6. Support Vector Machines (SVM)

Training set $(x_1, y_1), \ldots, (x_n, y_n)$

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

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

Determine a separating hyperplane.

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

a) $\{x \in \mathbb{R}^p \mid a^T x = 0\}$, $a \in \mathbb{R}^p$, $a \neq 0$

$(p-1)$-dim. subspace

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

(hyperplane)

c) Distance between 2 hyperplanes

$H_1 = \{a^T x - b_1 = 0\}$ $\quad H_2 = \{a^T x - b_2 = 0\}$

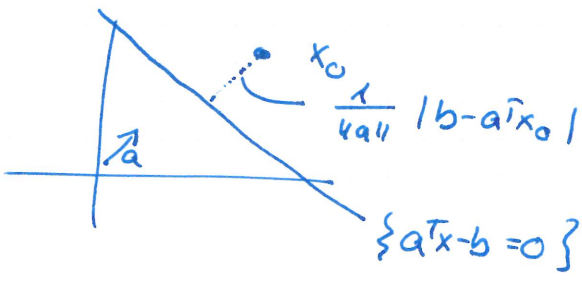
$$d(H_1, H_2) = \frac{1}{\|a\|} |b_2 - b_1|$$

d) Distance between a hyperplane

$H = \{a^T x - b = 0\}$ and $x_0 \in \mathbb{R}^p$:

$$d(H, x_0) = \frac{1}{\|a\|} |b - a^T x_0|$$

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



6.2. The optimal margin classifier

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

Assume there is a separating hyperplane.

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

Then $\exists \gamma \geq 0$

$$y_i = +1 \Rightarrow a^T x_i + b \geq \gamma$$

$$y_i = -1 \Rightarrow a^T x_i + b \leq -\gamma$$

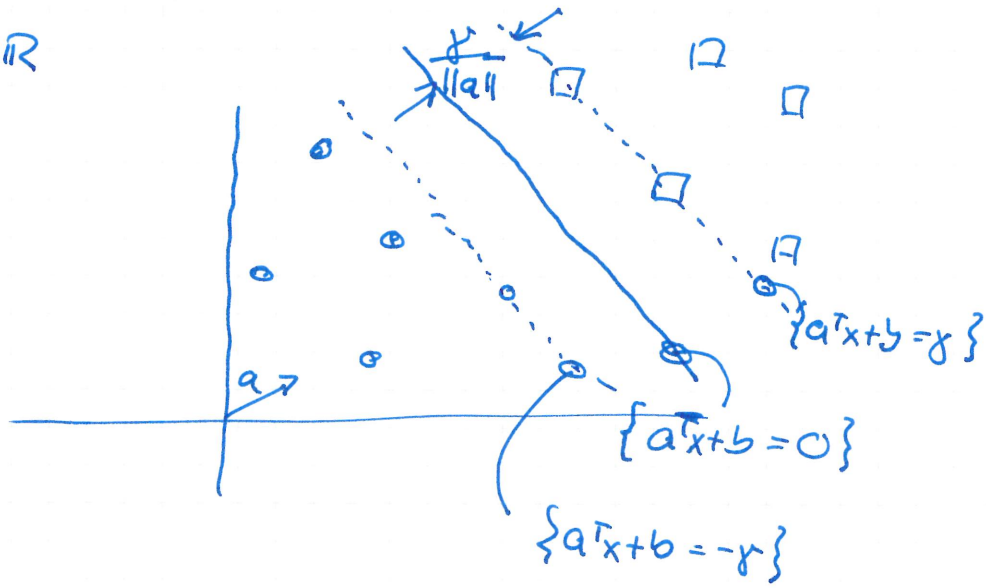
Hence

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

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

$$\max_{\gamma, a, b} \frac{\gamma}{\|a\|} \quad \text{s.t.} \quad y_i (a^T x_i + b) \geq \gamma$$

$$\gamma \geq 0, \; a \in \mathbb{R}^p, \; b \in \mathbb{R}$$



$$\max_{\gamma,a,b} \frac{\gamma}{\|a\|} \quad \text{s.t.} \quad y_i(a^T x_i + b) \geq \gamma$$

(not scale invariant)

$\Leftrightarrow$ (if γ, a, b is solution, then $\beta\gamma, \beta a, \beta b$, $\beta > 0$ is also a solution)

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

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

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

In summary (OMC):

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

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

Quadratic opt. problem with linear constraints, special case of convex optimization.

Use standard software.

- Assume a^* is an optimal solution to (OMC) and x_k some point with minimum margin, a <u>support</u> point.

  Then $y_k(a^{*T}x_k + b^*) = 1$

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

  $\Leftrightarrow$ $b^* = y_k - a^{*T}x_k$,

  the optimal b-value.

- The solution (a^*, b^*) is called <u>optimal margin classifier</u>.

- Use commercial software (QP) to solve (OMC)

More interesting things to do!

6.3 SVM and Lagrange Duality

Brief excursion on convex optimization.

- Convex optimization problem:

(P) minimize $f_0(x)$

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

$h_i(x) = 0$, $i = 1, \dots, r$

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 concave function.

- Lagrange dual problem:

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

s.t. $\lambda \geq 0$.

- Weak duality theorem:

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

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 strong duality, i.e., $g(\lambda^*, \nu^*) = f_0(x^*)$.
"Strong duality holds"; "the duality gap is zero".

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

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \qquad \text{(assumption)} \\ &= \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{r} \nu_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \underbrace{\lambda_i^*}_{\geq 0} \underbrace{f_i(x^*)}_{\leq 0} + \sum_{i=1}^{r} \nu_i^* \underbrace{h_i(x^*)}_{=0} \\ &\leq f_0(x^*) \end{aligned}$$

Collecting conditions so far

- 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 constraints)
3. $\lambda_i f_i(x) = 0$, $i = 1, \dots, m$ (complementary slackness)
4. $\nabla_x L(x, \lambda, \nu) = 0$

Th. 6.1. If Slater's conditions are satisfied (which holds if the constraints are linear) 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 hold. ⊥