SVM et machines à noyaux

Stéphane Canu
stephane.canu@litislab.eu

Ecole d'été du GRETSI - Peyresq

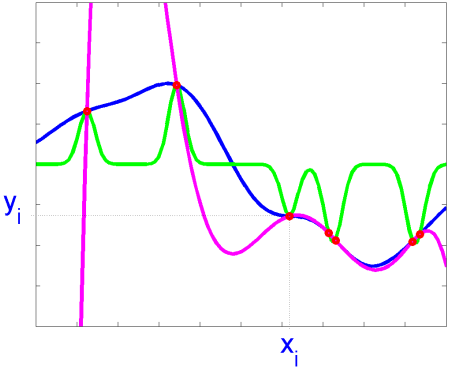
June 29, 2010

Plan

1. Kernel machines
   - Non sparse kernel machines
   - sparse kernel machines: SVM
   - Sparse kernel machines for regression: SVR
   - practical SVM

Interpolation splines

find out $f \in \mathcal{H}$ such that $f(\mathbf{x}_i) = y_i, \qquad i = 1, ..., n$



It is an ill posed problem

Interpolation splines: minimum norm interpolation

$$\begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i = 1, ..., n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f, \alpha) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i \big(f(\mathbf{x}_i) - y_i\big)$$

Interpolation splines: minimum norm interpolation

$$\begin{cases} \min_{f \in \mathcal{H}} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i = 1, ..., n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f, \alpha) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i \big(f(\mathbf{x}_i) - y_i\big)$$

optimality for f: $\quad \nabla_f L(f, \alpha) = 0 \quad \Leftrightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$

Interpolation splines: minimum norm interpolation

$$\begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i = 1, ..., n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f, \alpha) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i \big(f(\mathbf{x}_i) - y_i\big)$$

optimality for f: $\quad \nabla_f L(f, \alpha) = 0 \quad \Leftrightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$

dual formulation (remove f from the lagrangian):

$$Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \alpha_i y_i \qquad \text{solution:} \quad \max_{\alpha \in \mathbb{R}^n} Q(\alpha)$$

$$K\alpha = y$$

Representer theorem

Theorem (Representer theorem)

Let $\mathcal{H}$ be a RKHS with kernel $k(s,t)$. Let ℓ be a function from $\mathcal{X}$ to $\mathbb{R}$ (loss function) and Φ a non decreasing function from $\mathbb{R}$ to $\mathbb{R}$. If there exists a function f^* minimizing:

$$f^* = \underset{f \in \mathcal{H}}{\text{argmin}} \ \sum_{i=1}^{n} \ell\big(y_i, f(\mathbf{x}_i)\big) + \ \Phi\big(\|f\|_{\mathcal{H}}^2\big)$$

then there exists a vector $\alpha \in \mathbb{R}^n$ such that:

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

it can be generalized to the semi parametric case: $+\sum_{j=1}^{m} \ \beta_j \phi_j(\mathbf{x})$

Elements of a proof

1. $\mathcal{H}_s = span\{k(.,\mathbf{x}_1), ..., k(.,\mathbf{x}_i), ..., k(.,\mathbf{x}_n)\}$
2. orthogonal decomposition: $\mathcal{H} = \mathcal{H}_s \oplus \mathcal{H}_\perp \Rightarrow \forall f \in \mathcal{H}; f = f_s + f_\perp$
3. pointwise evaluation decomposition

$$\begin{aligned} f(\mathbf{x}_i) &= f_s(\mathbf{x}_i) + f_\perp(\mathbf{x}_i) \\ &= \langle f_s(.), k(.,\mathbf{x}_i)\rangle_{\mathcal{H}} + \underbrace{\langle f_\perp(.), k(.,\mathbf{x}_i)\rangle_{\mathcal{H}}}_{=0} \\ &= f_s(\mathbf{x}_i) \end{aligned}$$

4. norm decomposition $\qquad \|f\|^2_{\mathcal{H}} = \|f_s\|^2_{\mathcal{H}} + \underbrace{\|f_\perp\|^2_{\mathcal{H}}}_{\geq 0} \geq \|f_s\|^2_{\mathcal{H}}$
5. decompose the global cost

$$\begin{aligned} \sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i)) + \Phi(\|f\|^2_{\mathcal{H}}) &= \sum_{i=1}^n \ell(y_i, f_s(\mathbf{x}_i)) + \Phi(\|f_s\|^2_{\mathcal{H}} + \|f_\perp\|^2_{\mathcal{H}}) \\ &\geq \sum_{i=1}^n \ell(y_i, f_s(\mathbf{x}_i)) + \Phi(\|f_s\|^2_{\mathcal{H}}) \end{aligned}$$

6. $\boxed{\operatorname*{argmin}_{f\in\mathcal{H}} = \operatorname*{argmin}_{f\in\mathcal{H}_s}}$.

Smooting splines

introducing the error (the slack) $\xi = f(x_i) - y_i$

$$(\mathcal{S}) \quad \begin{cases} \displaystyle\min_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + \frac{1}{2\lambda}\sum_{i=1}^{n}\xi_i^2 \\ \text{such that} & f(x_i) = y_i + \xi_i, \qquad i = 1, n \end{cases}$$

three equivalent definitions

$$(\mathcal{S}') \quad \min_{f\in\mathcal{H}} \frac{1}{2}\sum_{i=1}^{n}\big(f(x_i) - y_i\big)^2 + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$$

$$\begin{cases} \displaystyle\min_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & \displaystyle\sum_{i=1}^{n}\big(f(\mathbf{x}_i) - y_i\big)^2 \le C' \end{cases} \qquad \begin{cases} \displaystyle\min_{f\in\mathcal{H}} & \displaystyle\sum_{i=1}^{n}\big(f(\mathbf{x}_i) - y_i\big)^2 \\ \text{such that} & \|f\|^2_{\mathcal{H}} \le C'' \end{cases}$$

using the representer theorem

$$(\mathcal{S}'') \quad \min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\|K\alpha - \mathbf{y}\|^2 + \frac{\lambda}{2}\alpha^\top K\alpha$$

solution:

$$(\mathcal{S}) \Leftrightarrow (\mathcal{S}') \Leftrightarrow (\mathcal{S}'') \Leftrightarrow (K + \lambda I)\alpha = \mathbf{y}$$

$\neq$ ridge regression:

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\|K\alpha - \mathbf{y}\|^2 + \frac{\lambda}{2}\alpha^\top\alpha$$

Kernel logistic regression

inspiration: the Bayes rule

$$D(\mathbf{x}) = \text{sign}\big(f(\mathbf{x}) + \alpha_0\big) \quad \Longrightarrow \quad \log\left(\frac{\mathbb{P}(Y=1|\mathbf{x})}{\mathbb{P}(Y=-1|\mathbf{x})}\right) = f(\mathbf{x}) + \alpha_0$$

probabilities:

$$\mathbb{P}(Y=1|\mathbf{x}) = \frac{\exp^{f(\mathbf{x})+\alpha_0}}{1+\exp^{f(\mathbf{x})+\alpha_0}} \qquad \mathbb{P}(Y=-1|\mathbf{x}) = \frac{1}{1+\exp^{f(\mathbf{x})+\alpha_0}}$$

Rademacher distribution

$$\mathcal{L}(x_i, y_i, f, \alpha_0) = \mathbb{P}(Y=1|\mathbf{x}_i)^{\frac{y_i+1}{2}}\left(1-\mathbb{P}(Y=1|\mathbf{x}_i)\right)^{\frac{1-y_i}{2}}$$

penalized likelihood

$$\begin{aligned} J(f, \alpha_0) &= -\sum_{i=1}^{n} \log\big(\mathcal{L}(x_i, y_i, f, \alpha_0)\big) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^{n} \log\left(1+\exp^{-y_i(f(\mathbf{x}_i)+\alpha_0)}\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \end{aligned}$$

Kernel logistic regression (2)

$$(\mathcal{R}) \quad \begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}\sum\limits_{i=1}^{n} \log\left(1 + \exp^{-\xi_i}\right) \\ \text{with} & \xi_i = y_i\left(f(\mathbf{x}_i) + \alpha_0\right), \qquad i = 1, n \end{cases}$$

Representer theorem

$$J(\alpha, \alpha_0) = \mathbb{1}^\top \log\left(\mathbb{1} + \exp^{\text{diag}(\mathbf{y})K\alpha + \alpha_0 \mathbf{y}}\right) + \frac{\lambda}{2}\,\alpha^\top K\alpha$$

gradient vector anf Hessian matrix:

$$\nabla_\alpha J(\alpha, \alpha_0) = K\big(\mathbf{y} - (2\mathbf{p} - \mathbb{1})\big) + \lambda K\alpha$$

$$H_\alpha J(\alpha, \alpha_0) = K\text{diag}\big(\mathbf{p}(\mathbb{1} - \mathbf{p})\big)K + \lambda K$$

solve the problem using Newton iterations

$$\alpha^{\text{new}} = \alpha^{\text{old}} + \left(K\text{diag}\big(\mathbf{p}(\mathbb{1} - \mathbf{p})\big)K + \lambda K\right)^{-1} K\big(\mathbf{y} - (2\mathbf{p} - \mathbb{1}) + \lambda\alpha\big)$$

Let's summarize

- pros
  - Universality
  - from $\mathcal{H}$ to $\mathbb{R}^n$ using the representer theorem
  - no (explicit) curse of dimensionality

- splines $\mathcal{O}(n^3)$ (can be reduced to $\mathcal{O}(n^2)$)

- logistic regression $\mathcal{O}(kn^3)$ (can be reduced to $\mathcal{O}(kn^2)$)

- no scalability!

sparsity comes to the rescue!

Roadmap

1. Kernel machines
    - Non sparse kernel machines
    - sparse kernel machines: SVM
    - Sparse kernel machines for regression: SVR
    - practical SVM

SVM: the separable case (no noise)

$$\begin{cases} \max\limits_{f,\alpha_0} & m \\ \text{with} & y_i\big(f(\mathbf{x}_i) + \alpha_0\big) \geq m \\ \text{and} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 = 1 \end{cases} \Leftrightarrow \begin{cases} \min\limits_{f,\alpha_0} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\ \text{with} & y_i\big(f(\mathbf{x}_i) + \alpha_0\big) \geq 1 \end{cases}$$

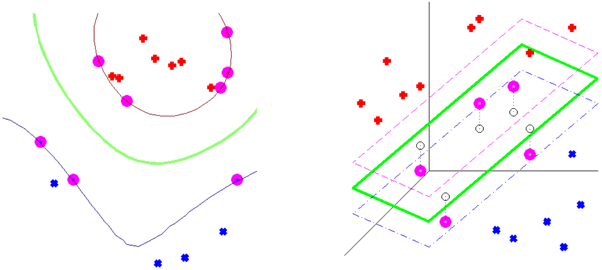
3 ways to represent function f

$$\underbrace{f(x)}_{\text{in the RKHS } \mathcal{H}} = \underbrace{\sum_{j=1}^{d} w_j \; \phi_j(\mathbf{x})}_{d \text{ features}} = \underbrace{\sum_{i=1}^{n} \alpha_i \; y_i \; k(\mathbf{x}, \mathbf{x}_i)}_{n \text{ data points}}$$

$$\begin{cases} \min\limits_{\mathbf{w},\alpha_0} & \frac{1}{2}\|\mathbf{w}\|_{\mathbb{R}^d}^2 = \frac{1}{2}\;\mathbf{w}^\top\mathbf{w} \\ \text{with} & y_i\big(\mathbf{w}^\top\phi(\mathbf{x}_i) + \alpha_0\big) \geq 1 \end{cases} \Leftrightarrow \begin{cases} \min\limits_{\alpha,\alpha_0} & \frac{1}{2}\;\alpha^\top K\alpha \\ \text{with} & y_i\big(\alpha^\top K(:,i) + \alpha_0\big) \geq 1 \end{cases}$$

using relevant features...

a data point becomes a function $\mathbf{x} \longrightarrow k(\mathbf{x}, \bullet)$



input space representation: x   feature space: k(x,.)

Representer theorem for SVM

$$\begin{cases} \min\limits_{f,\alpha_0} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{with} & y_i\big(f(\mathbf{x}_i)+\alpha_0\big) \geq 1 \end{cases}$$

Lagrangian

$$L(f,\alpha_0,\alpha) = \frac{1}{2}\|f\|^2_{\mathcal{H}} - \sum_{i=1}^{n} \alpha_i\big(y_i(f(\mathbf{x}_i)+\alpha_0)-1\big) \qquad \alpha \geq 0$$

optimility condition: $\nabla_f L(f,\alpha_0,\alpha) = 0 \Leftrightarrow f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i,\mathbf{x})$

Eliminate f from L: $\begin{cases} \|f\|^2_{\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \\ \sum_{i=1}^{n} \alpha_i y_i f(\mathbf{x}_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \end{cases}$

$$Q(\alpha_0,\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i\big(y_i\alpha_0 - 1\big)$$

Dual formulation for SVM

the intermediate function

$$Q(\alpha_0, \alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \alpha_0\Big(\sum_{i=1}^{n}\alpha_i y_i\Big) + \sum_{i=1}^{n}\alpha_i$$

$$\max_{\alpha}\min_{\alpha_0} \; Q(\alpha_0, \alpha)$$

α_0 can be seen as the Lagrange multiplier of the following (balanced) constaint $\sum_{i=1}^{n}\alpha_i y_i = 0$ which is also the optimality KKT condition on α_0

Dual formulation

$$\begin{cases} \displaystyle\max_{\alpha\in\mathbb{R}^n} & -\frac{1}{2}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i \\ \text{such that} & \displaystyle\sum_{i=1}^{n}\alpha_i y_i = 0 \\ \text{and} & 0 \le \alpha_i, \quad i = 1, n \end{cases}$$

SVM dual formulation

Dual formulation

$$\begin{cases} \max\limits_{\alpha \in \mathbb{R}^n} & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \alpha_i \\ \text{with} & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{and } 0 \leq \alpha_i, \quad i = 1, n \end{cases}$$

The dual formulation gives a quadratic program (QP)

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2} \alpha^\top G \alpha - \mathbb{1}^\top \alpha \\ \text{with} & \alpha^\top \mathbf{y} = 0 \quad \text{and } 0 \leq \alpha \end{cases}$$

with $G_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$

with the linear kernel $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}^\top \mathbf{x}_i) = \sum_{j=1}^{d} \beta_j x_j$
when d is small wrt. n primal may be interesting.

the general case: C-SVM

Primal formulation

$$
(\mathcal{P}) \begin{cases} \displaystyle\min_{f \in \mathcal{H}, \alpha_0, \xi \in \mathbb{R}^n} & \frac{1}{2}\|f\|^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i^p \\ \text{such that} & y_i\big(f(\mathbf{x}_i) + \alpha_0\big) \geq 1 - \xi_i, \ \ \xi_i \geq 0, \ \ i = 1, n \end{cases}
$$

C is the *regularization path* parameter (to be tuned)

$p = 1$, L_1 SVM

$$
\begin{cases} \displaystyle\max_{\alpha \in \mathbb{R}^n} & -\frac{1}{2}\alpha^\top H \alpha + \alpha^\top \mathbb{1} \\ \text{such that} & \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C \quad i = 1, n \end{cases}
$$

$p = 2$, L_2 SVM

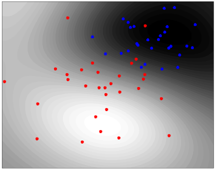
$$
\begin{cases} \displaystyle\max_{\alpha \in \mathbb{R}^n} & -\frac{1}{2}\alpha^\top \left(H + \frac{1}{C} I\right) \alpha + \alpha^\top \mathbb{1} \\ \text{such that} & \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \quad i = 1, n \end{cases}
$$

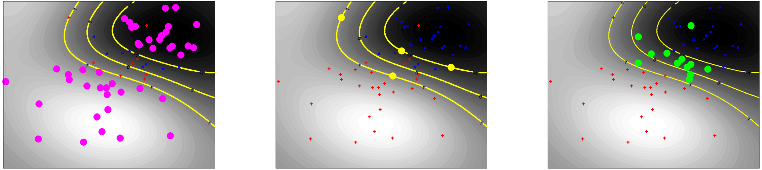
the regularization path: is the set of solutions $\alpha(C)$ when C varies

Data groups: illustration

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \alpha_0$$

$$D(x) = \text{sign}\big(f(\mathbf{x})\big)$$





useless data	important data	suspicious data
well classified	support	
$\alpha = 0$	$0 < \alpha < C$	$\alpha = C$

the regularization path: is the set of solutions $\alpha(C)$ when C varies

The importance of being support

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})$$

data point	α	constraint value	set
$\mathbf{x}_i$ *useless*	$\alpha_i = 0$	$y_i\big(f(\mathbf{x}_i) + \alpha_0\big) > 1$	I_0
$\mathbf{x}_i$ *support*	$0 < \alpha_i < C$	$y_i\big(f(\mathbf{x}_i) + \alpha_0\big) = 1$	I_α
$\mathbf{x}_i$ *suspicious*	$\alpha_i = C$	$y_i\big(f(\mathbf{x}_i) + \alpha_0\big) < 1$	I_C

Table: When a data point is « support » it lies exactly on the margin.

here lies the efficiency of the algorithm (and its complexity)!

sparsity: $\alpha_i = 0$

Two more ways to derivate SVM

Using the hinge loss

$$\min_{f\in\mathcal{H},\alpha_0\in\mathbb{R}} \frac{1}{p}\sum_{i=1}^{n}\max\big(0,1-y_i(f(\mathbf{x}_i)+\alpha_0)\big)^p + \frac{1}{2C}\|f\|^2$$

Minimizing the distance between the convex hulls

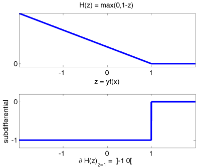
$$\begin{cases} \displaystyle\min_{\alpha} & \|u-v\|^2_{\mathcal{H}} \\ \text{with} & \displaystyle u(\mathbf{x}) = \sum_{\{i|y_i=1\}} \alpha_i k(\mathbf{x}_i,\mathbf{x}), \qquad v(\mathbf{x}) = \sum_{\{i|y_i=-1\}} \alpha_i k(\mathbf{x}_i,\mathbf{x}) \\ \text{and} & \displaystyle\sum_{\{i|y_i=1\}} \alpha_i = 1, \sum_{\{i|y_i=-1\}} \alpha_i = 1, \quad 0 \le \alpha_i \quad i=1,n \end{cases}$$

$$f(\mathbf{x}) = \frac{2}{\|u-v\|^2_{\mathcal{H}}}\big(u(\mathbf{x})-v(\mathbf{x})\big) \text{ and } \alpha_0 = \frac{\|u\|^2_{\mathcal{H}} - \|v\|^2_{\mathcal{H}}}{\|u-v\|^2_{\mathcal{H}}}$$

the regularization path: is the set of solutions $\alpha(C)$ when C varies

Regularization path for SVM

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \max(1 - y_i f(\mathbf{x}_i), 0) + \frac{\lambda_o}{2} \|f\|_{\mathcal{H}}^2$$

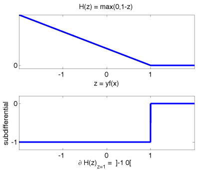


I_α is the set of support vectors s.t. $y_i f(\mathbf{x}_i) = 1$;

$$\partial_f J(f) = \sum_{i \in I_\alpha} \alpha_i y_i K(\mathbf{x}_i, \bullet) - \sum_{i \in I_1} y_i K(\mathbf{x}_i, \bullet) + \lambda_o f(\bullet) \quad \text{with} \quad \alpha_i \in \partial H(1) =]-1, 0[$$

Regularization path for SVM

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \max(1 - y_i f(\mathbf{x}_i), 0) + \frac{\lambda_o}{2} \|f\|^2_{\mathcal{H}}$$



I_α is the set of support vectors s.t. $y_i f(\mathbf{x}_i) = 1$;

$$\partial_f J(f) = \sum_{i \in I_\alpha} \alpha_i y_i K(\mathbf{x}_i, \bullet) - \sum_{i \in I_1} y_i K(\mathbf{x}_i, \bullet) + \lambda_o\, f(\bullet) \quad \text{with} \quad \alpha_i \in \partial H(1) =]-1, 0[$$

Let λ_n a value close enough to λ_o to keep the sets I_0, I_α and I_C unchanged

In particular at point $\mathbf{x}_j \in I_\alpha$ $\big(f_o(\mathbf{x}_j) = f_n(\mathbf{x}_j) = y_j\big)$: $\partial_f J(f)(\mathbf{x}_j) = 0$

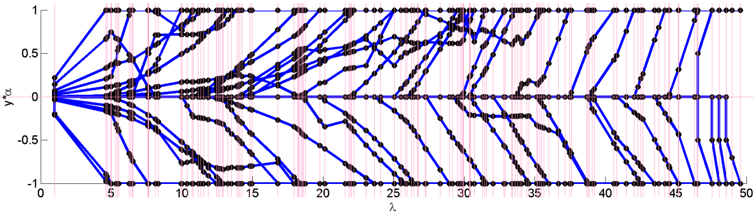
$$\begin{array}{rcl}
\sum_{i \in I_\alpha} \alpha_{io} y_i K(\mathbf{x}_i, \mathbf{x}_j) & = & \sum_{i \in I_1} y_i K(\mathbf{x}_i, \mathbf{x}_j) - \lambda_o\, y_j \\
\sum_{i \in I_\alpha} \alpha_{in} y_i K(\mathbf{x}_i, \mathbf{x}_j) & = & \sum_{i \in I_1} y_i K(\mathbf{x}_i, \mathbf{x}_j) - \lambda_n\, y_j \\
\hline
G(\alpha_n - \alpha_o) & = & (\lambda_o - \lambda_n)\mathbf{y} \qquad \text{avec} \qquad G_{ij} = y_i K(\mathbf{x}_i, \mathbf{x}_j)
\end{array}$$

$$\alpha_n = \alpha_o + (\lambda_o - \lambda_n)\mathbf{w}$$

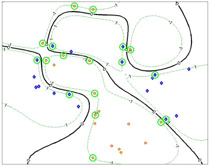
$$\mathbf{w} = (G)^{-1}\mathbf{y}$$

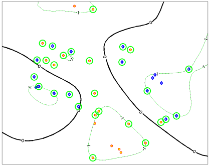
Example of regularization path

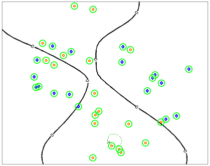
$$\alpha_i \in]-1, 0[\qquad y_i\alpha_i \in]-1, -1[\qquad \lambda = \frac{1}{C}$$



α_i estimation and data selection







How to choose ℓ and P to get linear *regularization path*?

the *path* is piecewise linear $\Leftrightarrow$ one is piecewise quadratic and the other is piecewise linear

the convex case [Rosset & Zhu, 07]

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \ \ell(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})$$

1. piecewise linearity: $\displaystyle \lim_{\varepsilon \to 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = \text{constant}$
2. optimality
$$\nabla \ell(\boldsymbol{\beta}(\lambda)) + \lambda \nabla P(\boldsymbol{\beta}(\lambda)) = 0$$
$$\nabla \ell(\boldsymbol{\beta}(\lambda + \varepsilon)) + (\lambda + \varepsilon) \nabla P(\boldsymbol{\beta}(\lambda + \varepsilon)) = 0$$
3. Taylor expension

$$\lim_{\varepsilon \to 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = \left[\nabla^2 \ell(\boldsymbol{\beta}(\lambda)) + \lambda \nabla^2 P(\boldsymbol{\beta}(\lambda))\right]^{-1} \nabla P(\boldsymbol{\beta}(\lambda))$$

$$\nabla^2 \ell(\boldsymbol{\beta}(\lambda)) = \text{constant} \quad \text{and} \quad \nabla^2 P(\boldsymbol{\beta}(\lambda)) = 0$$

Problems with Piecewise linear regularization path

L	P	*regression*	*classification*	*clustering*
L_2	L_1	Lasso/LARS	L1 L2 SVM	PCA L1
L_1	L_2	SVR	SVM	OC SVM
L_1	L_1	L1 LAD	L1 SVM	
		Danzig Selector		

Table: example of piecewise linear regularization path algorithms.

$$P: \quad L_p = \sum_{j=1}^{d} |\beta_j|^p$$

$$L: \quad L_p : |f(\mathbf{x}) - y|^p \quad \text{hinge } (yf(\mathbf{x}) - 1)_+^p$$

ε-insensitive
$$\begin{cases} 0 & \text{if } |f(\mathbf{x}) - y| < \varepsilon \\ |f(\mathbf{x}) - y| - \varepsilon & \text{else} \end{cases}$$

Huber's loss:
$$\begin{cases} |f(\mathbf{x}) - y|^2 & \text{if } |f(\mathbf{x}) - y| < t \\ 2t|f(\mathbf{x}) - y| - t^2 & \text{else} \end{cases}$$

K-Lasso (Kernel Basis pursuit)

The Kernel Lasso

$$(\mathcal{S}_1) \quad \left\{ \min_{\alpha \in \mathbb{R}^n} \quad \frac{1}{2}\|K\alpha - \mathbf{y}\|^2 + \lambda \sum_{i=1}^{n} |\alpha_i| \right.$$

- Typical parametric quadratic program (pQP) with $\alpha_i = 0$
- Piecewise linear regularization path

The dual:

$$(\mathcal{D}_1) \quad \left\{ \begin{array}{rl} \min\limits_{\alpha} & \frac{1}{2}\|K\alpha\|^2 \\ \text{such that} & K^\top (K\alpha - \mathbf{y}) \leq t \end{array} \right.$$

- The K-Danzig selector can be treated the same way
- require to compute $K^\top K$ - no more function f!

Support vector regression (SVR)

Lasso's dual adaptation:

$$\begin{cases} \min\limits_{\alpha} & \frac{1}{2}\|K\alpha\|^2 \\ \text{s. t.} & K^\top(K\alpha - \mathbf{y}) \leq t \end{cases} \qquad \begin{cases} \min\limits_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{s. t.} & |f(\mathbf{x}_i) - y_i| \leq t, \; i = 1, n \end{cases}$$

The support vector regression introduce slack variables

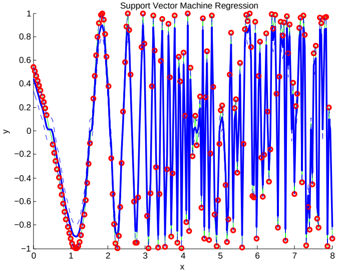
$$(SVR) \quad \begin{cases} \min\limits_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum|\xi_i| \\ \text{such that} & |f(\mathbf{x}_i) - y_i| \leq t + \xi_i \quad 0 \leq \xi_i \quad i = 1, n \end{cases}$$

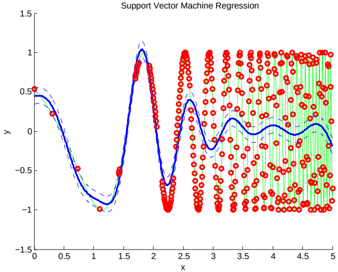
- a typical multi parametric quadratic program (mpQP)
- piecewise linear regularization path

$$\alpha(C, t) = \alpha(C_0, t_0) + (\frac{1}{C} - \frac{1}{C_0})\mathbf{u} + \frac{1}{C_0}(t - t_0)\mathbf{v}$$

- 2d Pareto's front (the tube width and the regularity)

Support vector regression illustration





C large

C small

- there exists other formulations such as LP SVR...

ν-SVM and other formulations...

$\nu \in [0, 1]$

$$(\nu) \begin{cases} \min\limits_{f,\alpha_0,\xi,m} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{np}\sum\limits_{i=1}^{n}\xi_i^p - \nu m \\ \text{with} & y_i\big(f(\mathbf{x}_i) + \alpha_0\big) \geq m - \xi_i, \quad i = 1, n, \\ \text{and} & m \geq 0, \quad \xi_i \geq 0, \quad i = 1, n, \end{cases}$$

for $p = 1$ the dual formulation is:

$$\begin{cases} \max\limits_{\alpha \in \mathbb{R}^n} & -\frac{1}{2}\alpha^\top G \alpha \\ \text{with} & \alpha^\top \mathbf{y} = 0 \text{ et } 0 \leq \alpha_i \leq \frac{1}{n} \quad i = 1, n \\ \text{and} & \nu \leq \alpha^\top \mathbb{I} \end{cases}$$

$$C = \frac{1}{m}$$

SVM with non symetric costs

problem in the primal

$$\begin{cases} \min\limits_{f\in\mathcal{H},\alpha_0,\xi\in\mathbb{R}^n} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + C^+ \sum\limits_{\{i|y_i=1\}} \xi_i^p + C^- \sum\limits_{\{i|y_i=-1\}} \xi_i^p \\ \text{with} & y_i\big(f(\mathbf{x}_i)+\alpha_0\big) \geq 1-\xi_i, \quad \xi_i \geq 0, \quad i=1,n \end{cases}$$

for $p=1$ the dual formulation is the following:

$$\begin{cases} \max\limits_{\alpha\in\mathbb{R}^n} & -\frac{1}{2}\alpha^\top G\alpha + \alpha^\top \mathbb{1} \\ \text{with} & \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C^+ \text{ or } C^- \quad i=1,n \end{cases}$$

Generalized SVM

$$\min_{f\in\mathcal{H},\alpha_0\in\mathbb{R}} \sum_{i=1}^{n} \max\big(0, 1 - y_i(f(\mathbf{x}_i)+\alpha_0)\big) + \frac{1}{C}\varphi(f) \qquad \varphi \text{ convex}$$

in particular $\varphi(f) = \|f\|_p^p$ with $p = 1$ leads to L1 SVM.

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n,\alpha_0,\xi} & \mathbb{1}^\top \boldsymbol{\beta} + C\mathbb{1}^\top \xi \\ \text{with} & y_i\Big(\sum\limits_{j=1}^{n} \alpha_j k(\mathbf{x}_i,\mathbf{x}_j) + \alpha_0\Big) \ge 1 - \xi_i, \\ \text{and} & -\beta_i \le \alpha_i \le \beta_i, \quad \xi_i \ge 0, \quad i = 1, n \end{cases}$$

with $\boldsymbol{\beta} = |\alpha|$. the dual is:

$$\begin{cases} \max\limits_{\gamma,\delta,\delta^*\in\mathbb{R}^{3n}} & \mathbb{1}^\top \gamma \\ \text{with} & \mathbf{y}^\top\gamma = 0, \quad \delta_i + \delta_i^* = 1 \\ & \sum_{j=1}^{n} \gamma_j k(\mathbf{x}_i,\mathbf{x}_j) = \delta_i - \delta_i^*, \quad i = 1, n \\ \text{and} & 0 \le \delta_i, 0 \le \delta_i^*, \ 0 \le \gamma_i \le C, \quad i = 1, n \end{cases}$$

Mangassarian, 2001

SVM reduction (reduced set method))

- objective: compile the model
- $f(x) = \sum_{i=1}^{n_s} \alpha_i k(\mathbf{x}_i, \mathbf{x}), n_s \ll n,$ $\quad n_s$ too big
- compiled model as the solution of: $g(\mathbf{x}) = \sum_{i=1}^{n_c} \beta_i k(\mathbf{z}_i, \mathbf{x}), n_c \ll n_s$
- $\beta, \mathbf{z}_i$ and c are tuned by minimizing:

$$\min_{\beta, \mathbf{z}_i} \|g - f\|_H^2$$

  where

$$\min_{\beta, \mathbf{z}_i} \|g - f\|_H^2 = \alpha^\top K_x \alpha + \boldsymbol{\beta}^\top K_z \boldsymbol{\beta} - 2\alpha^\top K_{xz} \boldsymbol{\beta}$$

  some authors advice $0,03 \leq \frac{n_c}{n_s} \leq 0,1$
- solve it by using use (stochastic) gradient (its a RBF problem)

Burges 1996, Ozuna 1997, Romdhani 2001

SVM and probabilities (1/2)

$\log \frac{\mathbb{P}(Y=1|\mathbf{x})}{\mathbb{P}(Y=-1|\mathbf{x})}$ as (almost) the same sign as $f(\mathbf{x})$

$$\log \frac{\mathbb{P}(Y=1|\mathbf{x})}{\mathbb{P}(Y=-1|\mathbf{x})} = a_1 f(\mathbf{x}) + a_2 \quad \mathbb{P}(Y=1|\mathbf{x}) = 1 - \frac{1}{1+\exp^{a_1 f(\mathbf{x})+a_2}}$$

a_1 et a_2 estimated using maximum likelihood

some facts

- SVM is universaly consistent (coverges towards the Bayes risk)
- SVM asymptotically implements the bayes rule
- but theoreticaly: no consistency towards conditional probabilities (due to the nature of sparsity)
- to estimate conditional probabilities on an interval (typicaly$[\frac{1}{2}-\eta, \frac{1}{2}+\eta]$) to spasness in this interval (all data points have to be support vectors)

, 99 ; Bartlett & Tewari, JMLR, 07

SVM and probabilities (2/2)

An alternative approach

$$g(\mathbf{x}) - \varepsilon^-(\mathbf{x}) \ \leq \ \mathbb{P}(Y = 1|\mathbf{x}) \ \leq \ g(\mathbf{x}) + \varepsilon^+(\mathbf{x})$$

with $g(\mathbf{x}) = \frac{1}{1+4^{-f(\mathbf{x})-\alpha_0}}$
non parametric functions ε^- and ε^+ have to verify:

$$\begin{aligned} g(\mathbf{x}) + \varepsilon^+(\mathbf{x}) &= \exp^{-a_1(1-f(\mathbf{x})-\alpha_0)_+ + a_2} \\ 1 - g(\mathbf{x}) - \varepsilon^-(\mathbf{x}) &= \exp^{-a_1(1+f(\mathbf{x})+\alpha_0)_+ + a_2} \end{aligned}$$

with $a_1 = \log 2$ and $a_2 = 0$

Grandvalet *et al.*, 07

logistic regression and the import vector machine

- Logistic regression is NON sparse
- kernalize it using the *dictionary* strategy
- Algorithm:
  - find the solution of the KLR using only a subset $\mathcal{S}$ of the data
  - build $\mathcal{S}$ iteratively using active constraint approach
- this trick brings sparsity
- it estimates probability
- it can naturally be generalized to the multiclass case

- efficent when uses:
  - a few import vectors
  - component-wise update procedure

- extention using L_1 KLR

Zhu & Hastie, 01 ; Keerthi *et. al.*, 02

Multiclass SVM

- one *vs* all: winner takes all
- one *vs* one:
  - max-wins voting
  - pairwise coupling: use probability
- global approach (size $c \times n$),
  - formal (differents variations)

$$\begin{cases} \min\limits_{f \in \mathcal{H}, \alpha_0, \xi \in \mathbb{R}^n} & \frac{1}{2} \sum\limits_{\ell=1}^{c} \|f_\ell\|_{\mathcal{H}}^2 + \frac{C}{p} \sum\limits_{i=1}^{n} \sum\limits_{\ell=1, \ell \neq y_i}^{c} \xi_{i\ell}^p \\ \text{with} & y_i \big(f_{y_i}(\mathbf{x}_i) + b_{y_i} - f_\ell(\mathbf{x}_i) - b_\ell \big) \geq 1 - \xi_{i\ell} \\ \text{and} & \xi_{i\ell} \geq 0 \text{ for } i = 1, ..., n; \quad \ell = 1, ..., c; \quad \ell \neq y_i \end{cases}$$

  non consistent estimator but practicaly usefull
  - structured outputs

approach	problem size	number of sub problems
all together	$n \times c$	1
1 vs. all	n	c
1 vs. 1	$\frac{2n}{c}$	$\frac{c(c-1)}{2}$

Multiclass SVM

- one *vs* all: winner takes all
- one *vs* one:
  - max-wins voting
  - pairwise coupling: use probability – best results
- global approach (size $c \times n$),
  - formal (differents variations)

$$\begin{cases} \displaystyle\min_{f \in \mathcal{H}, \alpha_0, \xi \in \mathbb{R}^n} & \displaystyle\frac{1}{2}\sum_{\ell=1}^{c} \|f_\ell\|_{\mathcal{H}}^2 + \frac{C}{p}\sum_{i=1}^{n}\sum_{\ell=1, \ell \neq y_i}^{c} \xi_{i\ell}^p \\ \text{with} & y_i\big(f_{y_i}(\mathbf{x}_i) + b_{y_i} - f_\ell(\mathbf{x}_i) - b_\ell\big) \geq 1 - \xi_{i\ell} \\ \text{and} & \xi_{i\ell} \geq 0 \text{ for } i = 1, ..., n; \quad \ell = 1, ..., c; \quad \ell \neq y_i \end{cases}$$

  non consistent estimator but practicaly usefull
  - structured outputs

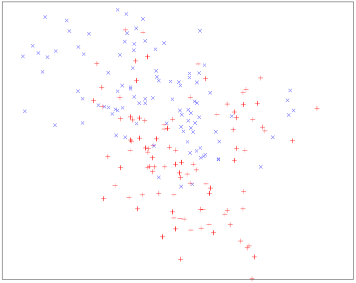
approach	problem size	number of sub problems
all together	$n \times c$	1
1 vs. all	n	c
1 vs. 1	$\frac{2n}{c}$	$\frac{c(c-1)}{2}$

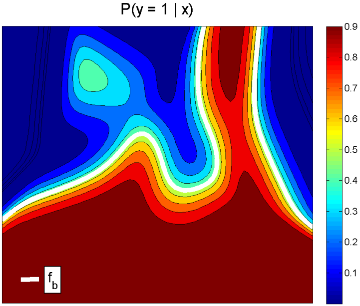
Duan & Keerti, 05 ; Bartlett 06

Roadmap

1. Kernel machines
    - Non sparse kernel machines
    - sparse kernel machines: SVM
    - Sparse kernel machines for regression: SVR
    - practical SVM

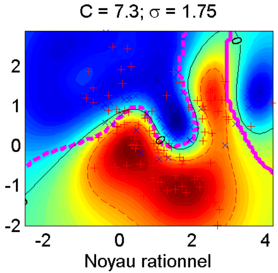
Mixture data

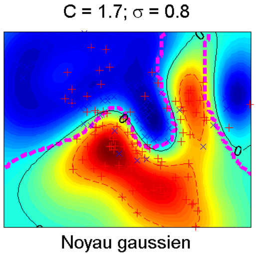




- x : 200×2
- y : 100 for each class
- mixturee model with 10 gaussians
- the bayes error is known

the kernel effect



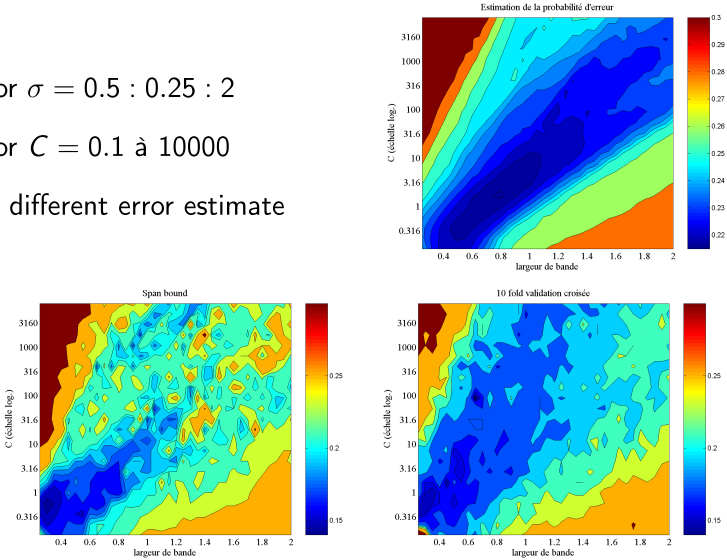


tuning C and σ : *grid search*

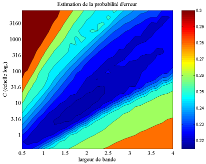
for $\sigma = 0.5 : 0.25 : 2$

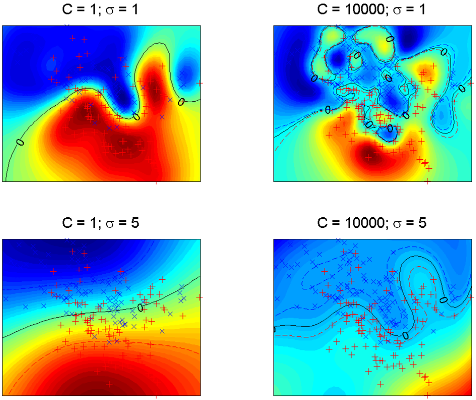
for $C = 0.1$ à 10000

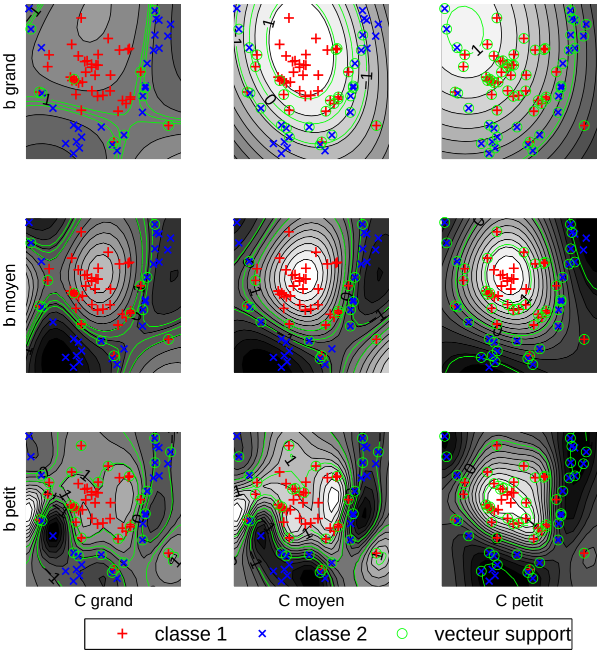
3 different error estimate



C and σ influence

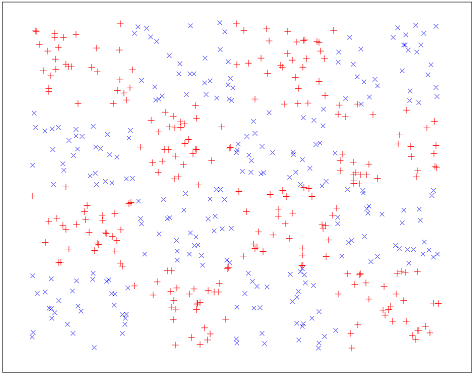




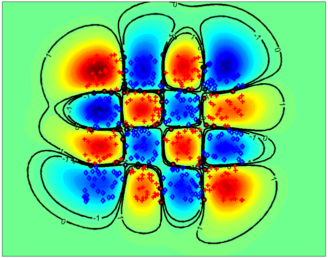

b grand
b moyen
b petit
C grand
C moyen
C petit
classe 1
classe 2
vecteur support

checker board

- 2 classes
- 500 examples
- separable

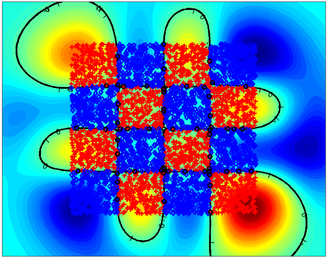


a separable case

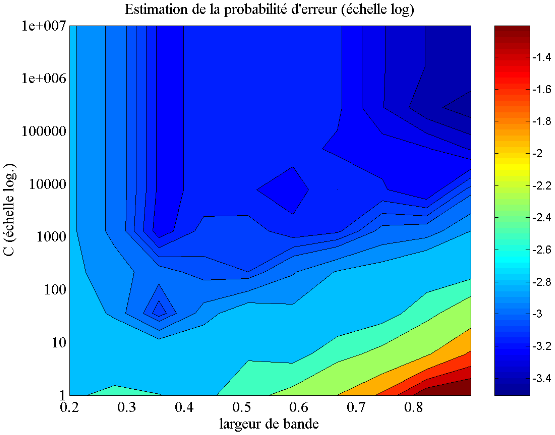


$n = 500$ data points

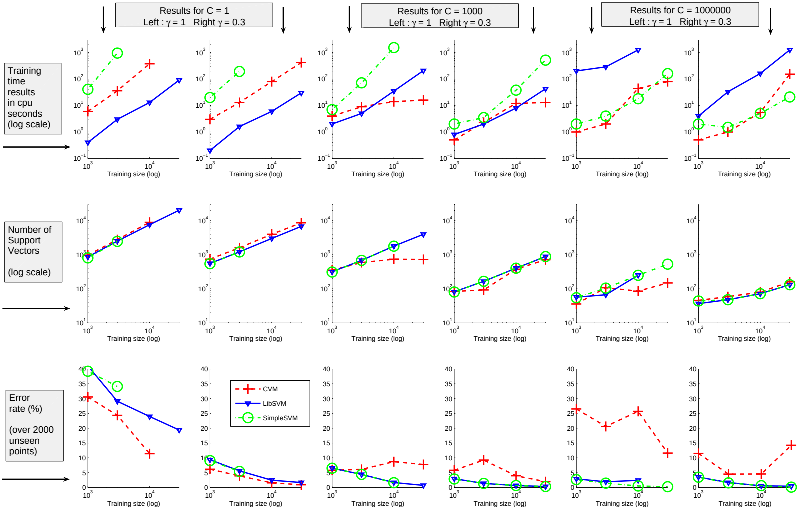
$n = 5000$ data points



tuning C and σ : *grid search*



empirical complexity



osli *et al* JMLR, 2007

Historical perspective on kernel machines

statistics

1960 Parzen, Nadaraya Watson

1970 Splines

1980 Kernels: Silverman, Hardle...

1990 sparsity: Donoho (pursuit), Tibshirani (Lasso)...

Statistical learning

1985 Neural networks:

- non linear - universal
- structural complexity
- non convex optimization

1992 Vapnik et. al.

- theory - regularization - consistancy
- convexity - Linearity
- Kernel - universality
- sparsity
- results: MNIST

what's new since 1995

- Applications
  - kernlisation $w^\top \mathbf{x} \rightarrow \langle f, k(\mathbf{x}, .)\rangle_{\mathcal{H}}$
  - kernel engineering
  - sturtured outputs
  - applications: image, text, signal, bio-info...

- Optimization
  - dual: mloss.org
  - regularization path
  - approximation
  - primal

- Statistic
  - proofs and bounds
  - model selection
    - span bound
    - multikernel: tuning (k and σ)

challenges: towards tough learning

- the size effect
  - ready to use: automatization
  - adaptative: on line context aware
  - beyond kenrels
- Automatic and adaptive model selection
  - variable selection
  - kernel tuning (k et σ)
  - hyperparametres: C, duality gap, λ
- $\mathbb{P}$ change
- Theory
  - non positive kernels
  - a more general representer theorem

biblio: kernel-machines.org

- John Shawe-Taylor and Nello Cristianini Kernel Methods for Pattern Analysis, Cambridge University Press, 2004
- Bernhard Schölkopf and Alex Smola. Learning with Kernels. MIT Press, Cambridge, MA, 2002.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning:. Data Mining, Inference, and Prediction, springer, 2001

- Léon Bottou, Olivier Chapelle, Dennis DeCoste and Jason Weston Large-Scale Kernel Machines (Neural Information Processing, MIT press 2007
- Olivier Chapelle, Bernhard Scholkopf and Alexander Zien, Semi-supervised Learning, MIT press 2006

- Vladimir Vapnik. Estimation of Dependences Based on Empirical Data. Springer Verlag, 2006, 2nd edition.
- Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.

- Grace Wahba. Spline Models for Observational Data. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics vol. 59, Philadelphia, 1990
- Alain Berlinet and Christine Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics,Kluwer Academic Publishers, 2003
- Marc Atteia et Jean Gaches , Approximation Hilbertienne - Splines, Ondelettes, Fractales, PUG, 1999