RWTH AACHEN UNIVERSITY

6.4. SVM - soft margins

(P) $\min_{a\in\mathbb{R}^p, b, \xi\in\mathbb{R}} \left\{ \frac{1}{2}\|a\|^2 + c\sum_{i=1}^n \xi_i \right\}$ (c parameters)

s.t. $y_i(a^T x_i + b) \geq 1 - \xi_i$

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

Consequences

$\xi_i > 1 \Rightarrow$ misclassification

$0 < \xi \leq 1 \Rightarrow$ correctly classified, but lies in the margin

(D) $\max_{\lambda} \left\{ \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \lambda_i \lambda_j x_i^T x_j \right\}$

s.t. $0 \leq \lambda_i \leq c, \quad i = 1,\dots,n$

$\sum_{i=1}^n \lambda_i y_i = 0$

Investigating complementary slackness

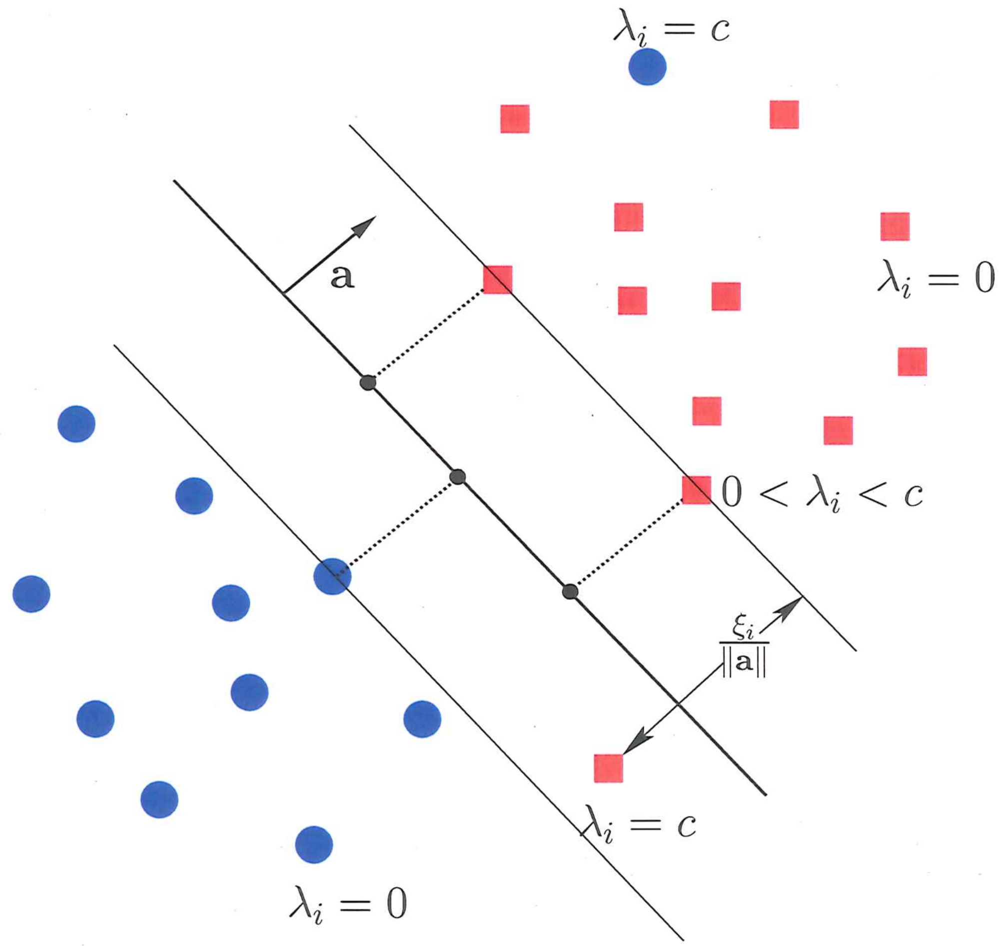
$0 < \lambda_i < c \Rightarrow y_i(a^T x_i + b) = 1$ (margin SVs)

$\lambda_i = c \Rightarrow y_i(a^T x_i + b) = 1 - \xi_i, \ \xi_i \geq 0$ (margin error)

$\lambda_i = 0 \Rightarrow y_i(a^T x_i + b) \geq 1$ (non SVs)

Hence, the solution is sparse, most of the points have $\lambda_i = 0$.

Thinking the Future
Zukunft denken



λi = c
λi = 0
a
0 < λi < c
ξi / ||a||
λi = c
λi = 0

6.5. The SMO-Algorithm

<u>S</u>equential <u>M</u>inimal <u>O</u>ptimization to solve the dual problem.

(D) —
$$\max_{\lambda} W(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j x_i^T x_j$$
$$\text{s.t. } 0 \le \lambda_i \le c$$
$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

Assume λ is a feasible point, i.e., λ satisfies the constraints.

Idea: "cyclic coordinatewise optimization" does not work, since, e.g.,
$$\lambda_1 y_1 = -\sum_{i=2}^{n} \lambda_i y_i \quad \text{or} \quad \lambda_1 = -y_1 \sum_{i=2}^{n} \lambda_i y_i$$
Hence each λ_j is determined by fixing $\lambda_i, i \ne j$.

Idea: update at least two λ_j simultaneously $\Rightarrow$ SMO alg.

(SMO) – <u>repeat</u>

1. Select a pair (i,j) to update next (the one which promises most progress)
2. Optimize $W(\lambda)$, w.r.t. λ_i and λ_j while holding λ_k, $k \ne i,j$, fixed

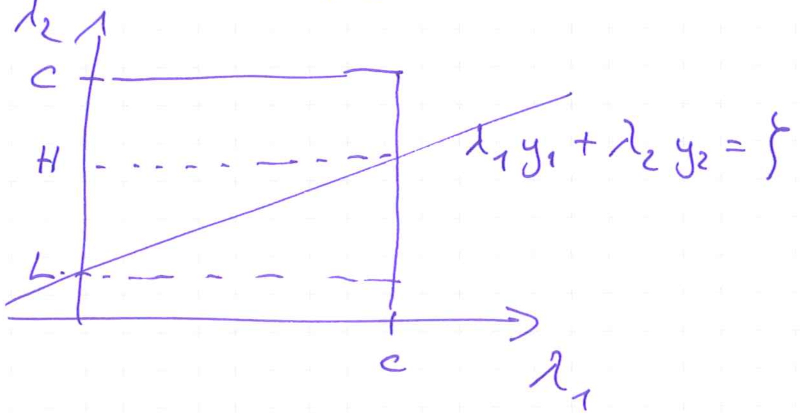
until convergence

Check KKT within a tolerance limit $\varepsilon = 0.01$ or 0.001, to verify convergence.

Optimize $W(\lambda)$ w.r.t. λ_1, λ_2 with $\lambda_3, \dots, \lambda_4$ fixed, λ feasible.

It holds

$$\lambda_1 y_1 + \lambda_2 y_2 = -\sum_{i=3}^{4} \lambda_i y_i = \zeta \ , \quad \zeta \text{ fixed.}$$



Derive $\quad 0 \leq \lambda_1, \lambda_2 \leq c \qquad (*)$

$\qquad\qquad L \leq \lambda_2 \leq H$

Moreover $\quad \lambda_1 = (\zeta - \lambda_2 y_2) y_1 \qquad (**)$

Hence: $\quad W(\lambda_1, \dots, \lambda_4) = W\big((\zeta - \lambda_2 y_2) y_1, \lambda_2, \underbrace{\lambda_3, \dots, \lambda_4}_{\text{fixed}}\big)$

is a quadratic function of λ_2.

It can be written as $\qquad$ (remember: $y_i \in \{-1, 1\}$)

$$\gamma_2 \lambda_2^2 + \gamma_1 \lambda_2 + \gamma_0 \quad , \ \gamma_0, \gamma_1, \gamma_2 \text{ appropriate}$$

Determine the max. by differentiation.

$$2\gamma_2 \lambda_2 + \gamma_1 \stackrel{!}{=} 0 \qquad \lambda_2 = -\frac{\gamma_1}{2\gamma_2}$$

with optimum solution $\lambda_2^{(r)}$ $\qquad$ (r: raw)

The final solution of (*) is

$$\lambda_2^{(c)} = \begin{cases} H, & \text{if } \lambda_2^{(r)} \geq H \\ \lambda_2^{(r)}, & \text{if } L \leq \lambda_2^{(r)} \leq H \\ L, & \text{if } \lambda_2^{(r)} < L \end{cases}.$$ (c: clipped)

λ_1 is computed from (**).

Still to clarify:

- What is the best choice of the next pair (i,j) to update.
- How to update the coefficients $\gamma_0, \gamma_1, \gamma_2$ in the run of SMO.
- The algorithm converges, however, the right choice of (i,j) in each step accelerates the rate of convergence.
- Osuna, Freund, Girosi (1997): generalization of SMO algorithm.

6.6. Kernels

Instead of applying SVM to the raw data ("attributes") x_i apply it to transformed data ("features") $\phi(x_i)$.

ϕ is called feature mapping.

Aim: achieve better separability.

$$(1) \quad \max_{\lambda} \; g(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^T x_j$$

$$\text{s.t.} \quad 0 \le \lambda_i \le c$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$g(\lambda)$ only depends on the inner products $x_i^T x_j$.

Substitute x_i by $\Phi(x_i)$ and use some inner product $\langle \cdot , \cdot \rangle$. Replace $x_i^T x_j$ by

$$\langle \Phi(x_i), \Phi(x_j) \rangle = K(x_i, x_j)$$

Remark: $K(x,y)$ is often easier to compute than $\Phi(x)$ itself.

Intuition: If $\Phi(x), \Phi(y)$ are close $\langle \Phi(x), \Phi(y) \rangle$ is large.

If $\Phi(x) \perp \Phi(y)$ then $\langle \Phi(x), \Phi(y) \rangle = 0$. Hence $K(x,y)$ measures how similar x and y are.

Needed: an inner product in some feature space $\{\Phi(x) \mid x \in \mathbb{R}^p\}$.

RWTH AACHEN UNIVERSITY

Example. 6.2.

$x, z \in \mathbb{R}^p$, $K(x,z) = \langle x, z\rangle^2 = \left(\sum_{i=1}^{p} x_i z_i\right)^2$

Question. Is there some ϕ such that $\langle x,z\rangle^2$ is an inner product in the feature space.

$p=2$: $x = (x_1, x_2)^T$, $z = (z_1, z_2)^T$

Use $\phi(x) = (x_1^2, x_2^2, x_1x_2, x_2x_1) : \mathbb{R}^2 \to \mathbb{R}^4$

$$\begin{aligned}\langle \phi(x), \phi(z)\rangle &= x_1^2 z_1^2 + x_2^2 z_2^2 + x_1x_2z_1z_2 + x_2x_1z_2z_1 \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 = \langle x,z\rangle^2 \quad \lrcorner\end{aligned}$$

Example. 6.3. (Gaussian Kernel)

$x, z \in \mathbb{R}^p$, $K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$

Question: $\exists$ feature mapping ϕ and a feature space with $\langle \cdot, \cdot \rangle$?

Def. 6.4. Kernel $K(x,z)$ is called <u>valid</u>, if there exists a feature function ϕ such that

$$K(x,z) = \langle \phi(x), \phi(z)\rangle \text{ for all } x, z \in \mathbb{R}^p. \quad \lrcorner$$

Thinking the Future
Zukunft denken

Theorem 6.5. (Mercer)

Given $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. K is a valid kernel if and only if for any $\{x_1, \ldots, x_n\}$ the kernel matrix

$$\left(K(x_i, x_j)\right)_{i,j=1,\ldots,n} \quad \text{is n.n.d.}$$

$\perp$

Proof. "$\Longrightarrow$"

$$K \text{ valid} \Rightarrow \exists \phi : K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \langle \phi(x_j), \phi(x_i) \rangle = K(x_j, x_i)$$

Moreover,

$$z^T \left(K(x_i, x_j)\right)_{ij} z = z^T \left(\langle \phi(x_i), \phi(x_j) \rangle\right)_{ij} z$$

$$= \sum_{k,l} z_k z_l \langle \phi(x_k), \phi(x_l) \rangle$$

$$= \left\langle \sum_k z_k \phi(x_k), \sum_l z_l \phi(x_l) \right\rangle \geq 0$$

Example 6 (Polynomial kernel)

$$K(x, z) = (x^T z + c)^d, \quad x, z \in \mathbb{R}^p, c \in \mathbb{R}, d \in \mathbb{N}, d \geq 2.$$

Feature space of dim $\binom{p+d}{d}$ containing all monomials of degree $\leq d$.

Determine $\phi(x) \to$ Ex. $\perp$

RWTH AACHEN UNIVERSITY

Kernels can also be constructed over infinite dimensional spaces, e.g., function space or probability distributions.
Provides a lot of modelling power.
The solution of the opt problem is still a convex problem with linear constraints.

Thinking the Future
Zukunft denken