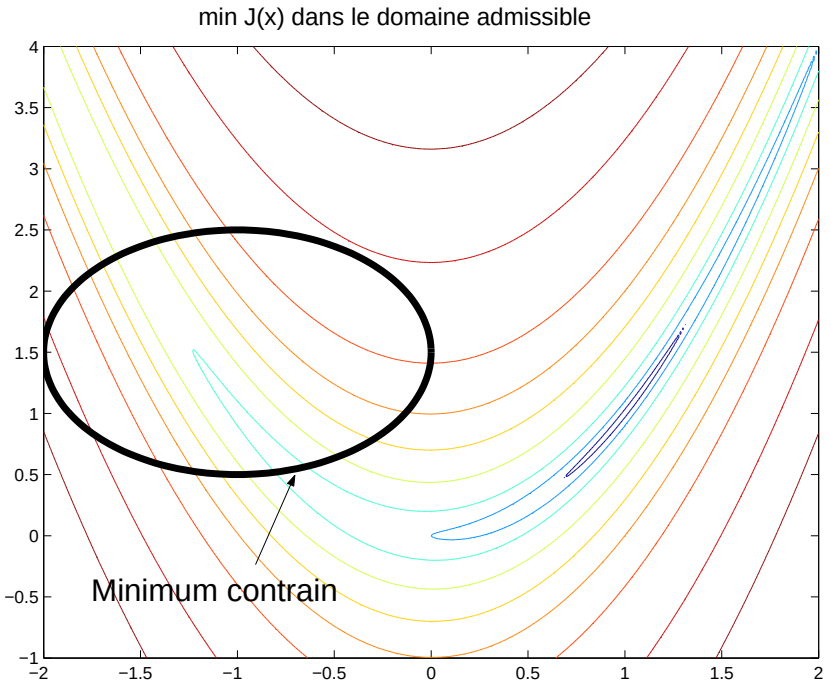
(2) Régularité - (4) Points « frontière »

Minimisation sous contraintes (cas séparable)



$$\left\{ \begin{array}{ll} & \min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{avec} & \quad y_i f(\mathbf{x}_i) > 1 \qquad i = 1, n \end{array} \right.$$

Minimisation sous contraintes (cas séparable)



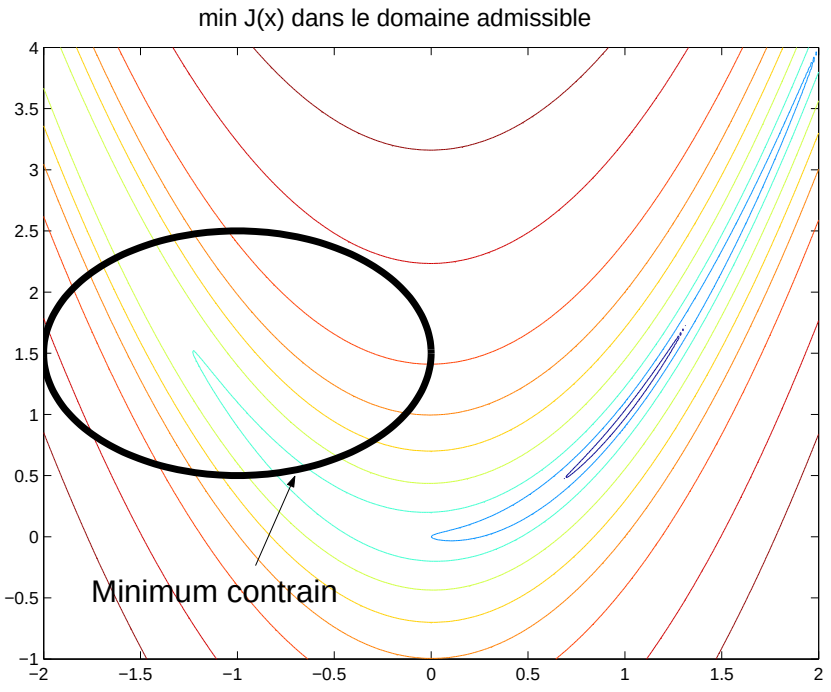
$$\left\{ \begin{array}{ll} & \min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 \qquad i = 1, n \end{array} \right.$$

$$\Leftrightarrow$$

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

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

Minimisation sous contraintes (cas séparable)



$$\left\{ \begin{array}{ll} & \min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{avec} & y_i f(\mathbf{x}_i) > 1 \qquad i = 1, n \end{array} \right.$$

$$\Leftrightarrow$$

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

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

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

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

dont on tire les conditions de Kuhn et Tucker :

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

conséquence pour f :

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

Reformulation dans l'espace des exemples

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

dont on tire les conditions de Kuhn et Tucker :

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

conséquence pour f :

$$f(\mathbf{x}) = \sum_{k=1}^{\infty} w_k \phi_k(\mathbf{x}) = \sum_{k=1}^{\infty} \left(\sum_{i=1}^{N} \lambda_i y_i \phi(\mathbf{x}_i) \right) \phi_k(\mathbf{x})$$

$$= \sum_{i=1}^{N} \underbrace{\lambda_i y_i}_{a_i} \underbrace{\sum_{k=1}^{\infty} \phi_k(\mathbf{x}) \phi(\mathbf{x}_i)}_{K_b(\mathbf{x}, \mathbf{x}_i)}$$

Stratégie

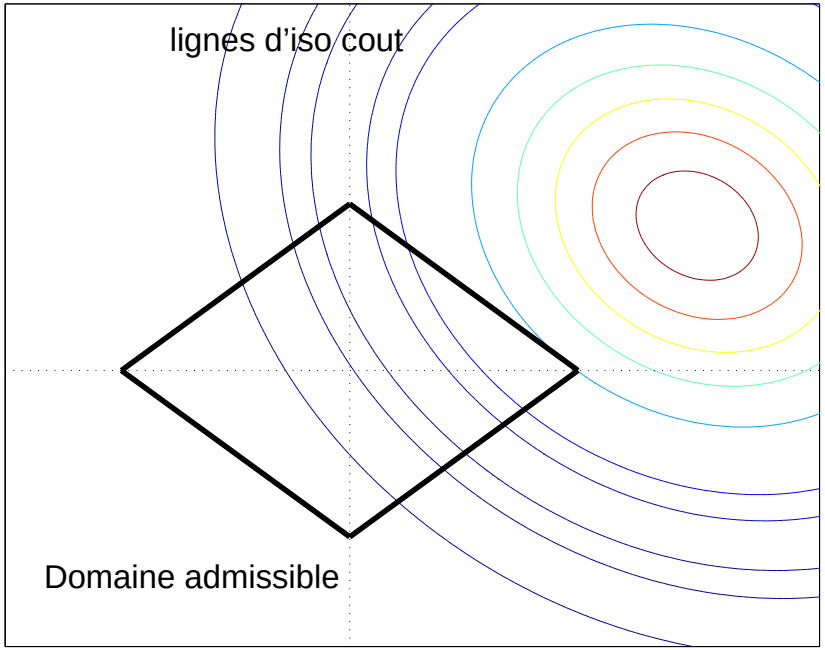
calcul de K

calcul des λ

calcul des $\mathbf{a}$

calcul de f

calcul des λ : problème Dual (2)



$$\begin{cases} \min\limits_{\lambda} & \frac{1}{2}\lambda^\top H\lambda + \mathbf{c}^\top\lambda \\ \text{avec} & \sum\limits_{i=1}^{N} \lambda_i y_i \varphi_j(x_i) = 0 \quad j = 1, m \\ \text{et} & 0 \leq \lambda_i \leq C \qquad i = 1, n \end{cases}$$

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.

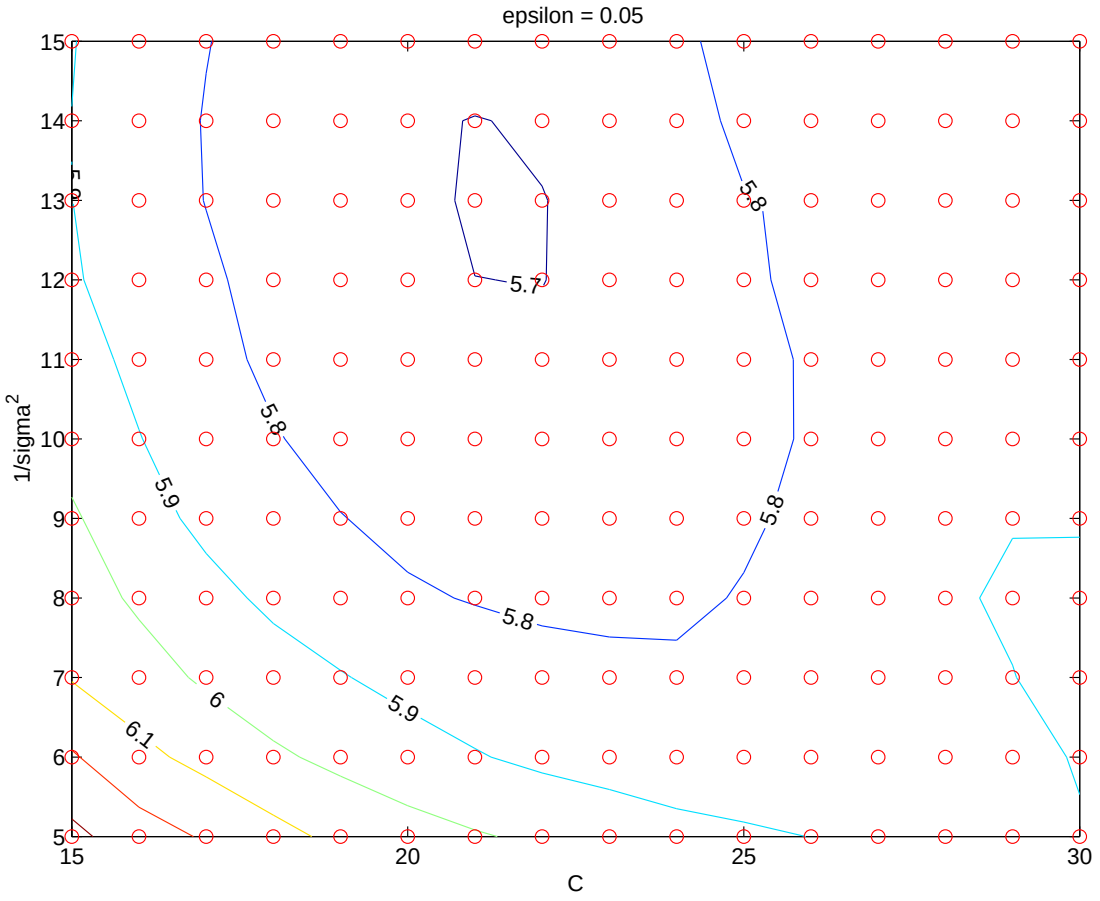
Reformulation de Girosi (97)

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

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

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

Solution pratique : Paramètres de régularisation



- b : métrique
- C : des erreurs ?

$\Rightarrow$ la *span bound*

Estimation du taux d'erreur en fonction des hyperparamètres.

SVMtoolbox : une démo ?

$[xsup, w, w0, tps, alpha] =$ svmclass$(Xapp, yi, C, lambda, kernel, kerneloption, 1, phi)$;

$ypred =$ svmval$(Xtest, xsup, w, w0, kernel, kerneloption)$;

Conclusion

- SVM les principes
  - Noyaux : unversel
  - Marge : minimum global unique
  - Parcimonieux : l'influence de chaque exemple

- SVM et les autres
  - SVM vs Réseaux de neurones (PMC) : optimisation
  - SVM vs Parzen : parcimonie ($L^2 \; vs \; L^1$)
  - SVM : des résultats (vision, génomique, texte)

- les questions sur les SVM
  - le multiclasse
  - comment choisir le noyau ?
    comment définir la distance entre deux points ?
  - noyaux non symétrique, non positifs ? d'autres critères ?
  - lissage et détection de ruptures

Références

- SVM
  - Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.
  - Nello Cristianini and John Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000
  - Kernel Machines, 2002
- Reconnaissance des formes statistiques
  - R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classifi cation (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