Support vector machines

To separate 2 sets of points (cat pictures vs dog pictures, or spam vs not spam emails, ...)

We aim to find a function

$\underline{f(x)}$: $f(x_i) > 0$ for $x_i \in$ Class 1

$\qquad f(x_i) < 0$ for $x_i \in$ Class 2

$i = 1, \dots, N \leftarrow$ sample size

Zero-level of f: $\{x \mid f(x) = 0\}$ separates, or discriminates 2 classes

Linear separation function $f(x)$:

$f(x) = a^T x - b$

We seek a hyperplane to separate 2 classes

$$\circledast \quad \begin{bmatrix} a^T x_i - b > 0 & x_i \in \text{Class } 1 \\ a^T x_i - b < 0 & x_i \in \text{Class } 2 \end{bmatrix} \overset{\text{feasibility}}{\Longleftrightarrow} \begin{bmatrix} a^T x_i - b \geq 1 & x_i \in \text{Class } 1 \\ a^T x_i - b \leq -1 & x_i \in \text{Class } 2 \end{bmatrix}$$

(since we can simultaneously rescale a and b)

Convex feasibility problem

$$\circledast\!\!\star \quad \begin{bmatrix} \min 0 \\ a \in \mathbb{R}^n, b \in \mathbb{R} \\ \tilde{X}^T \cdot a_b < 0 \end{bmatrix}, \text{ where}$$

$$\tilde{X} = \begin{array}{c} \overbrace{\quad x_1 \dots x_{N_1} \in \text{class } 1 \quad}\ \overbrace{\quad \tilde{x}_1 \dots \tilde{x}_{N_2} \in \text{class } 2 \quad} \\ \begin{bmatrix} -x_1 & -x_2 & \dots & -x_{N_1} & \tilde{x}_1 & \dots & \tilde{x}_{N_2} \\ +1 & +1 & \dots & +1 & -1 & \dots & -1 \end{bmatrix} \begin{array}{l} \}\, n \\ \ 1 \end{array} \\ \underbrace{\qquad\qquad\qquad}_{N_1} \ \underbrace{\qquad\qquad}_{N_2} \end{array} \qquad a_b = \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b \end{bmatrix}$$

(restating $\circledast$)

<u>When is it feasible?</u>

Geometrically, when convex hulls $\boxed{\text{conv}1 \cap \text{conv}2 = \emptyset}$ do not intersect

$\text{conv } 1 = \text{conv}\{x_i \mid x_i \in \text{Class } 1\}$

$\text{conv } 2 = \text{conv}\{x_i \mid x_i \in \text{Class } 2\}$

<u>Why?</u> This follows from the strong alternatives to linear inequalities (recall Farkas lemma $\rightarrow$ LP duality)

$$\boxed{\text{Thm} \quad \text{Let } A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^m \quad \exists x \in \mathbb{R}^n: \{Ax < b\} \ \underset{\text{strictly}}{\text{or}} \ \exists \lambda \in \mathbb{R}^m: \{\lambda \neq 0,\ \lambda \geq 0,\ A^T\lambda = 0,\ \lambda^T b \leq 0\}}$$

(Boyd & Vanderberge, §2.5; p 50 follows from a version of separability of convex sets theorem)

$\boxed{\text{Corollary}}$ (★) is infeasible $\iff$

$\boxed{\exists \lambda \in \mathbb{R}^N,\ \lambda \neq 0,\ \lambda \geq 0,\ \tilde{X}\lambda = 0,\quad 0 \leq 0}$

To restate, let $\lambda = [\underbrace{\bar{\lambda}}_{N_1} \mid \underbrace{\tilde{\lambda}}_{N_2}]$

$$\mathbb{1}_s = (\underbrace{1, \ldots, 1}_{s}) \in \mathbb{R}^s$$

Then $\sum_{i=1}^{N} \lambda_i \tilde{X}_i = 0 \iff \begin{cases} \sum_{i=1}^{N_1} \bar{\lambda}_i \cdot \bar{x}_{ki} = \sum_{i=1}^{N_2} \tilde{\lambda}_i \cdot \tilde{x}_{ki} \\ \sum_{i=1}^{N_1} \bar{\lambda}_i = \sum_{i=1}^{N_2} \tilde{\lambda}_i \end{cases}$

We could rescale all λ_i by the same number, e.g. making $\sum \bar{\lambda}_i = \sum \tilde{\lambda}_i = 1$

This implies the convex hull condition.

What if the problem is not feasible?

One could minimize the number of misclassified points instead

$$\begin{bmatrix} \min_{a,b,\eta} \|\eta\|_0 \\ y_i(a^T x_i - b) \geq 1 - \eta_i \\ \eta_i \geq 0 \end{bmatrix}$$

where $\eta \in \mathbb{R}^N$ $\begin{cases} \eta_i = 0 & \text{if a point is correctly classified} \\ \eta_i \in (0,1) & \text{if a point is still correctly classified, but "uncertain"} \\ \eta_i > 1 & \text{if a point is misclassified} \end{cases}$

and $y_i = +1$ if $x_i \in$ Class 1

$y_i = -1$ if $x_i \in$ Class 2

Non-convex since $\|\cdot\|_0$ is non-convex

$\|\cdot\|_0 \to \|\cdot\|_1$ relaxation (convex envelope)

$$\begin{bmatrix} \min_{a,b,\eta} \|\eta\|_1 \\ y_i(a_i^T x_i - b) \geq 1 - \eta_i \\ \eta_i \geq 0 \end{bmatrix}$$

Feasible case: seeking the most robust solution from many

To pick one of them, one could specify the optimization task:

(1)
$$\left[\begin{array}{l} \max_{a,b,t} \ t \\ y_i (a^T x_i - b) \geq t \\ \|a\| \leq 1 \end{array}\right]$$

"Max margin classifier"

← we maximize the "gap" between the data and the separating hyperplane

← normalization is required, otherwise $a \nearrow \infty$ increases $t \nearrow \infty$

1. Interpretation

This promotes robustness:
if a point's location is slightly ($<t$) off, it will be classified correctly

t maximizes the distance to $\{a^T x = b\}$

<u>Why?</u> <u>Lemma</u>
$$\boxed{\operatorname{dist}(v, a^T z = b) = \frac{|a^T v - b|}{\|a\|}}$$
$v \in \mathbb{R}^n$ is a point

Proof of the lemma follows from

Thm (Corollary of the optimality Thm: $\left[\begin{array}{l} \min f(x) \\ x \in \Omega \end{array}\right]$ convex, x is optimal $\iff \nabla f(x)^T (y-x) \geq 0 \quad \forall y \in \Omega$)

$\left[\begin{array}{l} \min f(x) \\ Ax = b \end{array}\right]$ f is convex, $A \in \mathbb{R}^{m \times n}$. For a feasible point x,
x is optimal $\iff \exists \mu: \nabla f(x) = A^T \mu$.

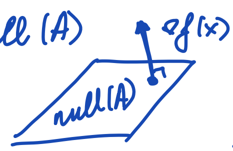
(Pf) Take $x: Ax = b$
x is optimal $\iff \nabla f^T(x)(y-x) \geq 0 \quad \forall y: Ay = b$

$\iff \forall z \in \operatorname{null}(A) \quad \nabla f^T(x) \cdot z \geq 0$

$\iff \forall z \in \operatorname{null}(A) \quad \nabla^T f(x) \cdot z \leq 0$ (consider $-z$, it is also in nullspace)

$\iff \nabla f(x)$ is orthogonal to $\operatorname{null}(A)$

$\iff \nabla f(x) \in$ row space of A



■

Pf) of the Lemma

Consider an optimization problem $\left[\begin{array}{l} \min \|v - z\|_2^2 \\ a^T z = b \end{array}\right]$

$$\begin{bmatrix} \min \sum_{i=1}^{n} v_i^2 - 2v_i z_i + z_i^2 \\ a^T z = b \end{bmatrix}$$

$\nabla f = 2z - 2v$

$\nabla^T f = a\mu$ (by Thm above)

$2(z-v) = a\mu$ $\|a\|^2 = a^T a$

$$\mu = \frac{2a^T(z-v)}{\|a\|^2}$$

Then,

$$\|v-z\| = \frac{\|a\|}{2}|\mu| = \frac{\|a\|}{2}\frac{2|a^T(z-v)|}{\|a\|^2}$$

$$= \frac{|b - a^T v|}{\|a\|^2}$$ ∎

2. Properties

Thm (Properties of the max margin classifier)

① Optimal value of (1) is achieved when $\|a\| = 1$

② Solution to (1) is unique

③ It is equivalent to the following problem

$$\begin{bmatrix} \min_{a,b} \|a\| \\ y_i(a^T x_i - b) \geq 1 \end{bmatrix} \quad (2)$$

(Proof) ① if (t^*, a^*, b^*) gives optimal solution and $\|a^*\| \in (0,1)$, then consider $a := \frac{a^*}{\|a^*\|}$, $b = \frac{b^*}{\|a^*\|}$, $t = \frac{t}{\|a^*\|} > t$

② Follows from ③: $\min \|a\| \iff \min \|a\|^2$ which is strictly convex, so min is unique in a.

if we have 2 optimal solutions (a, b^*) and (a, b^{**}), consider $b = \frac{b^* + b}{2}$

let $b^{**} < b^*$ (without loss of generality), then $b^{**} < b < b^*$

$y_i(a^T x_i - b^*) \geq 1 \quad \forall y_i = 1 \Rightarrow y_i(a^T x_i - b) > 1 \quad \forall y_i = 1$

$y_i(a^T x_i - b^*) \leq 1 \quad \forall y_i = -1 \Rightarrow y_i(a^T x_i - b) < 1 \quad \forall y_i = -1$

Then we can shrink both a and b to improve on the solution (exercise: check this formally by defining $\varepsilon := \min_i |1 - y_i(a^T x_i - b)|$

③ <u>Equivalence</u>: 2 problems are feasible at the same time

a)

If $a = 0$ for any feasible $(0, b, t)$ or $(0, b)$:

$-y_i b \geq t \quad \forall i.$

If there are indeed 2 classes ($y_i = +1$ and $y_j = -1$) this leads to a contradiction unless $t = 0$

So, both problems will have solution zero.

b) If not,

$(a, b) \rightarrow \left(\frac{a}{\|a\|}, \frac{b}{\|a\|}, \frac{1}{\|a\|}\right)$

$\left(\frac{a}{t}, \frac{b}{t}\right) \leftarrow (a, b, t)$

So, 2 problems are feasible together,

$t = \frac{1}{\|a\|}$,

so minimize $\|a\|$ is equivalent to maximize t. ∎

<u>Note</u>: problem (2) is easier for the analysis and has less variables

(e.g., uniqueness)

Infeasible case again: putting all together

① Support vector classifier

$$\begin{cases} \min_{a,b,\eta} \ \|a\| + \gamma \|\eta\|_1 \\ \text{such that } y_i (a^T x_i - b) \geq 1 - \eta_i \\ \eta_i \geq 0 \end{cases}$$

Minimizing (a) margin width and (b) the number of misclassifications at the same time

γ is a parameter (sets the relative importance of the goals (a) and (b))

Margin width: $\frac{2}{\|a\|}$ (easy exercise by using the material above)

Other approaches to handle non-separability:

② Bayesian: define a model based on observations, then via maximum likelihood (logistic modeling)

③ Non-linear separating functions (e.g., quadratic)