



**HAL**  
open science

# Understanding SVM (and associated kernel machines) through the development of a Matlab toolbox

Stephane Canu

► **To cite this version:**

Stephane Canu. Understanding SVM (and associated kernel machines) through the development of a Matlab toolbox. Engineering school. Introduction to Support Vector Machines (SVM), Sao Paulo, 2014, pp.33. cel-01003007

**HAL Id: cel-01003007**

**<https://cel.hal.science/cel-01003007>**

Submitted on 8 Jun 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Lecture 8: Multi Class SVM

Stéphane Canu  
[stephane.canu@litislab.eu](mailto:stephane.canu@litislab.eu)

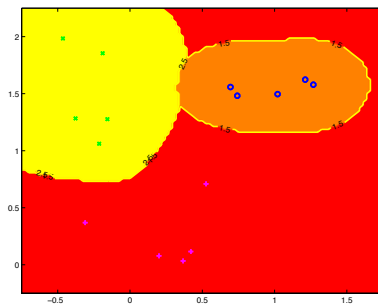
Sao Paulo 2014

April 10, 2014

# Roadmap

## 1 Multi Class SVM

- 3 different strategies for multi class SVM
- Multi Class SVM by decomposition
- Multi class SVM
- Coupling convex hulls



## 3 different strategies for multi class SVM

### 1 Decomposition approaches

- ▶ one vs all: winner takes all
- ▶ one vs one:
  - ★ max-wins voting
  - ★ pairwise coupling: use probability
- ▶  $c$  SVDD

### 2 global approach (size $c \times n$ ),

- ▶ formal (different variations)

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

non consistent estimator but practically useful

- ▶ structured outputs

### 3 A coupling formulation using the convex hulls

## 3 different strategies for multi class SVM

### 1 Decomposition approaches

- ▶ one vs all: winner takes all
- ▶ one vs one:
  - ★ max-wins voting
  - ★ pairwise coupling: use probability – best results
- ▶  $c$  SVDD

### 2 global approach (size $c \times n$ ),

- ▶ formal (different variations)

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

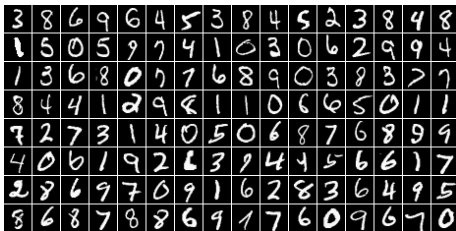
non consistent estimator but practically useful

- ▶ structured outputs

### 3 A coupling formulation using the convex hulls

## Multiclass SVM: complexity issues

- $n$  training data  
 $n = 60,000$  for MNIST
- $c$  class  
 $c = 10$  for MNIST

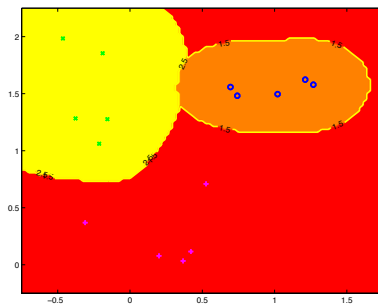


approach	problem size	number of sub problems	discrimination	rejection
<i>1 vs. all</i>	$n$	$c$	++	-
<i>1 vs. 1</i>	$\frac{2n}{c}$	$\frac{c(c-1)}{2}$	++	-
<i>c SVDD</i>	$\frac{n}{c}$	$c$	-	++
<i>all together</i>	$n \times c$	1	++	-
<i>coupling CH</i>	$n$	1	+	+

# Roadmap

## 1 Multi Class SVM

- 3 different strategies for multi class SVM
- **Multi Class SVM by decomposition**
- Multi class SVM
- Coupling convex hulls



# Multi Class SVM by decomposition

## One-Against-All Methods

→ winner-takes-all strategy

## One-vs-One: pairwise methods

→ max-wins voting

→ directed acyclic graph (DAG)

→ error-correcting codes

→ post process probabilities

Hierarchical binary tree for  
multi-class SVM

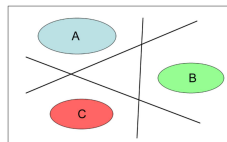


Figure 1: Diagram of binary OAA region boundaries on a basic problem

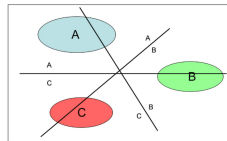
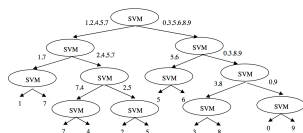


Figure 3: Diagram of pairwise SVM decision boundaries on a basic problem



<http://courses.media.mit.edu/2006fall/mas622j/Projects/aissen-project/>



## SVM and probabilities (Platt, 1999)

The decision function of the SVM is:  $\text{sign}(f(\mathbf{x}) + b)$

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

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

$a_1$  et  $a_2$  estimated using maximum likelihood **on new data**

$\max_{a_1, a_2} L$

with  $L = \prod_{i=1}^n \mathbb{P}(Y = 1|\mathbf{x}_i)^{y_i} + (1 - \mathbb{P}(Y = 1|\mathbf{x}_i))^{(1-y_i)}$

and  $\log L = \sum_{i=1}^n y_i \log(\mathbb{P}(Y = 1|\mathbf{x}_i)) + (1 - y_i) \log(1 - \mathbb{P}(Y = 1|\mathbf{x}_i))$   
 $= \sum_{i=1}^n y_i \log\left(\frac{\mathbb{P}(Y=1|\mathbf{x}_i)}{1 - \mathbb{P}(Y=1|\mathbf{x}_i)}\right) + \log(1 - \mathbb{P}(Y = 1|\mathbf{x}_i))$   
 $= \sum_{i=1}^n y_i (a_1(f(\mathbf{x}_i) + b) + a_2) - \log(1 + \exp^{a_1(f(\mathbf{x}_i)+b)+a_2})$   
 $= \sum_{i=1}^n y_i (\mathbf{a}^\top \mathbf{z}_i) - \log(1 + \exp^{\mathbf{a}^\top \mathbf{z}_i})$

Newton iterations:  $\mathbf{a}^{new} \leftarrow \mathbf{a}^{old} - H^{-1} \nabla \log L$

## SVM and probabilities (Platt, 1999)

$$\max_{\mathbf{a} \in \mathbb{R}^2} \log L = \sum_{i=1}^n y_i (\mathbf{a}^\top \mathbf{z}_i) - \log(1 + \exp^{\mathbf{a}^\top \mathbf{z}_i})$$

Newton iterations

$$\mathbf{a}^{new} \leftarrow \mathbf{a}^{old} - H^{-1} \nabla \log L$$

$$\begin{aligned} \nabla \log L &= \sum_{i=1}^n y_i \mathbf{z}_i - \frac{\exp^{\mathbf{a}^\top \mathbf{z}_i}}{1 + \exp^{\mathbf{a}^\top \mathbf{z}_i}} \mathbf{z}_i \\ &= \sum_{i=1}^n (y_i - \mathbb{P}(Y = 1 | \mathbf{x}_i)) \mathbf{z}_i = Z^\top (\mathbf{y} - \mathbf{p}) \end{aligned}$$

$$H = - \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top \mathbb{P}(Y = 1 | \mathbf{x}_i) (1 - \mathbb{P}(Y = 1 | \mathbf{x}_i)) = -Z^\top WZ$$

Newton iterations

$$\mathbf{a}^{new} \leftarrow \mathbf{a}^{old} + (Z^\top WZ)^{-1} Z^\top (\mathbf{y} - \mathbf{p})$$

## SVM and probabilities: practical issues

$$\mathbf{y} \longrightarrow \mathbf{t} = \begin{cases} 1 - \varepsilon_+ = \frac{n_+ + 1}{n_+ + 2} & \text{if } y_i = 1 \\ \varepsilon_- = \frac{1}{n_- + 2} & \text{if } y_i = -1 \end{cases}$$

- 1 in:  $X, \mathbf{y}, f$  / out:  $\mathbf{p}$
- 2  $\mathbf{t} \leftarrow$
- 3  $Z \leftarrow$
- 4 loop until convergence
  - 1  $\mathbf{p} \leftarrow 1 - \frac{1}{1 + \exp^{\mathbf{a}^\top \mathbf{z}}}$
  - 2  $W \leftarrow \text{diag}(\mathbf{p}(1 - \mathbf{p}))$
  - 3  $\mathbf{a}^{\text{new}} \leftarrow \mathbf{a}^{\text{old}} + (Z^\top W Z)^{-1} Z^\top (\mathbf{t} - \mathbf{p})$

## SVM and probabilities: pairwise coupling

From pairwise probabilities  $\mathbb{P}(c_\ell, c_j)$  to class probabilities  $p_\ell = \mathbb{P}(c_\ell|\mathbf{x})$

$$\min_{\mathbf{p}} \sum_{\ell=1}^c \sum_{j=1}^{\ell-1} \mathbb{P}(c_\ell, c_j)^2 (p_\ell - p_j)^2$$

$$\begin{pmatrix} Q & \mathbf{e} \\ \mathbf{e}^\top & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{with } Q_{\ell j} = \begin{cases} \mathbb{P}(c_\ell, c_j)^2 & \ell \neq j \\ \sum_i \mathbb{P}(c_\ell, c_i)^2 & \ell = j \end{cases}$$

The global procedure :

- 1  $(X_a, y_a, X_t, y_t) \leftarrow \text{split}(X, y)$
- 2  $(X_\ell, y_\ell, X_p, y_p) \leftarrow \text{split}(X_a, y_a)$
- 3 loop for all pairs  $(c_i, c_j)$  of classes
  - 1  $\text{model}_{i,j} \leftarrow \text{train\_SVM}(X_\ell, y_\ell, (c_i, c_j))$
  - 2  $\mathbb{P}(c_i, c_j) \leftarrow \text{estimate\_proba}(X_p, y_p, \text{model})$  % Platt estimate
- 4  $\mathbf{p} \leftarrow \text{post\_process}(X_t, y_t, \mathbb{P})$  % Pairwise Coupling

# SVM and probabilities

## Some facts

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

## 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  $\varepsilon^-$  and  $\varepsilon^+$  have to verify:

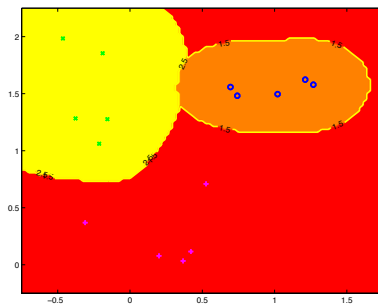
$$\begin{aligned}g(\mathbf{x}) + \varepsilon^+(\mathbf{x}) &= \exp^{-a_1(1-f(\mathbf{x})-\alpha_0)_+ + a_2} \\1 - g(\mathbf{x}) - \varepsilon^-(\mathbf{x}) &= \exp^{-a_1(1+f(\mathbf{x})+\alpha_0)_+ + a_2}\end{aligned}$$

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

# Roadmap

## 1 Multi Class SVM

- 3 different strategies for multi class SVM
- Multi Class SVM by decomposition
- **Multi class SVM**
- Coupling convex hulls



## Multi class SVM: the decision function

One hyperplane by class

$$f_{\ell}(\mathbf{x}) = \mathbf{w}_{\ell}^{\top} \mathbf{x} + b_{\ell} \quad \ell = 1, c$$

Winner takes all decision function

$$D(\mathbf{x}) = \underset{\ell=1,c}{\text{Argmax}} \left( \mathbf{w}_1^{\top} \mathbf{x} + b_1, \mathbf{w}_2^{\top} \mathbf{x} + b_2, \dots, \mathbf{w}_{\ell}^{\top} \mathbf{x} + b_{\ell}, \dots, \mathbf{w}_c^{\top} \mathbf{x} + b_c \right)$$

We can revisit the 2 classes case in this setting

$$c \times (d + 1) \text{ unknown variables } (\mathbf{w}_{\ell}, b_{\ell}); \ell = 1, c$$



# Multi class SVM: the optimization problem

The margin in the multidimensional case

$$m = \min_{\ell \neq y_i} (\mathbf{v}_{y_i}^\top \mathbf{x}_i - a_{y_i} - \mathbf{v}_\ell^\top \mathbf{x}_i + a_\ell) = \mathbf{v}_{y_i}^\top \mathbf{x}_i + a_{y_i} - \max_{\ell \neq y_i} (\mathbf{v}_\ell^\top \mathbf{x}_i + a_\ell)$$

The maximal margin multiclass SVM

$$\left\{ \begin{array}{l} \max_{\mathbf{v}_\ell, a_\ell} \quad m \\ \text{with} \quad \mathbf{v}_{y_i}^\top \mathbf{x}_i + a_{y_i} - \mathbf{v}_\ell^\top \mathbf{x}_i - a_\ell \geq m \quad \text{for } i = 1, n; \quad \ell = 1, c; \quad \ell \neq y_i \\ \text{and} \quad \frac{1}{2} \sum_{\ell=1}^c \|\mathbf{v}_\ell\|^2 = 1 \end{array} \right.$$

The multiclass SVM

$$\left\{ \begin{array}{l} \min_{\mathbf{w}_\ell, b_\ell} \quad \frac{1}{2} \sum_{\ell=1}^c \|\mathbf{w}_\ell\|^2 \\ \text{with} \quad \mathbf{x}_i^\top (\mathbf{w}_{y_i} - \mathbf{w}_\ell) + b_{y_i} - b_\ell \geq 1 \quad \text{for } i = 1, n; \quad \ell = 1, c; \quad \ell \neq y_i \end{array} \right.$$

# Multi class SVM: KKT and dual form: The 3 classes case

$$\left\{ \begin{array}{l} \min_{\mathbf{w}_\ell, b_\ell} \quad \frac{1}{2} \sum_{\ell=1}^3 \|\mathbf{w}_\ell\|^2 \\ \text{with} \quad \mathbf{w}_{y_i}^\top \mathbf{x}_i + b_{y_i} \geq \mathbf{w}_\ell^\top \mathbf{x}_i + b_\ell + 1 \quad \text{for } i = 1, n; \ell = 1, 3; \ell \neq y_i \end{array} \right.$$

$$\left\{ \begin{array}{l} \min_{\mathbf{w}_\ell, b_\ell} \quad \frac{1}{2} \|\mathbf{w}_1\|^2 + \frac{1}{2} \|\mathbf{w}_2\|^2 + \frac{1}{2} \|\mathbf{w}_3\|^2 \\ \text{with} \quad \mathbf{w}_1^\top \mathbf{x}_i + b_1 \geq \mathbf{w}_2^\top \mathbf{x}_i + b_2 + 1 \quad \text{for } i \text{ such that } y_i = 1 \\ \mathbf{w}_1^\top \mathbf{x}_i + b_1 \geq \mathbf{w}_3^\top \mathbf{x}_i + b_3 + 1 \quad \text{for } i \text{ such that } y_i = 1 \\ \mathbf{w}_2^\top \mathbf{x}_i + b_2 \geq \mathbf{w}_1^\top \mathbf{x}_i + b_1 + 1 \quad \text{for } i \text{ such that } y_i = 2 \\ \mathbf{w}_2^\top \mathbf{x}_i + b_2 \geq \mathbf{w}_3^\top \mathbf{x}_i + b_3 + 1 \quad \text{for } i \text{ such that } y_i = 2 \\ \mathbf{w}_3^\top \mathbf{x}_i + b_3 \geq \mathbf{w}_1^\top \mathbf{x}_i + b_1 + 1 \quad \text{for } i \text{ such that } y_i = 3 \\ \mathbf{w}_3^\top \mathbf{x}_i + b_3 \geq \mathbf{w}_2^\top \mathbf{x}_i + b_2 + 1 \quad \text{for } i \text{ such that } y_i = 3 \end{array} \right.$$

$$L = \frac{1}{2} (\|\mathbf{w}_1\|^2 + \|\mathbf{w}_2\|^2 + \|\mathbf{w}_3\|^2) - \alpha_{12}^\top (X_1(\mathbf{w}_1 - \mathbf{w}_2) + b_1 - b_2 - 1) - \alpha_{13}^\top (X_1(\mathbf{w}_1 - \mathbf{w}_3) + b_1 - b_3 - 1) - \alpha_{21}^\top (X_2(\mathbf{w}_2 - \mathbf{w}_1) + b_2 - b_1 - 1) - \alpha_{23}^\top (X_2(\mathbf{w}_2 - \mathbf{w}_3) + b_2 - b_3 - 1) - \alpha_{31}^\top (X_3(\mathbf{w}_3 - \mathbf{w}_1) + b_3 - b_1 - 1) - \alpha_{32}^\top (X_3(\mathbf{w}_3 - \mathbf{w}_2) + b_3 - b_2 - 1)$$

## Multi class SVM: KKT and dual form: The 3 classes case

$$L = \frac{1}{2} \|\mathbf{w}\|^2 - \alpha^\top (\mathcal{X} \mathcal{M} \mathbf{w} + \mathbf{A} \mathbf{b} - \mathbf{1})$$

with

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \end{pmatrix} \in \mathbb{R}^{3d} \quad \mathcal{M} = M \otimes I = \begin{pmatrix} I & -I & 0 \\ I & 0 & -I \\ -I & I & 0 \\ 0 & I & -I \\ -I & 0 & I \\ 0 & -I & I \end{pmatrix} \quad \begin{array}{l} \text{a } 6d \times 3d \text{ matrix} \\ \text{where} \\ I \text{ the identity matrix} \end{array}$$

and

$$\mathcal{X} = \begin{pmatrix} X_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & X_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & X_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & X_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & X_3 \end{pmatrix} \quad \begin{array}{l} \text{a } 2n \times 6d \text{ matrix} \\ \text{with input data} \\ X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} n \times d \end{array}$$

## Multi class SVM: KKT and dual form: The 3 classes case

KKT Stationality conditions =

$$\begin{aligned}\nabla_{\mathbf{w}}L &= \mathbf{w} - \mathcal{M}^T \mathcal{X}^T \alpha \\ \nabla_{\mathbf{b}}L &= \mathbf{A}^T \alpha\end{aligned}$$

The dual

$$\begin{aligned}\min_{\alpha \in \mathbb{R}^{2n}} \quad & \frac{1}{2} \alpha^T G \alpha - \mathbf{e}^T \alpha \\ \text{with} \quad & \mathbf{A} \mathbf{b} = 0 \\ \text{and} \quad & 0 \leq \alpha\end{aligned}$$

With

$$\begin{aligned}G &= \mathcal{X} \mathcal{M} \mathcal{M}^T \mathcal{X}^T \\ &= \mathcal{X} (M \otimes I) (M \otimes I)^T \mathcal{X}^T \\ &= \mathcal{X} (M M^T \otimes I) \mathcal{X}^T \\ &= (M M^T \otimes I) \cdot \mathcal{X} \mathcal{X}^T \\ &= (M M^T \otimes I) \cdot \mathbb{I} K \mathbb{I}^T\end{aligned} \quad \text{and} \quad M = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

## Multi class SVM and slack variables (2 variants)

- A slack for all (Vapnik & Blanz, Weston & Watkins 1998)

$$\left\{ \begin{array}{l} \min_{\mathbf{w}_\ell, b_\ell, \xi \in \mathbb{R}^{cn}} \quad \frac{1}{2} \sum_{\ell=1}^c \|\mathbf{w}_\ell\|^2 + C \sum_{i=1}^n \sum_{\ell=1, \ell \neq y_i}^c \xi_{i\ell} \\ \text{with} \quad \mathbf{w}_{y_i}^\top \mathbf{x}_i + b_{y_i} - \mathbf{w}_\ell^\top \mathbf{x}_i - b_\ell \geq 1 - \xi_{i\ell} \\ \text{and} \quad \xi_{i\ell} \geq 0 \quad \text{for } i = 1, n; \ell = 1, c; \ell \neq y_i \end{array} \right.$$

The dual

$$\begin{array}{l} \min_{\alpha \in \mathbb{R}^{2n}} \quad \frac{1}{2} \alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} \quad \mathbf{A} \mathbf{b} = 0 \\ \text{and} \quad 0 \leq \alpha \leq C \end{array}$$

- Max error, a slack per training data (Cramer and Singer, 2001)

$$\left\{ \begin{array}{l} \min_{\mathbf{w}_\ell, b_\ell, \xi \in \mathbb{R}^n} \quad \frac{1}{2} \sum_{\ell=1}^c \|\mathbf{w}_\ell\|^2 + C \sum_{i=1}^n \xi_i \\ \text{with} \quad (\mathbf{w}_{y_i} - \mathbf{w}_\ell)^\top \mathbf{x}_i \geq 1 - \xi_i \quad \text{for } i = 1, n; \ell = 1, c; \ell \neq y_i \\ \text{and} \quad \xi_i \geq 0 \quad \text{for } i = 1, n \end{array} \right.$$

## Multi class SVM and Kernels

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

The dual

$$\begin{array}{l} \min_{\alpha \in \mathbb{R}^{2n}} \quad \frac{1}{2} \alpha^{\top} G \alpha - \mathbf{e}^{\top} \alpha \\ \text{with} \quad \mathbf{A} \mathbf{b} = 0 \\ \text{and} \quad 0 \leq \alpha \leq C \end{array}$$

where  $G$  is the **multi class** kernel matrix

## Other Multi class SVM

Lee, Lin & Wahba, 2004

$$\left\{ \begin{array}{l} \min_{f \in \mathcal{H}} \quad \frac{\lambda}{2} \sum_{\ell=1}^c \|f_{\ell}\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n \sum_{\ell=1, \ell \neq y_i}^c (f_{\ell}(\mathbf{x}_i) + \frac{1}{c-1})_+ \\ \text{with} \quad \sum_{\ell=1}^c f_{\ell}(\mathbf{x}) = 0 \quad \forall \mathbf{x} \end{array} \right.$$

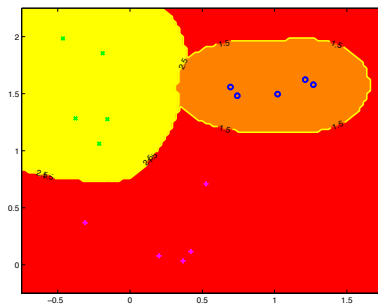
Structured outputs = Cramer and Singer, 2001

MSVMpack : A Multi-Class Support Vector Machine Package Fabien Lauer & Yann Guermeur

# Roadmap

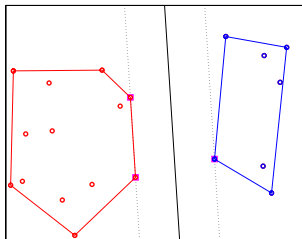
## 1 Multi Class SVM

- 3 different strategies for multi class SVM
- Multi Class SVM by decomposition
- Multi class SVM
- Coupling convex hulls





## One more way to derivate SVM



Minimizing the distance between the convex hulls

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

## The multi class case

$$\left\{ \begin{array}{l} \min_{\alpha} \sum_{\ell=1}^c \sum_{\ell'=1}^c \|u_{\ell} - u_{\ell'}\|^2 \\ \text{with } u_{\ell}(\mathbf{x}) = \sum_{\{i|y_i=\ell\}} \alpha_{i,\ell}(\mathbf{x}_i^{\top} \mathbf{x}), \quad \ell = 1, c \\ \text{and } \sum_{\{i|y_i=\ell\}} \alpha_{i,\ell} = 1, \quad 0 \leq \alpha_{i,\ell} \quad i = 1, n; \ell = 1, c \end{array} \right.$$

# Bibliography

- Estimating probabilities
  - ▶ Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Advances in large margin classifiers. MIT Press.
  - ▶ T. Lin, C.-J. Lin, R.C. Weng, A note on Platt's probabilistic outputs for support vector machines, Mach. Learn. 68 (2007) 267–276
  - ▶ <http://www.cs.cornell.edu/courses/cs678/2007sp/platt.pdf>
- Multiclass SVM
  - ▶ K.-B. Duan & S. Keerthi (2005). "Which Is the Best Multiclass SVM Method? An Empirical Study".
  - ▶ T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, JMLR. 5 (2004) 975–1005.
  - ▶ K. Crammer & Y. Singer (2001). "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines". JMLR 2: 265–292.
  - ▶ Lee, Y.; Lin, Y.; and Wahba, G. (2001). "Multicategory Support Vector Machines". Computing Science and Statistics 33.
  - ▶ [http://www.loria.fr/~guermeur/NN2008\\_M\\_SVM\\_YG.pdf](http://www.loria.fr/~guermeur/NN2008_M_SVM_YG.pdf)
  - ▶ <http://jmlr.org/papers/volume12/lauer11a/lauer11a.pdf>