Support Vector Machines

Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

May 23, 2008 / Statistical Learning Theory II

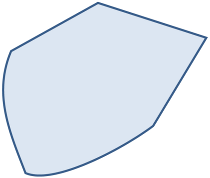
Outline

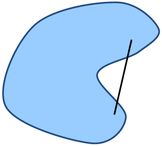
Convexity I

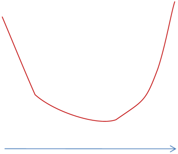
For the details on convex optimization, see [BV04].

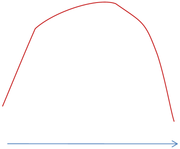
- Convex set:
  A set C in a vector space is convex if for every $x, y \in C$ and $t \in [0, 1]$
  $$tx + (1-t)y \in C.$$
- Convex function:
  Let C be a convex set. $f : C \to \mathbb{R}$ is called a convex function if for every $x, y \in C$ and $t \in [0, 1]$
  $$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$
- Concave function:
  Let C be a convex set. $f : C \to \mathbb{R}$ is called a concave function if for every $x, y \in C$ and $t \in [0, 1]$
  $$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y).$$

Convexity II


convex set


non-convex set


convex function


concave function

Convexity III

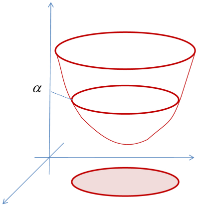
- Fact: If $f : C \to \mathbb{R}$ is a convex function, the set

$$\{x \in C \mid f(x) \leq \alpha\}$$

  is a convex set for every $\alpha \in \mathbb{R}$.
- If $f_t(x) : C \to \mathbb{R}$ ($t \in T$) are convex, then

$$f(x) = \sup_{t \in T} f_t(x)$$

  is also convex.



Convex optimization I

- A general form of convex optimization
  $f(x)$, $h_i(x)$ $(1 \leq i \leq \ell)$: $\mathcal{D} \to \mathbb{R}$, convex functions on $\mathcal{D} \subset \mathbb{R}^n$.
  $a_i \in \mathbb{R}^n, b_j \in \mathbb{R}$ $(1 \leq j \leq m)$.

$$\min_{x \in \mathcal{D}} f(x) \qquad \text{subject to} \begin{cases} h_i(x) \leq 0 & (1 \leq i \leq \ell), \\ a_j^T x + b_j = 0 & (1 \leq j \leq m). \end{cases}$$

  h_i: inequality constraints,
  $r_j(x) = a_j^T x + b_j$: linear equality constraints.

- Feasible set:

$$\mathcal{F} = \{x \in \mathcal{D} \mid h_i(x) \leq 0 \ (1 \leq i \leq \ell), r_j(x) = 0 \ (1 \leq j \leq m)\}.$$

  The above optimization problem is called feasible if $\mathcal{F} \neq \emptyset$.

- In convex optimization, there are no local minima. It is possible to find a minimizer numerically.

Convex optimization II

- Fact 1. The feasible set is a convex set.
- Fact 2. The set of minimizers

$$X_{opt} = \{x \in \mathcal{F} \mid f(x) = \inf\{f(y) \mid y \in \mathcal{F}\}\}$$

is convex.

proof. The intersection of convex sets is convex, which leads (1).
Let

$$p^* = \inf_{x \in \mathcal{F}} f(x).$$

Then,

$$X_{opt} = \{x \in \mathcal{D} \mid f(x) \leq p^*\} \cap \mathcal{F}.$$

Both sets in r.h.s. are convex. This proves (2) □

Examples

- Linear program (LP)

$$\min c^T x \qquad \text{subject to } \begin{cases} Ax = b, \\ Gx \preceq h.^1 \end{cases}$$

The objective function, the equality and inequality constraints are all linear.

- Quadratic program (QP)

$$\min \frac{1}{2} x^T P x + q^t x + r \qquad \text{subject to } \begin{cases} Ax = b, \\ Gx \preceq h, \end{cases}$$

where P is a positive semidefinite matrix.
The objective function is quadratic, while the equality and inequality constraints are linear.

[1] $Gx \preceq h$ denotes $g_j^T x \leq h_j$ for all j, where $G = (g_1, \ldots, g_m)^T$.

Lagrange duality I

- Consider an optimization problem (which may not be convex):

$$\text{(primal)} \qquad \min_{x \in \mathcal{D}} f(x) \qquad \text{subject to} \begin{cases} h_i(x) \le 0 & (1 \le i \le \ell), \\ r_j(x) = 0 & (1 \le j \le m). \end{cases}$$

- Lagrange dual function: $g : \mathbb{R}^\ell \times \mathbb{R}^m \to [-\infty, \infty)$

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu),$$

  where

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{\ell} \lambda_i h_i(x) + \sum_{j=1}^{m} \nu_j r_j(x).$$

  λ_i and ν_j are called Lagrange multipliers.
- g is a concave function.

Lagrange duality II

- Dual problem

$$(\text{dual}) \qquad \max g(\lambda, \nu) \qquad \text{subject to} \quad \lambda \succeq 0.$$

- The dual and primal problems have close connection.

Theorem (weak duality)

Let

$$p^* = \inf\{f(x) \mid h_i(x) \le 0 \ (1 \le i \le \ell), r_j(x) = 0 \ (1 \le j \le m)\}.$$

and

$$d^* = \sup\{g(\lambda, \nu) \mid \lambda \succeq 0, \nu \in \mathbb{R}^m\}.$$

Then,

$$d^* \le p^*.$$

The weak duality does not require the convexity of the primal optimization problem.

Lagrange duality III

- Proof. Let $\forall\lambda \succeq 0, \nu \in \mathbb{R}^m$.
  For any feasible point x,

  $$\sum_{i=1}^{\ell} \lambda_i h_i(x) + \sum_{j=1}^{m} \nu_j r_j(x) \leq 0.$$

  (The first part is non-positive, and the second part is zero.)
  This means for any feasible point x,

  $$L(x, \lambda, \nu) \leq f(x).$$

  By taking infimum,

  $$\inf_{x:feasible} L(x, \lambda, \nu) \leq p^*.$$

  Thus,

  $$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq \inf_{x:feasible} L(x, \lambda, \nu) \leq p^*$$

  for any $\lambda \succeq 0, \nu \in \mathbb{R}^m$.

Strong duality

We need some conditions to obtain the strong duality $d^* = p^*$.

- Convexity of the problem: f and h_i are convex, r_j are linear.
- Slater's condition

  There is $x \in \mathcal{D}$ such that

$$h_i(x) < 0 \quad (1 \leq \forall i \leq \ell), \qquad r_j(x) = 0 \quad (1 \leq \forall j \leq m).$$

Theorem (Strong duality)

Suppose the primal problem is convex, and Slater's condition holds. Then, there is $\lambda^* \geq 0$ and $\nu^* \in \mathbb{R}^m$ such that

$$g(\lambda^*, \nu^*) = d^* = p^*.$$

Proof is omitted (see [BV04] Sec.5.3.2.).

There are also other conditions to guarantee the strong duality.

Geometric interpretation of duality

- $\mathcal{G} = \{(h(x), r(x), f(x)) \in \mathbb{R}^\ell \times \mathbb{R}^m \times \mathbb{R} \mid x \in \mathbb{R}^n\}$.
- $p^* = \inf\{(u, 0, t) \in \mathcal{G} \mid u \leq 0\}$.
- For $\lambda \geq 0$ (neglecting ν),

$$g(\lambda) = \inf\{(\lambda, 1)^T(u, t) \mid (u, t) \in \mathcal{G}\}.$$

- $(\lambda, 1)^T(u, t) = g(\lambda)$ is a non-vertical hyperplane, which intersects with t-axis at $g(\lambda)$.
- Weak duality $d^* = \sup g(\lambda) \leq p^*$ is easy to see.
- Strong duality $d^* = p^*$ holds under some conditions.

Figure

Complementary slackness I

- Consider the (not necessarily convex) optimization problem:

$$\min f(x) \qquad \text{subject to } \begin{cases} h_i(x) \le 0 & (1 \le i \le \ell), \\ r_j(x) = 0 & (1 \le j \le m). \end{cases}$$

- Assumption: the optimum of the primal and dual problems are attained by x^* and (λ^*, ν^*), and they satisfy the strong duality;

$$g(\lambda^*, \nu^*) = f(x^*).$$

- Observation:

$$\begin{aligned} f(x^*) = g(\lambda^*, \nu^*) &= \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*) \qquad \text{[definition]} \\ &\le L(x^*, \lambda^*, \nu^*) \\ &= f(x^*) + \textstyle\sum_{i=1}^{\ell} \lambda_i^* h_i(x^*) + \sum_{j=1}^{m} \nu_j^* r_j(x^*) \\ &\le f(x^*) \qquad \text{[2nd} \le 0 \text{ and 3rd} = 0\text{]} \end{aligned}$$

The two inequalities are in fact equalities.

Complementary slackness II

- Consequence 1:

$$x^* \text{ minimizes } L(x, \lambda^*, \nu^*)$$

- Consequence 2:

$$\lambda_i^* h_i(x^*) = 0 \qquad \text{for all } i.$$

This is called complementary slackness.

Equivalently,

$$\lambda_i^* > 0 \quad \Rightarrow \quad h_i(x^*) = 0,$$

or

$$h_i(x^*) < 0 \quad \Rightarrow \quad \lambda_i^* = 0.$$

Solving primal problem via dual

If the dual problem is easier to solve, then we can sometimes solve the primal using the dual.

- Assumption: strong duality holds (e.g., Slater's condition), and we have the dual solution (λ^*, ν^*).
- The primal solution x^*, if it exists, should give $\min_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*)$ (previous slide).
- If a solution of

$$\min_{x \in \mathcal{D}} f(x) + \sum_{i=1}^{\ell} \lambda_i^* h_i(x) + \sum_{j=1}^{m} \nu_j^* r_j(x)$$

  is obtained and it is primary feasible, then it must be the primal solution.

KKT condition I

KKT conditions give useful relations between the primal and dual solutions.

- Consider the convex optimization problem. Assume $\mathcal{D}$ is open and $f(x)$, $h_i(x)$ are differentiable.

$$\min f(x) \qquad \text{subject to} \begin{cases} h_i(x) \leq 0 & (1 \leq i \leq \ell), \\ r_j(x) = 0 & (1 \leq j \leq m). \end{cases}$$

- x^* and (λ^*, ν^*): any optimal points of the primal and dual problems.
- Assume the strong duality holds.
- Since

$$x^* \text{ minimizes } L(x, \lambda^*, \nu^*),$$

$$\nabla f(x^*) + \textstyle\sum_{i=1}^{\ell} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{m} \nu_j^* \nabla r_j(x^*) = 0.$$

KKT condition II

The following are necessary conditions.

Karush-Kuhn-Tucker (KKT) conditions:

$h_i(x^*) \le 0 \quad (i = 1, \dots, \ell)$ [primal constraints]

$r_j(x^*) = 0 \quad (j = 1, \dots, m)$ [primal constraints]

$\lambda_i^* \ge 0 \quad (i = 1, \dots, \ell)$ [dual constraints]

$\lambda_i^* h_i(x^*) = 0 \quad (i = 1, \dots, \ell)$ [complementary slackness]

$\nabla f(x^*) + \sum_{i=1}^{\ell} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{m} \nu_j^* \nabla r_j(x^*) = 0.$

Theorem (KKT condition)

For a convex optimization problem with differentiable functions, x^* and (λ^*, ν^*) are the primal-dual solutions with strong duality if and only if they satisfy KKT conditions.

KKT condition III

- Proof.
  - x^* is primal-feasible by the first two conditions.
  - From $\lambda_i^* \geq 0$, $L(x, \lambda^*, \nu^*)$ is convex (and differentiable).
  - The last condition $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$ implies x^* is a minimizer.
  - It follows

$$\begin{aligned} g(\lambda^*, \nu^*) &= \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*) \qquad \text{[by definition]} \\ &= L(x^*, \lambda^*, \nu^*) \qquad [x^*\text{: minimizer}] \\ &= f(x^*) + \textstyle\sum_{i=1}^{\ell} \lambda_i^* h_i(x^*) + \sum_{j=1}^{m} \nu_j^* r_j(x^*) \\ &= f(x^*) \qquad \text{[complementary slackness and } r_j(x^*) = 0\text{].} \end{aligned}$$

  - Strong duality holds, and x^* and (λ^*, ν^*) must be the optimizers.

Example

- Quadratic minimization under equality constraints.

$$\min \frac{1}{2}x^T P x + q^T x + r \qquad \text{subject to} \quad Ax = b,$$

  where P is (strictly) positive definite.
- KKT conditions:

$$\begin{aligned} &Ax^* = b, && \text{[primal constraint]} \\ &\nabla_x L(x^*, \nu^*) = 0 \quad \Longrightarrow && Px^* + q + A^T \nu^* = 0 \end{aligned}$$

- The solution is given by

$$\begin{pmatrix} P & A^T \\ A & O \end{pmatrix} \begin{pmatrix} x^* \\ \nu^* \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}.$$

Lagrange dual of SVM I

- The QP for SVM can be solved in the primal form, but the dual form is easier.
- SVM: primal problem

$$\min_{w_i,b,\xi_i} \frac{1}{2}\sum_{i,j=1}^N w_i w_j k(X_i, X_j) + C\sum_{i=1}^N \xi_i,$$

$$\text{subj. to} \quad \begin{cases} Y_i(\sum_{j=1}^N k(X_i, X_j) w_j + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- Lagrange function of SVM:

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\sum_{i,j=1}^N w_i w_j k(X_i, X_j) + C\sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i(1 - Y_i f(X_i) - \xi_i) + \sum_{i=1}^N \beta_i(-\xi_i),$$

where $f(X_i) = \sum_{j=1}^N w_j k(X_i, X_j) + b$.

Lagrange dual of SVM II

SVM Dual problem:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j K_{ij} \quad \text{subj. to} \begin{cases} 0 \leq \alpha_i \leq C, \\ \sum_{i=1}^{N} \alpha_i Y_i = 0. \end{cases}$$

Notation: $K_{ij} = k(X_i, X_i)$.

Derivation.
Lagrange dual function

$$g(\alpha, \beta) = \min_{w,b,\xi} L(w, b, \xi, \alpha, \beta)$$

The minimizer (w^*, b^*, ξ^*) is given by

$$\begin{aligned} \nabla_w : & \quad \textstyle\sum_{j=1}^{n} K_{ij} w_j^* + \sum_{j=1}^{n} \alpha_j Y_j K_{ij} \quad (\forall i) \\ , \nabla_b : & \quad \textstyle\sum_{j=1}^{n} \alpha_j Y_j = 0, \\ \nabla_\xi : & \quad C - \alpha_i - \beta_i = 0 \quad (\forall i). \end{aligned}$$

Lagrange dual of SVM III

From these relations,

$$\frac{1}{2}\sum_{i,j=1}^{N} w_i^* w_j^* K_{ij} = \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j K_{ij},$$

$$\sum_{j=1}^{n} (C - \alpha_i - \beta_i) = 0,$$

$$\sum_{i=1}^{N} \alpha_i (1 - Y_i (\sum_j K_{ij} w_j^* + b)) = \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j K_{ij}.$$

Thus,

$$\begin{aligned} g(\alpha, \beta) &= L(w^*, b^*, \xi^*, \alpha, \beta) \\ &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j K_{ij}. \end{aligned}$$

From $\alpha_i \geq 0$ and $\beta_i \geq 0$, The constraints of α_i is given by

$$0 \leq \alpha_i \leq C \qquad (\forall i).$$

KKT conditions of SVM

KKT conditions

(1) $1 - Y_i f^*(X_i) - \xi_i^* \leq 0 \quad (\forall i)$,

(2) $-\xi_i^* \leq 0 \quad (\forall i)$,

(3) $\alpha_i^* \geq 0, \quad (\forall i)$,

(4) $\beta_i^* \geq 0, \quad (\forall i)$,

(5) $\alpha_i^*(1 - Y_i f^*(X_i) - \xi_i^*) = 0 \quad (\forall i)$,

(6) $\beta_i^* \xi_i^* = 0 \quad (\forall i)$,

(7) $\nabla_w : \quad \sum_{j=1}^n K_{ij} w_j^* + \sum_{j=1}^n \alpha_j^* Y_j K_{ij}$,
$\nabla_b : \quad \sum_{j=1}^n \alpha_j^* Y_j = 0$,
$\nabla_\xi : \quad C - \alpha_i^* - \beta_i^* = 0 \quad (\forall i)$.

Support vectors I

- Complementary slackness

$$\alpha_i^*(1 - Y_i f^*(X_i) - \xi_i^*) = 0 \qquad (\forall i),$$

$$(C - \alpha_i^*)\xi_i^* = 0 \qquad (\forall i).$$

- If $\alpha_i^* = 0$, then $\xi_i^* = 0$, and

$$Y_i f^*(X_i) \geq 1. \qquad \text{[well separated]}$$

- Support vectors
  - If $0 < \alpha_i^* < C$, then $\xi_i^* = 0$ and

$$Y_i f^*(X_i) = 1.$$

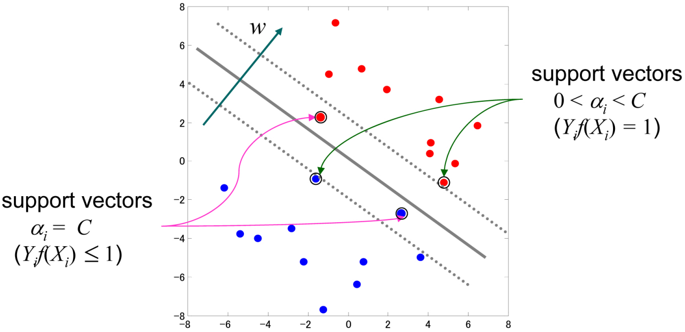
  - If $\alpha_i^* = C$,

$$Y_i f^*(X_i) \leq 1.$$

Support vectors II

- Sparse representation
  The optimum classifier is expressed only with the support vectors:

$$f(x) = \sum_{i:\text{support vector}} \alpha_i^* Y_i k(x, X_i) + b^*$$



How to solve b

- The optimum value of b is given by the complementary slackness.
- For any i with $0 < \alpha_i^* < C$,

$$Y_i\left(\sum_j k(X_i, X_j) Y_j \alpha_j^* + b\right) = 1.$$

- Use the above relation for any of such i, or take the average over all of such i.

1. A quick course on convex optimization
   - Convexity and convex optimization
   - Dual problem for optimization

2. Optimization in learning of SVM
   - Dual problem and support vectors
   - Sequential Minimal Optimization (SMO)
   - Other approaches

Computational problem in solving SVM

- The dual QP problem of SVM has N variables, where N is the sample size.
- If N is very large, say $N = 100,000$, the optimization is very hard.
- Some approaches have been proposed for optimizing subsets of the variables sequentially.
  - Chunking [Vap82]
  - Osuna's method [OFG]
  - Sequential minimal optimization (SMO) [Pla99]
  - SVMlight (`http://svmlight.joachims.org/`)

Sequential minimal optimization (SMO) I

- Solve small QP problems sequentially for a pair of variables (α_i, α_j).
- How to choose the pair? – Intuition from the KKT conditions is used.
  - After removing w, ξ, and β, the KKT conditions of SVM are equivalent to

$$0 \leq \alpha_i^* \leq C, \qquad \sum_{i=1}^{N} Y_i \alpha_i^* = 0,$$

$$(*) \begin{cases} \alpha_i^* = 0 & \Rightarrow \quad Y_i f^*(X_i) \geq 1, \\ 0 < \alpha_i^* < C & \Rightarrow \quad Y_i f^*(X_i) = 1, \\ \alpha_i^* = C & \Rightarrow \quad Y_i f^*(X_i) \leq 1. \end{cases}$$

    (shown later.)
  - The conditions can be checked for each data point.
  - Choose (i, j) such that at least one of them breaks the KKT conditions.

Sequential minimal optimization (SMO) II

The QP problem for (α_i, α_j) is analytically solvable!

- For simplicity, assume $(i, j) = (1, 2)$.
- Constraint of α_1 and α_2:

$$\alpha_1 + s_{12}\alpha_2 = \gamma, \qquad 0 \le \alpha_1, \alpha_2 \le C,$$

  where $s_{12} = Y_1 Y_2$ and $\gamma = \pm \sum_{\ell \ge 3} Y_\ell \alpha_\ell$ is constat.
- Objective function:

$$\alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 K_{11} - \frac{1}{2}\alpha_2^2 K_{22} - s_{12}\alpha_1\alpha_2 K_{12} \\ - Y_1\alpha_1 \sum_{j \ge 3} Y_j \alpha_j K_{1j} - Y_2 \alpha_2 \sum_{j \ge 3} Y_j \alpha_j K_{2j} + const.$$

- This optimization is a quadratic optimization of one variable on an interval. Directly solved.

KKT conditions revisited I

- β and w can be removed by

$$\begin{aligned} \nabla_\xi: \quad & \beta_i^* = C - \alpha_i^* = 0 \quad (\forall i), \\ \nabla_w: \quad & \textstyle\sum_{j=1}^n K_{ij} w_j^* = -\sum_{j=1}^n \alpha_j Y_j K_{ij} \quad (\forall i). \end{aligned}$$

- From (4) and (6),

$$\alpha_i^* \geq C, \qquad \xi_i^*(C - \alpha_i^*) = 0 \quad (\forall i).$$

- The KKT conditions are equivalent to
  (a) $1 - Y_i f^*(X_i) - \xi_i^* \leq 0 \quad (\forall i)$,
  (b) $\xi_i^* \geq 0 \quad (\forall i)$,
  (c) $0 \leq \alpha_i^* \leq C \quad (\forall i)$,
  (d) $\alpha_i^*(1 - Y_i f^*(X_i) - \xi_i^*) = 0 \quad (\forall i)$,
  (e) $\xi_i^*(C - \alpha_i^*) = 0 \quad (\forall i)$,
  (f) $\sum_{i=1}^N Y_i \alpha_i^* = 0$.

KKT conditions revisited II

- We can further remove ξ.
  - Case $\alpha_i^* = 0$:
    From (e), $\xi_i^* = 0$. Then, from (a), $Y_i f^*(X_i) \geq 1$.
  - Case $0 < \alpha_i^* < C$:
    From (e), $\xi_i^* = 0$. From (d), $Y_i f^*(X_i) = 1$.
  - Case $\alpha_i^* = C$:
    From (d), $\xi_i^* = 1 - Y_i f^*(X_i)$.

  Note in all cases, (a) and (b) are satisfied.
- The KKT conditions are equivalent to

$$0 \leq \alpha_i^* \leq C \quad (\forall i),$$

$$\sum_{i=1}^N Y_i \alpha_i^* = 0,$$

$$\begin{cases} \alpha_i^* = 0 & \Rightarrow \quad Y_i f^*(X_i) \geq 1, \quad (\xi_i^* = 0) \\ 0 < \alpha_i^* < C & \Rightarrow \quad Y_i f^*(X_i) = 1, \quad (\xi_i^* = 0) \\ \alpha_i^* = C & \Rightarrow \quad Y_i f^*(X_i) \leq 1, \quad (\xi_i^* = 1 - Y_i f^*(X_i)). \end{cases}$$

Recent studies (not a compete list).

- Solution in primal.
  - O. Chapelle [Cha07]
  - T. Joachims, SVM^{perf} [Joa06]
  - S. Shalev-Shwartz et al. [SSSS07]
- Online SVM.
  - Tax and Laskov [TL03]
  - LaSVM [BEWB05]
    `http://leon.bottou.org/projects/lasvm/`
- Parallel computation
  - Cascade SVM [GCB$^+$05]
  - Zanni et al [ZSZ06]
- Others
  - Column generation technique for large scale problems [DBS02]

References I

Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou.
Fast kernel classifiers with online and active learning.
Journal of Machine Learning Research, 6:1579–1619, 2005.

Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.
http://www.stanford.edu/ boyd/cvxbook/.

Olivier Chapelle.
Training a support vector machine in the primal.
Neural Computation, 19:1155–1178, 2007.

Ayhan Demiriz, Kristin P. Bennett, and John Shawe-Taylor.
Linear programming boosting via column generation.
Machine Learning, 46(1-3):225–254, 2002.

References II

Hans Peter Graf, Eric Cosatto, Léon Bottou, Igor Dourdanovic, and Vladimir Vapnik.
Parallel support vector machines: The Cascade SVM.
In Lawrence Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.

Thorsten Joachims.
Training linear svms in linear time.
In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

Edgar Osuna, Robert Freund, and Federico Girosi.
An improved training algorithm for support vector machines.
In *Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing (IEEE NNSP 1997)*, pages 276–285.

References III

- John Platt.
  Fast training of support vector machines using sequential minimal optimization.
  In Bernhard Schölkopf, Cristopher Burges, and Alexander Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press, 1999.

- Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro.
  Pegasos: Primal estimated sub-gradient solver for svm.
  In *Proc. International Conerence of Machine Learning*, 2007.

- D.M.J. Tax and P. Laskov.
  Online svm learning: from classification to data description and back.
  In *Proceddings of IEEE 13th Workshop on Neural Networks for Signal Processing (NNSP2003)*, pages 499–508, 2003.

References IV

Vladimir N. Vapnik.
Estimation of Dependences Based on Empirical Data.
Springer-Verlag, 1982.

Luca Zanni, Thomas Serafini, and Gaetano Zanghirati.
Parallel software for training large scale support vector machines on multiprocessor systems.
Journal of Machine Learning Research, 7:1467–1492, 2006.