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

Stephane Canu

Lecture 8: Multi Class SVM

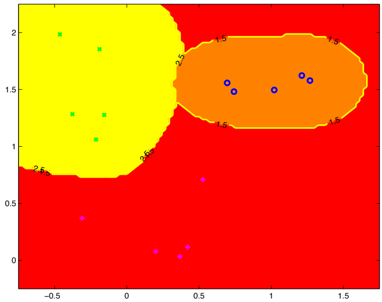
Stéphane Canu
stephane.canu@litislab.eu

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)

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

     non consistent estimator but practically useful
   - structured outputs
3. A coupling formulation using the convex hulls

3 different strategies for multi class SVM

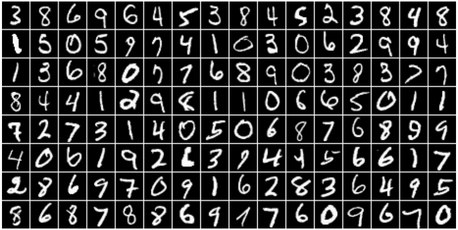
1. Decomposition approaches
   - one *vs* all: winner takes all
   - one *vs* one:
     - ★ max-wins voting
     - ★ pairwise coupling: use probability – best results
   - c SVDD
2. global approach (size $c \times n$),
   - formal (different variations)

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

   non consistent estimator but practically useful
   - structured outputs
3. A coupling formulation using the convex hulls

Multiclass SVM: complexity issues

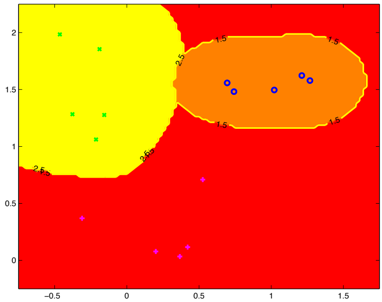
- n training data
  $n = 60,000$ for MNIST
- c class
  $c = 10$ for MNIST



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

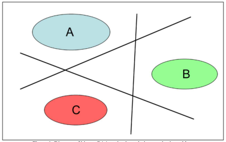


Figure 1: Diagram of binary OAA region boundaries on a basic problem

One-vs-One: pairwise methods
→ max-wins voting
→ directed acyclic graph (DAG)
→ error-correcting codes
→ post process probabilities

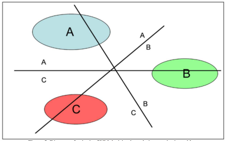
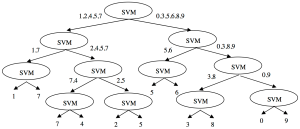


Figure 5: Diagram of pairwise SVM decision boundaries on a basic problem

Hierarchical binary tree for multi-class SVM



http://courses.media.mit.edu/2006fall/mas622j/Projects/aisen-project/

SVM and probabilities (Platt, 1999)

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

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

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

a_1 et a_2 estimated using maximum likelihood on new data

$$\max_{a_1,a_2} L$$

$$\begin{aligned}
\text{with} \quad L &= \prod_{i=1}^{n} \mathbb{P}(Y=1|\mathbf{x}_i)^{y_i} + (1-\mathbb{P}(Y=1|\mathbf{x}_i))^{(1-y_i)} \\
\text{and} \quad \log L &= \textstyle\sum_{i=1}^{n} \; y_i \log(\mathbb{P}(Y=1|\mathbf{x}_i)) + (1-y_i) log(1-\mathbb{P}(Y=1|\mathbf{x}_i)) \\
&= \textstyle\sum_{i=1}^{n} \; y_i \log\big(\frac{\mathbb{P}(Y=1|\mathbf{x}_i)}{1-\mathbb{P}(Y=1|\mathbf{x}_i)}\big) + \log(1-\mathbb{P}(Y=1|\mathbf{x}_i)) \\
&= \textstyle\sum_{i=1}^{n} \; y_i\big(a_1(f(\mathbf{x}_i)+b)+a_2\big) - \log(1+\exp^{a_1(f(\mathbf{x}_i)+b)+a_2}) \\
&= \textstyle\sum_{i=1}^{n} \; y_i\big(\mathbf{a}^\top \mathbf{z}_i\big) - \log(1+\exp^{\mathbf{a}^\top \mathbf{z}_i})
\end{aligned}$$

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

SVM and probabilities (Platt, 1999)

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

Newton iterations

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

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

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

Newton iterations

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

SVM and probabilities: practical issues

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

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. $W \leftarrow diag\big(\mathbf{p}(1 - \mathbf{p})\big)$
    3. $\mathbf{a}^{new} \leftarrow \mathbf{a}^{old} + (Z^\top W Z)^{-1} Z^\top (\mathbf{t} - \mathbf{p})$

SVM and probabilities: pairwise coupling

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

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

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

The global procedure :

1. $(Xa, ya, Xt, yt) \leftarrow split(X, y)$
2. $(X\ell, y\ell, Xp, yp) \leftarrow split(Xa, ya)$
3. loop for all pairs (c_i, c_j) of classes
    1. $model_{i,j} \leftarrow train\_SVM(X\ell, y\ell, (c_i, c_j))$
    2. $\mathbb{P}(c_i, c_j) \leftarrow estimate\_proba(Xp, yp, model)$ % Platt estimate
4. $\mathbf{p} \leftarrow post\_process(Xt, yt, \mathbb{P})$ % Pairwise Coupling

Wu, Lin & Weng, 2004, Duan & Keerti, 05

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)

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:

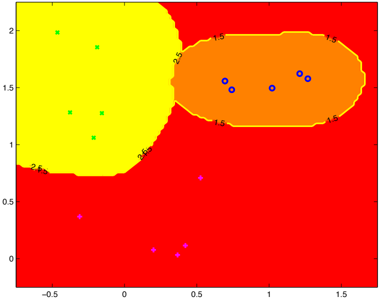
$$\begin{aligned} g(\mathbf{x}) + \varepsilon^+(\mathbf{x}) &= \exp^{-a_1(1-f(\mathbf{x})-\alpha_0)_+ + a_2} \\ 1 - g(\mathbf{x}) - \varepsilon^-(\mathbf{x}) &= \exp^{-a_1(1+f(\mathbf{x})+\alpha_0)_+ + a_2} \end{aligned}$$

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

Grandvalet *et al.*, 07

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

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

Winner takes all decision function

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

We can revisit the 2 classes case in this setting

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

Multi class SVM: the optimization problem

The margin in the multidimensional case

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

The maximal margin multiclass SVM

$$\begin{cases} \max\limits_{\mathbf{v}_\ell, a_\ell} & m \\ \text{with} & \mathbf{v}_{y_i}^\top \mathbf{x}_i + a_{y_i} - \mathbf{v}_\ell^\top \mathbf{x}_i - a_\ell \geq m \qquad \text{for } i = 1, n; \;\; \ell = 1, c; \;\; \ell \neq y_i \\ \text{and} & \frac{1}{2} \sum\limits_{\ell=1}^{c} \|\mathbf{v}_\ell\|^2 = 1 \end{cases}$$

The multiclass SVM

$$\begin{cases} \min\limits_{\mathbf{w}_\ell, b_\ell} & \frac{1}{2} \sum\limits_{\ell=1}^{c} \|\mathbf{w}_\ell\|^2 \\ \text{with} & \mathbf{x}_i^\top (\mathbf{w}_{y_i} - \mathbf{w}_\ell) + b_{y_i} - b_\ell \geq 1 \qquad \text{for } i = 1, n; \;\; \ell = 1, c; \;\; \ell \neq y_i \end{cases}$$

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

$$\begin{cases} \min\limits_{\mathbf{w}_\ell, b_\ell} & \frac{1}{2}\sum\limits_{\ell=1}^{3} \|\mathbf{w}_\ell\|^2 \\ \text{with} & \mathbf{w}_{y_i}^\top \mathbf{x}_i + b_{y_i} \geq \mathbf{w}_\ell^\top \mathbf{x}_i + b_\ell + 1 \qquad \text{for } i = 1, n; \;\; \ell = 1, 3; \;\; \ell \neq y_i \end{cases}$$

$$\begin{cases} \min\limits_{\mathbf{w}_\ell, b_\ell} & \frac{1}{2}\|\mathbf{w}_1\|^2 + \frac{1}{2}\|\mathbf{w}_2\|^2 + \frac{1}{2}\|\mathbf{w}_3\|^2 & \\ \text{with} & \mathbf{w}_1^\top \mathbf{x}_i + b_1 \geq \mathbf{w}_2^\top \mathbf{x}_i + b_2 + 1 & \text{for } i \text{ such that } y_i = 1 \\ & \mathbf{w}_1^\top \mathbf{x}_i + b_1 \geq \mathbf{w}_3^\top \mathbf{x}_i + b_3 + 1 & \text{for } i \text{ such that } y_i = 1 \\ & \mathbf{w}_2^\top \mathbf{x}_i + b_2 \geq \mathbf{w}_1^\top \mathbf{x}_i + b_1 + 1 & \text{for } i \text{ such that } y_i = 2 \\ & \mathbf{w}_2^\top \mathbf{x}_i + b_2 \geq \mathbf{w}_3^\top \mathbf{x}_i + b_3 + 1 & \text{for } i \text{ such that } y_i = 2 \\ & \mathbf{w}_3^\top \mathbf{x}_i + b_3 \geq \mathbf{w}_1^\top \mathbf{x}_i + b_1 + 1 & \text{for } i \text{ such that } y_i = 3 \\ & \mathbf{w}_3^\top \mathbf{x}_i + b_3 \geq \mathbf{w}_2^\top \mathbf{x}_i + b_2 + 1 & \text{for } i \text{ such that } y_i = 3 \end{cases}$$

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

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

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

with

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \end{pmatrix} \in \mathbb{R}^{3d} \qquad \mathcal{M} = M \otimes I = \begin{pmatrix} I & -I & 0 \\ I & 0 & -I \\ -I & I & 0 \\ 0 & I & -I \\ -I & 0 & I \\ 0 & -I & I \end{pmatrix}$$

a $6d \times 3d$ matrix
where
I the identity matrix

and

$$\mathcal{X} = \begin{pmatrix} X_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & X_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & X_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & X_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & X_3 \end{pmatrix}$$

a $2n \times 6d$ matrix
with input data

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \; n \times d$$

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

KKT Stationality conditions =

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

The dual

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

With

$$\begin{aligned}
G &= \mathcal{X}\mathcal{M}\mathcal{M}^\top \mathcal{X}^\top \\
&= \mathcal{X}(M \otimes I)(M \otimes I)^\top \mathcal{X}^\top \\
&= \mathcal{X}(MM^\top \otimes I)\mathcal{X}^\top \\
&= (MM^\top \otimes I) .\times XX^\top \\
&= (MM^\top \otimes I) .\times \mathbb{I}\, K\, \mathbb{I}^\top
\end{aligned}
\qquad \text{and} \qquad
M = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

Multi class SVM and slack variables (2 variants)

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

$$\begin{cases} \min\limits_{\mathbf{w}_\ell, b_\ell, \xi \in \mathbb{R}^{cn}} & \frac{1}{2}\sum\limits_{\ell=1}^{c} \|\mathbf{w}_\ell\|^2 + C \sum\limits_{i=1}^{n} \sum\limits_{\ell=1, \ell \neq y_i}^{c} \xi_{i\ell} \\ \text{with} & \mathbf{w}_{y_i}^\top \mathbf{x}_i + b_{y_i} - \mathbf{w}_\ell^\top \mathbf{x}_i - b_\ell \geq 1 - \xi_{i\ell} \\ \text{and} & \xi_{i\ell} \geq 0 \qquad \text{for } i = 1, n; \;\; \ell = 1, c; \;\; \ell \neq y_i \end{cases}$$

The dual

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

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

$$\begin{cases} \min\limits_{\mathbf{w}_\ell, b_\ell, \xi \in \mathbb{R}^{n}} & \frac{1}{2}\sum\limits_{\ell=1}^{c} \|\mathbf{w}_\ell\|^2 + C \sum\limits_{i=1}^{n} \xi_i \\ \text{with} & (\mathbf{w}_{y_i} - \mathbf{w}_\ell)^\top \mathbf{x}_i \geq 1 - \xi_i \qquad \text{for } i = 1, n; \;\; \ell = 1, c; \;\; \ell \neq y_i \\ \text{and} & \xi_i \geq 0 \qquad \text{for } i = 1, n \end{cases}$$

Multi class SVM and Kernels

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

The dual

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

where G is the multi class kernel matrix

Other Multi class SVM

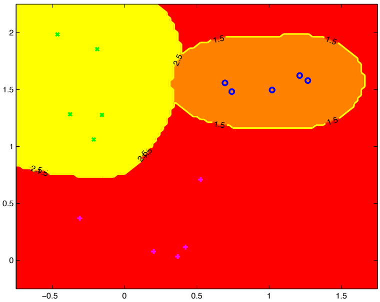
Lee, Lin & Wahba, 2004

$$\begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{\lambda}{2} \sum\limits_{\ell=1}^{c} \|f_\ell\|_{\mathcal{H}}^2 + \frac{1}{n} \sum\limits_{i=1}^{n} \sum\limits_{\ell=1, \ell \neq y_i}^{c} (f_\ell(\mathbf{x}_i) + \frac{1}{c-1})_+ \\ \text{with} & \sum\limits_{\ell=1}^{c} f_\ell(\mathbf{x}) = 0 \qquad \forall \mathbf{x} \end{cases}$$

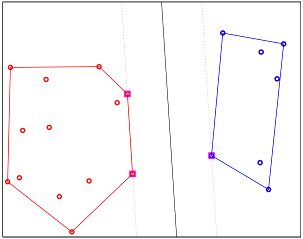
Structured outputs = Cramer and Singer, 2001
MSVMpack : A Multi-Class Support Vector Machine Package Fabien Lauer & Yann Guermeur

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

$$\left\{\begin{array}{ll} \min\limits_{\alpha} & \|u-v\|^2 \\ \text{with} & u(\mathbf{x}) = \sum\limits_{\{i|y_i=1\}} \alpha_i(\mathbf{x}_i^\top \mathbf{x}), \qquad v(\mathbf{x}) = \sum\limits_{\{i|y_i=-1\}} \alpha_i(\mathbf{x}_i^\top \mathbf{x}) \\ \text{and} & \sum\limits_{\{i|y_i=1\}} \alpha_i = 1, \sum\limits_{\{i|y_i=-1\}} \alpha_i = 1, \quad 0 \le \alpha_i \quad i = 1, n \end{array}\right.$$

The multi class case

$$
\begin{cases}
\min_{\alpha} & \sum_{\ell=1}^{c} \sum_{\ell'=1}^{c} \|u_\ell - u_{\ell'}\|^2 \\
\text{with} & u_\ell(\mathbf{x}) = \sum_{\{i | y_i = \ell\}} \alpha_{i,\ell} (\mathbf{x}_i^\top \mathbf{x}), \qquad \ell = 1, c \\
\text{and} & \sum_{\{i | y_i = \ell\}} \alpha_{i,\ell} = 1, \quad 0 \le \alpha_{i,\ell} \quad i = 1, n; \ell = 1, c
\end{cases}
$$

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://jmlr.org/papers/volume12/lauer11a/lauer11a.pdf`