

Lecture 5: SVM as a kernel machine

Stéphane Canu
stephane.canu@litislab.eu

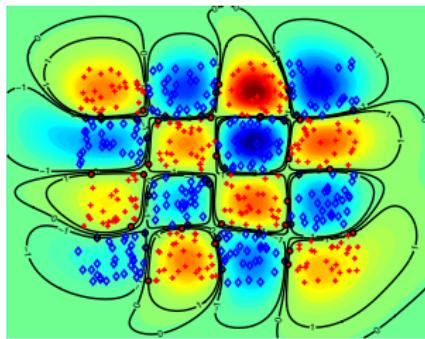
Sao Paulo 2014

April 26, 2014

Plan

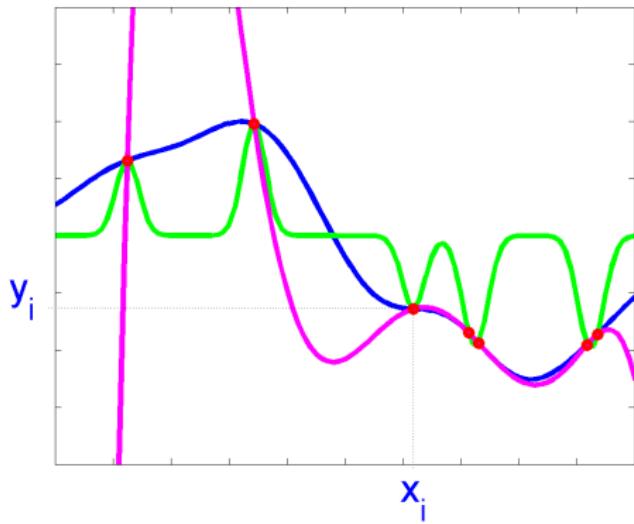
1 Kernel machines

- Non sparse kernel machines
- Sparse kernel machines: SVM
 - SVM: variations on a theme
- Sparse kernel machines for regression: SVR



Interpolation splines

find out $f \in \mathcal{H}$ such that $f(x_i) = y_i, \quad i = 1, \dots, n$



It is an ill posed problem

Interpolation splines: minimum norm interpolation

$$\left\{ \begin{array}{l} \min_{f \in \mathcal{H}} \frac{1}{2} \|f\|_{\mathcal{H}}^2 \\ \text{such that } f(\mathbf{x}_i) = y_i, \quad i = 1, \dots, n \end{array} \right.$$

The lagrangian (α_i Lagrange multipliers)

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

Interpolation splines: minimum norm interpolation

$$\left\{ \begin{array}{l} \min_{f \in \mathcal{H}} \frac{1}{2} \|f\|_{\mathcal{H}}^2 \\ \text{such that } f(\mathbf{x}_i) = y_i, \quad i = 1, \dots, n \end{array} \right.$$

The lagrangian (α_i Lagrange multipliers)

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

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

Interpolation splines: minimum norm interpolation

$$\left\{ \begin{array}{l} \min_{f \in \mathcal{H}} \frac{1}{2} \|f\|_{\mathcal{H}}^2 \\ \text{such that } f(\mathbf{x}_i) = y_i, \quad i = 1, \dots, n \end{array} \right.$$

The lagrangian (α_i Lagrange multipliers)

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

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

dual formulation (remove f from the lagrangian):

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

$$K\alpha = y$$

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:

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

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 $\mathcal{H}_s = \text{span}\{k(., \mathbf{x}_1), \dots, k(., \mathbf{x}_i), \dots, k(., \mathbf{x}_n)\}$
- 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
- ⑤ decompose the global cost

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \underbrace{\|f_{\perp}\|_{\mathcal{H}}^2}_{\geq 0} \geq \|f_s\|_{\mathcal{H}}^2$$

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

⑥

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} = \underset{f \in \mathcal{H}_s}{\operatorname{argmin}}$$

Smoothing splines

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

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

3 equivalent definitions

$$\left\{ \begin{array}{l} (S') \quad \min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n (f(x_i) - y_i)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \\ \text{such that } \sum_{i=1}^n (f(x_i) - y_i)^2 \leq C' \\ (S'') \quad \min_{f \in \mathcal{H}} \sum_{i=1}^n (f(x_i) - y_i)^2 \\ \text{such that } \|f\|_{\mathcal{H}}^2 \leq C'' \end{array} \right.$$

using the representer theorem

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

solution:

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

\neq ridge regression:

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

Kernel logistic regression

inspiration: the Bayes rule

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

probabilities:

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

Rademacher distribution

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

penalized likelihood

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

Kernel logistic regression (2)

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

Representer theorem

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

gradient vector and Hessian matrix:

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

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

solve the problem using Newton iterations

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

Let's summarize

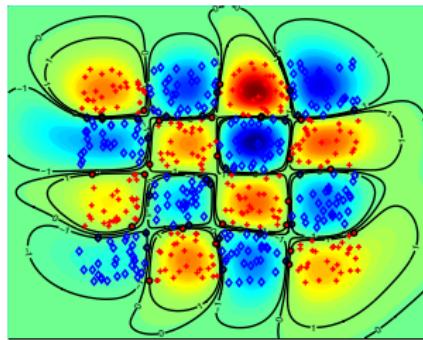
- pros
 - ▶ Universality
 - ▶ from \mathcal{H} to \mathbb{R}^n using the representer theorem
 - ▶ no (explicit) curse of dimensionality
- splines $\mathcal{O}(n^3)$ (can be reduced to $\mathcal{O}(n^2)$)
- logistic regression $\mathcal{O}(kn^3)$ (can be reduced to $\mathcal{O}(kn^2)$)
- no scalability!

sparsity comes to the rescue!

Roadmap

1 Kernel machines

- Non sparse kernel machines
- Sparse kernel machines: SVM
 - SVM: variations on a theme
- Sparse kernel machines for regression: SVR



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

$$\left\{ \begin{array}{l} \max_{f,b} m \\ \text{with } y_i(f(\mathbf{x}_i) + b) \geq m \\ \text{and } \|f\|_{\mathcal{H}}^2 = 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \min_{f,b} \frac{1}{2} \|f\|_{\mathcal{H}}^2 \\ \text{with } y_i(f(\mathbf{x}_i) + b) \geq 1 \end{array} \right.$$

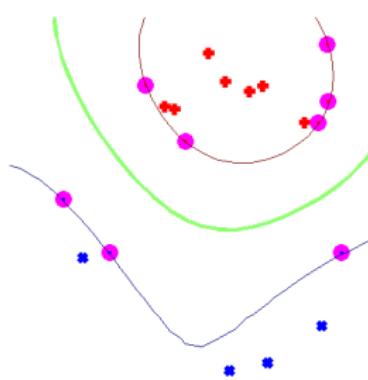
3 ways to represent function f

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

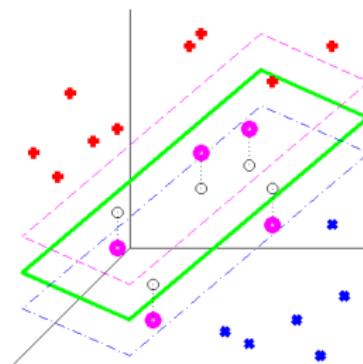
$$\left\{ \begin{array}{l} \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_{\mathbb{R}^d}^2 = \frac{1}{2} \mathbf{w}^\top \mathbf{w} \\ \text{with } y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \min_{\alpha, b} \frac{1}{2} \alpha^\top K \alpha \\ \text{with } y_i(\alpha^\top K(:, i) + b) \geq 1 \end{array} \right.$$

using relevant features...

a data point becomes a function $x \longrightarrow k(x, \cdot)$



input space representation: x



feature space: $k(x, \cdot)$

Representer theorem for SVM

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

Lagrangian

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

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

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

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

Dual formulation for SVM

the intermediate function

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

$$\max_{\alpha} \min_b Q(b, \alpha)$$

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

Dual formulation

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

SVM dual formulation

Dual formulation

$$\left\{ \begin{array}{l} \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n \alpha_i \\ \text{with } \sum_{i=1}^n \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i, \quad i = 1, n \end{array} \right.$$

The dual formulation gives a quadratic program (QP)

$$\left\{ \begin{array}{l} \min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top G \alpha - \mathbf{1}^\top \alpha \\ \text{with } \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha \end{array} \right.$$

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

$$(\mathcal{P}) \left\{ \begin{array}{l} \min_{f \in \mathcal{H}, b, \xi \in \mathbb{R}^n} \frac{1}{2} \|f\|^2 + \frac{C}{p} \sum_{i=1}^n \xi_i^p \\ \text{such that } y_i(f(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, n \end{array} \right.$$

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

$p = 1$, L_1 SVM

$$\left\{ \begin{array}{l} \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2} \alpha^\top G \alpha + \alpha^\top \mathbb{1} \\ \text{such that } \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C \quad i = 1, n \end{array} \right.$$

$p = 2$, L_2 SVM

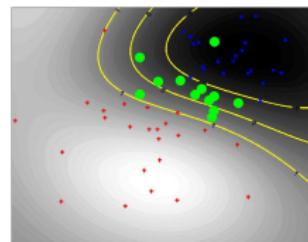
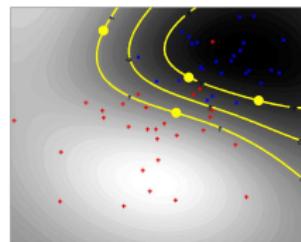
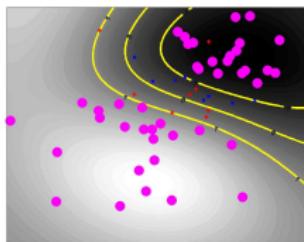
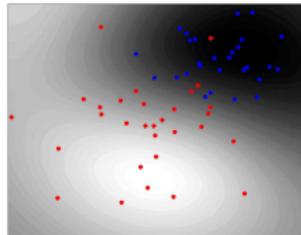
$$\left\{ \begin{array}{l} \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2} \alpha^\top (G + \frac{1}{C} I) \alpha + \alpha^\top \mathbb{1} \\ \text{such that } \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C \quad i = 1, n \end{array} \right.$$

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

Data groups: illustration

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

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



useless data
well classified
 $\alpha = 0$

important data
support
 $0 < \alpha < C$

suspicious data
 $\alpha = C$

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

$$\left\{ \begin{array}{l} \min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top G \alpha - \alpha^\top \mathbb{I} \\ \text{such that} \quad \alpha^\top \mathbf{y} = 0 \quad i = 1, n \\ \text{and} \quad 0 \leq \alpha_i \quad i = 1, n \end{array} \right. \quad \left\{ \begin{array}{l} G\alpha - \mathbb{I} - \beta + b\mathbf{y} = 0 \\ \alpha^\top \mathbf{y} = 0 \\ 0 \leq \alpha_i \quad i = 1, n \\ 0 \leq \beta_i \quad i = 1, n \\ \alpha_i \beta_i = 0 \quad i = 1, n \end{array} \right.$$

$$\begin{array}{|c|c|} \hline G_a & G_i^\top \\ \hline \hline G_i & G_0 \\ \hline \end{array}$$

G

$$\alpha_a - \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ \beta_0 \end{pmatrix} + b \begin{pmatrix} \mathbf{y}_a \\ \mathbf{y}_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\alpha - \mathbb{I} - \beta + b \mathbf{y} = 0$$

$$(1) \quad G_a \alpha_a - \mathbb{I}_a + b\mathbf{y}_a = 0$$

$$(2) \quad G_i \alpha_a - \mathbb{I}_0 - \beta_0 + b\mathbf{y}_0 = 0$$

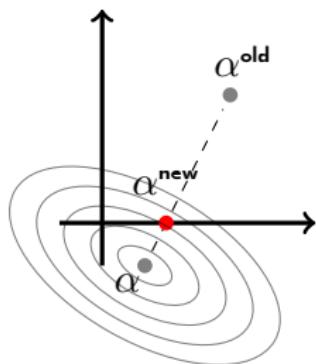
- ➊ solve (1) (find α together with b)
- ➋ if $\alpha < 0$ move it from I_α to I_0
goto 1
- ➌ else solve (2)
if $\beta < 0$ move it from I_0 to I_α
goto 1

The active set method for SVM (2)

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

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

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

Using the hinge loss

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

Minimizing the distance between the convex hulls

$$\left\{ \begin{array}{l} \min_{\alpha} \|u - v\|_{\mathcal{H}}^2 \\ \text{with } u(\mathbf{x}) = \sum_{\{i|y_i=1\}} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \quad v(\mathbf{x}) = \sum_{\{i|y_i=-1\}} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \\ \text{and } \sum_{\{i|y_i=1\}} \alpha_i = 1, \quad \sum_{\{i|y_i=-1\}} \alpha_i = 1, \quad 0 \leq \alpha_i \quad i = 1, n \end{array} \right.$$

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

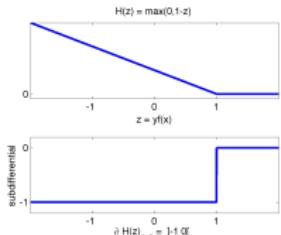
the regularization path: is the set of solutions $\alpha(C)$ when C varies

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. $y_i f(\mathbf{x}_i) = 1$;

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



Regularization path for SVM

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

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

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

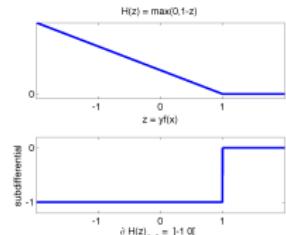
Let λ_n a value close enough to λ_o to keep the sets I_0, I_α and I_C unchanged

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

$$\begin{aligned} \sum_{i \in I_\alpha} \gamma_{io} y_i K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i \in I_\alpha} y_i K(\mathbf{x}_i, \mathbf{x}_j) - \lambda_o y_j \\ \sum_{i \in I_\alpha} \gamma_{in} y_i K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i \in I_\alpha} y_i K(\mathbf{x}_i, \mathbf{x}_j) - \lambda_n y_j \\ G(\gamma_n - \gamma_o) &= (\lambda_o - \lambda_n)\mathbf{y} \quad \text{avec} \quad G_{ij} = y_i K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

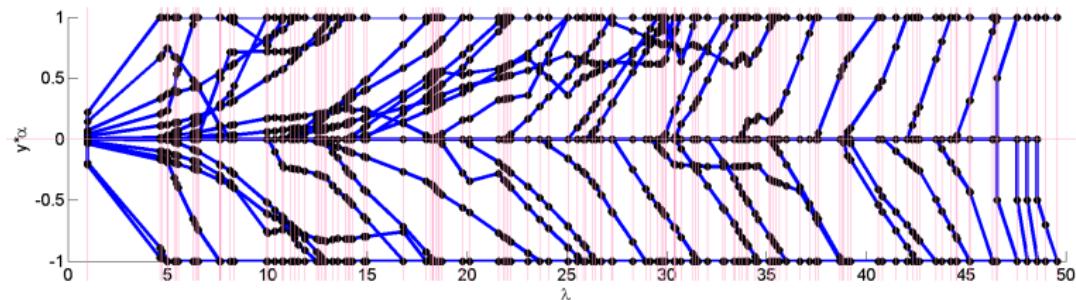
$$\gamma_n = \gamma_o + (\lambda_o - \lambda_n)\mathbf{w}$$

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

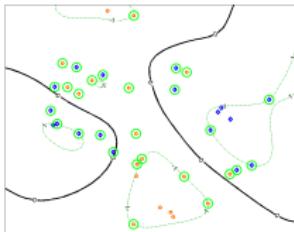
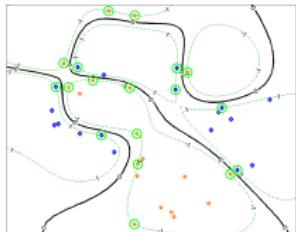


Example of regularization path

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



$\gamma_i = -\frac{1}{C} \alpha_i$; performing together 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]

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

① piecewise linearity: $\lim_{\varepsilon \rightarrow 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = \text{constant}$

② optimality

$$\nabla \ell(\beta(\lambda)) + \lambda \nabla P(\beta(\lambda)) = 0$$

$$\nabla \ell(\beta(\lambda + \varepsilon)) + (\lambda + \varepsilon) \nabla P(\beta(\lambda + \varepsilon)) = 0$$

③ Taylor expansion

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

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

Problems with Piecewise linear regularization path

L	P	<i>regression</i>	<i>classification</i>	<i>clustering</i>
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 Danzig Selector	L1 SVM	

Table : example of piecewise linear regularization path algorithms.

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

ε -insensitive

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

Huber's loss:

$$\begin{cases} |f(\mathbf{x}) - y|^2 & \text{if } |f(\mathbf{x}) - y| < t \\ 2t|f(\mathbf{x}) - y| - t^2 & \text{else} \end{cases}$$

SVM with non symmetric costs

problem in the primal

$$\left\{ \begin{array}{ll} \min_{f \in \mathcal{H}, b, \xi \in \mathbb{R}^n} & \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C^+ \sum_{\{i | y_i=1\}} \xi_i^p + C^- \sum_{\{i | y_i=-1\}} \xi_i^p \\ \text{with} & y_i(f(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, n \end{array} \right.$$

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

$$\left\{ \begin{array}{ll} \max_{\alpha \in \mathbb{R}^n} & -\frac{1}{2} \alpha^\top G \alpha + \alpha^\top \mathbb{1} \\ \text{with} & \alpha^\top \mathbf{y} = 0 \text{ and } 0 \leq \alpha_i \leq C^+ \text{ or } C^- \quad i = 1, n \end{array} \right.$$

ν -SVM and other formulations...

$$\nu \in [0, 1]$$

$$(\nu) \left\{ \begin{array}{ll} \min_{f, b, \xi, m} & \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \frac{1}{np} \sum_{i=1}^n \xi_i^p - \nu m \\ \text{with} & y_i(f(\mathbf{x}_i) + b) \geq m - \xi_i, \quad i = 1, n, \\ \text{and} & m \geq 0, \quad \xi_i \geq 0, \quad i = 1, n, \end{array} \right.$$

for $p = 1$ the dual formulation is:

$$\left\{ \begin{array}{ll} \max_{\alpha \in \mathbb{R}^n} & -\frac{1}{2} \alpha^\top G \alpha \\ \text{with} & \alpha^\top \mathbf{y} = 0 \text{ et } 0 \leq \alpha_i \leq \frac{1}{n} \quad i = 1, n \\ \text{and} & \nu \leq \alpha^\top \mathbb{1} \end{array} \right.$$

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

Generalized SVM

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

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

$$\left\{ \begin{array}{l} \min_{\alpha \in \mathbb{R}^n, b, \xi} \mathbb{I}^\top \beta + C \mathbb{I}^\top \xi \\ \text{with } y_i \left(\sum_{j=1}^n \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq 1 - \xi_i, \\ \text{and } -\beta_i \leq \alpha_i \leq \beta_i, \quad \xi_i \geq 0, \quad i = 1, n \end{array} \right.$$

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

$$\left\{ \begin{array}{l} \max_{\gamma, \delta, \delta^* \in \mathbb{R}^{3n}} \mathbb{I}^\top \gamma \\ \text{with } \mathbf{y}^\top \gamma = 0, \quad \delta_i + \delta_i^* = 1 \\ \quad \sum_{j=1}^n \gamma_j k(\mathbf{x}_i, \mathbf{x}_j) = \delta_i - \delta_i^*, \quad i = 1, n \\ \text{and } 0 \leq \delta_i, 0 \leq \delta_i^*, \quad 0 \leq \gamma_i \leq C, \quad i = 1, n \end{array} \right.$$

K-Lasso (Kernel Basis pursuit)

The Kernel Lasso

$$(S_1) \quad \left\{ \begin{array}{ll} \min_{\alpha \in \mathbf{R}^n} & \frac{1}{2} \|K\alpha - \mathbf{y}\|^2 + \lambda \sum_{i=1}^n |\alpha_i| \end{array} \right.$$

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

The dual:

$$(D_1) \quad \left\{ \begin{array}{ll} \min_{\alpha} & \frac{1}{2} \|K\alpha\|^2 \\ \text{such that} & K^\top(K\alpha - \mathbf{y}) \leq t \end{array} \right.$$

- 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:

$$\left\{ \begin{array}{ll} \min_{\alpha} & \frac{1}{2} \|K\alpha\|^2 \\ \text{s. t.} & K^\top(K\alpha - \mathbf{y}) \leq t \end{array} \right. \quad \left\{ \begin{array}{ll} \min_{f \in \mathcal{H}} & \frac{1}{2} \|f\|_{\mathcal{H}}^2 \\ \text{s. t.} & |f(\mathbf{x}_i) - y_i| \leq t, \quad i = 1, n \end{array} \right.$$

The support vector regression introduce slack variables

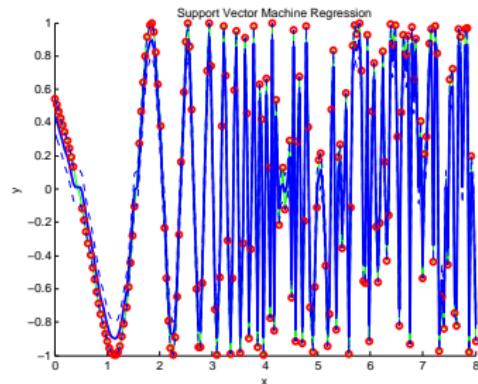
$$(SVR) \quad \left\{ \begin{array}{ll} \min_{f \in \mathcal{H}} & \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum |\xi_i| \\ \text{such that} & |f(\mathbf{x}_i) - y_i| \leq t + \xi_i \quad 0 \leq \xi_i \quad i = 1, n \end{array} \right.$$

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

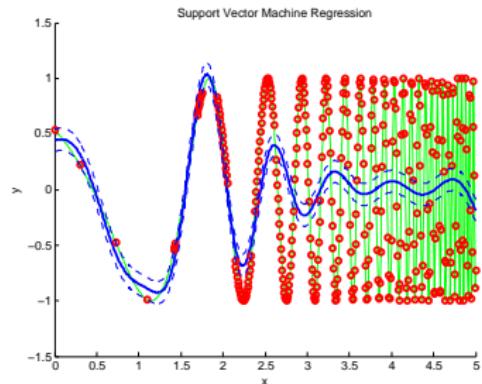
$$\alpha(C, t) = \alpha(C_0, t_0) + \left(\frac{1}{C} - \frac{1}{C_0} \right) \mathbf{u} + \frac{1}{C_0} (t - t_0) \mathbf{v}$$

- 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...

SVM reduction (reduced set method))

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

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

where

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

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)

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
- efficient when uses:
 - ▶ a few import vectors
 - ▶ component-wise update procedure
- extention using L_1 KLR

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 $w^\top \mathbf{x} \rightarrow \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$
- ▶ 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 kernels
- Automatic and adaptive model selection
 - ▶ variable selection
 - ▶ kernel tuning (k et σ)
 - ▶ hyperparametres: C , duality gap, λ
- IP 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