Application to SVM:

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

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

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

Lagrangian:

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

$$\nabla_a L(a,b,\lambda) = a - \sum_{i=1}^n \lambda_i y_i x_i = 0$$

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

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

Dual function

$$g(\lambda) = L(a^*, b^*, \lambda)$$

$$= \frac{1}{2}\|a^*\|^2 - \sum_{i=1}^n \lambda_i (y_i (a^{*T} + b^*) - 1$$

$$= \sum_{i=1}^n \lambda_i + \frac{1}{2}\Big(\sum_i \lambda_i y_i x_i^T\Big)\Big(\sum_i \lambda_i y_i x_i\Big)$$

$$- \sum_i \lambda_i y_i \Big(\sum_j \lambda_j y_j x_j\Big)^T x_i - \underbrace{\sum_i \lambda_i y_i b^*}_{=0}$$

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

Dual problem:

$$(D)\quad \left|\begin{array}{l} \max\limits_{\lambda} \left\{ g(\lambda) = \sum\limits_{i=1}^{4} \lambda_i - \frac{1}{2} \sum\limits_{i,j} y_i y_j \lambda_i \lambda_j x_i^T x_j \right\} \\ \text{s.t.} \quad \lambda_i \geq 0 \ , \ i = 1, \dots, 4 \\ \qquad \sum\limits_{i=1}^{4} \lambda_i y_i = 0 \end{array}\right.$$

If λ_i^* is the optimum of (D), then $a^* = \sum\limits_{i=1}^{4} \lambda_i^* y_i x_i$

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

Slater's condition is satisfied, strong duality holds.

Complementary slackness follows from KKT:

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

Hence,

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

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

$\lambda_i^* > 0$ for the support points, those which have smallest distance ~~zero~~ to the sep. 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 $k \in \mathcal{S}_+$, $\ell \in \mathcal{S}_-$ (*) Ex

Application to SVM:

- Training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$
- Determine λ^*, a^*, b^* from (D) and (*)
- 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 $y = -1$.

Remark:

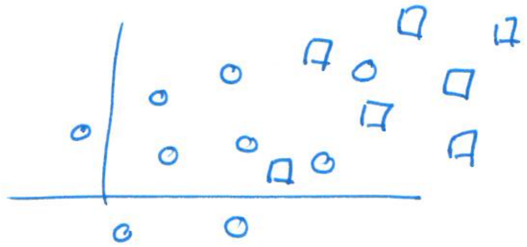
a) $|\mathcal{S}|$ is much smaller than n.

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

6.4. Non-Separability and Robustness

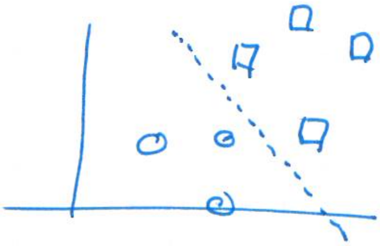
Assumption by now: there is a sep. hyperplane.
What happens if not.

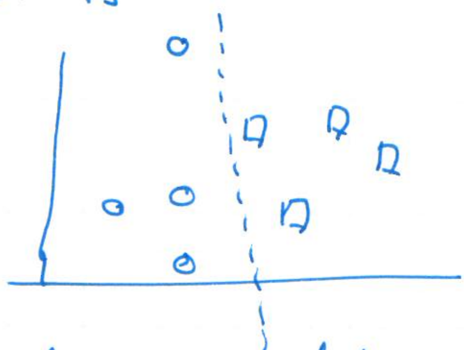
Ex.



no sep. hyperplane

Ex. Sensitivity to outliers





Outlier causes a drastic swing of the sep. hyperplane.
Both problems are addressed by the foll. approach:
l_1 - regularization:

$$(P) \quad \left| \begin{array}{l} \min\limits_{a,b} \frac{1}{2}\|a\|^2 + c \sum\limits_{i=1}^{n} \xi_i \\ \text{s.t. } y_i(a^T x_i + b) \geq 1 - \xi_i, \; i=1,..,n \\ \qquad \xi_i \geq 0, \; i=1,..,n \end{array} \right.$$

For the optimal a^*, b^*
Admitted that margins are less than $\frac{1}{\|a^*\|}$,
i.e., $y_i(a^{*T}x_i + b^*) \leq 1$.
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 penalty.

Lagrangian for (P):

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

λ, γ are the Lagrangian multipliers.

Analogous to the above obtain the dual problem

$$(D) \quad \left| \begin{array}{l} \max\limits_{\lambda} \ \sum\limits_{i=1}^{n}\lambda_i - \frac{1}{2}\sum\limits_{i,j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^T x_j \\ \text{s.t.} \quad 0 \le \underline{\lambda_i \le c}, \ i = 1,\dots,n \\ \qquad \sum\limits_{i=1}^{n}\lambda_i y_i = 0 \end{array} \right.$$

new

Let λ_i^* be the optimal solution of (D)

Let $S = \{i \mid \lambda_i^* > 0\}$. Then

$$a^* = \sum_{i\in S}\lambda_i^* y_i x_i \quad \text{is the optimum } a.$$

Complementary slackness.

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

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

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

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

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

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

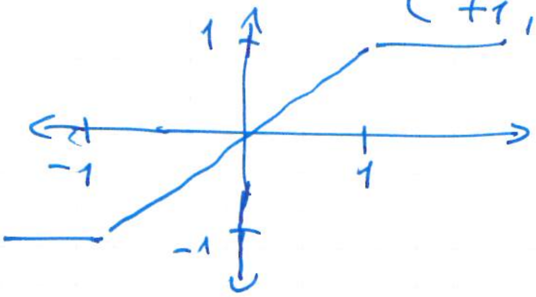
$$= \sum_{i \in S} \lambda_i^* y_i x_i^T x + b^* = d(x)$$

- Hard classifier:

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

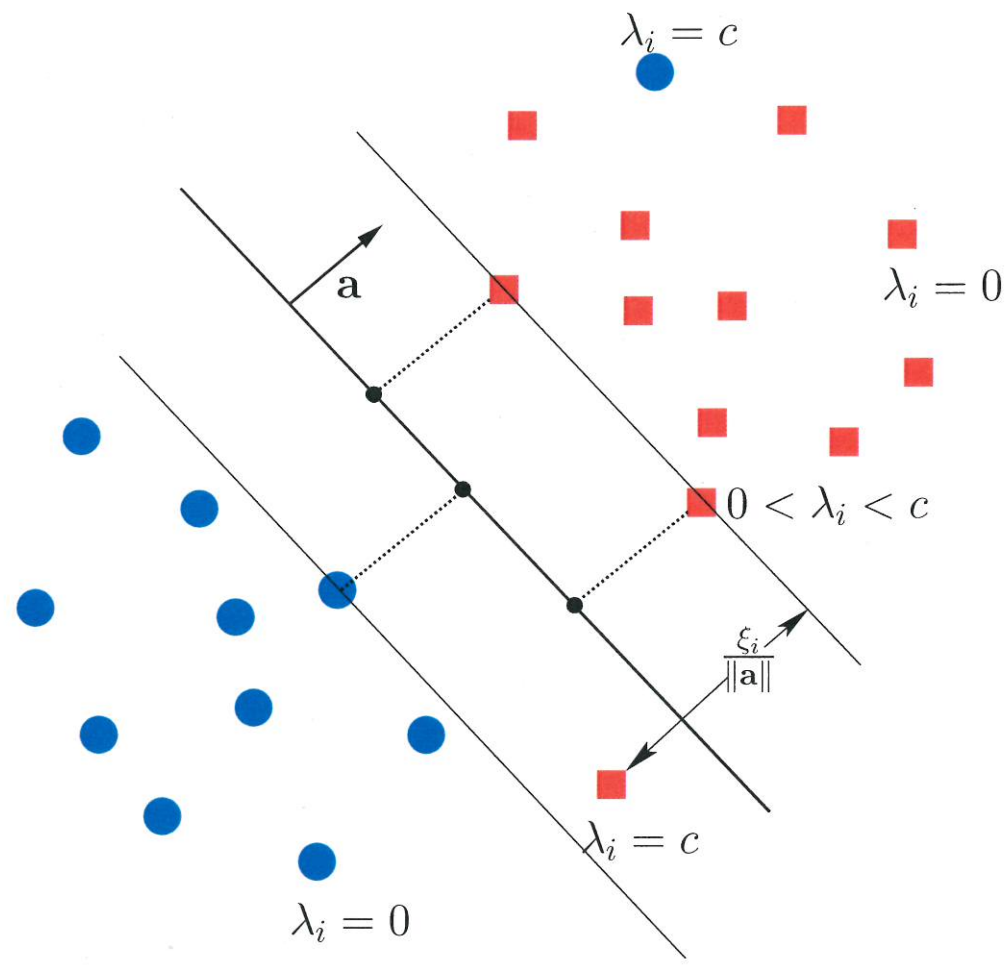
- Soft classifier

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



$d(x)$ is a real number in $[-1, 1]$ if $a^{*T}x + b^* \in [-1, 1]$, if x is residing in the "overlapping" area.

Again: the decision only depends on the inner products between x_i, $i \in S$, and x.



$\lambda_i = c$
$\mathbf{a}$
$\lambda_i = 0$
$0 < \lambda_i < c$
$\frac{\xi_i}{\|\mathbf{a}\|}$
$\lambda_i = c$
$\lambda_i = 0$