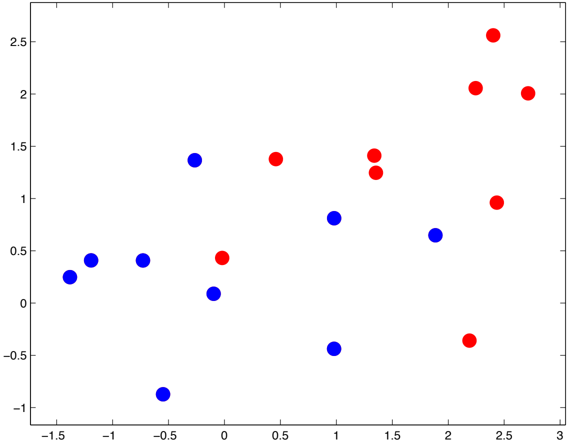
Lecture 3: Linear SVM with slack variables

Stéphane Canu
stephane.canu@litislab.eu

Sao Paulo 2014

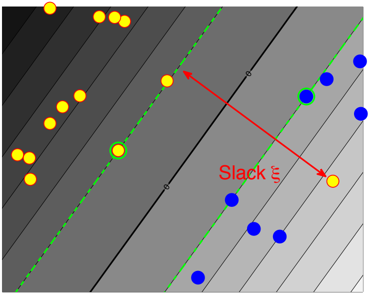
March 23, 2014

The non separable case



Road map

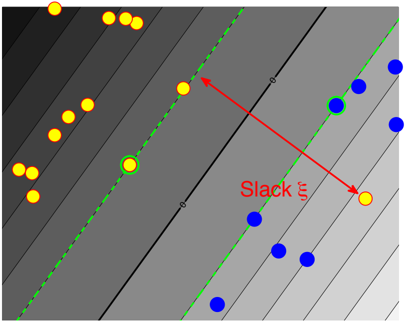
1. Linear SVM
   - The non separable case
   - The C (L1) SVM
   - The L2 SVM and others "variations on a theme"
   - The hinge loss



The non separable case: a bi criteria optimization problem

Modeling potential errors: introducing slack variables ξ_i

$$(x_i, y_i) \qquad \begin{cases} \text{no error:} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \Rightarrow & \xi_i = 0 \\ \text{error:} & & \xi_i = 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \end{cases}$$



$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \min\limits_{\mathbf{w},b,\xi} & \frac{C}{p}\sum_{i=1}^{n} \xi_i^p \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

Our hope: almost all $\xi_i = 0$

Bi criteria optimization and dominance

$$\begin{cases} L(\mathbf{w}) &= \dfrac{1}{p}\displaystyle\sum_{i=1}^{n} \xi_i^p \\ P(\mathbf{w}) &= \|\mathbf{w}\|^2 \end{cases}$$

Dominance

$\mathbf{w}_1$ dominates $\mathbf{w}_2$

if $L(\mathbf{w}_1) \leq L(\mathbf{w}_2)$ and $P(\mathbf{w}_1) \leq P(\mathbf{w}_2)$

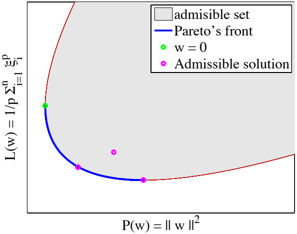


Figure: dominated point (red), non dominated point (purple) and Pareto front (blue).

Pareto front (or Pareto Efficient Frontier)

it is the set of all nondominated solutions

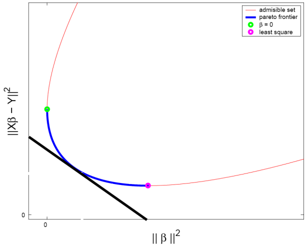
Pareto frontier

$\Leftrightarrow$

Regularization path

3 equivalent formulations to reach Pareto's front

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; \frac{1}{p}\sum_{i=1}^{n}\xi_i^p + \lambda \|\mathbf{w}\|^2$$

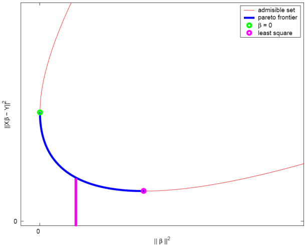


it works for CONVEX criteria!

3 equivalent formulations to reach Pareto's front

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; \frac{1}{p}\sum_{i=1}^{n} \xi_i^p + \lambda \, \|\mathbf{w}\|^2$$

$$\begin{cases} \displaystyle\min_{\mathbf{w}} \; \frac{1}{p}\sum_{i=1}^{n} \xi_i^p \\ \text{with } \; \|\mathbf{w}\|^2 \leq k \end{cases}$$



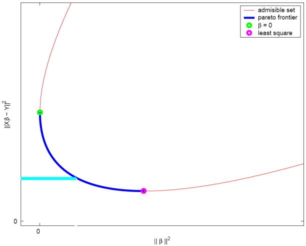
it works for CONVEX criteria!

3 equivalent formulations to reach Pareto's front

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; \frac{1}{p}\sum_{i=1}^{n} \xi_i^p + \lambda \, \|\mathbf{w}\|^2$$

$$\left\{ \begin{array}{l} \displaystyle\min_{\mathbf{w}} \; \frac{1}{p}\sum_{i=1}^{n} \xi_i^p \\ \text{with} \;\; \|\mathbf{w}\|^2 \leq k \end{array} \right.$$

$$\left\{ \begin{array}{l} \displaystyle\min_{\mathbf{w}} \; \|\mathbf{w}\|^2 \\ \text{with} \;\; \displaystyle\frac{1}{p}\sum_{i=1}^{n} \xi_i^p \leq k' \end{array} \right.$$



it works for CONVEX criteria!

The non separable case

Modeling potential errors: introducing slack variables ξ_i

$$(x_i, y_i) \qquad \begin{cases} \text{no error:} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \Rightarrow \quad \xi_i = 0 \\ \text{error:} & \qquad\qquad\qquad\qquad\quad \xi_i = 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \end{cases}$$

Minimizing also the slack (the error), for a given $C > 0$

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i^p \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \quad & i = 1, n \\ & \xi_i \geq 0 & i = 1, n \end{cases}$$

Looking for the saddle point of the lagrangian with the Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i^p - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n} \beta_i \xi_i$$

The KKT$(p = 1)$

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum_{i=1}^{n}\xi_i^p - \sum_{i=1}^{n}\alpha_i\big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\beta_i\xi_i$$

stationarity	$\mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i = 0$ and $\sum_{i=1}^{n}\alpha_i\ y_i = 0$	
	$C - \alpha_i - \beta_i = 0$	$i = 1, \dots, n$
primal admissibility	$y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1$	$i = 1, \dots, n$
	$\xi_i \geq 0$	$i = 1, \dots, n$
dual admissibility	$\alpha_i \geq 0$	$i = 1, \dots, n$
	$\beta_i \geq 0$	$i = 1, \dots, n$
complementarity	$\alpha_i\Big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\Big) = 0$	$i = 1, \dots, n$
	$\beta_i\xi_i = 0$	$i = 1, \dots, n$

Let's eliminate β!

KKT ($p = 1$)

stationarity $\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$ and $\sum_{i=1}^{n} \alpha_i \; y_i = 0$

primal admissibility $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$ $\quad i = 1, \dots, n$

$\xi_i \geq 0$ $\quad i = 1, \dots, n;$

dual admissibility $\alpha_i \geq 0$ $\quad i = 1, \dots, n$

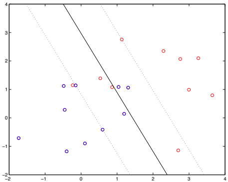
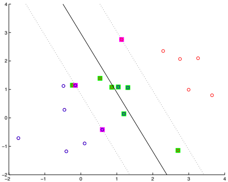
$C - \alpha_i \geq 0$ $\quad i = 1, \dots, n;$

complementarity $\alpha_i \Big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i \Big) = 0 \quad i = 1, \dots, n$

$(C - \alpha_i)\, \xi_i = 0 \quad i = 1, \dots, n$

sets	I_0	$I_{\mathcal{A}}$	I_C
α_i	0	$0 < \alpha < C$	C
β_i	C	$C - \alpha$	0
ξ_i	0	0	$1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)$
	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1$	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1$
	useless	usefull (support vec)	suspicious

The importance of being support

  .

data point	α	constraint value	set
$\mathbf{x}_i$ *useless*	$\alpha_i = 0$	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1$	I_0
$\mathbf{x}_i$ *support*	$0 < \alpha_i < C$	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$	I_α
$\mathbf{x}_i$ *suspicious*	$\alpha_i = C$	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1$	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: $\alpha_i = 0$

Optimality conditions ($p = 1$)

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\beta_i\xi_i$$

Computing the gradients: $\begin{cases} \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \displaystyle\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial\mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \displaystyle\sum_{i=1}^{n}\alpha_i\; y_i \\ \nabla_{\xi_i}\mathcal{L}(\mathbf{w}, b, \alpha) & = C - \alpha_i - \beta_i \end{cases}$

- no change for $\mathbf{w}$ and b
- $\beta_i \geq 0$ and $C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i \leq C$

The dual formulation:

$$\begin{cases} \displaystyle\min_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \leq C \qquad i = 1, n \end{cases}$$

SVM primal vs. dual

Primal

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

- $d + n + 1$ unknown
- $2n$ constraints
- classical QP
- to be used when n is too large to build G

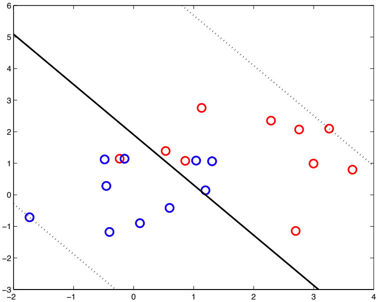
Dual

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

- n unknown
- G Gram matrix (pairwise influence matrix)
- $2n$ box constraints
- easy to solve
- to be used when n is not too large

The smallest C

C small $\Rightarrow$ all the points are in I_C: $\alpha_i = C$



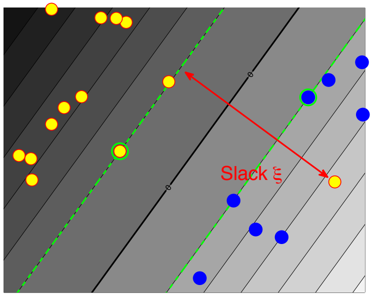
$$-1 \leq f_j \ = \ C \ \sum_{i=1}^{n} y_i(\mathbf{x}_i^\top \mathbf{x}_j) + b \leq 1$$

$$f_M = max(f) \qquad f_m = min(f)$$

$$C_{max} = \frac{2}{f_M - f_m}$$

Road map

1. Linear SVM
   - The non separable case
   - The C (L1) SVM
   - The L2 SVM and others "variations on a theme"
   - The hinge loss



L2 SVM: optimality conditions ($p = 2$)

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big)$$

Computing the gradients: $\begin{cases} \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \displaystyle\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial\mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \displaystyle\sum_{i=1}^{n}\alpha_i\ y_i \\ \nabla_{\xi_i}\mathcal{L}(\mathbf{w}, b, \alpha) & = C\xi_i - \alpha_i \end{cases}$

- no need of the positivity constraint on ξ_i
- no change for $\mathbf{w}$ and b
- $C\xi_i - \alpha_i = 0 \ \Rightarrow\ \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\xi_i = -\frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2$

The dual formulation:

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top(G + \frac{1}{C}I)\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \le \alpha_i \qquad\qquad i = 1, n \end{cases}$$

SVM primal vs. dual

Primal

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

- $d + n + 1$ unknown
- n constraints
- classical QP
- to be used when n is too large to build G

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top(G + \frac{1}{C}I)\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix is regularized
- n box constraints
- easy to solve
- to be used when n is not too large

One more variant: the ν SVM

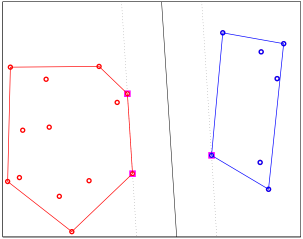
$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{with} & \min\limits_{i=1,n} |\mathbf{v}^\top \mathbf{x}_i + a| \geq m \\ & \|\mathbf{v}\|^2 = k \end{cases}$$

$$\begin{cases} \min\limits_{\mathbf{v},a} & \frac{1}{2}\|\mathbf{v}\|^2 - \nu\ m + \sum_{i=1}^n \xi_i \\ \text{with} & y_i(\mathbf{v}^\top \mathbf{x}_i + a) \geq m - \xi_i \\ & \xi_i \geq 0,\ \ m \geq 0 \end{cases}$$

The dual formulation:

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

The convex hull formulation



Minimizing the distance between the convex hulls

$$\begin{cases} \min\limits_{\alpha} & \|u - v\| \\ \text{with} & u = \sum\limits_{\{i|y_i=1\}} \alpha_i \mathbf{x}_i, \qquad v = \sum\limits_{\{i|y_i=-1\}} \alpha_i \mathbf{x}_i \\ \text{and} & \sum\limits_{\{i|y_i=1\}} \alpha_i = 1, \sum\limits_{\{i|y_i=-1\}} \alpha_i = 1, \quad 0 \le \alpha_i \le C \quad i = 1, n \end{cases}$$

$$\mathbf{w}^\top \mathbf{x} = \frac{2}{\|u - v\|}\left(u^\top \mathbf{x} - v^\top \mathbf{x}\right) \text{ and } b = \frac{\|u\| - \|v\|}{\|u - v\|}$$

SVM with non symetric costs

Problem in the primal ($p = 1$)

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

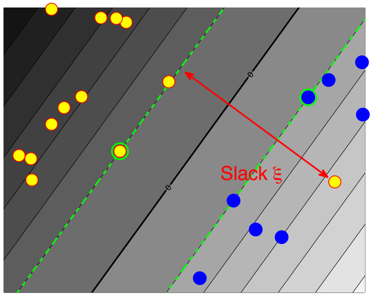
for $p = 1$ the dual formulation is the following:

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

It generalizes to any cost (useful for unbalanced data)

Road map

1. Linear SVM
   - The non separable case
   - The C (L1) SVM
   - The L2 SVM and others "variations on a theme"
   - The hinge loss



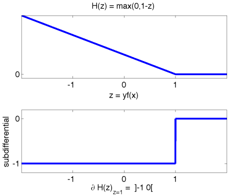
Eliminating the slack but not the possible mistakes

$$\begin{cases} \displaystyle\min_{\mathbf{w},b,\xi\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

Introducing the hinge loss

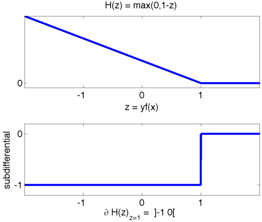
$$\xi_i = \max\big(1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b), 0\big)$$

$$\min_{\mathbf{w},b} \frac{1}{2} \; \|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max\big(0, 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b)\big)$$



Back to $d + 1$ variables, but this is no longer an explicit QP

Ooops! the notion of sub differential



Definition (Sub gradient)

a subgradient of $J : \mathbb{R}^d \longmapsto \mathbb{R}$ at f_0 is any vector $g \in \mathbb{R}^d$ such that

$$\forall f \in \mathcal{V}(f_0), \qquad J(f) \;\geq\; J(f_0) + g^\top (f - f_0)$$

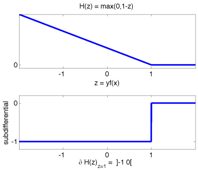
Definition (Subdifferential)

$\partial J(f)$, the subdifferential of J at f is the set of all subgradients of J at f .

$$\mathbb{R}^d = \mathbb{R} \qquad J_3(x) = |x| \qquad \partial J_3(0) = \{g \in \mathbb{R} \mid -1 < g < 1\}$$
$$\mathbb{R}^d = \mathbb{R} \qquad J_4(x) = \max(0, 1 - x) \qquad \partial J_4(1) = \{g \in \mathbb{R} \mid -1 < g < 0\}$$

Regularization path for SVM

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \max(1 - y_i \mathbf{w}^\top \mathbf{x}_i, 0) + \frac{\lambda_o}{2} \|\mathbf{w}\|^2$$

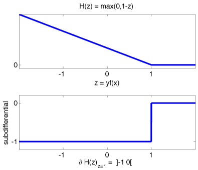


I_α is the set of support vectors s.t. $y_i \mathbf{w}^\top \mathbf{x}_i = 1$;

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

Regularization path for SVM

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \max(1 - y_i \mathbf{w}^\top \mathbf{x}_i, 0) + \frac{\lambda_o}{2} \|\mathbf{w}\|^2$$



I_α is the set of support vectors s.t. $y_i \mathbf{w}^\top \mathbf{x}_i = 1$;

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

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

In particular at point $\mathbf{x}_j \in I_\alpha$ $(\mathbf{w}_o^\top \mathbf{x}_j = \mathbf{w}_n^\top \mathbf{x}_j = y_j)$: $\partial_{\mathbf{w}} J(\mathbf{w})(\mathbf{x}_j) = 0$

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

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

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

Solving SVM in the primal

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

- What for: Yahoo!, Twiter, Amazon, Google (Sibyl), Facebook. . . : Big data
  Data-intensive machine learning systems

- "on terascale datasets, with trillions of features,1 billions of training examples and millions of parameters in an hour using a cluster of 1000 machines"

- How: hybrid online+batch approach adaptive gradient updates (stochastic gradient descent)

- Code available: `http://olivier.chapelle.cc/primal/`

A Reliable Effective Terascale Linear Learning System

Alekh Agarwal* ALEKHA@MICROSOFT.COM
Microsoft Research
New York, NY

Olivier Chapelle OLIVIER@CHAPELLE.CC
Criteo
Palo Alto, CA

Miroslav Dudík MDUDIK@MICROSOFT.COM
Microsoft Research
New York, NY

John Langford JCL@MICROSOFT.COM
Microsoft Research
New York, NY

Editor:

Abstract

We present a system and a set of techniques for learning linear predictors with convex losses on terascale datasets, with trillions of features,[1] billions of training examples and millions of parameters in an hour using a cluster of 1000 machines. Individually none of the component techniques are new, but the careful synthesis required to obtain an efficient implementation is. The result is, up to our knowledge, the most scalable and efficient linear learning system reported in the literature (as of 2011 when our experiments were conducted). We describe and thoroughly evaluate the components of the system, showing the importance of the various design choices.

Solving SVM in the primal

$$
\begin{aligned}
J(\mathbf{w}, b) &= \quad \tfrac{1}{2}\,\|\mathbf{w}\|_2^2 \quad + \tfrac{C}{2}\,\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0\big)^2 \\
&= \quad \tfrac{1}{2}\,\|\mathbf{w}\|_2^2 \quad + \tfrac{C}{2}\,\xi^\top \xi \\
& \qquad\qquad\qquad \text{with} \qquad \xi_i = \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0\big) \\
\nabla_{\mathbf{w}} J(\mathbf{w}, b) &= \qquad \mathbf{w} \qquad - C\,\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0\big)\; y_i \mathbf{x}_i \\
&= \qquad \mathbf{w} \qquad - C\,(\mathrm{diag}(\mathbf{y}) X)^\top \xi \\
H_{\mathbf{w}} J(\mathbf{w}, b) &= \qquad I_d \qquad + C\,\sum_{i \notin I_0} \mathbf{x}_i \mathbf{x}_i^\top
\end{aligned}
$$

Optimal step size ρ in the Newton direction:

$$
\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \rho\, H_{\mathbf{w}}^{-1} \nabla_{\mathbf{w}} J(\mathbf{w}^{\text{old}}, b^{\text{old}})
$$

The hinge and other loss

Square hinge: (huber/hinge) and Lasso SVM

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_1 + \; C\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0\big)^p$$

Penalized Logistic regression (Maxent)

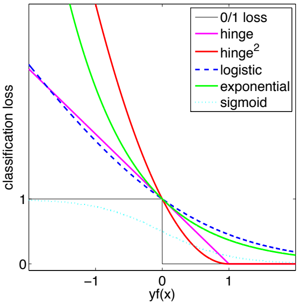
$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_2^2 - \; C\sum_{i=1}^{n} \log\big(1 + \exp^{-2y_i(\mathbf{w}^\top \mathbf{x}_i + b)}\big)$$

The exponential loss (commonly used in boosting)

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_2^2 + \; C\sum_{i=1}^{n} \exp^{-y_i(\mathbf{w}^\top \mathbf{x}_i + b)}$$

The sigmoid loss

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_2^2 - \; C\sum_{i=1}^{n} \tanh\big(y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big)$$



Choosing the data fitting term and the penalty

For a given C: controling the tradeoff between loss and penalty

$$\min_{\mathbf{w},b} \quad \text{pen}(\mathbf{w}) + \; C \sum_{i=1}^{n} \text{Loss}\big(y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big)$$

For a long list of possible penalties:

A Antoniadis, I Gijbels, M Nikolova, Penalized likelihood regression for generalized linear models with non-quadratic penalties, 2011.

A tentative of classification:

- convex/non convex
- differentiable/non differentiable

What are we looking for

- consistency
- efficiency $\longrightarrow$ sparcity

Conclusion: variables or data point?

- seeking for a universal learning algorithm
  - no model for $\mathbb{P}(\mathbf{x}, y)$
- the linear case: data is separable
  - the non separable case
- double objective: minimizing the error together with the regularity of the solution
  - multi objective optimisation
- dualiy : variable – example
  - use the primal when $d < n$ (in the liner case) or when matrix G is hard to compute
  - otherwise use the dual
- universality = nonlinearity
  - kernels

Bibliography

- C. Cortes & V. Vapnik, *Support-vector networks*, Machine learning, 1995
- J. Bi & V. Vapnik, *Learning with rigorous SVM*, COLT 2003
- T. Hastie, S. Rosset, R. Tibshirani, J. Zhu, *The entire regularization path for the support vector machine*, JMLR, 2004
- P. Bartlett, M. Jordan, J. McAuliffe, *Convexity, classification, and risk bounds*, JASA, 2006.
- A. Antoniadis, I. Gijbels, M. Nikolova, Penalized likelihood regression for generalized linear models with non-quadratic penalties, 2011.
- A Agarwal, O Chapelle, M Dudík, J Langford, A reliable effective terascale linear learning system, 2011.

- `informatik.unibas.ch/fileadmin/Lectures/FS2013/CS331/Slides/my_SVM_without_b.pdf`
- `http://ttic.uchicago.edu/~gregory/courses/ml2010/lectures/lect12.pdf`
- `http://olivier.chapelle.cc/primal/`