

Optimization Theory

Reference: Cristianini, Shawe-Taylor: An introduction to SVM

1. Type of problem:

$$\Omega \subset \mathbb{R}^n \quad \text{minimize } f(a) \quad a \in \Omega$$

← objective function

- f^* : "value" of the problem
- solution a^* : $\forall a \in \Omega \quad f(a) \geq f(a^*)$
← global minimum
- local minimum a^* : $\exists \varepsilon > 0 \quad \forall a \in \Omega \quad \|a - a^*\| < \varepsilon \Rightarrow f(a) \geq f(a^*)$
- if f convex: a^* local minimum $\Rightarrow a^*$ global minimum
- Hessian positive semi-definite $\Rightarrow f$ convex
- Ω convex: $\Leftrightarrow \alpha a + (1-\alpha)b \in \Omega$
 $\forall a, b \in \Omega$
- good: f and Ω convex

Solution:

$$f \in C^1 \quad \frac{\partial f(a^*)}{\partial a} = 0 \quad \Leftrightarrow \quad a^* \text{ solution}$$

↑
(if f convex)

gradient

2. Type of problem:

$$\Omega \subset \mathbb{R}^n \quad \text{minimize } f(a) \quad a \in \Omega$$

$$\text{s.t. } h_i(a) = 0 \quad i = 1, \dots, m$$

constraints

• feasible region: $\{a \in \Omega : h_i(a) = 0 \quad \forall i = 1, \dots, m\}$

Solution: Lagrange Theory: Combine f and h_i to one function!

$$L(a, \beta) := f(a) + \sum_{i=1}^m \beta_i h_i(a)$$

↑
Lagrangian,
Lagrange Function

← Lagrange Multipliers

$$f, h_i \in C^1 \quad \exists \beta^* \in \mathbb{R}^m \quad \begin{cases} \frac{\partial L(a^*, \beta^*)}{\partial a} = 0 \\ \frac{\partial L(a^*, \beta^*)}{\partial \beta} = 0 \end{cases} \quad \Leftrightarrow \quad a^* \text{ solution}$$

(if $L(\cdot, \beta^*)$ convex)

• 2. condition gives constraints

• 1. condition gives $\frac{\partial f(a^*)}{\partial a} + \sum_{i=1}^m \beta_i \frac{\partial h_i(a^*)}{\partial a} = 0$

Thus $\frac{\partial f(a^*)}{\partial a}$ must not be 0, but in the span of $\frac{\partial h_i(a^*)}{\partial a}$

So each further improvement would violate a constraint.

• β_i^* expresses the sensitivity of the solution f^* to h_i ;

• $L(a^*, \beta^*) = f(a^*) = f^*$ ← not necessarily feasible

• (a^*, β^*) saddle point: $\forall a \in \Omega \quad L(a, \beta^*) \geq L(a^*, \beta^*)$

$\forall \beta \quad L(a^*, \beta) \leq L(a^*, \beta^*)$

• dual function: $\Theta(\beta) := \inf_{a \in \Omega} L(a, \beta)$

3. Type of problem:

$$\begin{array}{l} \Omega \subset \mathbb{R}^n \\ \text{minimize } f(a) \quad a \in \Omega \\ \text{s.t. } g_i(a) \leq 0 \quad i = 1, \dots, k \\ \quad \quad h_i(a) = 0 \quad i = 1, \dots, m \end{array}$$

inequality constraints
 equality constraints

- feasible region: $\{a \in \Omega : g_i(a) \leq 0 \ \forall i \ \wedge \ h_i(a) = 0 \ \forall i\}$
- Linear program: f linear, g_i and h_i linear
- quadratic program: f quadratic, g_i and h_i linear
- g_i active: $\Leftrightarrow g_i(a^*) = 0$
- application to SVM: f quadratic, convex
 g_i and h_i linear
 $\Omega = \mathbb{R}^n$ convex quadratic program

Solution: Generalized Lagrange Function:

$$L(a, \alpha, \beta) := f(a) + \sum_{i=1}^k \alpha_i g_i(a) + \sum_{i=1}^m \beta_i h_i(a)$$

Dual Function:

$$\theta(\alpha, \beta) := \inf_{a \in \Omega} L(a, \alpha, \beta)$$

↑ dual variables

Standard technique to compute the dual Function:

Eliminate the primal variables:

$$\underbrace{\frac{\partial L(a, \dots)}{\partial a} = 0}_{\text{"stationarity"}} \Rightarrow a = \dots \xrightarrow{\text{substitute}} L(a, \dots) =: \theta(\alpha, \beta)$$

Dual Problem:

$$\begin{array}{l} \text{maximize } \theta(\alpha, \beta) \\ \text{s.t. } \alpha \geq 0 \end{array}$$

Weak Duality Theorem:

$$\begin{array}{l} a \in \Omega \quad \text{feasible for primal problem} \\ (\alpha, \beta) \quad \text{feasible for dual problem} \\ \Rightarrow \theta(\alpha, \beta) \leq f(a) \end{array}$$

Proof: $\theta(\alpha, \beta) \leq L(a, \alpha, \beta) \leq f(a)$

Corollary:

$$\left. \begin{array}{l} a^* \text{ feasible for primal problem} \\ (\alpha^*, \beta^*) \text{ feasible for dual problem} \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} a^* \text{ solution of primal problem} \\ (\alpha^*, \beta^*) \text{ solution of dual problem} \\ \alpha_i^* g_i(a^*) = 0 \quad \forall i = 1, \dots, k \end{array} \right.$$

no duality gap

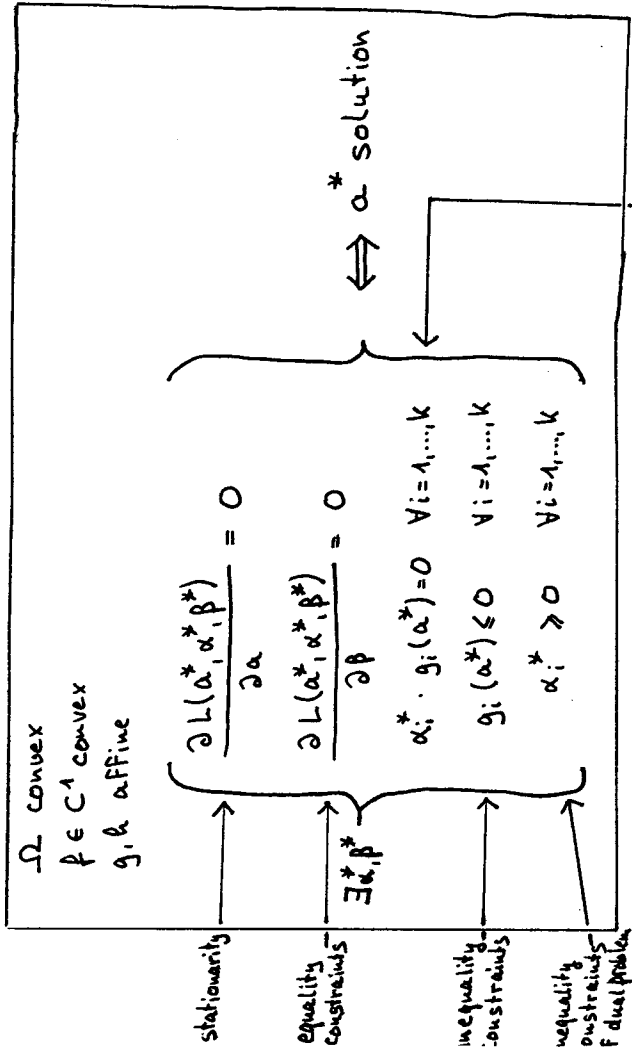
- Primal and dual problem not always have the same value!
 If NO: duality gap If YES: difference can be used as convergence criteria
- If $(\alpha^*, \alpha^*, \beta^*)$ is a saddle point of L , then a^* and (α^*, β^*) are solutions of the primal resp. dual problem, and there is no duality gap.

Strong Duality Theorem:

$$\left. \begin{array}{l} \Omega \text{ convex} \\ g, h \text{ affine} \end{array} \right\} \Rightarrow \text{no duality gap}$$

$\left(\begin{array}{c} g_1 \\ \vdots \\ g_k \end{array} \right) \left(\begin{array}{c} h_1 \\ \vdots \\ h_m \end{array} \right)$ — g affine: $\Leftrightarrow \exists A, b \quad g(a) = A \cdot a + b$

Solution: Kuhn-Tucker Theorem



• KKT condition: g_i inactive $\Rightarrow \alpha_i^* = 0$

• g_i active: α_i^* expresses sensitivity

g_i inactive: perturbation has no effect

• geometric interpretation: two possible positions of α^* with respect to g_i :

- 1) Inside the feasible region: $\alpha_i = 0, g_i$ inactive, solution as in 1. Type
- 2) On the border of the feasible region: $g_i(\alpha^*) = 0, g_i$ active, solution as in 2. Type

Examples

① minimize $f(x_1, x_2) = x_1^2 + x_2^2 + 1$

$\frac{\partial f(x_1, x_2)}{\partial x} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \stackrel{!}{=} 0 \Rightarrow x_1^* = x_2^* = 0$

② minimize $f(x_1, x_2) = x_1^2 + x_2^2 + 1$
 s.t. $x_1 + x_2 - 1 = 0$

$L(x, \beta) = x_1^2 + x_2^2 + 1 + \beta(x_1 + x_2 - 1)$

$\frac{\partial L(x, \beta)}{\partial x} = \begin{pmatrix} 2x_1 + \beta \\ 2x_2 + \beta \end{pmatrix} \stackrel{!}{=} 0 \Rightarrow x_1 = x_2 = -\frac{\beta}{2}$

Substitution in constraint: $-\frac{\beta}{2} - \frac{\beta}{2} - 1 = 0 \Rightarrow \beta^* = -1 \Rightarrow x_1^* = x_2^* = \frac{1}{2}$

Or: Substitution in L: $L(x, \beta) = -\frac{\beta^2}{2} - \beta + 1 =: \theta(\beta)$

$f^* = f(x^*) = \frac{3}{2}$ and $\theta(\beta)$ maximal at $\beta^* = -1$ with $\theta(\beta^*) = \frac{3}{2}$

③ minimize $f(x_1, x_2) = x_1^2 + x_2^2 + 1$
 s.t. $x_1 + x_2 - 1 \geq 0$
 $-x_1 - x_2 + 1 \leq 0$

$L(x, \alpha) = x_1^2 + x_2^2 + 1 + \alpha(-x_1 - x_2 + 1)$

$\frac{\partial L(x, \alpha)}{\partial x} = \begin{pmatrix} 2x_1 - \alpha \\ 2x_2 - \alpha \end{pmatrix} \stackrel{!}{=} 0 \Rightarrow x_1 = x_2 = \frac{\alpha}{2}$

Substitution in KKT condition: $\alpha(-x_1 - x_2 + 1) = \alpha(1 - \alpha) \stackrel{!}{=} 0$

$\alpha = 0$ not possible because $g(\alpha) = 1 - \alpha > 0$

$\Rightarrow \alpha^* = 1 \Rightarrow x_1^* = x_2^* = \frac{1}{2}$

\uparrow constraint active

Substitution in L: $L(x, \alpha) = -\frac{\alpha^2}{2} + \alpha + 1 =: \theta(\alpha)$
 $f^* = f(x^*) = \frac{3}{2}$ and $\theta(\alpha)$ maximal for $\alpha^* = 1$ with $\theta(\alpha^*) = \frac{3}{2}$

(4)

$$\begin{aligned} & \text{minimize} && f(x_1, x_2) = x_1^2 + x_2^2 + 1 \\ & \text{s.t.} && x_1 + x_2 - 1 \leq 0 \end{aligned}$$

$$L(x, \alpha) = x_1^2 + x_2^2 + 1 + \alpha (x_1 + x_2 - 1)$$

$$\frac{\partial L(x, \alpha)}{\partial x} = \begin{pmatrix} 2x_1 + \alpha \\ 2x_2 + \alpha \end{pmatrix} \stackrel{!}{=} 0$$

Substitution in KKT condition: $\alpha \cdot (x_1 + x_2 - 1) = \alpha(-\alpha - 1) \stackrel{!}{=} 0$
 $\alpha = -1$ not possible because $\alpha < 0$
 $\Rightarrow \alpha = 0 \Rightarrow x_1^* = x_2^* = 0$
 \uparrow constraint inactive

$$\text{Substitution in } L: L(x, \alpha) = -\frac{\alpha^2}{2} - \alpha + 1 = \theta(\alpha)$$

$\theta(\alpha)$ restricted to $\alpha \geq 0$ has its maximum at $\alpha = 0$
 $f(x^*) = \theta(\alpha^*) = 1$

$$\Rightarrow x_1 = x_2 = -\frac{\alpha}{2}$$

Support Vector Machines

Problem: Classification (2 classes)

Training Data: $(y_1, z_1), \dots, (y_e, z_e)$
 \uparrow data point $\in \mathbb{R}^m$ \uparrow class $\{ \pm 1 \}$

Wanted: Separating Hyperplane $(a, b): \{ y \in \mathbb{R}^m : \langle a, y \rangle + b = 0 \}$ with $\|a\| = 1$

Decision Function: $g(y) := \langle a, y \rangle + b$ $\begin{cases} \geq 0 : \text{predict class } + \\ < 0 : \text{predict class } - \end{cases}$

Maximal Margin Classifier

Condition: $z_i = +1: \langle a, y_i \rangle + b \geq \gamma \leftarrow \text{margin}$

$z_i = -1: \langle a, y_i \rangle + b \leq -\gamma$

Together: $z_i \cdot (\langle a, y_i \rangle + b) \geq \gamma \quad i = 1, \dots, e$
 Objective Function: maximize γ

Reformulation: $\alpha_{\text{new}} := \frac{1}{\gamma} \cdot a \quad b_{\text{new}} := \frac{1}{\gamma} \cdot b$
 $\Rightarrow \| \alpha_{\text{new}} \| = \frac{1}{\gamma}$

Optimization Problem:
 (quadratic program)

$$\begin{aligned} & \text{minimize} && \langle a, a \rangle \\ & \text{s.t.} && z_i \cdot (\langle a, y_i \rangle + b) \geq 1 \end{aligned}$$

Generalized Lagrange Function:

$$L(a, b, \alpha) = \frac{1}{2} \langle a, a \rangle - \sum_{i=1}^e \alpha_i [z_i \cdot (\langle a, y_i \rangle + b) - 1]$$

primal vars

$$\frac{\partial L(a, b, \alpha)}{\partial a} = a - \sum_{i=1}^e z_i \alpha_i y_i \stackrel{!}{=} 0$$

$$\frac{\partial L(a, b, \alpha)}{\partial b} = \sum_{i=1}^e z_i \alpha_i \stackrel{!}{=} 0 \quad \textcircled{*}$$

Substituting $a = \sum_{i=1}^{\ell} z_i \alpha_i y_i$ in $L(a, b, \alpha)$ and using \otimes gives:

$$L(a, b, \alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} z_i z_j \alpha_i \alpha_j \langle y_i, y_j \rangle =: W(\alpha)$$

depends only on the scalar product
 dual function
 (primal variables a and b are eliminated)

Dual Problem:
 (quadratic program)

$$\begin{array}{ll} \text{maximize} & W(\alpha) \\ \text{s.t.} & \alpha_i \geq 0 \quad \forall i=1, \dots, \ell \\ & \sum_{i=1}^{\ell} z_i \alpha_i = 0 \end{array}$$

Here we need a quadratic solver.
 If α^* is solution of the dual problem, then $\alpha^* = \sum_{i=1}^{\ell} z_i \alpha_i^* y_i$ is solution of the primal problem.

b^* can be computed from the constraints of the primal problem:

$$\left. \begin{array}{l} \min_{z_i: z_i=+1} \langle a, y_i \rangle + b = +1 \\ \min_{z_i: z_i=-1} \langle a, y_i \rangle + b = -1 \end{array} \right\} \Rightarrow b^* = -\frac{1}{2} \left[\min_{z_i: z_i=+1} \langle a, z_i \rangle + \max_{z_i: z_i=-1} \langle a, y_i \rangle \right]$$

(α^*, b^*) is the solution, and the margin is $\gamma = \frac{1}{\|\alpha^*\|_2}$

KKT conditions: $\alpha_i^* \cdot [z_i \cdot g(y_i) - 1] = 0 \quad \forall i=1, \dots, \ell$

Support Vectors: $SV := \{i : g(y_i) = \pm 1\}$

All $\alpha_i^* = 0$ except for $i \in SV$.

Decision Function:

$$g(y) = \langle \alpha^*, y \rangle + b^* = \sum_{i=1}^{\ell} z_i \alpha_i^* \langle y_i, y \rangle + b^*$$

$$\text{Short Sum} = \sum_{i \in SV} z_i \alpha_i^* \langle y_i, y \rangle + b^* =: g(y, \alpha^*, \beta^*)$$

depends only on the scalar product

dual representation

$$\text{Margin: } \gamma = \frac{1}{\|\alpha^*\|_2} = \frac{1}{\sqrt{\sum_{i \in SV} \alpha_i^*}}$$

Soft Margin Optimization

Primal Problem:

$$\begin{array}{ll} \text{minimize (over } \xi, a, b) & \langle a, a \rangle + C \cdot \sum_{i=1}^n \xi_i \\ \text{s.t.} & z_i \cdot (\langle a, y_i \rangle + b) \geq 1 - \xi_i \quad i=1, \dots, n \\ & \xi_i \geq 0 \end{array}$$

$$L(a, b, \xi, \alpha, \tau) = \frac{1}{2} \langle a, a \rangle + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [z_i (\langle a, y_i \rangle + b) - 1 + \xi_i] - \sum_{i=1}^n \tau_i \xi_i$$

$$\frac{\partial L(\dots)}{\partial a} = a - \sum_{i=1}^n z_i \alpha_i y_i \stackrel{!}{=} 0 \Rightarrow a = \sum_{i=1}^n z_i \alpha_i y_i$$

$$\frac{\partial L(\dots)}{\partial \xi_i} = C - \alpha_i - \tau_i \stackrel{!}{=} 0$$

$$\frac{\partial L(\dots)}{\partial b} = - \sum_{i=1}^n z_i \alpha_i \stackrel{!}{=} 0$$

$$L(a, b, \xi, \alpha, \tau) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n z_i z_j \alpha_i \alpha_j \langle y_i, y_j \rangle + (\alpha_i + \tau_i) \sum_{i=1}^n \xi_i - \sum_{i=1}^n \sum_{j=1}^n \alpha_i z_i \alpha_j z_j \langle y_i, y_j \rangle + \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j z_i z_j \langle y_i, y_j \rangle$$

$$:= W(\alpha)$$

Soft Margin Optimization

$$C = \alpha_i + r_i \Rightarrow \alpha_i \leq C$$

$$r_i \geq 0$$

KKT Conditions:

$$\begin{cases} \alpha_i [z_i (\langle a, y_i \rangle + b) - 1 + \xi_i] = 0 \\ (C - \alpha_i) \cdot \xi_i = 0 \end{cases}$$

3 Types of Constraints:

$\alpha_i = 0$	not a support vector
$\alpha_i \in]0, C[$	support vector
$\alpha_i = C \Leftrightarrow r_i = 0 \Leftrightarrow \xi_i \neq 0$	

Dual Problem:

$$\begin{aligned} &\text{maximize } w(\alpha) \\ &\text{s.t. } \sum_{i=1}^n z_i \alpha_i = 0 \\ &0 \leq \alpha_i \leq C \quad i=1, \dots, n \end{aligned}$$

Find solution α^* . $\alpha^* = \sum_{i=1}^n \alpha_i^* z_i y_i$

$$g(y) = \sum_{i=1}^n z_i \alpha_i^* \langle y_i, y \rangle + b^*$$

choose b^* such that $\forall i \in]0, C[$

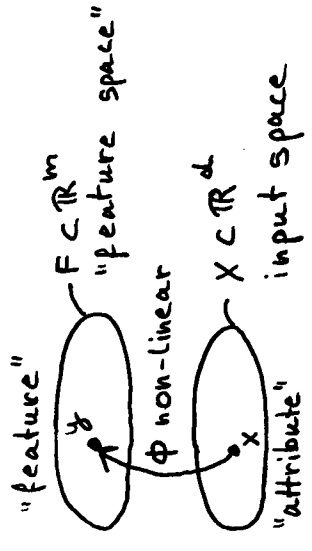
$$z_i \cdot g(y_i) = 1$$

$$\text{Margin: } \delta = \frac{1}{\|\alpha^*\|} = \frac{1}{\sqrt{\sum_{i,j \in S} z_i z_j \alpha_i^* \alpha_j^* \langle y_i, y_j \rangle}}$$

BOX CONSTRAINT

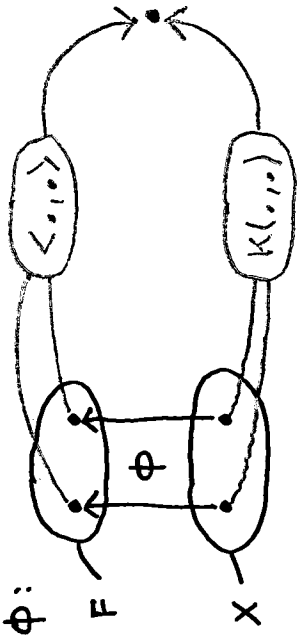
Kernel-Induced Feature Space

Preprocessing:



To improve separability: $m \gg d$

Kernel of ϕ :



$$K: X^2 \rightarrow \mathbb{R}$$

$$(x, z) \mapsto \langle \phi(x), \phi(z) \rangle$$

Define K (ϕ is defined implicitly)

Conditions on K to be kernel of some ϕ ?

Kernel-Induced Feature Spaces

$X = \{x_1, \dots, x_n\}$ Finite
 $K: X^2 \rightarrow \mathbb{R}$ Symmetric
 K kernel \Leftrightarrow Matrix $K := (K(x_i, x_j))_{i,j=1}^n$
 is positive semi-definite

How to build Φ from K :

K symmetric $\Rightarrow \exists$ Orthogonal $K = V \Lambda V^T$

Eigen values of K : λ_t

Eigen vectors of K : $v_t = (v_{ti})_{i=1}^n$

$\Phi(x_i) := (\underbrace{\sqrt{\lambda_t}}_{\geq 0} \cdot v_{ti})_{t=1}^n$

Application to SVM:

$$W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n z_i z_j \alpha_i \alpha_j K(x_i, x_j)$$

$$g(\Phi(x)) = \sum_{i \in S^+} z_i \alpha_i^* \cdot K(x_i, x) + b^*$$

Kernel-Induced Feature Spaces

Examples:

① $\Phi: (x_1, x_2) \mapsto (x_1^2, x_2^2, x_1 x_2)$
 (monomials of degree 2)

② $K(x, z) = \langle x, z \rangle \quad \Phi = \text{Id}$

③ $K(x, z) = \langle x, z \rangle^2$

$n=2$: $\Phi: (x_1, x_2) \mapsto (x_1^2, x_2^2, 2x_1 x_2)$

④ $K(x, z) = (\langle x, z \rangle + C)^d$ \leftarrow parameters
 polynomial kernel

⑤ $K(x, z) = \exp\left(-\frac{\|x-z\|^2}{\sigma^2}\right)$ \leftarrow parameter
 Gauss kernel

Radial Basis Function (RBF) kernel
 translation invariant

parameters too large \rightarrow OVERFITTING