6.3 SVM and Lagrange Duality

- Convex optimization problem:

$$(P) \quad \min f_0(x)$$
$$\text{s.t. } f_i(x) \le 0, \; i = 1, \dots, m$$
$$h_i(x) = 0, \; i = 1, \dots, r$$

f_0, f_i are convex, h_i are linear

- Lagrangian (prime function):

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

- Lagrangian dual function:

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

$$\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom}(f_i) \cap \bigcap_{i=1}^{r} \operatorname{dom}(h_i)$$

is a concave function.

- Lagrangian dual problem

$$(D) \quad \max g(\lambda, \nu)$$
$$\text{s.t. } \lambda \ge 0.$$

- Weak duality theorem

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

λ^*, ν^* opt. solutions of (D), x^* opt. sol. of (P)

- Strong duality:

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

- If the constraints are linear then "Slater's condition" holds which implies that $g(\lambda^*, \nu^*) = f_0(x^*)$, "strong duality holds" or "the duality gap is zero".

- Karush-Kuhn-Tucker conditions (KKT)

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

Th. 6.1. If Slater's condition is satisfied (which is the case for linear constraints) then strong duality holds. If f_i, h_i are differentiable then for x^*, (λ^*, ν^*) to be primal and dual optimal it is nec. & suff. that the KKT conditions hold.

Application to SVM.

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

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

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

Lagrangian:

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

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

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

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

Dual function

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

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

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

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

Dual problem

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

$$\text{s.t.} \quad \lambda_i \geq 0, \; i = 1, \dots, N$$

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

If λ_i^* is the optimum ~~value~~ argument of (D), then $a^* = \sum_{i=1}^{N} \lambda_i^* y_i x_i$ and

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

Complementary slackness

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

Hence

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

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

$\lambda_i^* > 0$ for the support vectors.

After having solved (D) the support vectors are known.

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

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

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

$b^* = ~~y_k (a^{*T} x_k + b^*)~~ = y_k - a^{*T} x_k$

for some k such that $\lambda_k^* > 0$, i.e., $k \in S$

Application:

- Training set $(x_1, y_1) \dots (x_4, y_4)$
- Determine λ^* and a^*, b^* from (D) and above
- New point x. Find class label $y \in \{-1, +1\}$.
  Compute $a^{*T}x + b^* = \left(\sum_{i \in S} \lambda_i^* y_i x_i\right)^T x + b^*$

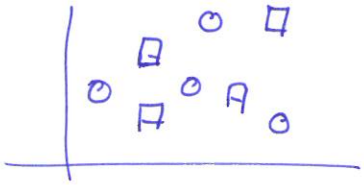
$$= \sum_{i \in S} \lambda_i^* y_i x_i^T x + b^* = d(x)$$

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

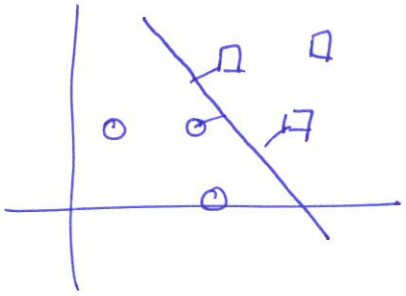
Remarks:

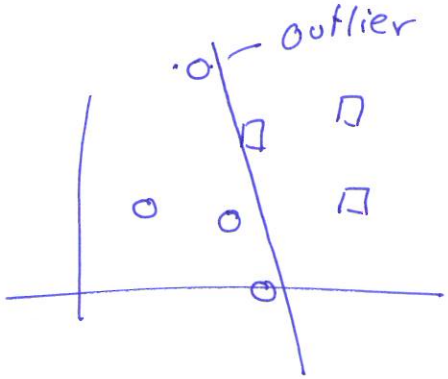
- $|S|$ is normally much ~~smaller~~ less than n.
- Decision only depends on $x_i^T x$, the inner products of with support vectors, $i \in S$.

6.4. Non-separability and Robustness



no linear separation





ℓ_1 - regularization

$$(P) \quad \min_{a,b,\xi} \ \frac{1}{2}\|a\|^2 + c\sum_{i=1}^{n}\xi_i$$

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

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

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

Lagrangian for (P):

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

λ, γ are Lagrangian multipliers.

Analagous to the above steps

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

$$\text{s.t.} \quad 0 \le \lambda_i \le C \quad \text{(new)}$$

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

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

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

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

Complementary slackness:

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

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

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

If $0 < \lambda_k < C$ for some k (x_k is a support vector)

$b^* = y_k - a^{*T} x_k$ is the optimal b.

(by solving (*))

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

Compute $a^{*T}x + b^* = \left(\sum_{i=1}^{N} \lambda_i^* y_i x_i\right)^T x + b^*$

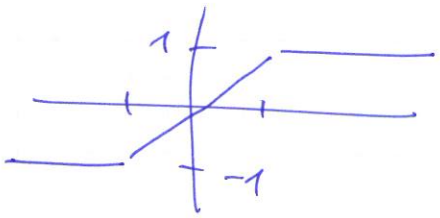
$$= \sum_{i=1}^{N} \lambda_i^* y_i x_i^T x + b^* = d(x)$$

- Hard decision

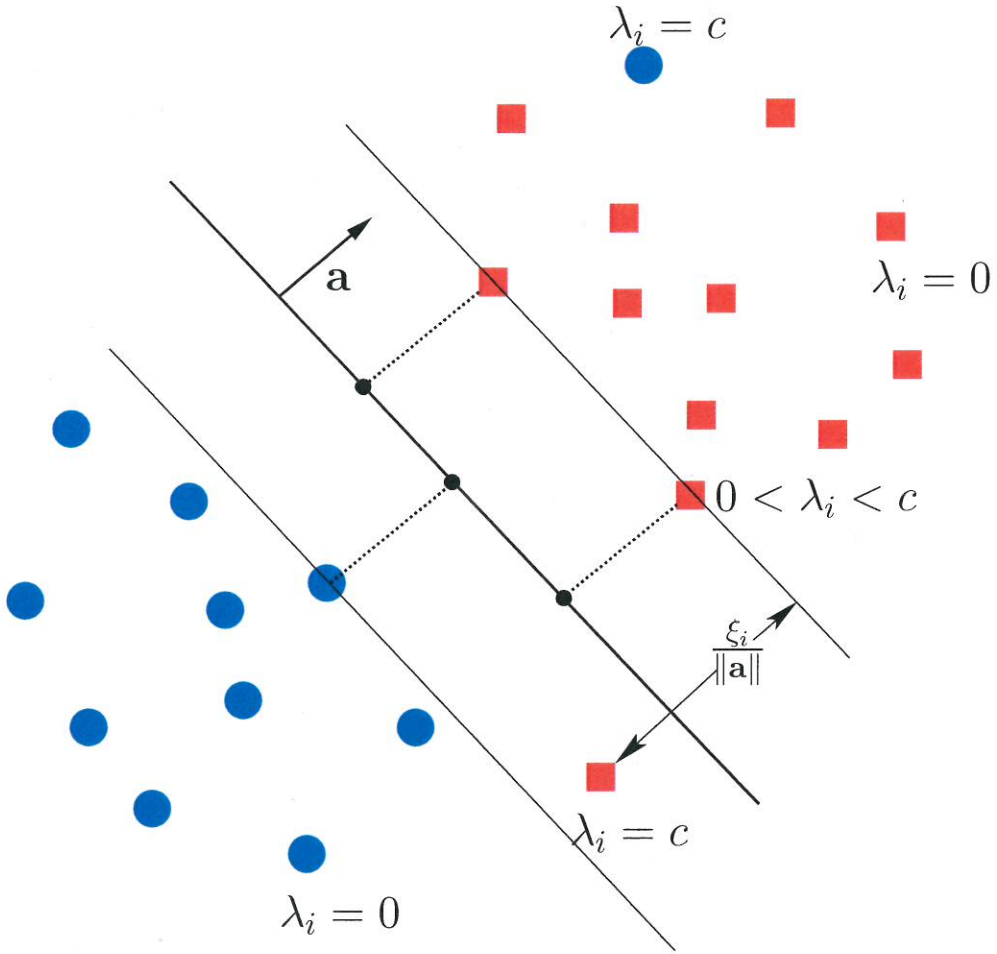
  Decide $y = 1$ if $d(x) \geq 0$, otherwise $y = -1$

- Soft classifier

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



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


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