6. Support Vector Machines (SVM)

(introduced 1992 by Boser, Guyon, Vapnik)

- SVM's are supervised learning methods for classification, regression, outlier detection, ...
- SVMs one of the best "off-the-shelf" methods.
- Very flexible by the use of kernels.
- SMO (sequ. min. opt.) is an efficient implementation.
- Effective in high-dim. spaces.
- Effective if no. of dim. > no. of samples
- Uses a subset of points in the decision fct. (called support vectors).

Given a training set $(x_1, y_1), \dots, (x_n, y_n)$

$x_i \in \mathbb{R}^p$: data points

$y_i \in \{-1, +1\}$: class membership (2 classes/groups)

Assume that the sets are linearly separable, i.e., there exists a hyperplane H s.t.

$\{x_i \mid y_i = 1\}$ and $\{x_i \mid y_i = -1\}$ are separable by H.

Which H to choose?

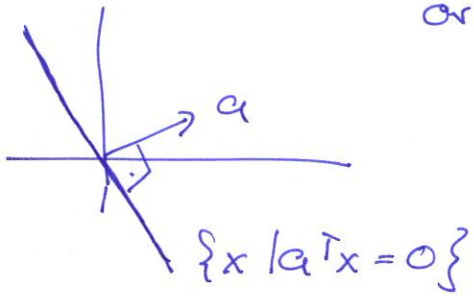
→ The one which maximizes the min. distance from the hyperplane.

6.1. Hyperplanes and Margins

Representing hyperplanes in $\mathbb{R}^p$:

a) Given $a \in \mathbb{R}^p$

$\{x \in \mathbb{R}^p \mid a^T x = 0\}$ is the $(p-1)$-dim. space orth. to a.



b) Given $a \in \mathbb{R}^p$, $b \in \mathbb{R}$. Consider

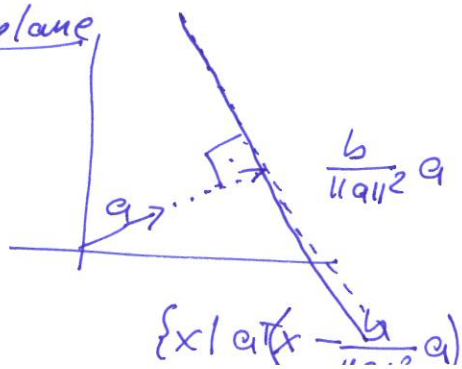
$\{x \in \mathbb{R}^p \mid a^T x - b = 0\}$

It holds

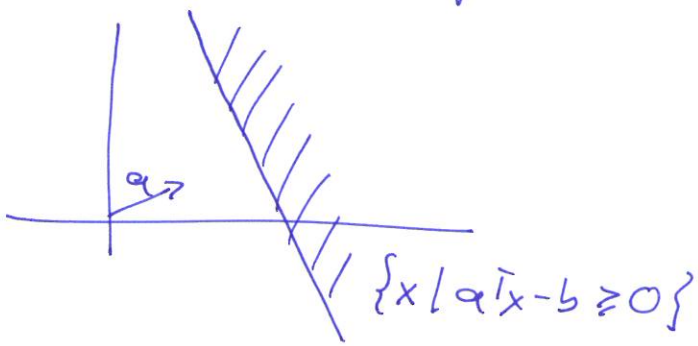
$$a^T x - b = 0 \Leftrightarrow a^T x - \underbrace{\frac{a^T a}{\|a\|^2}}_{=1} b = 0$$

$$\Leftrightarrow a^T \left(x - \frac{b}{\|a\|^2} a \right) = 0$$

Hence: $\{x \mid a^T x - b = 0\}$ is a linear space shifted by $\frac{b}{\|a\|^2} a$, a hyperplane



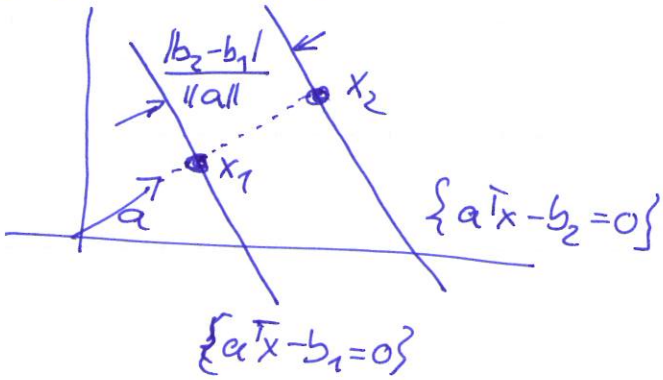
$\{x \mid a^T x \gtrless b\}$ is called a half-space:



c) Given $a \in \mathbb{R}^p$, $b_1, b_2 \in \mathbb{R}$

Distance between $H_1 = \{a^T x - b_1 = 0\}$ and

$H_2 = \{a^T x - b_2 = 0\}$



Pick x_1 and x_2

$x_1 = \lambda_1 a$ $\qquad$ $x_2 = \lambda_2 a$

$a^T x_1 - b_1 = 0$ $\qquad$ $a^T x_2 - b_2 = 0$

Then

$\lambda_1 a^T a - b_1 = 0$

$\lambda_1 \|a\|^2 - b_1 = 0$

$\lambda_1 = \frac{b_1}{\|a\|^2}$ $\qquad$ $\lambda_2 = \frac{b_2}{\|a\|^2}$

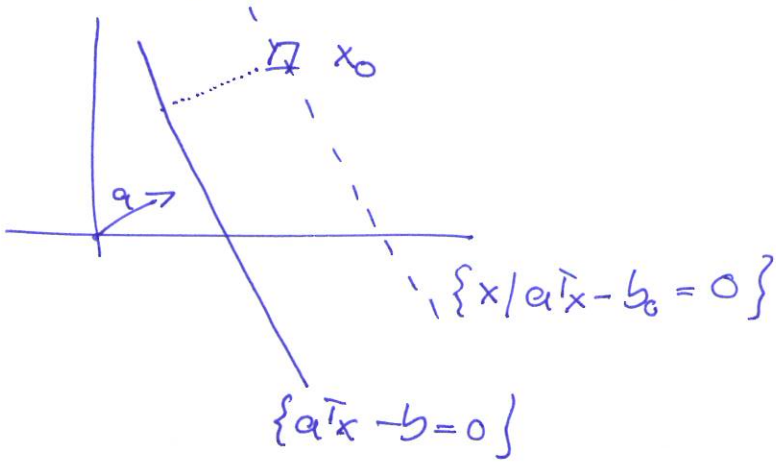
$\|x_2 - x_1\| = \|\lambda_2 a - \lambda_1 a\| = |\lambda_2 - \lambda_1| \|a\|$

$= \left\| \frac{b_2}{\|a\|^2} - \frac{b_1}{\|a\|^2} \right\| \|a\| = \underbrace{\frac{1}{\|a\|} |b_2 - b_1|}$

distance between the hyperplanes

d) Given $a \in \mathbb{R}^p, b \in \mathbb{R}, x_0 \in \mathbb{R}^p$

Distance between $H = \{x / a^T x - b = 0\}$ and point x_0.



Consider auxiliary hyperplane containing x_0:

$H_0 = \{x / a^T x - b_0 = 0\} = \{x / a^T x - a^T x_0 = 0\}$

$(b_0 = a^T x_0$ since $a^T x_0 - b_0 = 0\}$

By c), the distance between H and H_0 is

$$\frac{1}{\|a\|} |b - a^T x_0| .$$

This distance is called <u>margin of x_0</u>.

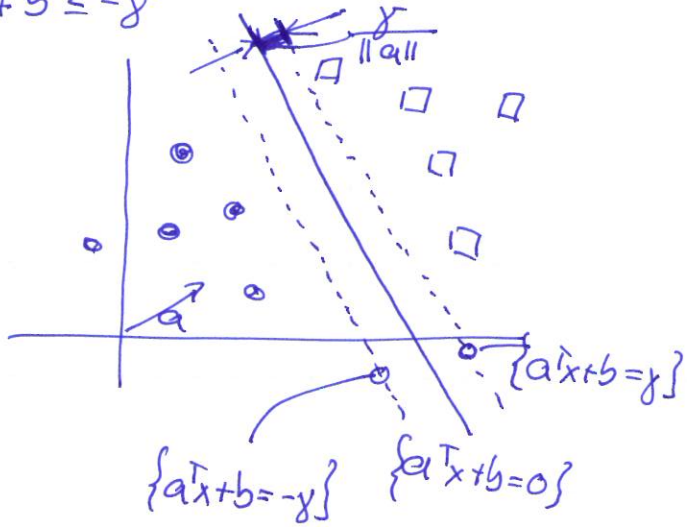
6.2. The optimal margin classifier

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

Assume there exists a separating hyperplane.

$$\{x \mid a^T x + b = 0\}.$$

Then $y_i = +1 \Rightarrow a^T x_i + b \geq \gamma$
$y_i = -1 \Rightarrow a^T x_i + b \leq -\gamma$
for some $\gamma \geq 0$



Hence:

$y_i (a^T x_i + b) \geq \gamma$
for some $\gamma \geq 0$
for all $i = 1, \dots, n$

Objective: Find a hyperplane $\{x \mid a^T x + b = 0\}$ such that the min. margin is maximum.

$$\max_{\gamma, a, b} \frac{\gamma}{\|a\|} \quad \text{s.t. } y_i (a^T x_i + b) \geq \gamma \ \forall i = 1, \dots, n$$
$$\gamma \geq 0, a \in \mathbb{R}^p, b \in \mathbb{R}$$

(not scale invariant, with γ, a, b also $2\gamma, 2a, 2b$ is a solution)

$\Leftrightarrow$ $\min_{\gamma, a, b,\ \gamma \geq 0,\ a \in \mathbb{R}^p,\ b \in \mathbb{R}} \left\| \frac{a}{\gamma} \right\|$ s.t. $y_i \left(\left(\frac{a^T}{\gamma} \right) x_i + \frac{b}{\gamma} \right) \geq 1$

$\Leftrightarrow$ $\min_{a \in \mathbb{R}^p,\ b \in \mathbb{R}} \|a\|$ s.t. $y_i (a^T x + b) \geq 1$

$\Leftrightarrow$ $\min_{a, b} \frac{1}{2} \|a\|^2$ s.t. $y_i (a^T x + b) \geq 1$

In summary:

(OMC)

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

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

Quadr. opt. problem with linear constraints.

Special case of a convex opt. problem.

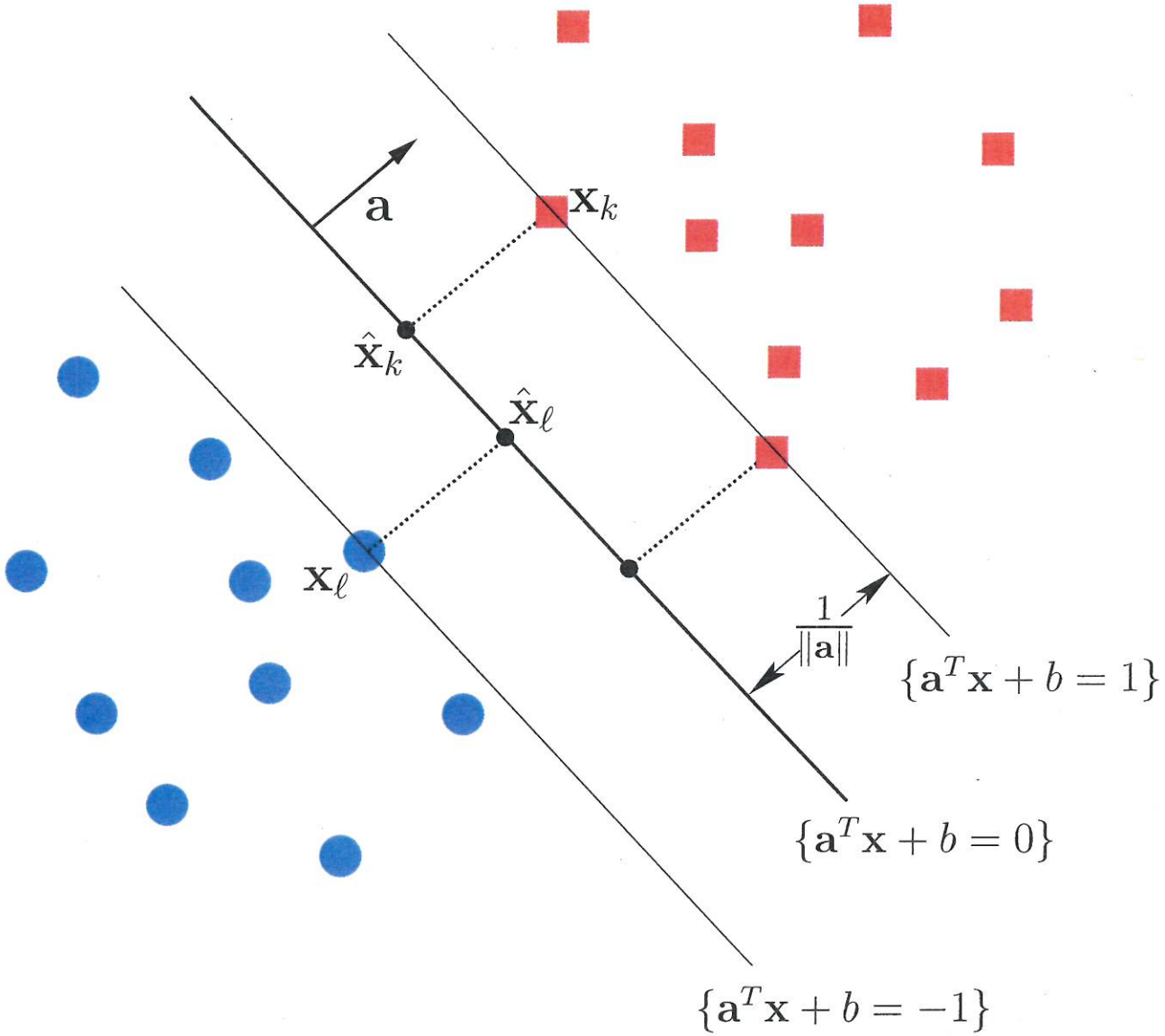
- Assume that a^* is an optimum solution of (OMC) and x_k some point with minimum margin. Then

$$y_k (a^{*T} x_k + b^*) = 1$$
$$\Leftrightarrow a^{*T} x_k + b^* = y_k \quad (\text{since } y_k^2 = 1)$$
$$\Leftrightarrow b^* = y_k - a^{*T} x_k$$

Hence, $b^* = y_k - a^{*T} x_k$ is the opt. b-value.

- The (a^*, b^*) is called <u>optimal margin classifier</u>.
- Use commercial or public domain generic quadratic programming (QP) to solve OMC.

Fig 4


a
$\mathbf{x}_k$
$\hat{\mathbf{x}}_k$
$\hat{\mathbf{x}}_\ell$
$\mathbf{x}_\ell$
$\frac{1}{\|\mathbf{a}\|}$
$\{\mathbf{a}^T\mathbf{x} + b = 1\}$
$\{\mathbf{a}^T\mathbf{x} + b = 0\}$
$\{\mathbf{a}^T\mathbf{x} + b = -1\}$