Kernel Machines

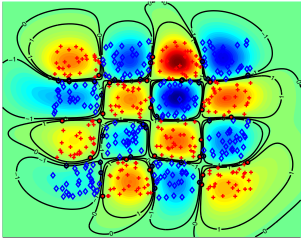
Stéphane Canu, Gilles Gasso
{stephane.canu, gilles.gasso}@insa-rouen.fr

DM_2 2016

February 1, 2017

Plan

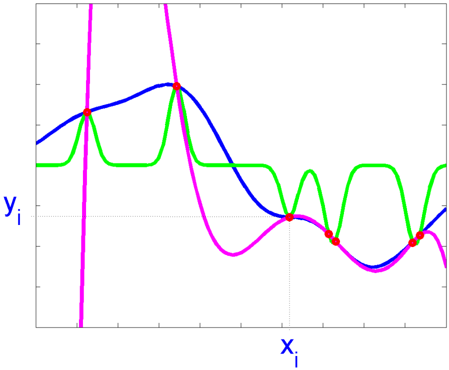
1. Non sparse kernel machines
   - Interpolation problem
   - From Regression to classification: kernel logistic regression
2. Sparse kernel machines: SVM
   - Sparsity of kernel SVM for classification
     - SVM: variations on a theme
   - Sparse kernel machines for regression



Splines interpolation

Interpolation and regression problems

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



It is an ill posed problem

Interpolation splines: minimum norm interpolation

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

Remark: we assume $\mathcal{H}$ is a Hilbert space induced by reproducing kernel k

The lagrangian ($\alpha_i \in \mathbb{R}$ are Lagrange multipliers)

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

Interpolation splines: minimum norm interpolation

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

Remark: we assume $\mathcal{H}$ is a Hilbert space induced by reproducing kernel k

The lagrangian ($\alpha_i \in \mathbb{R}$ are Lagrange multipliers)

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

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

Interpolation splines: minimum norm interpolation

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

Remark: we assume $\mathcal{H}$ is a Hilbert space induced by reproducing kernel k

The lagrangian ($\alpha_i \in \mathbb{R}$ are Lagrange multipliers)

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

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

dual formulation (remove f from the lagrangian):

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

$$K\boldsymbol{\alpha} = y$$

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:

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

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

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

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

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

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

Smoothing splines

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

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

one equivalent formulation

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

using the representer theorem

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

solution: $\quad (\mathcal{S}) \Leftrightarrow (\mathcal{S}') \quad \Leftrightarrow \quad \boldsymbol{\alpha} = (K + \lambda I)^{-1}\mathbf{y}$

Remark: this is different from ridge regression problem

$$\min_{\alpha \in \mathbb{R}^n} \ \frac{1}{2}\|K\boldsymbol{\alpha} - \mathbf{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} \quad \text{with} \quad \text{solution} \quad \boldsymbol{\alpha} = (K^\top K + \lambda I)^{-1}K^\top \mathbf{y}$$

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

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

penalized log-likelihood

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

Kernel logistic regression (2)

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

Representer theorem leads to

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

gradient vector and Hessian matrix:

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

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

solve the problem using Newton iterations

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

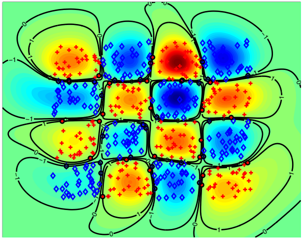
Let's summarize

- Kernel machines: 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!

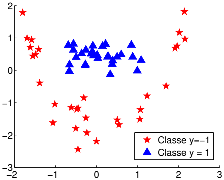
sparsity comes to the rescue!

Roadmap

1. Non sparse kernel machines
   - Interpolation problem
   - From Regression to classification: kernel logistic regression
2. Sparse kernel machines: SVM
   - Sparsity of kernel SVM for classification
     - SVM: variations on a theme
   - Sparse kernel machines for regression



SVM in a RKHS: the separable case (no noise)



- dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1}^n$
- problem : learn a non linear SVM

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

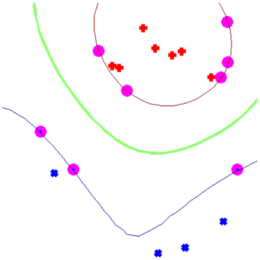
3 ways to represent function f

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

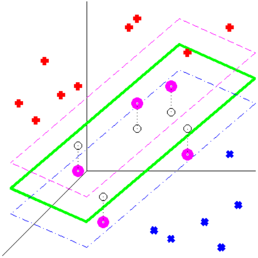
$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|_{\mathbb{R}^d}^2 = \frac{1}{2}\,\mathbf{w}^\top\mathbf{w} \\ \text{with} & y_i\big(\mathbf{w}^\top\phi(\mathbf{x}_i) + b\big) \geq 1 \end{cases} \Leftrightarrow \begin{cases} \min\limits_{\boldsymbol{\alpha},b} & \frac{1}{2}\,\boldsymbol{\alpha}^\top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

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

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

Lagrangian

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

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

(ii) $\nabla_f L(f,b,\boldsymbol{\alpha}) = 0 \Leftrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$

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

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

Dual formulation for SVM

the intermediate function

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

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

b 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 b

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)

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

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

case of linear kernel: $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}^\top \mathbf{x}_i) = \sum_{j=1}^{d} \beta_j x_j$

remark: when d is small wrt. n solving the primal may be interesting.

the general case: C-SVM

Primal formulation

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

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

Dual formulation

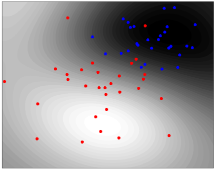
$$\begin{cases} \displaystyle\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} & -\frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbb{1} \\ \text{such that} & \boldsymbol{\alpha}^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C \quad i=1,n \end{cases}$$

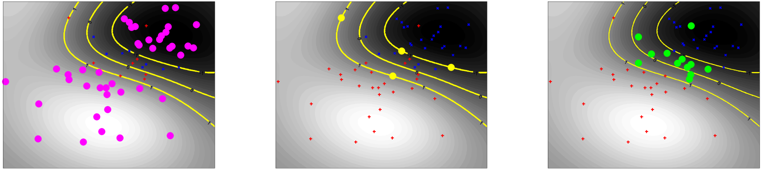
remark: regularization path is the set of solutions $\boldsymbol{\alpha}(C)$ when C varies

Data groups: illustration

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

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





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*	$\alpha_i = 0$	$y_i\big(f(\mathbf{x}_i) + b\big) > 1$	I_0
$\mathbf{x}_i$ *support*	$0 < \alpha_i < C$	$y_i\big(f(\mathbf{x}_i) + b\big) = 1$	I_α
$\mathbf{x}_i$ *suspicious*	$\alpha_i = C$	$y_i\big(f(\mathbf{x}_i) + b\big) < 1$	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: $\alpha_i = 0$

The active set method for SVM (1)

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbb{1} \\ \text{such that} & \boldsymbol{\alpha}^\top \mathbf{y} = 0 \quad i = 1, n \\ \text{and} & 0 \le \alpha_i \quad i = 1, n \end{cases}$$

$$\begin{cases} G\boldsymbol{\alpha} - \mathbb{1} - \beta + b\mathbf{y} = 0 \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0 \\ 0 \le \alpha_i \quad i = 1, n \\ 0 \le \beta_i \quad i = 1, n \\ \alpha_i\beta_i = 0 \quad i = 1, n \end{cases}$$

$$\underbrace{\begin{bmatrix} G_a & G_i^\top \\ G_i & G_0 \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} \alpha_a \\ 0 \end{bmatrix}}_{\boldsymbol{\alpha}} - \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\mathbb{1}} - \underbrace{\begin{bmatrix} 0 \\ \beta_0 \end{bmatrix}}_{\beta} + b \underbrace{\begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_0 \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{0}$$

$$\begin{array}{ll} (1) & G_a\boldsymbol{\alpha}_a - \mathbb{1}_a + b\mathbf{y}_a = 0 \\ (2) & G_i\boldsymbol{\alpha}_a - \mathbb{1}_0 - \beta_0 + b\mathbf{y}_0 = 0 \end{array}$$

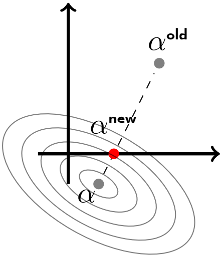
1. solve (1) (find α together with b)
2. if $\alpha_j < 0$ move it from I_α to I_0
   goto 1
3. else solve (2)
   if $\beta_j < 0$ move it from I_0 to I_α
   goto 1

The active set method for SVM (2)

Function $(\alpha, b, I_\alpha) \leftarrow$Solve_QP_Active_Set$(G, \mathbf{y})$

% Solve $min_\alpha \quad 1/2\alpha^\top G\alpha - \mathbb{1}^\top\alpha$
% s.t. $\quad 0 \leq \alpha$ and $\mathbf{y}^\top\alpha = 0$

$(I_\alpha, I_0, \alpha) \leftarrow$ initialization
while The_optimal_is_not_reached **do**
$(\alpha, b) \leftarrow$ solve $\begin{cases} G_a\boldsymbol{\alpha}_a - \mathbb{1}_a + b\mathbf{y}_a \\ \mathbf{y}_a^\top\boldsymbol{\alpha}_a \end{cases} = 0$
if $\exists i \in I_\alpha$ such that $\alpha_i < 0$ **then**
$\alpha \leftarrow$ projection($\boldsymbol{\alpha}_a, \boldsymbol{\alpha}$)
move i from I_α to I_0
else if $\exists j \in I_0$ such that $\beta_j < 0$ **then**
use $\beta_0 = \mathbf{y}_0(K_i\boldsymbol{\alpha}_a + b\mathbb{1}_0) - \mathbb{1}_0$
move j from I_0 to I_α
else
The_optimal_is_not_reached $\leftarrow$ FALSE
end if
end while



Projection step of the active constraints algorithm

```
d   = alpha - alphaold;
alpha   = alpha + t * d;
```

Caching Strategy

Save space and computing time by computing only the needed parts of kernel matrix G

Two more ways to get SVM

Using the hinge loss

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

Minimizing the distance between the convex hulls

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

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

SVM with non symmetric costs

problem in the primal

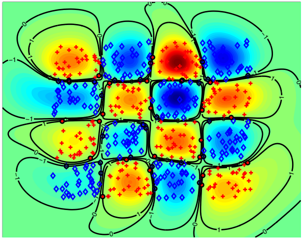
$$\begin{cases} \min\limits_{f\in\mathcal{H},b,\xi\in\mathbb{R}^n} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C^+ \sum\limits_{\{i|y_i=1\}} \xi_i + C^- \sum\limits_{\{i|y_i=-1\}} \xi_i \\ \text{with} & y_i\big(f(\mathbf{x}_i)+b\big) \geq 1-\xi_i, \;\; \xi_i \geq 0, \;\; i=1,n \end{cases}$$

dual formulation

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

Roadmap

1. Non sparse kernel machines
   - Interpolation problem
   - From Regression to classification: kernel logistic regression
2. Sparse kernel machines: SVM
   - Sparsity of kernel SVM for classification
     - SVM: variations on a theme
   - Sparse kernel machines for regression



K-Lasso (Kernel Basis pursuit)

The Kernel Lasso : non-linear regression with sparsity penalization

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

- Typical parametric quadratic program (pQP)
- The ℓ_1 norm $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^n |\alpha_i|$ induces sparsity of the solution $\boldsymbol{\alpha}$ (i.e. some $\alpha_i = 0$)

The dual:

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

- require to compute $K^\top K$!

Obtained regression function

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

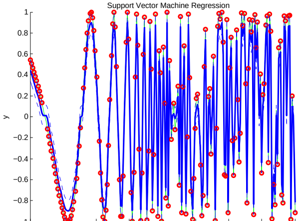
Support vector regression (SVR)

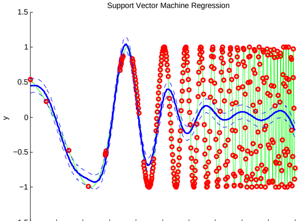
Regression with absolute value error $\begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{s. t.} & |f(\mathbf{x}_i) - y_i| \leq t, \; i = 1, n \end{cases}$

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)





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

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

  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

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

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
  - approximation
  - primal
- Statistics
  - proofs and bounds
  - model selection
    - span bound
    - multikernel: tuning (k and σ)

challenges

- 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)
  - hyper-parameters: 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