Lecture 3. Nonsmooth optimization for classification problems in Machine Learning

M. Gaudioso,*

*DIMES-Universitá della Calabria and ICAR-CNR, Rende (CS), Italia

EUROPT Summer School

August 30th– September 1st, 2021

Machine Learning and Optimization

- Optimization for Machine Learning;
- Machine Learning for Optimization.

Our point of view: Machine Learning as a source of many
non trivial
optimization problems.

Outline

Classification as decision making

To make a decision about **class membership** of an individual (*sample*).

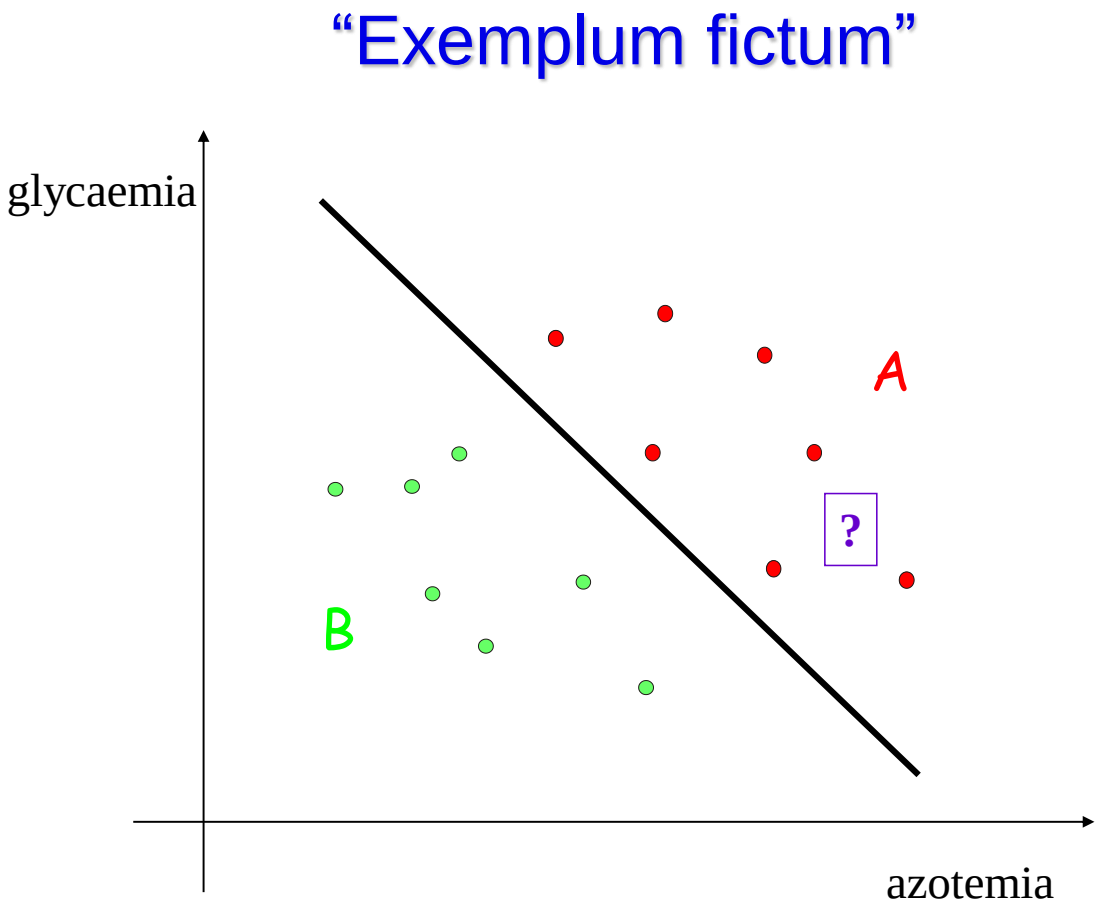
- The decision is taken on the basis of **similarities** among samples;
- The samples are characterized by their **attributes** (*features*);
- Two or more classes are considered.

Classification problems

- **Supervised classification**: A set of "labelled" samples is available. The objective is to predict class membership (the "label") of newly incoming samples;
- **Unsupervised classification**: All samples are "unlabelled". The objective is to group ("cluster") them;
- **Semisupervised classification**: Both "labelled" and "unlabelled" samples are available. The objective is to predict the "label" of newly incoming samples.

Outline

The idea



fictum.pdf

Modelling the binary supervised classification

To **discriminate** between two finite point sets in the n-dimensional space by a separating surface.

- Two disjoint, finite point sets

$$\mathcal{A} = \{a_1, \ldots, a_m\}, \qquad \text{with } a_i \in \mathbb{R}^n,\ i = 1, \ldots, m$$

$$\mathcal{B} = \{b_1, \ldots, b_k\}, \qquad \text{with } b_l \in \mathbb{R}^n,\ l = 1, \ldots, k.$$

- The points are labelled

Linear separability

The sets $\mathcal{A}$ and $\mathcal{B}$ are **linearly separable** if and only if there exists a hyperplane

$$H(w', \gamma') \triangleq \{x \in \mathbb{R}^n \mid w'^\top x = \gamma'\}, \text{ with } w' \in \mathbb{R}^n \text{ and } \gamma' \in \mathbb{R},$$

such that

- $a_i^\top w' < \gamma', \quad i = 1, \ldots, m$
- $b_l^\top w' > \gamma', \quad l = 1, \ldots, k.$

or, equivalently, iff there exists $H(w, \gamma) \triangleq \{x \in \mathbb{R}^n \mid w^\top x = \gamma\}$ s.t.

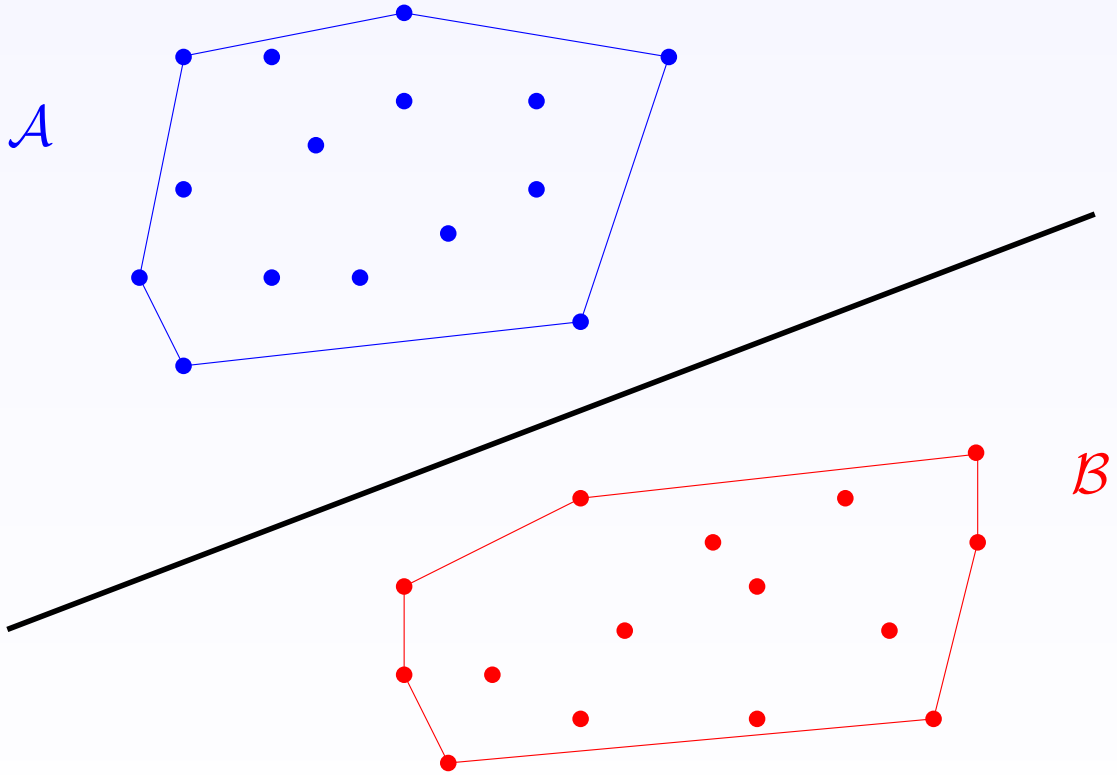
- $a_i^\top w \leq \gamma - 1, \quad i = 1, \ldots, m$
- $b_l^\top w \geq \gamma + 1, \quad l = 1, \ldots, k.$

Remark: $\mathcal{A}$ and $\mathcal{B}$ are linearly separable if and only if

$$\text{conv}(\mathcal{A}) \cap \text{conv}(\mathcal{B}) = \emptyset.$$

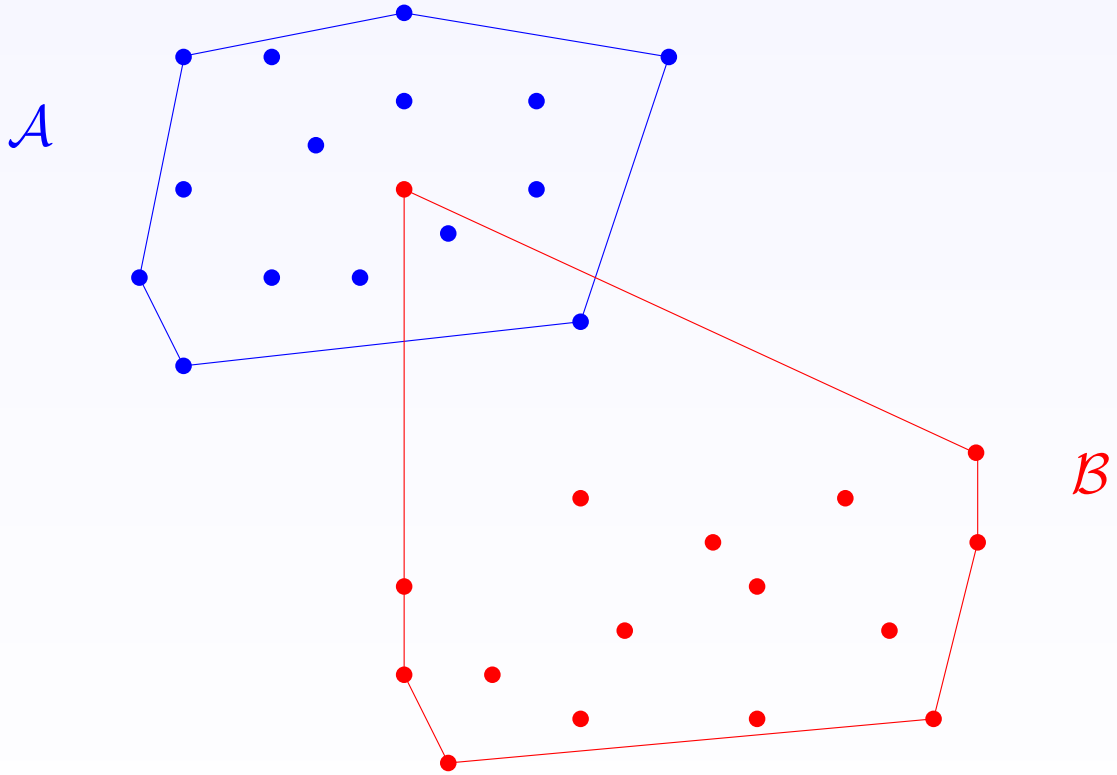
The geometry of linear separability-1

$$\mathrm{conv}(\mathcal{A}) \cap \mathrm{conv}(\mathcal{B}) = \emptyset.$$



The geometry of linear separability-2

$$\mathrm{conv}(\mathcal{A}) \cap \mathrm{conv}(\mathcal{B}) \neq \emptyset.$$



The separation margin

Consider the three hyperplanes:

$$H^-(w,\gamma) \triangleq \{x \in \mathbb{R}^n \mid w^\top x = \gamma - 1\};$$

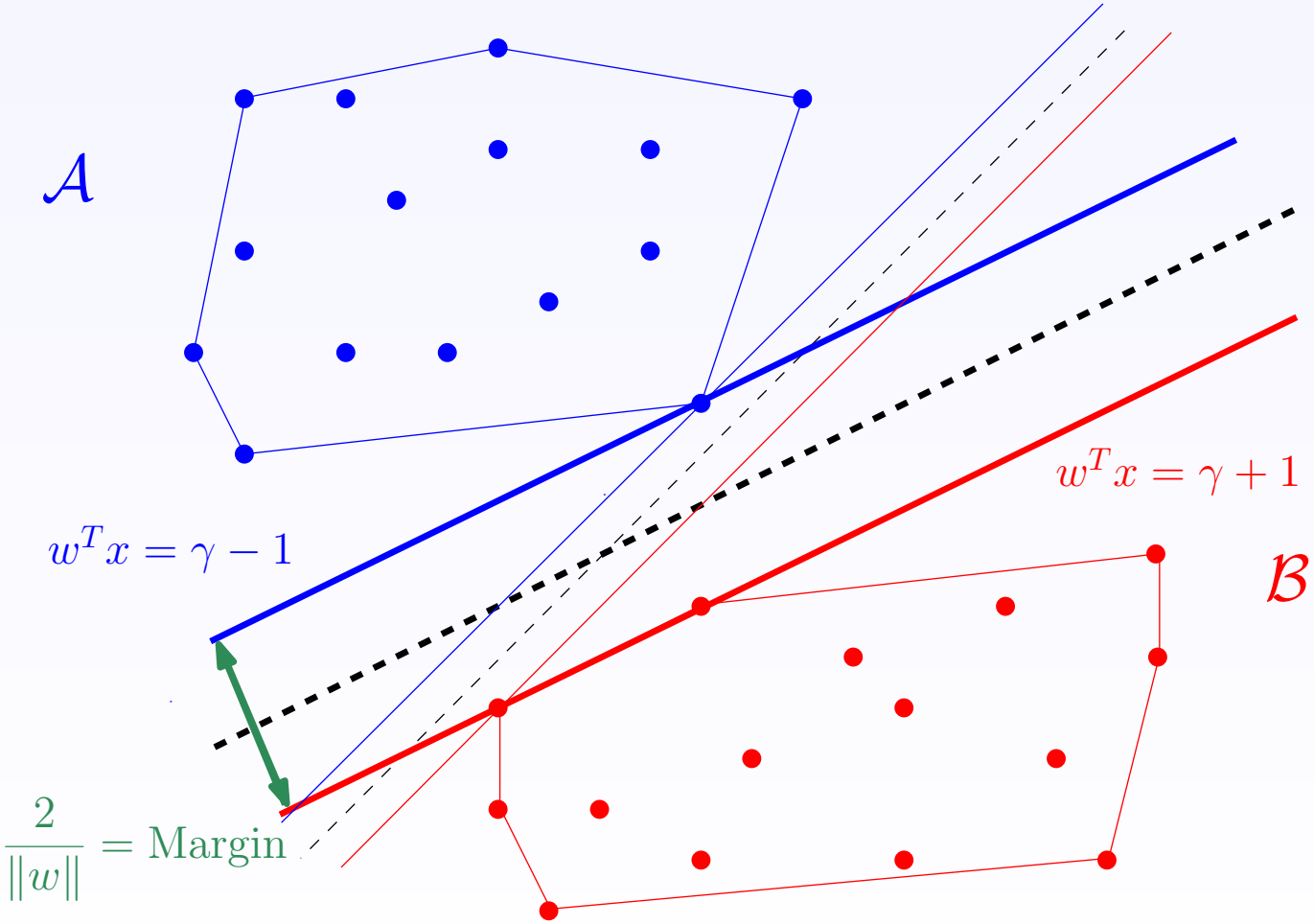
$$H(w,\gamma) \triangleq \{x \in \mathbb{R}^n \mid w^\top x = \gamma\};$$

$$H^+(w,\gamma) \triangleq \{x \in \mathbb{R}^n \mid w^\top x = \gamma + 1\}.$$

They are parallel and the distance between H^- and H^+ is the separation margin (width of the *strip*) .
It can be proved that the value of the separation margin is $\frac{2}{\|w\|}$
(Hint: Take any $\bar{x} \in H^-$ and solve $min_{y \in H^+} \|y - \bar{x}\|^2$).

It is important to have a large margin, that is a large strip separating the sets $\mathcal{A}$ and $\mathcal{B}$!

SVM: maximizing the margin



The error function and the Support Vector Machine (SVM) approach

How calculate a large margin separating hyperplane (if any!)? Define the classification error function depending on the hyperplane.

- Point a_i is correctly classified by hyperplane $H(w,\gamma)$ iff $a_i^\top w - \gamma + 1 \leq 0$ (equivalently, iff $e_i(w,\gamma) = \max(0, a_i^\top w - \gamma + 1) = 0$)
- Point b_l is correctly classified by hyperplane $H(w,\gamma)$ iff $-b_l^\top w + \gamma + 1 \leq 0$,(equivalently, iff $e_l(w,\gamma) = \max(0, -b_l^\top w + \gamma + 1) = 0$)
- $e_i(w,\gamma) \geq 0$ ($e_l(w,\gamma) \geq 0$) is the classification error provided by H w.r.t. point a_i (b_l).

(Total) error function $e(w,\gamma) = \sum_{i=1}^{m} e_i(w,\gamma) + \sum_{l=1}^{k} e_l(w,\gamma)$

The Support Vector Machine (SVM) model:

$$\min_{w,\gamma} \|w\|^2 + Ce(w,\gamma)$$

Properties of the SVM problem

- A convex optimization problem;
- The objective function is the (weighted) sum of a quadratic term and a piecewise affine, convex function;
- Can be transformed into a convex quadratic problem (by introducing additional variables);
- The dual is usually solved.

Outline

The problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$f : \mathbb{R}^n \mapsto \mathbb{R}$ is a DC function.

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is a DC function if there exist convex functions $f_1,\ f_2 : \mathbb{R}^n \to \mathbb{R}$ such that:

$$f(x) = f_1(x) - f_2(x),\ \forall x \in \mathbb{R}^n.$$

A multiextremal problem!

Linearization approach

Generate a sequence $\{x_{k+1}\}$ by solving the convex problems (P_k) obtained by linearizing f_2 starting from the stability center y_k:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} c_k(x) \quad (P_k),$$

where

$$c_k(x) \triangleq f_1(x) - \left(f_2(y_k) - g_k^{(2)T}(x - y_k)\right),$$

with $g_k^{(2)} \in \partial f_2(y_k)$.

Remark

$c_k(x) \geq f(x)$, $\forall x \in \mathbb{R}^n$ *and* $c_k(y_k) = f(y_k)$.

Double bundle

Generate two separate cutting plane models for f_1 and f_2

$$\tilde{f}_1^{(k)}(x) \triangleq \max_{i=1,\dots,I} \left(f_1(x_i) + g_i^{(1)T}(x - x_i)\right), \quad g_i^{(1)} \in \partial f_1(x_i),$$

$$\tilde{f}_2^{(k)}(x) \triangleq \max_{j=1,\dots,J} \left(f_2(x_j) + g_j^{(2)T}(x - x_j)\right), \quad g_j^{(2)} \in \partial f_2(x_j),$$

and adopt as a model:

$$\tilde{f}^{(k)}(x) \triangleq \tilde{f}_1^{(k)}(x) - \tilde{f}_1^{(k)}(x).$$

Then set

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \tilde{f}^{(k)}(x)$$

Remark

$\tilde{f}^{(k)}(x)$ is still the difference of two piecewise affine convex functions.

More about double bundle

Minimization of $\tilde{f}^{(k)}(x)$ can be achieved by solving J convex problems.

Hint. Write

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \tilde{f}^{(k)}(x) &= \min_{x\in\mathbb{R}^n} \left(\tilde{f}_1^{(k)}(x) + \min_{j=1,\dots,J}(-f_2(x_j) - g_j^{(2)T}(x-x_j)\right) = \\ &= \min_{x\in\mathbb{R}^n} \min_{j=1,\dots,J} \left(\tilde{f}_1^{(k)}(x) - f_2(x_j) - g_j^{(2)T}(x-x_j)\right),\end{aligned}$$

and then change the order of the two minimizations.

The new approach (1)

- $\tilde{f}_1^{(k)}(x) - f_1(x) \leq 0, \ \forall x$ and $\tilde{f}_2^{(k)}(x) - f_2(x) \leq 0, \ \forall x$
- Let y_k the stability center and assume

$$\tilde{f}^{(k)}(y_k+d) \triangleq \tilde{f}_1^{(k)}(y_k+d) - \tilde{f}_2^{(k)}(y_k+d) < f_1(y_k) - f_2(y_k) = f(y_k),$$

while

$$f(y_k+d) = f_1(y_k+d) - f_2(y_k+d) > f_1(y_k) - f_2(y_k) = f(y_k).$$

This implies

$$\tilde{f}_1^{(k)}(y_k+d) - f_1(y_k+d) < \tilde{f}_2^{(k)}(y_k+d) - f_2(y_k+d) \leq 0.$$

Remark

Roughly speaking: $\tilde{f}_1^{(k)}$ *is a worse approximation of* f_1 *than* $\tilde{f}_2^{(k)}$ *of* f_2.

The new approach (2)

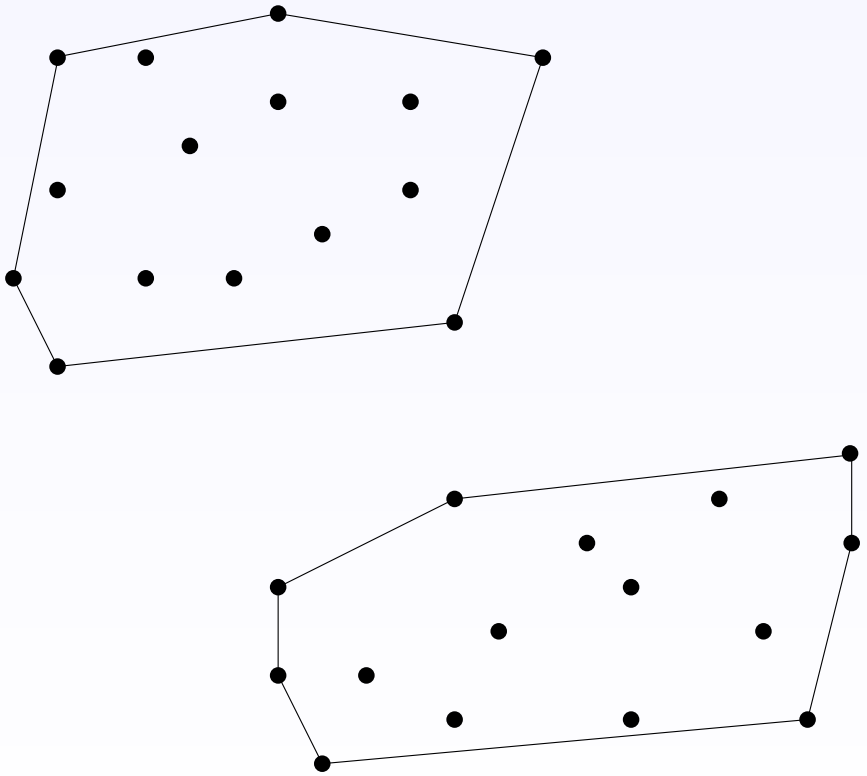
The rationale

- Maintain two bundles but try to be parsimonious.
- Whenever no descent is achieved invest only on getting a better approximation of f_1 (see previous Remark).
- Keep as much as possible the bundle size to 1.
- Solve at each iteration one or, at most, two convex problems.
- Adopt a more sophisticated DC model only for testing purposes.

Outline

A clustering problem

The points are now unlabelled and we want to group them into two clusters.



Cluster optimization

Just one finite point set: $\mathcal{A} = \{a_1, \ldots, a_m\}$ with $a_i \in \mathbb{R}^n, \ \forall i$.
Look for two optimal centroids c_1^* and c_2^* by solving:

$$\min_{c_1,c_2} d(c_1, c_2) = \min_{c_1,c_2} \sum_{i=1}^{m} \underbrace{\min(\|a_i - c_1\|, \|a_i - c_2\|)}_{\text{distance of } a_i \text{ from its closest centroid}}$$

Since $\min(x, y) = (x + y) - \max(x, y)$, we write:

$$d(c_1, c_2) = \sum_{i=1}^{m} \Big(\underbrace{\|a_i - c_1\| + \|a_i - c_2\|}_{\text{convex}} - \underbrace{\max(\|a_i - c_1\|, \|a_i - c_2\|)}_{\text{convex}} \Big)$$

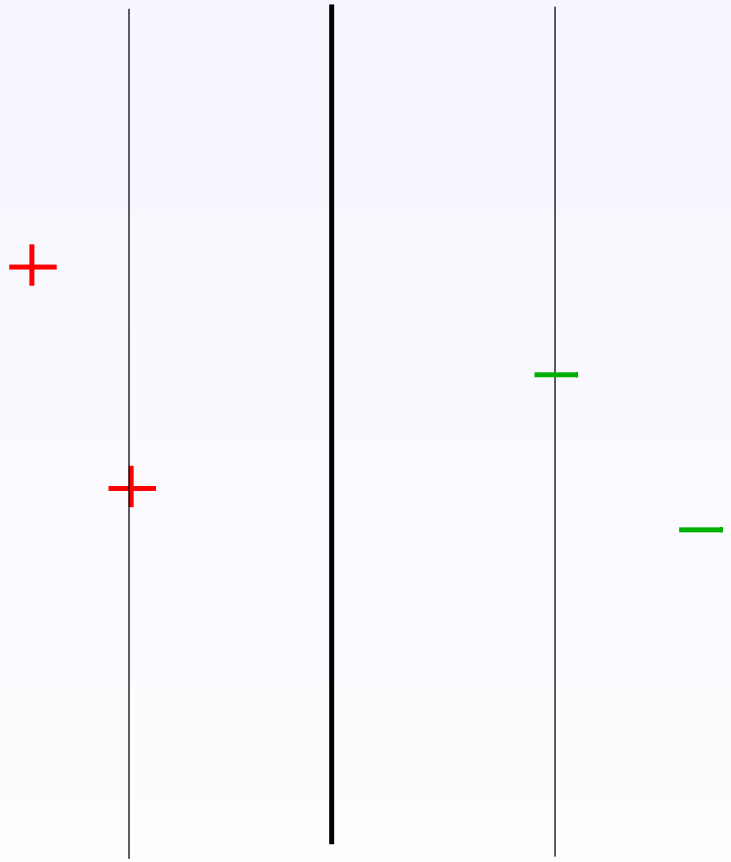
- Nonconvex nonsmooth optimization problem;
- Difference of Convex (DC) type.

Outline

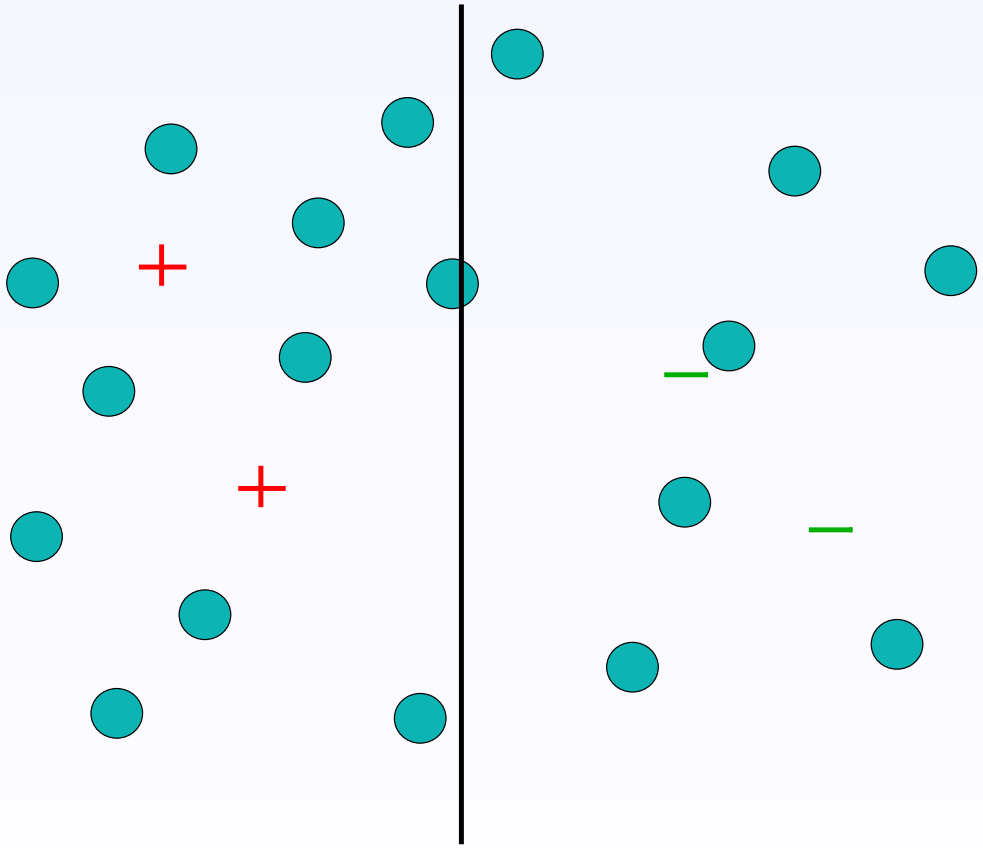
Semi-supervised Classification

- Sometimes labelling is fairly expensive (e.g. Medical diagnosis, web categorization, text processing...);
- **semi-supervised classification**: the prediction is based on (few) labelled and (many) unlabelled samples;
- The TSVM (Transductive Support Vector Machine) technique is the semi-supervised version of the SVM approach;
- Create a no-man's land in the separating strip.

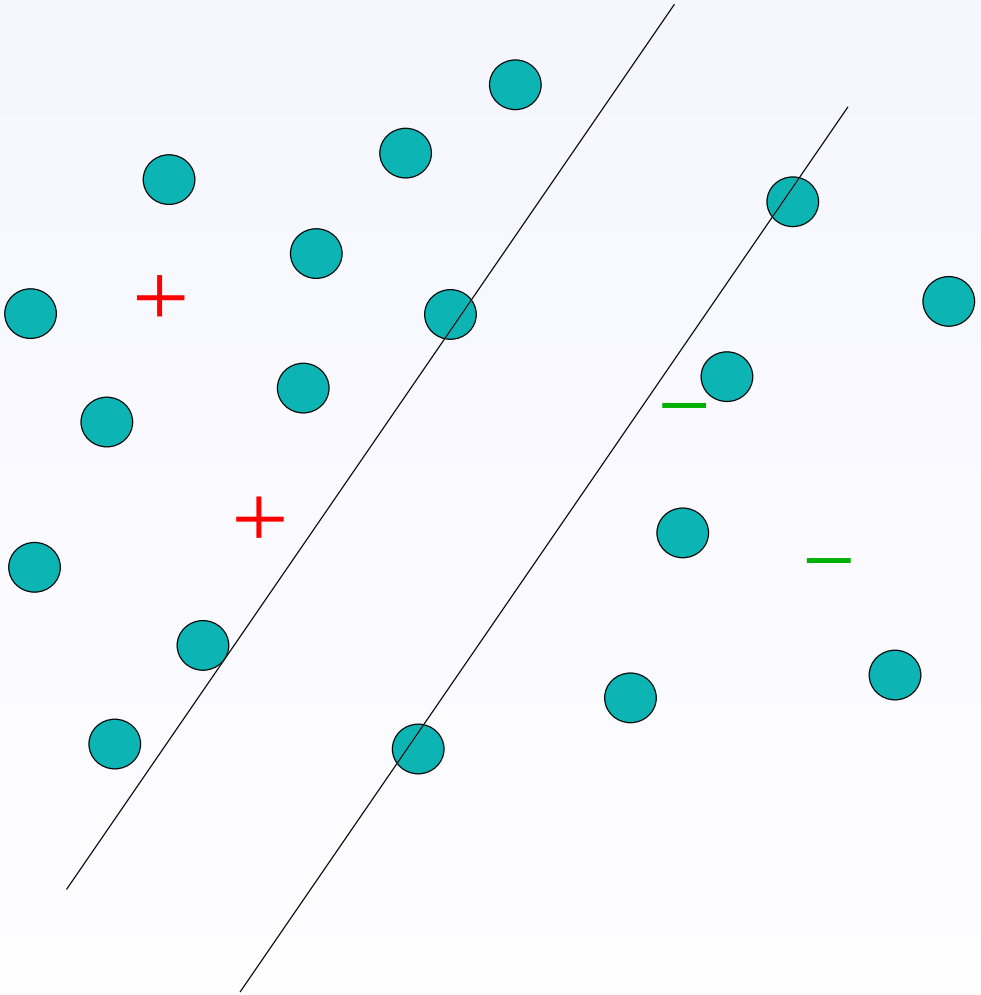
TSVM: an example 1



TSVM: an example 2



TSVM: an example 3



TSVM

Three point sets: $\mathcal{A} = \{a_1, \ldots, a_m\}$ and $\mathcal{B} = \{b_1, \ldots, b_k\}$ labelled; $\mathcal{C} = \{c_1, \ldots, c_r\}$ unlabelled.

- Look for a large margin separating hyperplane, no point, possibly, in the middle;
- Nonconvex nonsmooth optimization problem (Difference of Convex type).

$$\min_{w,\gamma} \underbrace{\|w\|^2 + C_1(e(w,\gamma))}_{\text{as in SVM}} + C_2 \sum_{h=1}^{r} \underbrace{\max(0, 1 - |c_i^\top w - \gamma|)}_{> 0 \text{ if } c_i \text{ is in the strip}}$$

In fact

$$\max(0, 1 - |c_i^\top w - \gamma|) = \underbrace{\max(0, |c_i^\top w - \gamma| - 1)}_{\text{convex}} + \underbrace{(1 - |c_i^\top w - \gamma|)}_{\text{concave}}$$

Outline

Nonsmoothness

Use of different types of nonlinear surfaces:

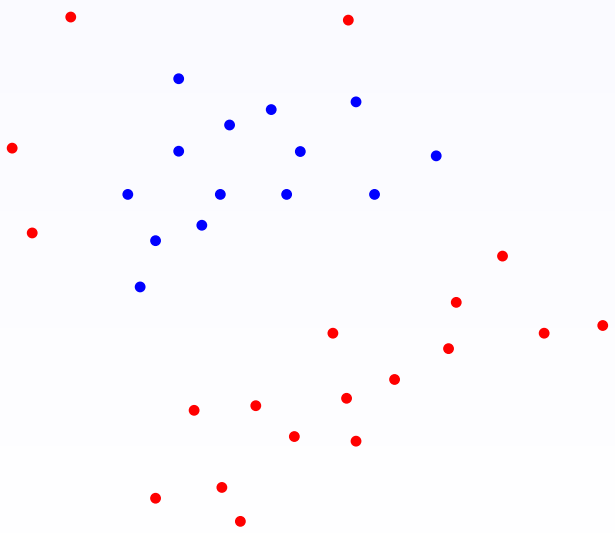
- Polyhedral;
- Spherical;
- Ellipsoidal;
- Conical.

Outline

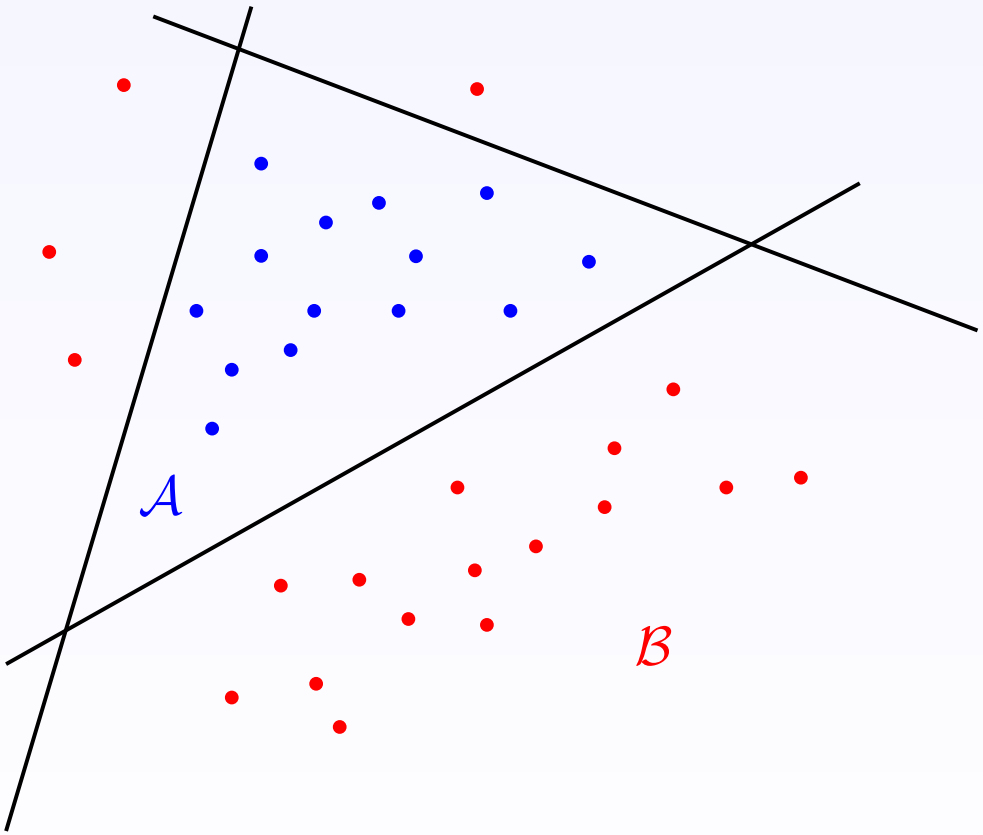
h-Polyhedral separability definition

The set $\mathcal{A}$ is h-**polyhedrally separable** from $\mathcal{B}$ if and only if there exists a set of h hyperplanes $\{w^{(j)}, \gamma_j\}$, with $w^{(j)} \in \mathbb{R}^n$ and $\gamma_j \in \mathbb{R}$, $j = 1, \ldots, h$, such that

- $a_i^T w^{(j)} \leq \gamma_j - 1$, for all $i = 1, \ldots, p$ and for all $j = 1, \ldots, h$
- for all $l = 1, \ldots, q$, there exists an index $j \in \{1, \ldots, h\}$ such that $b_l^T w^{(j)} \geq \gamma_j + 1$.



Pictorial representation

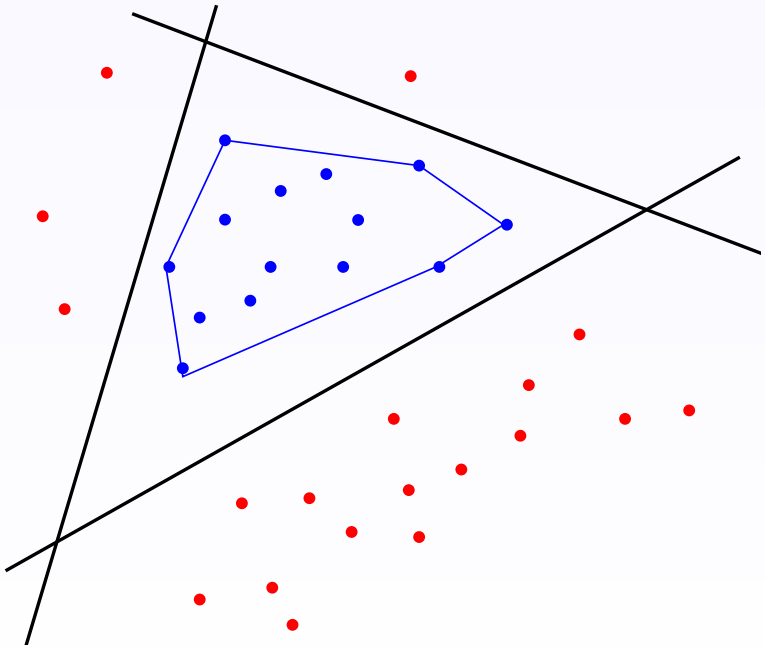


Property

Theorem

The set $\mathcal{A}$ is h-polyhedrally separable from $\mathcal{B}$, with $h \leq |\mathcal{B}|$, if and only if

$$\mathrm{conv}(\mathcal{A}) \cap \mathcal{B} = \emptyset.$$



A mathematical programming model

Once h has been fixed, the polyhedral separation of A and B is obtained by minimizing the following error function (h fixed):

$$z(\mathbf{w},\gamma) \triangleq \frac{1}{|A|}\sum_{i=1}^{p}\max_{j=1,\ldots,h}\{0, a_i^T w^{(j)} - \gamma_j + 1\} +$$

$$\frac{1}{|B|}\sum_{l=1}^{q}\max\{0, \min_{j=1,\ldots,h}[-b_l^T w^{(j)} + \gamma_j + 1]\}$$

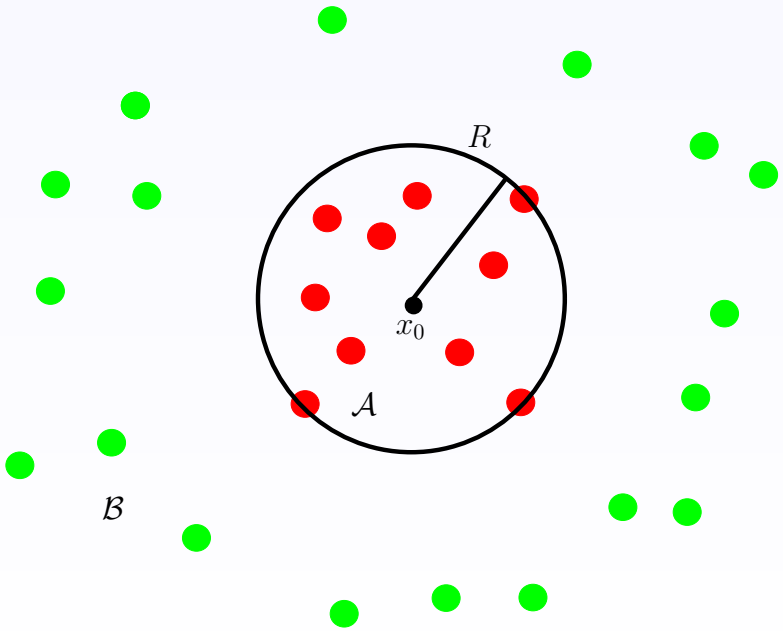
where $\mathbf{w} \triangleq [w^{(1)}, w^{(2)}, \ldots, w^{(h)}]$ and $\gamma^T \triangleq [\gamma_1, \gamma_2, \ldots, \gamma_h]$.

Function z is nondifferentiable; it is nonconvex and nonconcave, but can be put in DC form.

Outline

The model

- Find a sphere enclosing all points of $\mathcal{A}$ and no points of $\mathcal{B}$.
- The role of the two sets $\mathcal{A}$ and $\mathcal{B}$ is not symmetric.
- $\mathcal{B} \cap \mathrm{conv}(\mathcal{A}) = \emptyset$ is necessary (but not sufficient) condition for the existence.



Problem definition

$\mathcal{A}$ and $\mathcal{B}$ are spherically separated by $S(x_0, R)$ if:

$$\|a_i - x_0\|^2 \leq R^2, \quad a_i \in \mathcal{A},\ i = 1, ..., m$$

and

$$\|b_l - x_0\|^2 \geq R^2, \quad b_l \in \mathcal{B},\ l = 1, ..., k$$

Look for a "small" separating sphere by minimizing the error function $h(x_0, R)$:

$$h(x_0, R) \triangleq \quad R^2 + Cw(x_0, R) = \quad R^2 + C\sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - R^2\} + C\sum_{l=1}^{k} \max\{0, R^2 - \|b_l - x_0\|^2\}.$$

$C > 0$ tradeoff parameter between the radius and the classification error.

A DC problem again

Outline

The model

$\mathcal{A}$ and $\mathcal{B}$ are ellipsoidally separable if and only if there exists an ellipsoid

$$E(Q; x_0) := \{x \in \mathbb{R}^n | (x - x_0^T Q(x - x_0) \leq 1\},$$

with x_0 center of the ellipsoid and Q symmetric and positive definite matrix, such that

1. $(a_i - x_0)TQ(a_i - x_0) \leq 1$ for all points $a_i \in \mathcal{A}$ $(i = 1, ..., m)$ and
2. $(b_l - x_0)TQ(b_l - x_0) > 1$ for all points $b_l \in \mathcal{B}$ $(l = 1, ..., k)$.

The approach

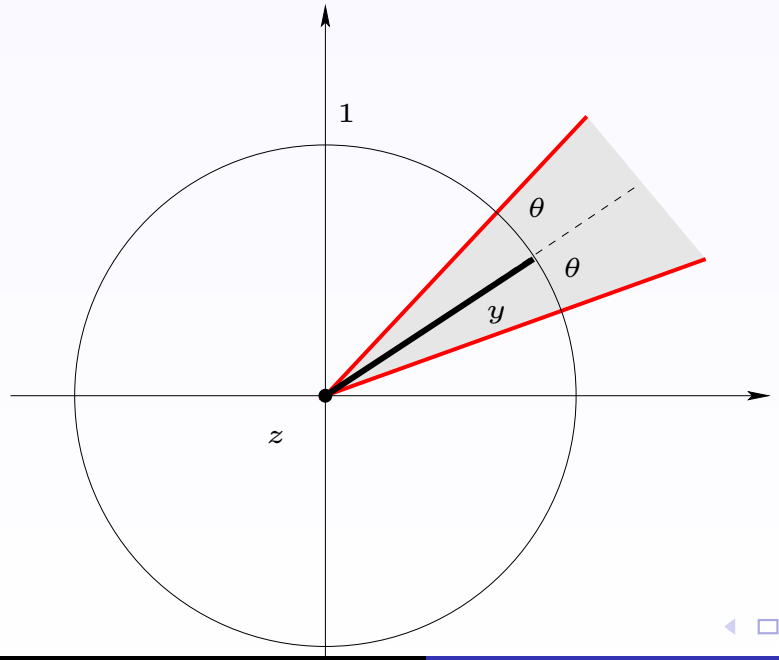
- Define the error function;
- State the problem in SDP form;
- Apply an ad hoc heuristic based on rank one correction of the current positive definite matrix defining the ellipsoid

Outline

Revolution Cone

$$\Gamma(z, y, s) = \{x \in \mathbb{R}^n : s\,\|x - z\| - y^T(x - z) = 0\}$$

Apex: $z \in Z \subseteq \mathbb{R}^n$.
Axis: $y \in \mathbb{S}_n$ (unit sphere of $\mathbb{R}^n$).
Aperture coefficient: $s \in [0, 1]$. Note that $\arccos s$ corresponds to the half-aperture angle of the cone.



$\|y\| = 1$
$s = \arccos\theta$

The Conic Separation problem

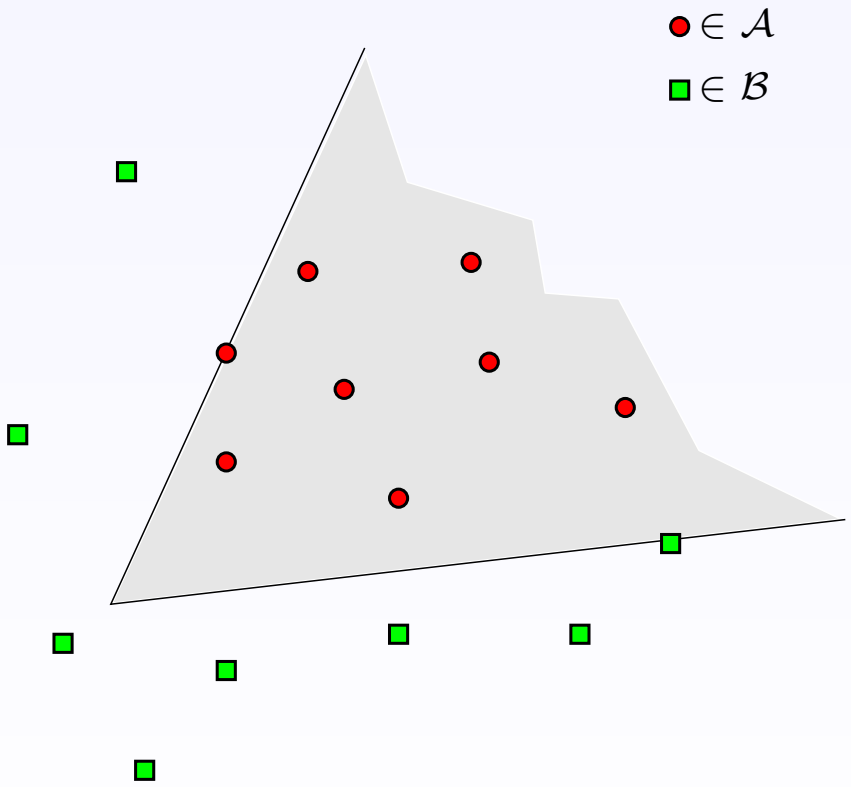
$$(\mathrm{CS}) \quad \begin{cases} \text{Find a triplet } (z, y, s) \in Z \times \mathbb{S}_n \times [0,1] \text{ such that} \\ \mathcal{A} \subseteq \{x \in \mathbb{R}^n : s\|x - z\| - y^T(x - z) \leq 0\}, \\ \mathcal{B} \subseteq \{x \in \mathbb{R}^n : s\|x - z\| - y^T(x - z) \geq 0\}. \end{cases}$$

$$\Downarrow$$

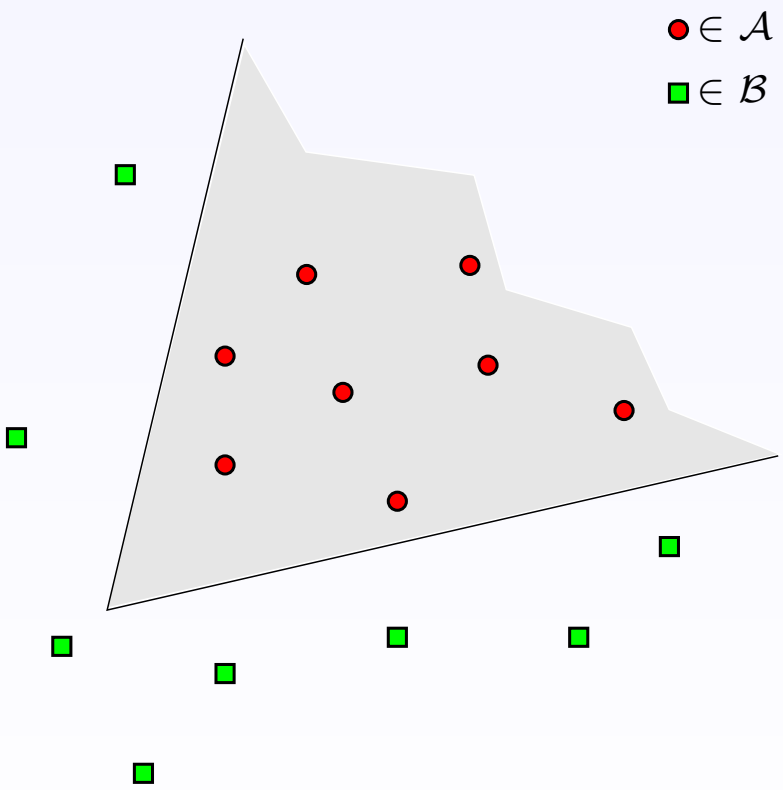
Find a solution $(z, y, s) \in Z \times \mathbb{S}_n \times [0,1]$ to the following nonlinear inequality system

$$(\mathrm{CS}_s) \quad \begin{cases} s\|a_i - z\| - y^T(a_i - z) \leq 0 & i = 1, \ldots, p, \\ s\|b_l - z\| - y^T(b_l - z) \geq 0 & , \ l = 1, \ldots, q. \end{cases}$$

Separator cones for the pair $(\mathcal{A}, \mathcal{B})$



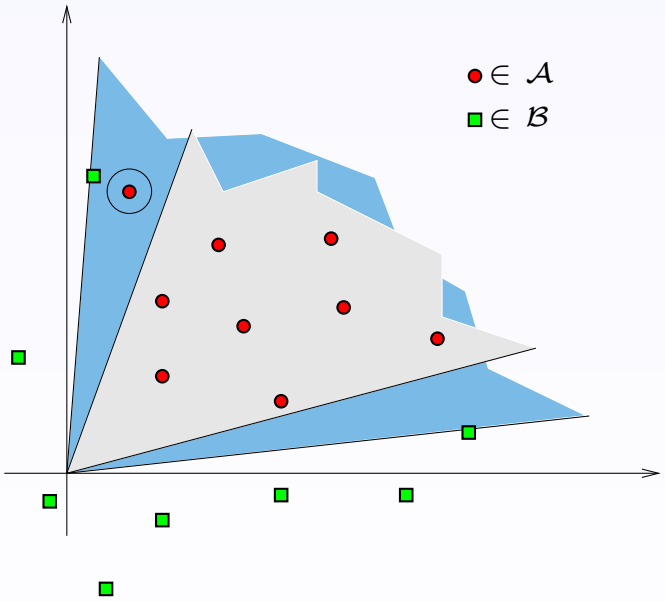
(a) Separator cone



(b) Strict separator cone

Geometric justification

A largest angle homogeneous separator, say $\Gamma_0(y_*, s_*)$, is better than an arbitrary homogeneous separator $\Gamma_0(y, s)$, because it reduces possibility of false-negative outcome.



Computing a largest angle homogeneous separator

Homogeneous (apex $z = 0$) vs. non-homogeneous case.

$$
(\mathrm{LAH}) \quad \begin{cases} \text{minimize } s \\ (y, s) \in \mathbb{R}^n \times \mathbb{R} \\ \|y\| = 1 \\ 0 \leq s \leq 1 \\ -a_i^T y + \|a_i\| s \leq 0 \quad i = 1, \ldots, p \\ b_j^T y - \|b_l\| s \ \leq 0 \quad l = 1, \ldots, q. \end{cases}
$$

.

- It is a nonconvex optimization problem;
- Constraint $\|y\| = 1$ can be replaced by $\|y\| \geq 1$;
- Introduction of the error function.

Outline

Outline

2. Current topics
   - Feature Selection

Feature Selection (FS) and Sparse Optimization in SVM

Finding the minimum number of attributes capable to ensure a reasonable classification correctness.

- Minimize the number of non zero components of vector w;
- Minimize the classification error.

Introduce the ℓ_0 pseudo-norm :

$$\|w\|_0 = number\, of\, non\, zero\, components\, of\, w$$

$$\min_{w,\gamma} Ce(w,\gamma) + \|w\|_0 \qquad (P_0),$$

The ℓ_0 pseudo-norm and the sparse optimization problem

Properties of function $x \mapsto \|x\|_0$:

i) $\|x\|_0 = 0 \quad \Leftrightarrow \quad x = 0$;

ii) $\|x + y\|_0 \leq \|x\|_0 + \|y\|_0$;

iii) $\|\alpha x\|_0 \neq |\alpha| \|x\|_0$;

iv) is *l.s.c.*, that is $\lim \inf_{x_k \to x} \|x_k\|_0 \geq \|x\|_0$ when $x_k \to x$;

v) $(\|.\|_p)^p \to \|.\|_0$ when $p \to 0$.

The sparse optimization problem:

$$f_0^* = \min_{x \in \mathbb{R}^n} f(x) + \|x\|_0 \qquad (P_0),$$

with $f : R^n \to R$, $n \geq 2$, convex and not necessarily differentiable

Sparse optimization and polyhedral norms

The k-norm of x, ($\|x\|_{[k]}$) is the sum of k maximal components (in modulus) of x, $k = 1, \ldots, n$.

The following hold:

i) $\|x\|_\infty = \|x\|_{[1]} \leq \ldots \leq \|x\|_{[k]} \leq \ldots \|x\|_{[n]} = \|x\|_1$;

ii) $\|x\|_0 \leq k \;\; \Rightarrow \;\; \|x\|_1 - \|x\|_{[s]} = 0, \;\; k \leq s \leq n.$

In particular it is, $1 \leq k \leq n$,

$$\|x\|_0 \leq k \;\; \Leftrightarrow \;\; \|x\|_1 - \|x\|_{[k]} = 0,$$

which allows us to replace $\|x\|_0 \leq k$ with a constraint in terms of difference of norms , which is DC.

The MINLP k-norm formulation

Introduce the binary variables y_k, $k = 1, \ldots, n$ and define

$$
\begin{aligned}
& f_I^* = \min_{x,y} f(x) + \sum_{k=1}^{n} y_k \\
& \|x\|_1 - \|x\|_{[k]} \leq M y_k, , \quad k = 1, \ldots, n \\
& y_k \in \{0, 1\}, \quad k = 1, \ldots, n.
\end{aligned}
$$

At the optimum it is:

$$\|x\|_1 - \|x\|_{[k]} = 0 \quad \Leftrightarrow \quad y_k = 0.$$

Continuous relaxation of the k-norm formulation

Replace the binary constraint $y_k \in \{0,1\}$ with $0 \leq y_k \leq 1$, $k = 1, \ldots, n$.

At the optimum the constraints $\|x\|_1 - \|x\|_{[k]} \leq M y_k, \;\; k = 1, \ldots, n$ are satisfied by equality, then it is:

$$y_k = \frac{1}{M}(\|x\|_1 - \|x\|_{[k]}).$$

The relaxation is then:

$$\min_{x \in \mathbb{R}^n} f(x) + \frac{1}{M} \sum_{k=1}^{n} \big(\|x\|_1 - \|x\|_{[k]}\big),$$

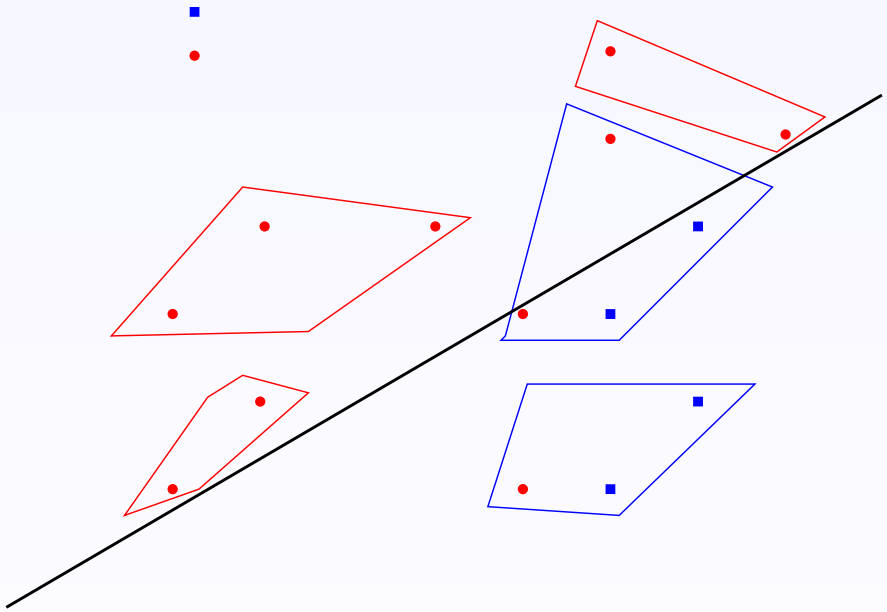
with penalization of $\|x\|_0 > k$, for $k = 1, \ldots, n$. Note that the problem is DC.

Outline

2. Current topics
   - Multiple Instance Learning

Bag classification

To classify bags of samples



Example: each bag is an image representing a set of animals. To discriminate the images containing a crocodile from those with no crocodile.

Bags and samples: the MIL paradigm I

Data:

- k negative and m positive subsets of samples (bags) with index sets I^- and I^+;
- Each sample $x_j \in \mathbb{R}^n$ belongs to exactly one bag;

Find a hyperplane

$$H(w, \gamma) \triangleq \{x \in \mathbb{R}^n | w^\top x = \gamma\}$$

which separates the two sets of bags in the following MIL sense:

- all the negative bags are entirely confined in the interior of one of the two halfspaces generated by H;
- each positive bag has at least one of its samples falling into the interior of the other halfspace.

Bags and samples: the MIL paradigm II

Formally $H(w, \gamma)$ is a separating hyperplane if and only if:

$$\begin{array}{lll} \text{for each } i \in I^- & \text{it is } w^\top x_j \leq \gamma - 1, & \text{for each } j \in J_i^-, \\ \text{for each } i \in I^+ & \text{it is } w^\top x_j \geq \gamma + 1, & \text{for at least one } \ j \in J_i^+. \end{array}$$

where $\{J_1^-, \ldots, J_k^-\}$ and $\{J_1^+, \ldots, J_m^+\}$ are the instance index sets of the negative and positive bags, respectively.

Bags and samples: the MIL paradigm III

Error function $e_i^-(w,\gamma)$ in classifying the negative bag $i \in I^-$:

$$e_i^-(w,\gamma) \triangleq \sum_{j \in J_i^-} \max\left\{0, w^\top x_j - \gamma + 1\right\}, \quad i = 1, \ldots, k,$$

A positive bag $i \in I^+$ is badly classified if $w^\top x_j < \gamma + 1, \forall j \in J_i^+$, that is if $\min_{j \in J_i^+}\{-w^\top x_j + \gamma + 1\} > 0$. Thus:

$$e_i^+(w,\gamma) \triangleq \max\left\{0, \min_{j \in J_i^+}\{-w^\top x_j + \gamma + 1\}\right\}, \quad i = 1, \ldots, m.$$

Overall error function $e(w,\gamma)$:

$$\sum_{i \in I^-} \sum_{j \in J_i^-} \max\left\{0, w^\top x_j - \gamma + 1\right\} + \sum_{i \in I^+} \max\left\{0, \min_{j \in J_i^+}\{-w^\top x_j + \gamma + 1\}\right\}$$

$e(w,\gamma)$ is nonsmooth, nonconvex and DC.

Outline

2. Current topics
   - Adversarial machine learning

The Adversarial ML in the classification framework

- Two players in action: the Defender and the Attacker.
- The Defender tries to design a classifier as much as possible robust.
- The Attacker tries to introduce small modifications in data to mislead the classifier.

An example: Calculating an adversarial sample in the SVM framework

Given a binary classifier $F : \mathbb{R}^n \mapsto \{-1, 1\}$ and any $x \in \mathbb{R}^n$ find the *smallest* (in the ℓ_0 pseudo-norm) perturbation $\delta \in \mathbb{R}^n$ such that

$$F(x + \delta) \neq F(x)$$

In SVM take $a \in \mathcal{A}$ and consider a separating hyperplane $H(w, \gamma)$ which correctly classifies a, ($a^\top w < \gamma - 1$). Look for $\delta \in \mathbb{R}^n$ such that $(a + \delta)^\top w \geq \gamma - 1$, that is $w^\top \delta \geq \rho_a = -a^\top w + \gamma - 1 \geq 0$. Thus solve:

$$\begin{array}{l} \min_\delta \|\delta\|_0 \\ \text{s.t.} \\ w^\top \delta \geq \rho_a \\ \|\delta\| \leq \Delta. \end{array}$$

References

A. Astorino, A. Fuduli, M. Gaudioso and E. Gorgone, *Classification problems and nonlinear optimization: separation by means of nonlinear surfaces*,Technical Note, DIMES, Unical, 2017.