given $(x_1,y_1),\dots,(x_n,y_n)$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1,1\}$

- (OMC) $\min_{a\in\mathbb{R}^p, b\in\mathbb{R}} \frac{1}{2}\|a\|^2$

  s.t. $y_i(a^T x_i + b) \geq 1 \quad \forall i=1,\dots,n$

## 6.3 SVM and Lagrange Duality

- (P) $\min f_0(x)$

  s.t. $f_i(x) \leq 0, \quad i=1,\dots,m$

  $h_i(x) = 0, \quad i=1,\dots,\cancel{q}\,r$

  $f_0, f_i$ convex, $h_i$ linear

- Lagrangian:

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{\cancel{q}\,r} \nu_i h_i(x)$$

- Lagr. dual

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

(D) $\max g(\lambda,\nu)$

s.t. $\lambda_i \geq 0$

- Weak duality:

$$g(\lambda^*,\nu^*) \leq f_0(x^*)$$

$\lambda^*, \nu^*, x^*$ opt. solutions

- Strong duality:

$$g(\lambda^*,\nu^*) = f_0(x^*)$$

- Slater's condition $\Rightarrow$ strong duality

  $f_i$ linear $\Rightarrow$ Slater's cond. holds

[RECAP]

- Karush-Kuhn-Tucker conditions (KKT)

1. $f_i(x) \leq 0$, $i = 1, \dots, m$ (primal constraints)
   $h_i(x) = 0$, $i = 1, \dots, r$
2. $\lambda \geq 0$ (dual constraints)
3. $\lambda_i f_i(x) = 0$ (complementar slackness)
4. $\nabla_x L(x, \lambda, \nu) = 0$

Th. 6.1. If Slater's condition ~~a~~ is satisfied (which is the case if the constraints are affine) then strong duality holds.
If in addition $f_i, h_i$ are differentiable then for $x^*$, $(\lambda^*, \nu^*)$ to be primal and dual optimal it is necessary and sufficient that the KKT conditions holds.

Application to SVM

Given training set $\{(x_1, y_1), \dots, (x_4, y_4)\}$

$$x_i \in \mathbb{R}^p, \quad y_i \in \{-1, 1\}$$

(P)
$$\min_{a \in \mathbb{R}^p,\, b \in \mathbb{R}} \quad \frac{1}{2}\|a\|^2$$
$$\text{s.t.} \quad y_i(a^T x_i + b) \geq 1, \quad i = 1, \dots, 4$$

Langrangian:

$$L(a, b, \lambda) = \frac{1}{2}\|a\|^2 - \sum_{i=1}^{4} \lambda_i \left(y_i(a^T x_i + b) - 1\right)$$

$$\frac{\partial}{\partial a} L(a, b, \lambda) = a - \sum_{i=1}^{4} \lambda_i y_i x_i \overset{!}{=} 0$$

$$\Rightarrow \quad a^* = \sum_{i=1}^{4} \lambda_i y_i x_i$$

$$\frac{d}{db} L(a, b, \lambda) = \sum_{i=1}^{4} \lambda_i y_i \overset{!}{=} 0$$

$$\Rightarrow \quad \sum_{i=1}^{4} \lambda_i y_i = 0$$

Dual function:

$$g(\lambda) = L(a^*, b^*, \lambda) = \frac{1}{2}\|a^*\|^2 - \sum_{i=1}^{4} \lambda_i \left(y_i(a^{*T} x_i + b^*) - 1\right)$$

$$= \sum_{i=1}^{4} \lambda_i + \frac{1}{2}\left(\sum_{i=1}^{4} \lambda_i y_i x_i\right)^T \left(\sum_{i=1}^{4} \lambda_i y_i x_i\right)$$

$$- \sum_{i=1}^{4} \lambda_i y_i \left(\sum_{j=1}^{4} \lambda_j y_j x_j\right)^T x_i - \underbrace{\sum_{i=1}^{4} \lambda_i y_i}_{=0} b^*$$

$$= \sum_{i=1}^{4} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j x_i^T x_j$$

Thinking the Future
Zukunft denken

Dual problem

(D) $\max_{\lambda} g(\lambda) = \sum_{i=1}^{4} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j x_i^T x_j$

s.t. $\lambda_i \geq 0$

$\sum_{i=1}^{4} \lambda_i y_i = 0$

If $\lambda_i^*$ is the solution of (D), then $a^* = \sum_{i=1}^{4} \lambda_i^* y_i x_i$

and $b^* = y_k - a^{*T} x_k$, $x_k$ some supp. vector.

Slater's condition is satisfied, strong duality holds.

Complementary slackness (from KKT for opt. $\lambda^*$):

$$\lambda_i^* \left( y_i (a^{*T} x_i + b^*) - 1 \right) = 0 \quad , i = 1, \dots, 4$$

Hence

$$\lambda_i^* > 0 \Rightarrow y_i (a^{*T} x_i + b^*) = 1$$

$$\lambda_i^* = 0 \Rightarrow y_i (a^{*T} x_i + b^*) \geq 1$$

$\lambda_i^* > 0$ indicates supporting vectors, those which have smallest distance to the separating hyperplane.

Let $\mathcal{S} = \{ i \mid \lambda_i^* > 0 \}$, $\mathcal{S}_+ = \{ i \in \mathcal{S} \mid y_i = +1 \}$

$\mathcal{S}_- = \{ i \in \mathcal{S} \mid y_i = -1 \}$

Then $a^* = \sum_{i \in \mathcal{S}} \lambda_i^* y_i x_i$

$[\, b^* = -\frac{1}{2} a^{*T} (x_k + x_\ell)$ where $k \in \mathcal{S}_+$, $\ell \in \mathcal{S}_- \,]$

Application to SVM:

- Training set $\{ (x_1, y_1), \dots, (x_n, y_n) \}$
- Determine $\lambda^*$, $a^*$, $b^*$
- New point $x$. Find class label $y \in \{-1, 1\}$.

Compute $a^{*T} x + b^* = \left( \sum_{i \in \mathcal{S}} \lambda_i^* y_i x_i \right)^T x + b^*$

$$= \sum_{i \in \mathcal{S}} \lambda_i^* y_i x_i^T x + b^* = d(x)$$

Predict $y = 1$, if $d(x) \geq 0$, otherwise $y = -1$.

Remarks:

a) $|\mathcal{S}|$ is normally much less than $n$.

b) The decision only depends on the inner products $x_i^T x$ for support-vector $x_i$, $i \in \mathcal{S}$.

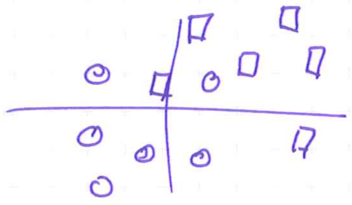
RWTH AACHEN UNIVERSITY

## 6.4. Non-Separability and Robustness

By now: assumption that $\exists$ separating hyperplane.

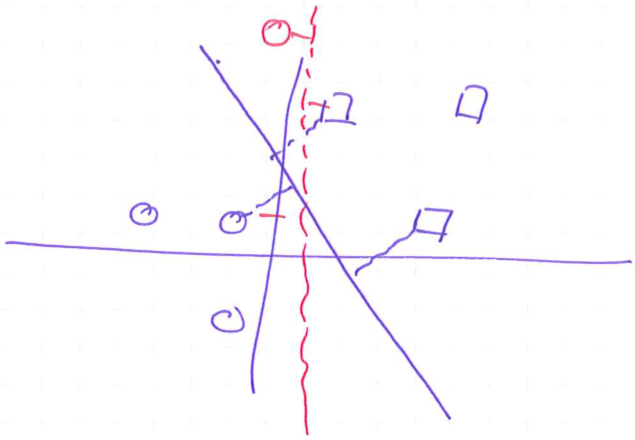
What happens if not?

Example:



Points are not linearly separable.

The optimum margin classifier is sensitive to outliers.



Outlier causes a drastic swing of the OMC.

Both problems are addressed by the following approach: $\ell_1$ - regularization

(P)

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

s.t. $$y_i(a^T x_i + b) \geq 1 - \xi_i, \quad i = 1,\dots,n$$

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

For the optimal solution $a^*, b^*$

It is allowed that margins are less than $\frac{1}{\|a^*\|}$, i.e.

$$y_i({a^*}^T x_i + b^*) \leq 1.$$

Thinking the Future
Zukunft denken

If $y_i(a^{*T}x_i + b^*) = 1 - \xi_i$, $\xi_i > 0$,
then a cost of $c\xi_i$ is paid.
Parameter $c$ controls the balance between the two goals in (P).

Lagrangian for (P):

$$L(a, b, \xi, \lambda, \gamma) = \frac{1}{2}\|a\|^2 + c\sum_{i=1}^{n} \xi_i$$
$$- \sum_{i=1}^{n} \lambda_i \left(y_i(a^T x_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{n} \gamma_i \xi_i$$

$\lambda, \gamma$ are Lagrangian multipliers.

Analogously to the above obtain the dual problem:

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

Let $\lambda_i^*$ be the optimum solution of (D). As before:
Let $\mathcal{S} = \{i \mid \lambda_i^* > 0\}$ (determines the support vectors)
Then $a^* = \sum_{i \in \mathcal{S}} \lambda_i^* y_i x_i$ is the optimum $a$.

Complementary slackness conditions are:

$$\lambda_i = 0 \Rightarrow y_i(a^{*T}x_i + b^*) \ge 1$$
$$\lambda_i = c \Rightarrow y_i(a^{*T}x_i + b^*) \le 1$$
$$0 < \lambda_i < c \Rightarrow y_i(a^{*T}x_i + b^*) = 1$$

RWTH AACHEN UNIVERSITY

If $0 < \lambda_k < c$ for some $k$ ($x_k$ a support vector)
then $b^* = y_k - a^{*T} x_k$ is opt. $b$.

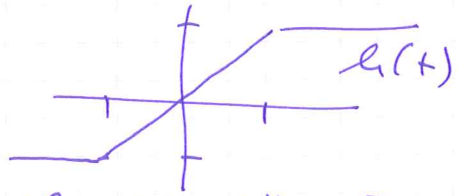
To classify a new point $x \in \mathbb{R}^p$:

Compute $a^{*T}x + b^* = \left(\sum_{i \in S} \lambda_i^* y_i x_i\right)^T x + b^*$

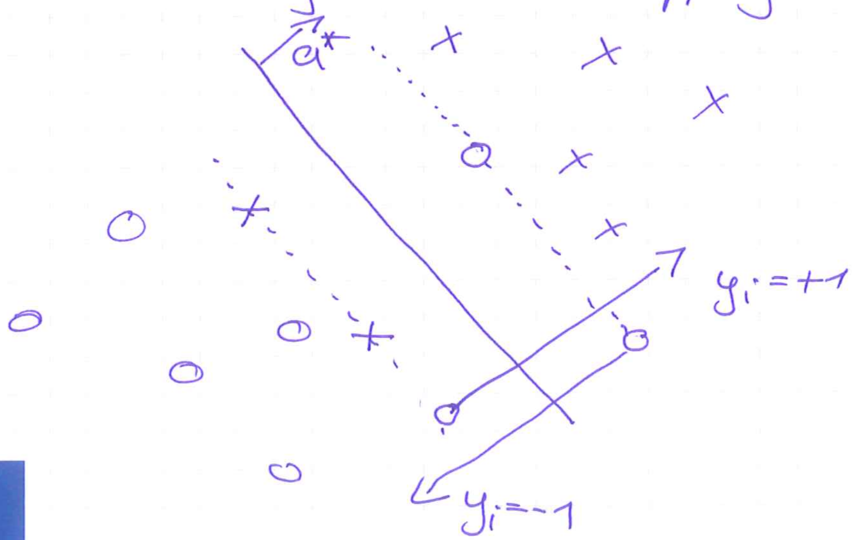
$$= \sum_{i \in S} \lambda_i^* y_i \, x_i^T x + b^* = d(x)$$

- Hard classifier:
  Decide $y = 1$ if $d(x) \geq 0$, otherwise $y = -1$.

- Soft classifier:
  $$d(x) = h\left(a^{*T}x + b^*\right) \text{ where } h(t) = \begin{cases} -1, & t < -1 \\ t, & -1 \leq t \leq +1 \\ +1, & t > 1 \end{cases}$$




$d(x)$ a real no in $[-1, +1]$ if $a^{*T}x + b^* \in [-1, 1]$,
if $x$ is residing in the overlapping area.




Thinking the Future
Zukunft denken

Both classifiers only depend on the inner products $x_i^T x = \langle x_i, x \rangle$ with support vector $x_i$, $i \in S$.