Optimization for Machine Learning

Antonio Fuduli

Dipartimento di Matematica e Informatica - Università della Calabria
Email: antonio.fuduli@unical.it
Home page: http://www.mat.unical.it/∼fuduli

January 19, 2023

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Sign of square matrices

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Sign of square matrices

Definition (Positive semidefinite matrix)

A matrix $A \in \mathbb{R}^{n \times n}$ is positive semidefinite if

$$x^T Ax \geq 0 \quad \text{for any } x \in \mathbb{R}^n.$$

Definition (Positive definite matrix)

A matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if

$$x^T Ax > 0 \quad \text{for any } x \in \mathbb{R}^n, \text{ such that } x \neq 0.$$

Sign of square matrices

Definition (Negative semidefinite matrix)

A matrix $A \in \mathbb{R}^{n \times n}$ is negative semidefinite if

$$x^T Ax \leq 0 \quad \text{for any } x \in \mathbb{R}^n.$$

Definition (Negative definite matrix)

A matrix $A \in \mathbb{R}^{n \times n}$ is negative definite if

$$x^T Ax < 0 \quad \text{for any } x \in \mathbb{R}^n, \text{ such that } x \neq 0.$$

NOTE: In all the other cases the matrix A is said to be indefinite.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Sign of square matrices: characterizations

NOTE 1: A matrix $A \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if all the eigenvalues are ≥ 0.

NOTE 2: A matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if and only if all the eigenvalues are > 0.

NOTE 3: A matrix $A \in \mathbb{R}^{n \times n}$ is negative semidefinite if and only if all the eigenvalues are ≤ 0.

NOTE 4: A matrix $A \in \mathbb{R}^{n \times n}$ is negative definite if and only if all the eigenvalues are < 0.

Eigenvectors and eigenvalues

Definition (Eigenvector and eigenvalue)

Letting $A \in \mathbb{R}^{n \times n}$, $\lambda \in \mathbb{R}$ and $x \in R^n$ such that $x \neq 0$, the vector x is an eigenvector of A and λ is the corresponding eigenvalue if

$$Ax = \lambda x.$$

Computing the eigenvalues

$$Ax = \lambda x, \quad x \neq 0$$

$$\Downarrow$$

$$(Ax - \lambda x) = 0, \quad x \neq 0$$

$$\Downarrow$$

$$(A - \lambda I)x = 0, \quad x \neq 0$$

$$\Downarrow$$

The columns of the matrix $A - \lambda I$ are linearly dependent, i.e. the matrix $A - \lambda I$ is singular, i.e.

$$\underbrace{det(A - \lambda I) = 0}_{\text{characteristic polynomial}} \qquad .$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA
E INFORMATICA

Norm of a vector

Definition

The norm, denoted by $\|\cdot\|$, is a map

$$\|\cdot\| : \mathbb{R}^n \mapsto \mathbb{R}_+$$

such that

1. $$\|x\| = 0 \Rightarrow x = 0;$$
2. $$\|x+y\| \le \|x\| + \|y\| \quad \text{for any } x, y \in \mathbb{R}^n;$$
3. $$\|\alpha x\| = |\alpha|\|x\| \quad \text{for any } \alpha \in \mathbb{R} \text{ and } x \in \mathbb{R}^n.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Some norms

1. L_1-norm: $\|x\|_1 = \sum_{j=1}^{n} |x_j|$;
2. L_2-norm (Euclidean): $\|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2}$;
3. L_∞-norm: $\|x\|_\infty = \max_{1 \le j \le n} |x_j|$.

NOTE 1: If not differently specified, by $\|x\|$ we mean the Euclidean norm of vector x.
NOTE 2: The norm is a convex function.
NOTE 3: The norm is a nonsmooth function. In fact, in case $n = 1$, we have $\|x\|_1 = \|x\|_2 = \|x\|_\infty = |x|$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A note on the Euclidean norm

1.
$$x \in \mathbb{R}^n \Rightarrow \|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2}$$

2. Since the Euclidean norm is a nonsmooth function, we adopt the following trick:
$$\min_x \|x\|_2 \Leftrightarrow \min_x \frac{1}{2}\|x\|_2^2.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

PART II

ELEMENTS OF NONLINEAR PROGRAMMING

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The optimization problems

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Optimization problems: some definitions

$$P \begin{cases} \min\limits_{x} & f(x) \\ & x \in X, \end{cases}$$

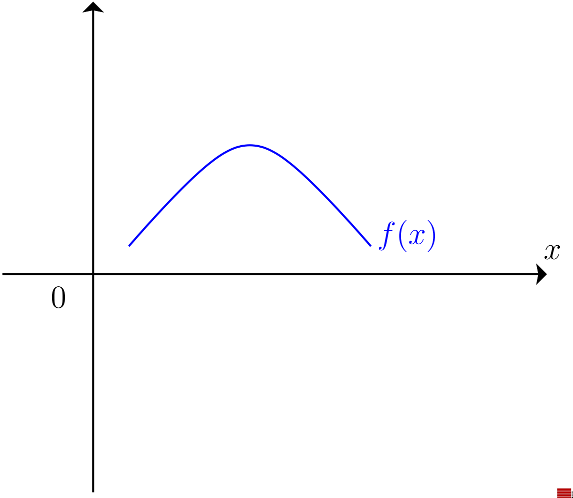
where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $X \subseteq \mathbb{R}^n$.

NOTE:

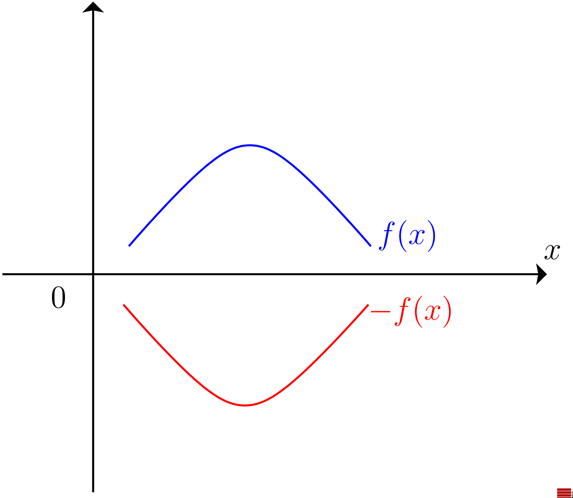
$$\begin{cases} \max\limits_{x} & f(x) \\ & x \in X, \end{cases} \Leftrightarrow \begin{cases} -\min\limits_{x} & -f(x) \\ & x \in X, \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

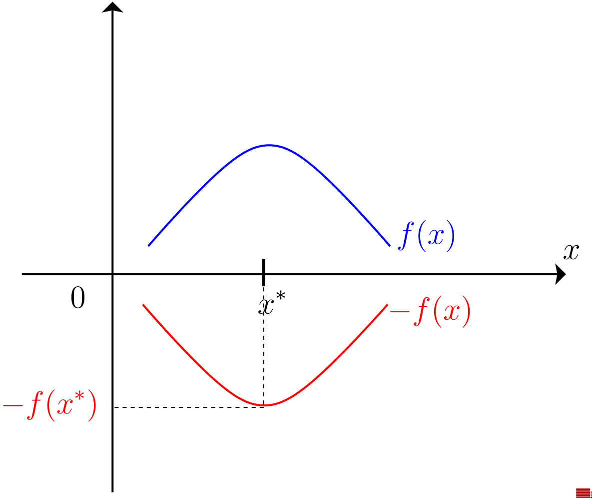
Optimization problems: some definitions



Optimization problems: some definitions

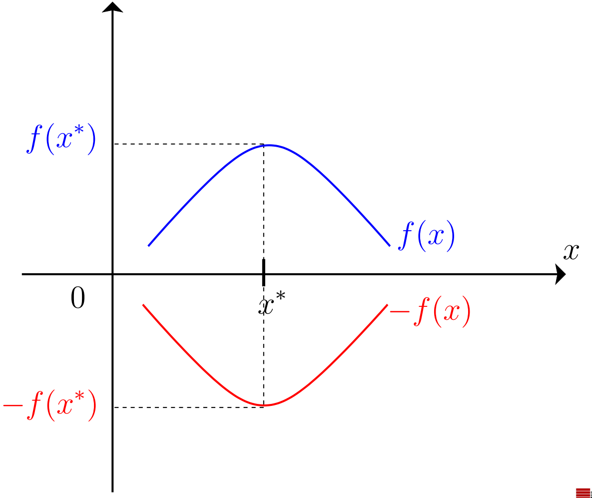


Optimization problems: some definitions



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Optimization problems: some definitions



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Global and local minima

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Global and local minima

$$P \begin{cases} \min_{x} & f(x) \\ & x \in X \end{cases}$$

Definition (Global minimum)

A point x^* is a global minimum for P if

- $x^* \in X$;
- $f(x^*) \leq f(x) \quad$ for any $x \in X$.

Global and local minima

$$P \begin{cases} \min\limits_{x} & f(x) \\ & x \in X, \end{cases}$$

Definition (Local minimum)

A point x^* is a local minimum for P if

- $x^* \in X$;
- there exists a neighbourhood N of x^*, such that $f(x^*) \leq f(x)$ for any $x \in N \cap X$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Global and local minima

$$P \begin{cases} \min\limits_{x} & f(x) \\ & x \in X \end{cases}$$

Definition (Strict local minimum)

A point x^* is a strict local minimum for P if

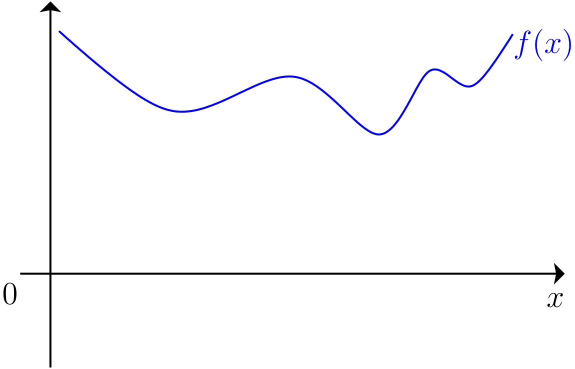
- $x^* \in X$;
- there exists a neighbourhood N of x^*, such that $f(x^*) < f(x)$ for any $x \in N \cap X$, with $x \neq x^*$.

NOTE: x^* is a global minimum $\Rightarrow$ x^* is a local minimum.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Global and local minima

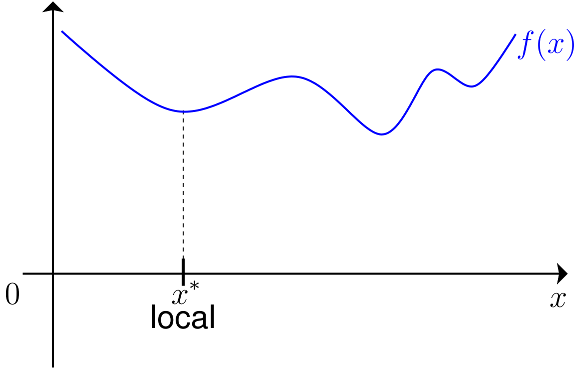
$$P \left\{ \min_{x} \quad f(x) \right.$$



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Global and local minima

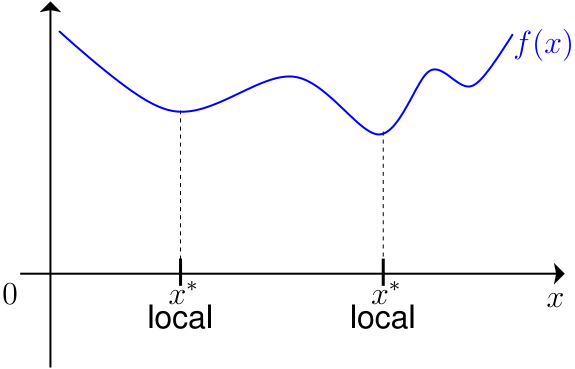
$$P \left\{ \min_{x} \quad f(x) \right.$$



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

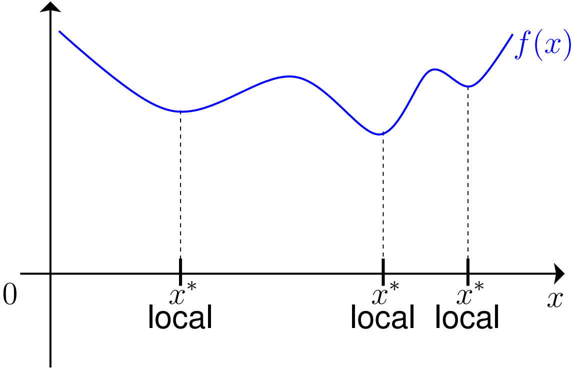
Global and local minima

$$P \left\{ \min_{x} \quad f(x) \right.$$



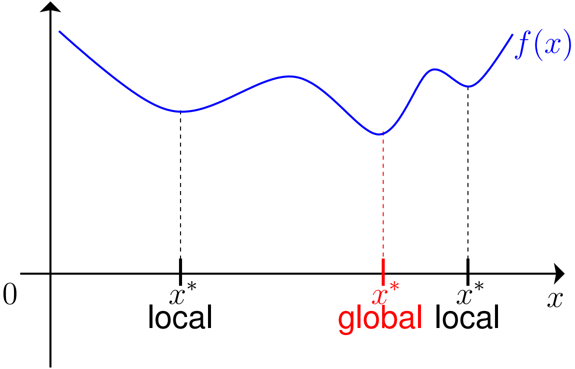
Global and local minima

$$P \left\{ \min_{x} \quad f(x) \right.$$



Global and local minima

$$P \left\{ \min_x \quad f(x) \right.$$



Convexity

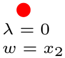
Convex combination of two vectors

Definition (Convex combination of two vectors)

Let $x_1, x_2 \in R^n$. The convex combination of x_1 and x_2 is the vector

$$w = \lambda x_1 + (1-\lambda)x_2,$$

with $\lambda \in [0,1]$.



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex combination of two vectors

Definition (Convex combination of two vectors)

Let $x_1, x_2 \in R^n$. The convex combination of x_1 and x_2 is the vector

$$w = \lambda x_1 + (1 - \lambda) x_2,$$

with $\lambda \in [0, 1]$.

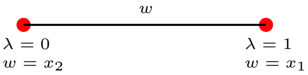
Convex combination of two vectors

Definition (Convex combination of two vectors)

Let $x_1, x_2 \in R^n$. The convex combination of x_1 and x_2 is the vector

$$w = \lambda x_1 + (1-\lambda) x_2,$$

with $\lambda \in [0,1]$.



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex functions

Definition (Convex function)

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if for any $x_1, x_2 \in \mathbb{R}^n$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2),$$

for any $\lambda \in [0, 1]$.

Definition (Strictly convex function)

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is strictly convex if for any $x_1, x_2 \in \mathbb{R}^n$, with $x_1 \neq x_2$,

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2),$$

for any $\lambda \in (0, 1)$.

NOTE: The sum of convex functions is a convex function.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex functions

Definition (Concave function)

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is concave if $-f(x)$ is convex.

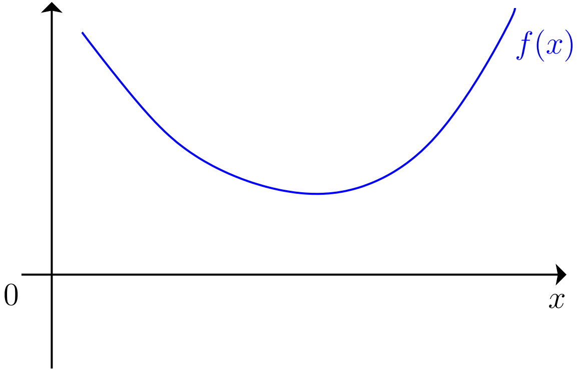
Definition (Strictly concave function)

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is strictly concave if $-f(x)$ is strictly convex.

NOTE: A linear function is at the same time convex and concave.

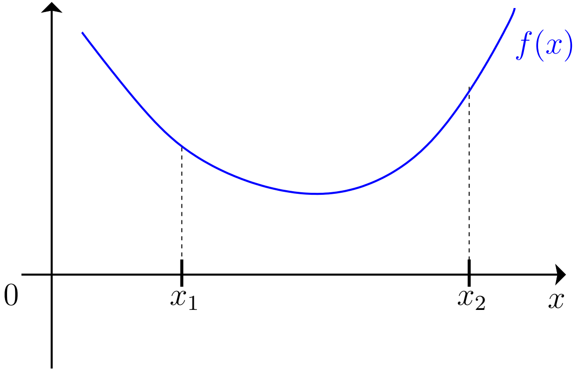
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex functions

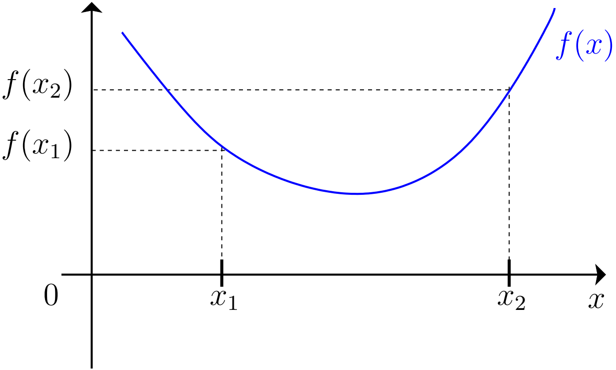


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex functions

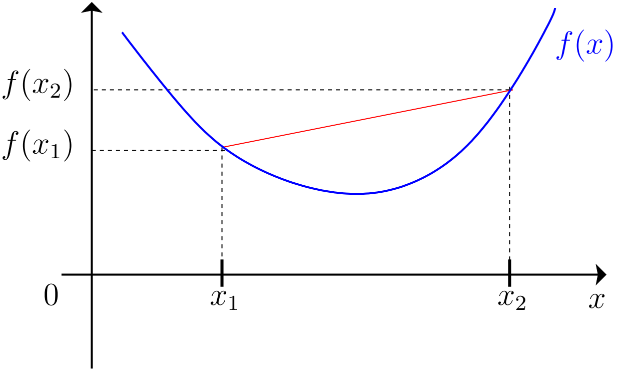


Convex functions



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

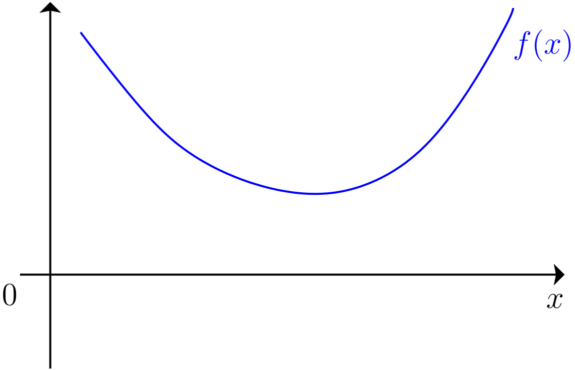
Convex functions



$$\underbrace{f(\lambda x_1 + (1-\lambda)x_2)}_{\text{arc}} \le \underbrace{\lambda f(x_1) + (1-\lambda) f(x_2)}_{\text{chord}}, \text{ for any } \lambda \in [0,1]$$

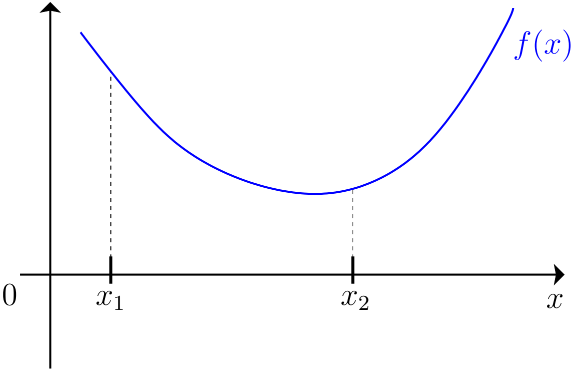
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex functions

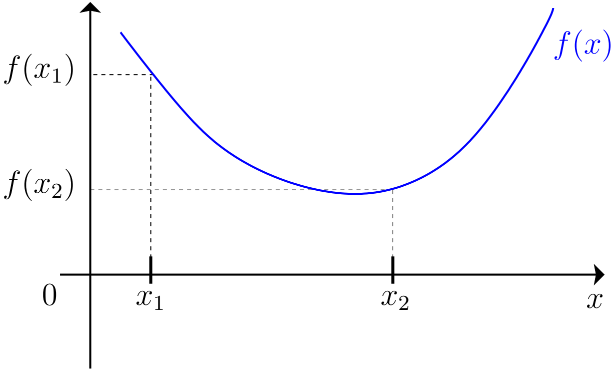


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

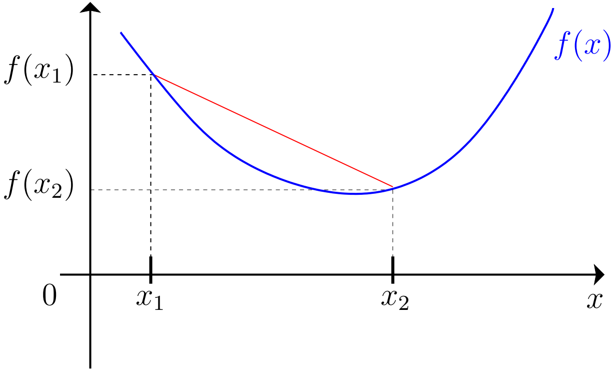
Convex functions



Convex functions

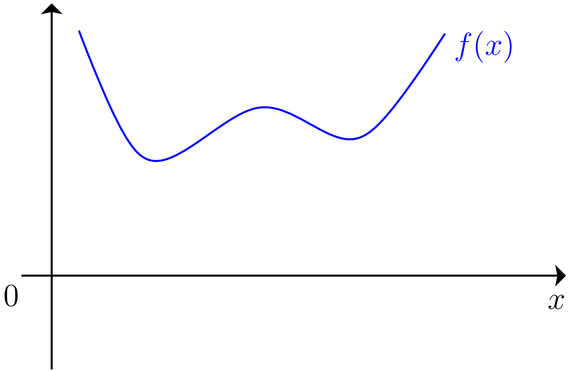


Convex functions



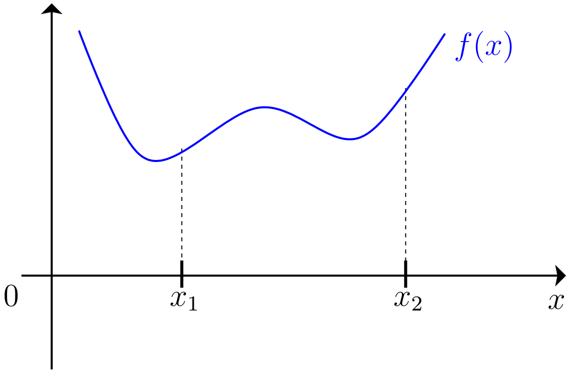
$$\underbrace{f(\lambda x_1 + (1-\lambda)x_2)}_{\text{arc}} \leq \underbrace{\lambda f(x_1) + (1-\lambda) f(x_2)}_{\text{chord}}, \text{ for any } \lambda \in [0,1]$$

Convex functions

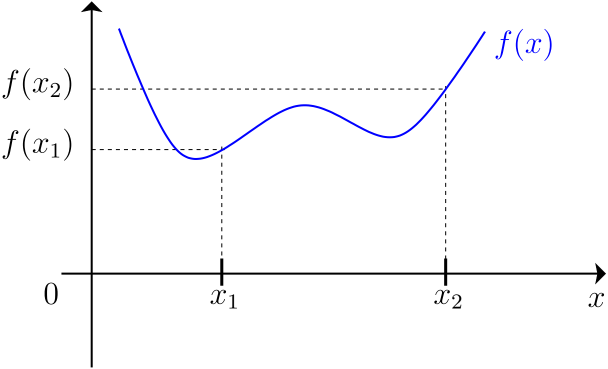


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

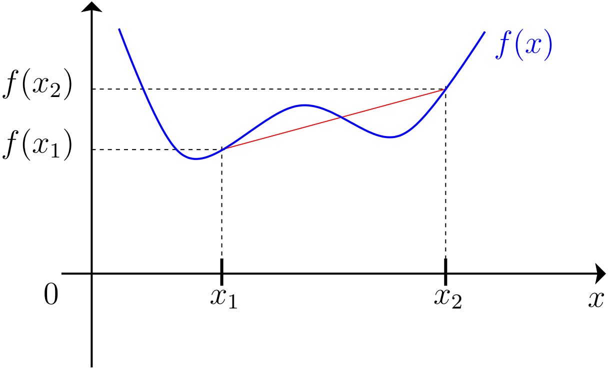
Convex functions



Convex functions



Convex functions



$f(\lambda x_1 + (1-\lambda)x_2) \not\leq \lambda f(x_1) + (1-\lambda) f(x_2)$, for any $\lambda \in [0,1]$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex sets

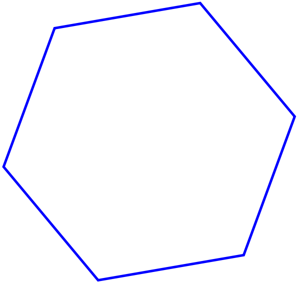
Definition (Convex set)

A set $X \subseteq \mathbb{R}^n$ is convex if for any $x_1, x_2 \in X$, the vector

$$w = \lambda x_1 + (1-\lambda)x_2 \in X$$

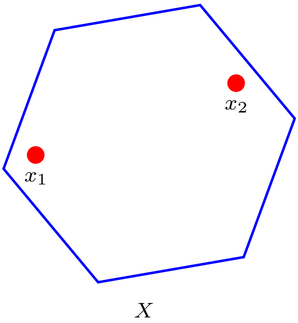
for any $\lambda \in [0, 1]$.

Convex sets



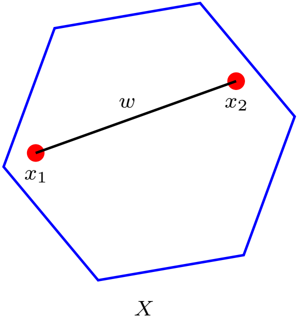
X

Convex sets



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

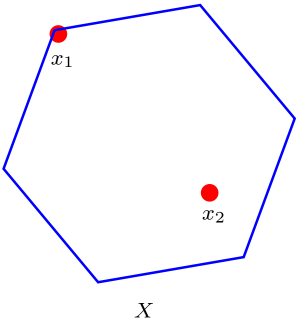
Convex sets



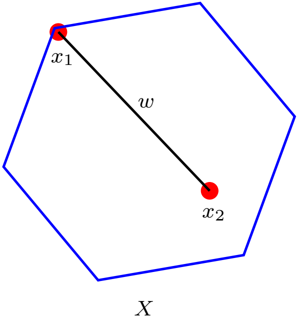
$w = \lambda x_1 + (1-\lambda)x_2 \in X$ for any $\lambda \in [0, 1]$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Convex sets

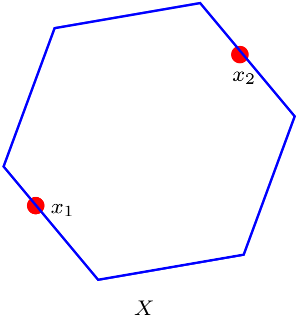


Convex sets

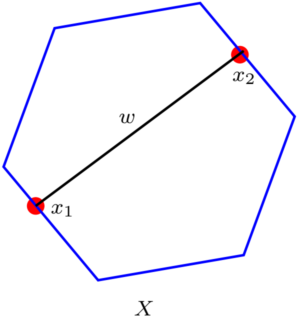


$w = \lambda x_1 + (1 - \lambda) x_2 \in X$ for any $\lambda \in [0, 1]$

Convex sets

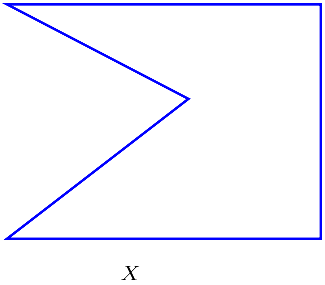


Convex sets

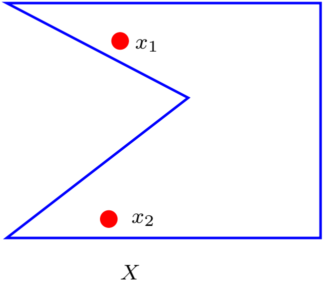


$w = \lambda x_1 + (1-\lambda)x_2 \in X$ for any $\lambda \in [0,1]$

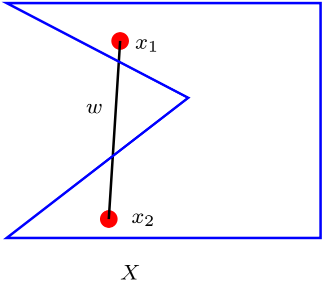
Convex sets



Convex sets



Convex sets



$w = \lambda x_1 + (1-\lambda)x_2 \notin X$ for any $\lambda \in [0,1]$

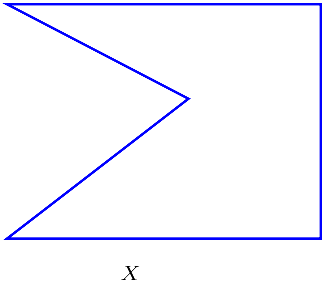
Convex sets

Definition (Convex hull)

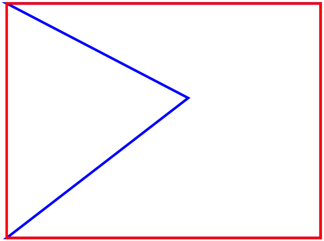
Given a set $X \subset \mathbb{R}^n$, the convex hull of X is the smallest convex set containing X. It is indicated by $\mathrm{conv}(X)$.

NOTE: If X is convex, then $\mathrm{conv}(X) = X$.

Convex hull

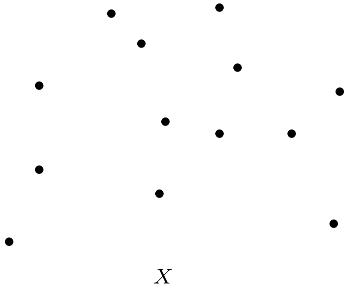


Convex hull

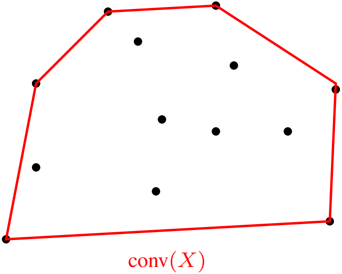


$\text{conv}(X)$

Convex hull



Convex hull



$$P \begin{cases} \min\limits_{x} & f(x) \\ & x \in X, \end{cases}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $X \subseteq \mathbb{R}^n$.

NOTE: If f is a convex function and X is a convex set, then P is a convex program.

Optimality conditions

The unconstrained case: optimality conditions

If $X = \mathbb{R}^n$, then we have the following unconstrained optimization problem:

$$P \left\{ \min_{x \in \mathbb{R}^n} \quad f(x) \right.$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$.

The unconstrained case: optimality conditions

Assumption: $f \in C^2$, i.e. the first and second order derivatives exist and are continuous.

$$\text{Gradient: } \nabla f(x) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The unconstrained case: optimality conditions

Assumption: $f \in C^2$, i.e. the first and second order derivatives exist and are continuous.

$$\text{Hessian matrix: } \nabla^2 f(x) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

.

The unconstrained case: optimality conditions

$$P \left\{ \min_{x} \quad f(x) \right.$$

Theorem (First order necessary condition)

x^* is a local minimum $\Rightarrow \nabla f(x^*) = 0$.

NOTE 1: We call x^* a stationary point if $\nabla f(x^*) = 0$.
NOTE 2: If f is convex then x^* is a global minimum $\Leftrightarrow \nabla f(x^*) = 0$.

Theorem (Second order necessary condition)

x^* is a local minimum $\Rightarrow \nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

The unconstrained case: optimality conditions

$$P \left\{ \min_{x} \quad f(x) \right.$$

Theorem (Second order sufficient condition)

If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite $\Rightarrow$ x^* is a strict local minimum.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The constrained case: optimality conditions

$$P \begin{cases} \min\limits_{x} \quad f(x) \\ \qquad \left.\begin{array}{ll} g_i(x) = 0 & i \in E \\ g_i(x) \geq 0 & i \in I \end{array}\right\} \triangleq X \text{ (feasible region)} \end{cases}$$

where $f, g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i \in E \cup I$.

NOTE 1: If $f(x) = \frac{1}{2}x^T Hx + c^T x$, with $H \in \mathbb{R}^{n\times n}$ is symmetric, and all the functions g_i are linear, then P is a quadratic program.

NOTE 2: If function f is linear (i.e. $f = c^T x$) and all the functions g_i are linear, then P is a linear program.

The constrained case: optimality conditions

$$P \begin{cases} \min\limits_{x} \quad f(x) \\ \qquad \left.\begin{array}{ll} g_i(x) = 0 & i \in E \\ g_i(x) \geq 0 & i \in I \end{array}\right\} \stackrel{\triangle}{=} X \text{ (feasible region)} \end{cases}$$

where $f, g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i \in E \cup I$.

SOME PRELIMINARIES

Definition (Active constraint)

Given a point $\bar{x} \in X$, a constraint g_i, $i \in E \cup I$, is active at $\bar{x}$ if $g_i(\bar{x}) = 0$.

$A(\bar{x}) \stackrel{\triangle}{=}$ index set of the active constraints at $\bar{x}$.

The constrained case: optimality conditions

$$P \begin{cases} \min\limits_{x} & f(x) \\ & \left.\begin{array}{ll} g_i(x) = 0 & i \in E \\ g_i(x) \geq 0 & i \in I \end{array}\right\} \triangleq X \text{ (feasible region)} \end{cases}$$

where $f, g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i \in E \cup I$.

$$A(\bar{x}) \triangleq \text{index set of the active constraints at } \bar{x}.$$

Definition (Linear Independence Constraint Qualification - LICQ)

Given a point $\bar{x} \in X$, we say that the Linear Independence Constraint Qualification holds at $\bar{x}$, if the set

$$\{\nabla g_i(\bar{x}) \mid i \in A(\bar{x})\}$$

is linearly independent.

The constrained case: optimality conditions

$$P \begin{cases} \min\limits_{x} & f(x) \\ & \left.\begin{array}{ll} g_i(x) = 0 & i \in E \\ g_i(x) \geq 0 & i \in I \end{array}\right\} \triangleq X \text{ (feasible region)} \end{cases}$$

where $f, g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i \in E \cup I$.

Definition (Lagrangian function)

The Lagrangian function of problem P is the following:

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in E} \lambda_i g_i(x) - \sum_{i \in I} \lambda_i g_i(x),$$

with $\lambda_i \geq 0, \quad i \in I$.

NOTE: The variables λ_i, $i \in E \cup I$ are the Lagrangian multipliers.

The constrained case: optimality conditions

Assumptions: $f \in C^1$; $g_i \in C^1, \; i \in E \cup I$

Theorem (Karush Kuhn Tucker conditions - KKT)

Let x^* be a local mimimum of P and let LICQ hold at x^*. Then there exist λ^* such that

$$KKT \begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) = 0 & & \\ g_i(x^*) = 0 & i \in E & \leftarrow \textit{ feasibility} \\ g_i(x^*) \geq 0 & i \in I & \leftarrow \textit{ feasibility} \\ \lambda_i^* \geq 0 & i \in I & \\ \lambda_i^* g_i(x^*) = 0 & i \in E \cup I & \leftarrow \textit{ complementarity conditions.} \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The constrained case: optimality conditions

NOTE

$$\mathcal{L}(x,\lambda) = f(x) - \sum_{i\in E}\lambda_i g_i(x) - \sum_{i\in I}\lambda_i g_i(x)$$

$$\nabla_x \mathcal{L}(x^*,\lambda^*) = 0$$

$$\Updownarrow$$

$$\nabla f(x^*) - \sum_{i\in E}\lambda_i^* \nabla g_i(x^*) - \sum_{i\in I}\lambda_i^* \nabla g_i(x^*) = 0$$

$$\Updownarrow$$

$$\nabla f(x^*) = \sum_{i\in E}\lambda_i^* \nabla g_i(x^*) + \sum_{i\in I}\lambda_i^* \nabla g_i(x^*).$$

$\Updownarrow$ since $\lambda_i^* g_i(x^*) = 0 \quad i \in E \cup I$

$$\nabla f(x^*) = \sum_{i\in A(x^*)}\lambda_i^* \nabla g_i(x^*).$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

The Wolfe dual

The Wolfe dual (Wolfe, 1961 [Wol61])

Definition (Dual)

Given an optimization problem (called primal), its dual is another optimization problem associated to the primal (by means of suitable rules).

The Wolfe dual (Wolfe, 1961 [Wol61])

PRIMAL

$$P\begin{cases} \min\limits_{x} \quad f(x) \\ \left.\begin{array}{ll} g_i(x) = 0 & i \in E \\ g_i(x) \geq 0 & i \in I \end{array}\right\} \text{feasible region } X \end{cases}$$

where $f, g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i \in E \cup I$.

Assumptions: $f \in C^1$; $g_i \in C^1,\ i \in E \cup I$

The Wolfe dual of problem P is defined as follows:

$$D\begin{cases} \max\limits_{x,\lambda} \quad \mathcal{L}(x,\lambda) \\ \nabla_x \mathcal{L}(x,\lambda) = 0 \\ \lambda_i \geq 0, \quad i \in I \end{cases}$$

The Wolfe dual of a linear program

Theorem

Given the problem

$$P \begin{cases} \min\limits_{x} & c^T x \\ & Ax \geq b \\ & x \geq 0, \end{cases}$$

the Wolfe dual of P is the ordinary dual.

Proof.

$$\mathcal{L}(x,\lambda,\mu) = \overbrace{c^T x + \lambda^T(b - Ax) - \mu^T x}^{(c - A^T\lambda - \mu)^T x + \lambda^T b} \quad \text{and} \quad \nabla_x \mathcal{L} = c - A^T\lambda - \mu$$

$$\Downarrow$$

$$D \begin{cases} \max\limits_{x,\lambda,\mu} & \overbrace{(c - A^T\lambda - \mu)^T}^{0} x + \lambda^T b \\ & c - A^T\lambda - \mu = 0 \\ & \lambda, \mu \geq 0. \end{cases} \Leftrightarrow D \begin{cases} \max\limits_{x,\lambda} & \lambda^T b \\ & \underbrace{c - A^T\lambda \geq 0}_{A^T\lambda \leq c} \\ & \lambda \geq 0. \end{cases} \Leftrightarrow D \begin{cases} \max\limits_{x,\lambda} & \lambda^T b \\ & A^T\lambda \leq c \\ & \lambda \geq 0, \end{cases}$$

□

DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

Some notions on the algorithms

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Sketch of the algorithms

- x_0: starting point;
- x_1, x_2,...: next iterates;
- unconstrained case: stop in case a stationary point is generated;
- constrained case: stop in case a KKT point is generated.

Line search methods

- $$x_{k+1} = x_k + \alpha_k d_k,$$

  where $\alpha_k > 0$ is the stepsize and d_k is the search direction.
- Once a search direction d_k is computed, the stepsize α_k is determined by solving the following univariate problem:

$$LS \left\{ \min_{\alpha} f(x_k + \alpha d_k) \right.$$

- Exact line search if problem LS is exactly solved.
- Inexact line search if problem LS is approximately solved.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Trust region methods

- $$x_{k+1} = x_k + d_k,$$

  where d_k is the search direction, obtained by solving the following problem:

  $$TR \left\{ \min_d m_k(x_k + d), \right.$$

  where m_k is a "model function", well approximating f in a neighbourhood of x_k.
- Generally:

  $$TR \left\{ \begin{array}{ll} \min_d & m_k(x_k + d) \\ & \|d\| \leq \Delta_k, \end{array} \right.$$

  with d_k being the radius of the trust region.

The unconstrained case: descent directions

$$P \left\{ \min_{x} \quad f(x), \right.$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $f \in C^2$.

Definition (Descent direction)

Let $\bar{x} \in \mathbb{R}^n$. The vector $\bar{d}$ is a descent direction for problem P at $\bar{x}$ if there exists $\bar{\alpha} > 0$ such that

$$f(\bar{x} + \alpha \bar{d}) < f(\bar{x}), \qquad \text{for any } \alpha \in]0, \bar{\alpha}].$$

NOTE 1: If $\nabla f(\bar{x})^T \bar{d} < 0$, then $\bar{d}$ is a descent direction at $\bar{x}$.

NOTE 2: In case f is convex, if $\bar{d}$ is a descent direction at $\bar{x}$, then $\nabla f(\bar{x})^T \bar{d} < 0$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The unconstrained case: the steepest descent method

By the Taylor theorem:

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d$$

$$\Downarrow$$

$$P_k \left\{ \min_d \quad f(x_k) + \nabla f(x_k)^T d \right.$$

$$\Downarrow$$

$$P_k \left\{ \begin{array}{ll} \min\limits_d & \nabla f(x_k)^T d \\ & \|d\| = 1 \end{array} \right.$$

$$\Downarrow$$

$$P_k \left\{ \begin{array}{ll} \min\limits_d & \nabla f(x_k)^T d \\ & \frac{1}{2}\|d\|^2 = \frac{1}{2} \end{array} \right.$$

The unconstrained case: the steepest descent method

$$\mathcal{L}(d,\lambda) = \nabla f(x_k)^T d - \lambda\left(\frac{1}{2}\|d\|^2 - \frac{1}{2}\right)$$

$$\Downarrow$$

$$\nabla_d \mathcal{L}(d,\lambda) = \nabla f(x_k) - \lambda d = 0 \overbrace{\Rightarrow}^{if \lambda \neq 0} d = \frac{\nabla f(x_k)}{\lambda}$$

$$\frac{1}{2}\|d\|^2 = \frac{1}{2} \Rightarrow \|d\|^2 = 1 \Rightarrow \frac{\|\nabla f(x_k)\|^2}{\lambda^2} = 1 \Rightarrow \lambda = -\|\nabla f(x_k)\|$$

$$\Downarrow$$

$$d = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|},$$

which is a descent direction.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The unconstrained case: Newton method

By the Taylor theorem:

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$$

$\Downarrow$ Assumption: $\nabla^2 f(x_k)$ positive definite

$$P_k \left\{ \min_d \quad \underbrace{f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d}_{m_k(d)} \right.$$

$$\Downarrow$$

The unconstrained case: Newton method

$$\nabla m_k(d) = \nabla f(x_k) + \nabla^2 f(x_k) d$$

$$\Downarrow$$

$$d = -\nabla^2 f(x_k)^{-1} \nabla f(x_k),$$

which is a descent direction.

NOTE:

- $m_k(0) = f(x_k)$;
- $\nabla m_k(0) = \nabla f(x_k)$;
- $\nabla^2 m_k(0) = \nabla^2 f(x_k)$.

PART III

THE LAGRANGIAN RELAXATION

Relaxed problems

Relaxed problems

Definition (Relaxed problem)

Given

$$P \begin{cases} \min\limits_x & f(x) \\ & x \in X_P \end{cases} \quad \text{and} \quad R \begin{cases} \min\limits_x & g(x) \\ & x \in X_R, \end{cases}$$

R is a relaxed problem with respect to P if

1. $X_R \supseteq X_P$;
2. $g(x) \leq f(x)$, for any $x \in X_P$.

Relaxed problems

Theorem (Properties of the relaxed problems)

1. *R infeasible $\Rightarrow$ P infeasible.*
2. *Let x_P^* be an optimal solution to P and x_R^* be an optimal solution to R. Then $g(x_R^*) \leq f(x_P^*)$.*
3. *Let x_R^* be an optimal solution to R. If $x_R^* \in X_P$ and $f(x_R^*) = g(x_R^*)$, then x_R^* is optimal to P.*

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Relaxed problems

Proof.

1. $X_R = \emptyset \Rightarrow X_P = \emptyset$.
2. $$g(x_R^*) \le g(x) \quad \text{for any } x \in X_R$$
   $$\Downarrow$$
   $$g(x_R^*) \le g(x) \quad \text{for any } x \in X_P.$$
   Moreover, by the definition of relaxed problem:
   $$g(x) \le f(x) \quad \text{for any } x \in X_P.$$
   Combining the last two inequalities, we have:
   $$g(x_R^*) \le f(x) \quad \text{for any } x \in X_P$$
   and then:
   $$g(x_R^*) \le f(x_P^*).$$
3. As above,
   $$\underbrace{g(x_R^*)}_{=f(x_R^*)} \le f(x) \quad \text{for any } x \in X_P.$$

□

The Lagrangian relaxation problem

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The Lagrangian relaxation problem

Let ILP be the following integer program:

$$ILP \left\{ \begin{array}{ll} \min\limits_x & c^T x \\ & \left. \begin{array}{c} \overbrace{Ax \geq b}^{m \text{ difficult constraints}} \\ Bx \geq d \\ x \geq 0 \\ x \text{ int} \end{array} \right\} \triangleq X \text{ (feasible region)} \end{array} \right.$$

The Lagrangian relaxation problem

Definition (Lagrangian relaxation)

Let $\lambda \in \mathbb{R}^m$ such that $\lambda \geq 0$. The Lagrangian relaxation of ILP, with respect to the constraints $Ax \geq b$, is the following problem:

$$
LR(\lambda) \left\{ \begin{array}{ll} z^*_{LR}(\lambda) = \min\limits_x & \overbrace{c^T x - \lambda^T (Ax - b)}^{\mathcal{L}(x,\lambda)} \\ & \left. \begin{array}{c} Bx \geq d \\ x \geq 0 \\ x \text{ int} \end{array} \right\} \triangleq X_{LR} \end{array} \right.
$$

The Lagrangian relaxation problem

Theorem

For any $\lambda \geq 0$, $LR(\lambda)$ is a relaxed problem with respect to ILP.

Proof.

1. $X_{LR} \supseteq X$.
2. Let $\bar{x} \in X$. Then:

$$A\bar{x} \geq b \Rightarrow A\bar{x} - b \geq 0$$

$$\Downarrow$$

$$\underbrace{c^T\bar{x} - \underbrace{\lambda^T}_{\geq 0}\underbrace{(A\bar{x} - b)}_{\geq 0}}_{g(\bar{x})} \leq \underbrace{c^T\bar{x}}_{f(\bar{x})}$$

□

The Lagrangian relaxation problem

Theorem

Let $x^*_{LR}(\lambda)$ be an optimal solution to $LR(\lambda)$. If $x^*_{LR}(\lambda) \in X$ and $\lambda^T(Ax^*_{LR}(\lambda) - b) = 0$, then $x^*_{LR}(\lambda)$ is optimal to ILP.

Proof.

See property 3 of the theorem relative to the properties of the relaxed problems. $\square$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The Lagrangian dual

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The Lagrangian dual

The Lagrangian dual of ILP is the following problem:

$$LD \left\{ z^*_{LD} = \max_{\lambda \geq 0} z_{LR}(\lambda), \right.$$

i.e.

$$LD \left\{ \begin{array}{l} z^*_{LD} = \max_{\lambda \geq 0} \min_{x} \quad \overbrace{c^T x - \lambda^T (Ax - b)}^{\mathcal{L}(x,\lambda)} \\ \qquad\qquad\qquad\qquad \left. \begin{array}{c} Bx \geq d \\ x \geq 0 \\ x \text{ int} \end{array} \right\} \stackrel{\triangle}{=} X_{LR} \end{array} \right.$$

The Lagrangian dual

i.e.

$$LD \begin{cases} z^*_{LD}(\lambda) = \max\limits_{\lambda \geq 0} \min\limits_{x} & \overbrace{c^T x - \lambda^T (Ax - b)}^{\mathcal{L}(x,\lambda)} \\ & x \in \text{conv}(X_{LR}) \end{cases}$$

Theorem

z^*_{LD} is the optimal objective function value of the following problem:

$$\overline{LD} \begin{cases} z^*_{LD} = \min\limits_{x} & c^T x \\ & x \in \text{conv}(X_{LR}) \cap X_b, \end{cases}$$

where

$$X_b \triangleq \{x \in \mathbb{R}^n \mid Ax \geq b\}.$$

The Lagrangian dual and the continuous relaxation

On one hand:

$$\overline{LD} \begin{cases} z^*_{LD} = \min\limits_x & c^T x \\ & x \in \operatorname{conv}(X_{LR}) \cap X_b. \end{cases}$$

On the other hand, the continuous relaxation of ILP is:

$$LP \begin{cases} z^*_{LP} = \min\limits_x & c^T x \\ & \overbrace{Ax \geq b}^{X_b} \\ & \left.\begin{array}{c} Bx \geq d \\ x \geq 0 \end{array}\right\} \triangleq X_{d0} \supseteq \operatorname{conv}(X_{LR}), \end{cases}$$

As a consequence, LP is a relaxed problem with respect to $\overline{LD}$. Then:

$$z^*_{LP} \leq z^*_{LD}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The integrality property

We say that the integrality property holds if the extreme points of X_{d0} are integer. In such case:

$$\mathrm{conv}(X_{RL}) = X_{d0}.$$

As a consequence, LP and $\overline{LD}$ coincide and $z^*_{LP} = z^*_{LD}$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The Lagrangian dual of a linear program

Theorem

Given the problem

$$P \begin{cases} \min\limits_{x} & c^T x \\ & Ax \geq b \\ & x \geq 0, \end{cases}$$

the Lagrangian dual of P is the ordinary dual.

Proof.

Relaxing the constraints $Ax \geq b$, the Lagrangian dual of P is

$$LD \left\{ \max_{\lambda \geq 0} \lambda^T b + \min_{x \geq 0} (c - A^T \lambda)^T x. \right.$$

If there exists j such that $c_j - A_j^T \lambda < 0$, with A_j being the jth column of A, then the min-problem is unbounded. Then we need to impose $c - A^T \lambda \geq 0$ and in such case $x^*_{LR}(\lambda) = 0$. As a consequence:

$$LD \begin{cases} \max\limits_{\lambda \geq 0} & \lambda^T b \\ & A^T \lambda \leq c. \end{cases}$$

□

PART IV

NUMERICAL OPTIMIZATION AND MACHINE LEARNING

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Introduction to Machine Learning

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Machine Learning

Definition (Arthur Samuel (1901-1990), 1959)

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Definition (Tom Mitchell, Machine Learning, McGraw Hill, 1997)

The field of Machine Learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine Learning and pattern classification

- A relevant part of Machine Learning is constituted by the pattern classification, whose objective is to categorize different objects into two or more classes, on the basis of their similarities.
- From the mathematical point of view, the objects can be represented as vectors of n real numbers (points in $\mathbb{R}^n$), where each number describes a feature of the object (feature vector).
- Constructing a classifier means to generate one or more surfaces, which separate the objects into two ore more different classes.
- The generation of the surfaces is performed by learning from some objects (training set) whose class is known (for example on the basis of the experience).
- Why? The aim is to predict the class of any new object, after training the classifier on the training set.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: an example

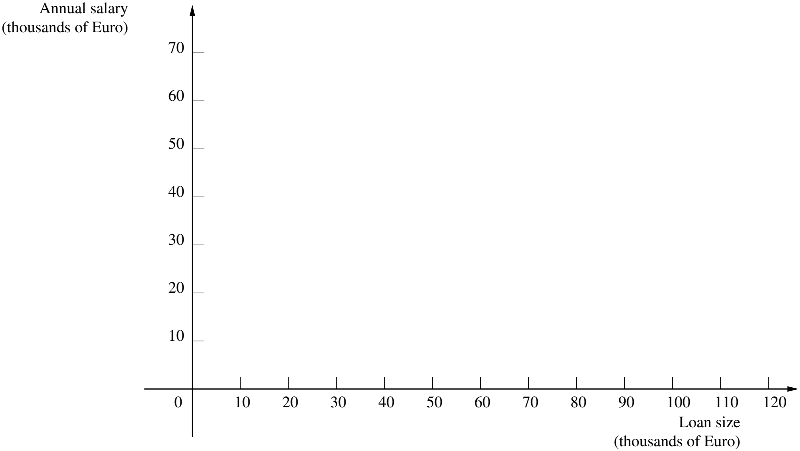
The aim is to predict the class of any new object (after training the classifier on the training set).

EXAMPLE

- A bank needs a criterion to decide whether to loan money or not.
- Starting from the past experience, the analyst tries to analyze the data relative to the past clients on the basis of their salary and of the size of the loan (two features, i.e. $n = 2$).

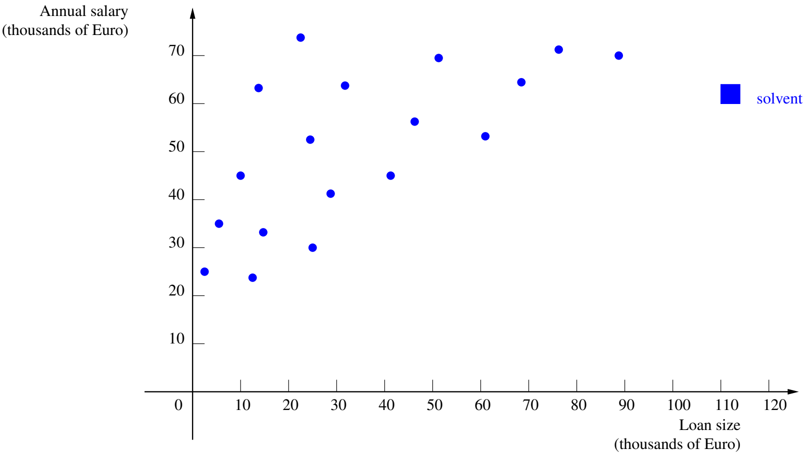
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: an example



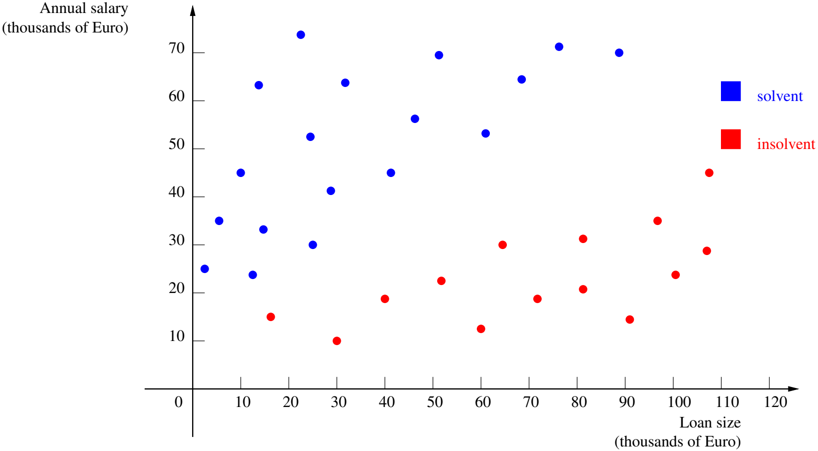
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: an example



18 past solvent clients

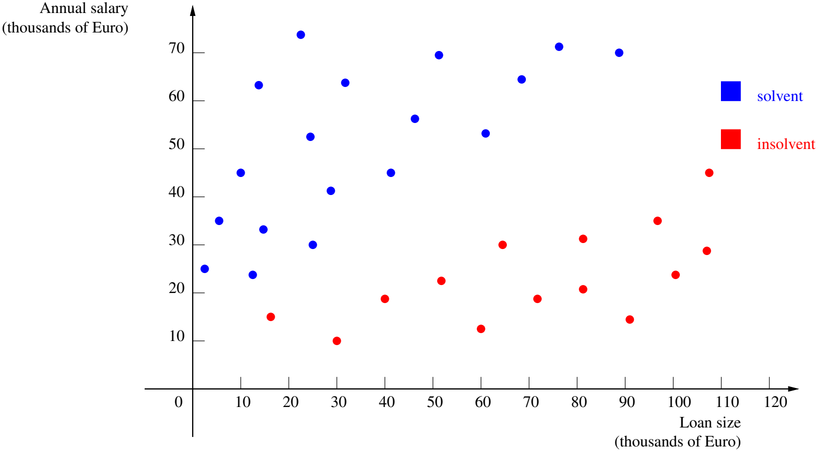
Pattern classification: an example



18 past solvent clients and
14 past insolvent ones.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

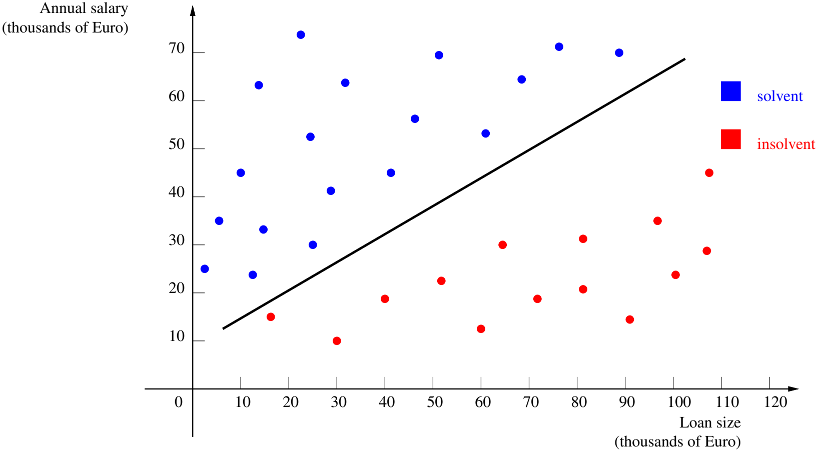
Pattern classification: an example



TRAINING SET:
18 past solvent clients and
14 past insolvent ones.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

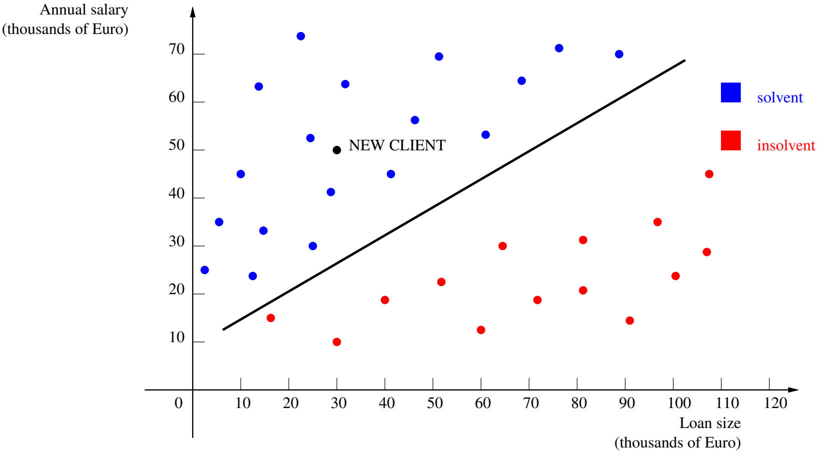
Pattern classification: an example



TRAINING SET:
18 past solvent clients and
14 past insolvent ones.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

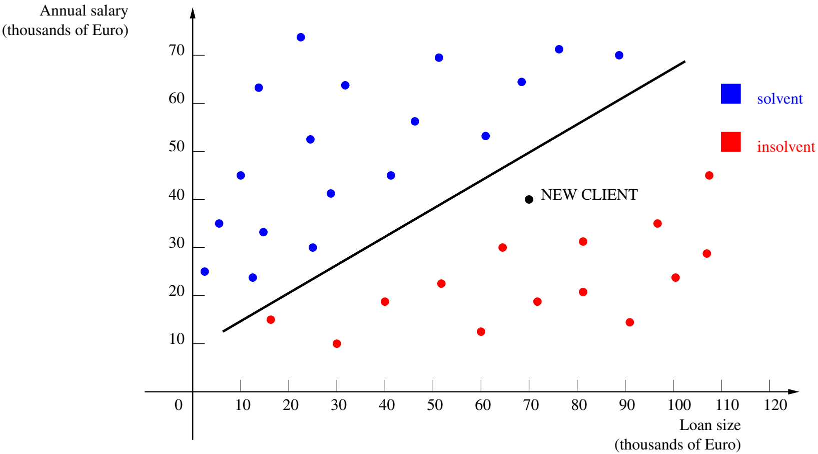
Pattern classification: an example



NEW CLIENT:
30.000 euros (loan size)
50.000 euros (annual salary).

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: an example



NEW CLIENT:
70.000 euros (loan size)
40.000 euros (annual salary).

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: an example

In the example, given the separating hyperplane,

$$H(v,\gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | v^T x = \gamma\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

for classifying the new client $\bar{x}$, we have used the following decision function:

$$\text{sign}(v^T \bar{x} - \gamma),$$

i.e.

$$\text{if } v^T\bar{x} - \gamma \begin{cases} \geq 0, & \text{the client is classified as solvent} \\ < 0, & \text{the client is classified as insolvent} \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Pattern classification: some applications

- Text and web classification.
- Object recognition of machine vision.
- Gene expression profile analysis.
- DNA and protein analysis.
- Medical diagnosis.

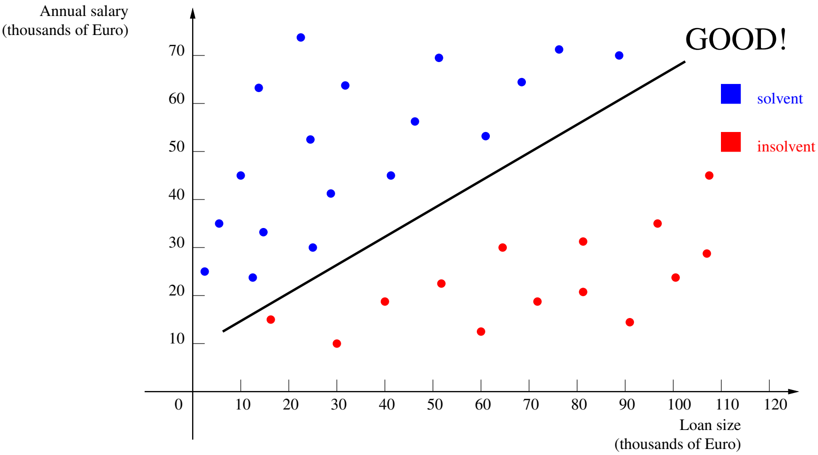
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Optimization in machine learning

Question:

Where does optimization intervene?

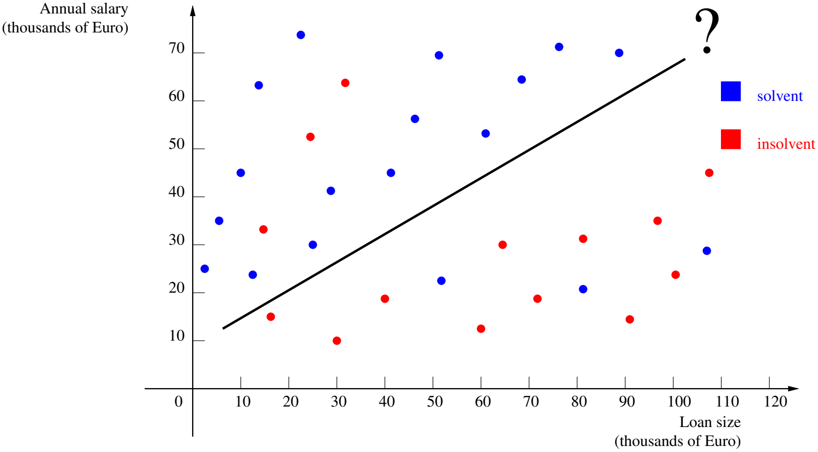
Machine Learning and Optimization



Separable case by a hyperplane.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Machine Learning and Optimization



This set of points is not separable by a hyperplane!!

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Answer:

To minimize a measure of the number of misclassified points.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Wordplay...

Classification of classification approaches...

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Classification of classification approaches...

At each client, we have "attached" a label:

$$\text{client} \rightarrow \begin{cases} \text{solvent} \\ \text{insolvent} \end{cases}$$

$$\Downarrow$$

SUPERVISED CLASSIFICATION

$$\Downarrow$$

On the basis of the labelled objects, we would like to predict the class of any new future object.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Supervised, unsupervised and semisupervised classification

- **Supervised classification**: on the basis of the labelled objects, we would like to predict the class of any new future object.
- **Unsupervised classification**: we have only unlabelled objects that we would like to cluster on the basis of their similarities.
- **Semisupervised classification**: on the basis of the labelled and unlabelled objects, we would like to predict the class of the unlabelled objects.

PART V

BINARY SUPERVISED CLASSIFICATION

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Binary supervised classification

In the binary classification, we would like to discriminate only between two classes of objects (points in $\mathbb{R}^n$).
We have two nonempty, disjoint, finite point sets in $\mathbb{R}^n$:

- $$\mathcal{A} = \{a_1, \ldots, a_m\}, \quad \text{with } a_i \in \mathbb{R}^n,\ i = 1, \ldots, m$$
- $$\mathcal{B} = \{b_1, \ldots, b_k\}, \quad \text{with } b_l \in \mathbb{R}^n,\ l = 1, \ldots, k.$$
- The objective is to construct a criterion for discriminating between the elements of the two sets. Then the classifier can be utilized for classifying any new object point $\bar{x} \in \mathbb{R}^n$ as a point belonging to the set $\mathcal{A}$ or, alternatively, to the set $\mathcal{B}$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Linear separation

Linear separation (Mangasarian, 1965 [Man65])

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

$$H(v,\gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | v^T x = \gamma\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

such that

  - $$v^T a_i \geq \gamma + 1, \quad i = 1, \ldots, m$$

    and

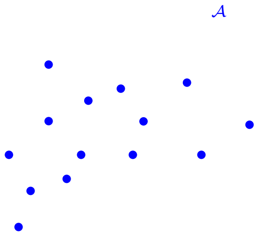
  - $$v^T b_l \leq \gamma - 1, \quad l = 1, \ldots, k.$$

- **NOTE**: $\mathcal{A}$ and $\mathcal{B}$ are linearly separable if and only if

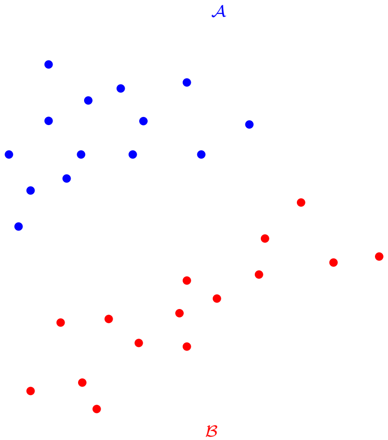
$$\text{conv}(\mathcal{A}) \cap \text{conv}(\mathcal{B}) = \emptyset.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

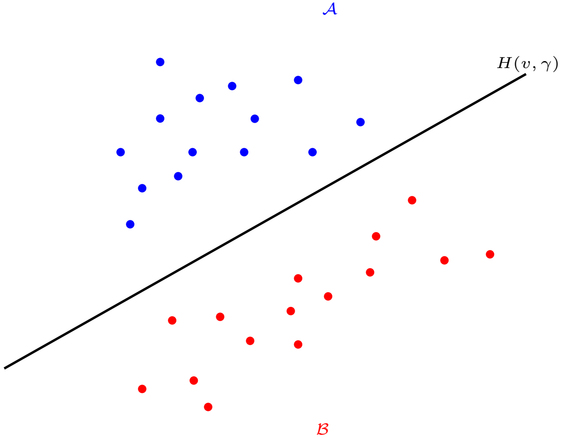
Linear separation: first example



Linear separation: first example

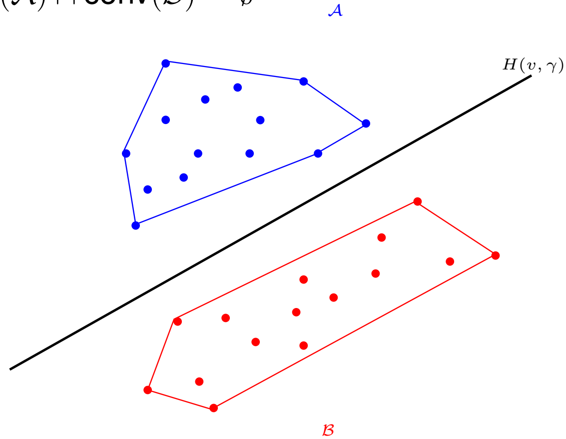


Linear separation: first example



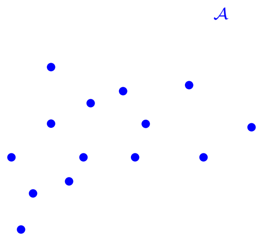
Linear separation: first example

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

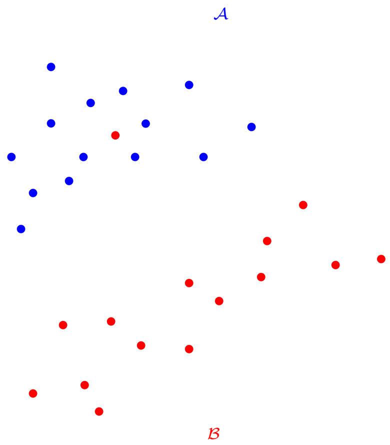


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Linear separation: second example

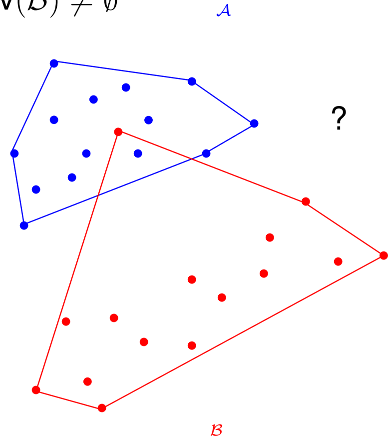


Linear separation: second example



Linear separation: second example

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



Linear separation: error function

What can we do when $\mathcal{A}$ and $\mathcal{B}$ are not linearly separable?

- A point $a_i \in \mathcal{A}$ is correctly classified if

$$v^T a_i \geq \gamma + 1, \text{ i.e. if } v^T a_i - \gamma - 1 \geq 0$$

- As a consequence, a point $a_i \in \mathcal{A}$ is misclassified if

$$v^T a_i - \gamma - 1 < 0, \text{i.e. if } -v^T a_i + \gamma + 1 > 0.$$

- Then, for a point $a_i \in \mathcal{A}$, the classification error is

$$\max\{0, -v^T a_i + \gamma + 1\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Linear separation: error function

- A point $b_l \in \mathcal{B}$ is correctly classified if

$$v^T b_l \leq \gamma - 1, \text{ i.e. if } v^T b_l - \gamma + 1 \leq 0.$$

- As a consequence, a point $b_l \in \mathcal{B}$ is misclassified if

$$v^T b_l - \gamma + 1 > 0.$$

- Then, for a point $b_l \in \mathcal{B}$, the classification error is

$$\max\{0, v^T b_l - \gamma + 1\}.$$

- Then we minimize the following classification error function:

$$f(v,\gamma) \stackrel{\triangle}{=} \frac{1}{m}\sum_{i=1}^{m} \max\{0, -v^T a_i + \gamma + 1\} + \frac{1}{k}\sum_{l=1}^{k} \max\{0, v^T b_l - \gamma + 1\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Linear separation

$$f(v,\gamma) \triangleq \frac{1}{m}\sum_{i=1}^{m} \overbrace{\max\{0, -v^T a_i + \gamma + 1\}}^{\xi_i} + \frac{1}{k}\sum_{l=1}^{k} \overbrace{\max\{0, v^T b_l - \gamma + 1\}}^{\psi_l}.$$

- Function f is a convex nonsmooth function;
- Minimizing f corresponds to solving the following linear program:

$$\begin{cases} \min\limits_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{m}\sum_{i=1}^{m}\xi_i + \frac{1}{k}\sum_{l=1}^{k}\psi_l & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \end{cases}$$

Polyhedral separation

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Polyhedral separation - (Megiddo, 1988 [Meg88])

- The set $\mathcal{A}$ is h-polyhedrally separable from $\mathcal{B}$ if there exists a set of h hyperplanes

  $$H(v_j, \gamma_j) \triangleq \{x \in \mathbb{R}^n | v_j^T x = \gamma_j\}, \text{ with } v_j \in \mathbb{R}^n \text{ and } \gamma_j \in \mathbb{R}, j = 1, \dots, h,$$

  such that

  - $$v_j^T a_i \le \gamma_j - 1, \quad i = 1, \dots, m, \quad j = 1, \dots, h$$

    and
  - for any $l = 1, \dots, k$, there exists an index $j \in \{1, \dots, h\}$ such that

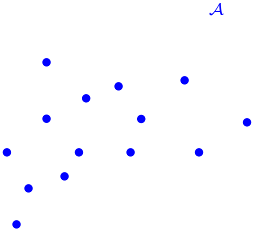
    $$v_j^T b_l \ge \gamma_j + 1.$$

- **NOTE**: $\mathcal{A}$ is h-polyhedrally separable from $\mathcal{B}$ if and only if

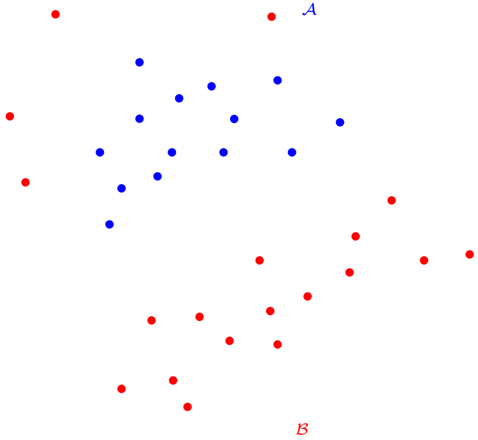
  $$\text{conv}(\mathcal{A}) \cap \mathcal{B} = \emptyset.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

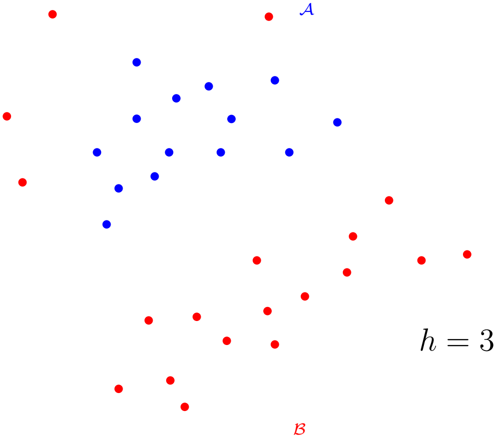
Polyhedral separation



Polyhedral separation: first example

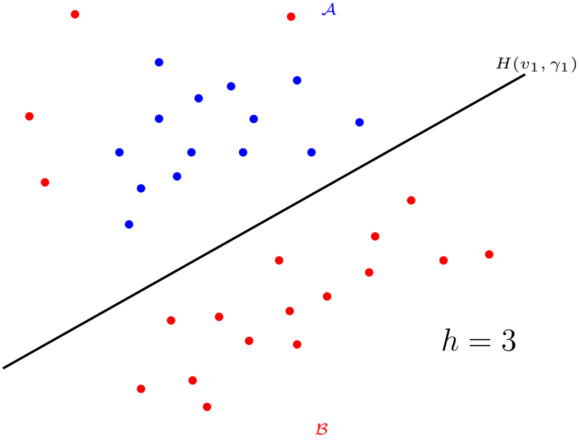


Polyhedral separation: first example

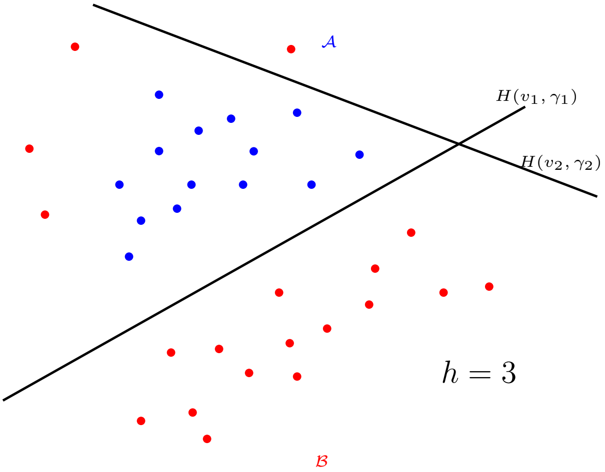


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Polyhedral separation: first example

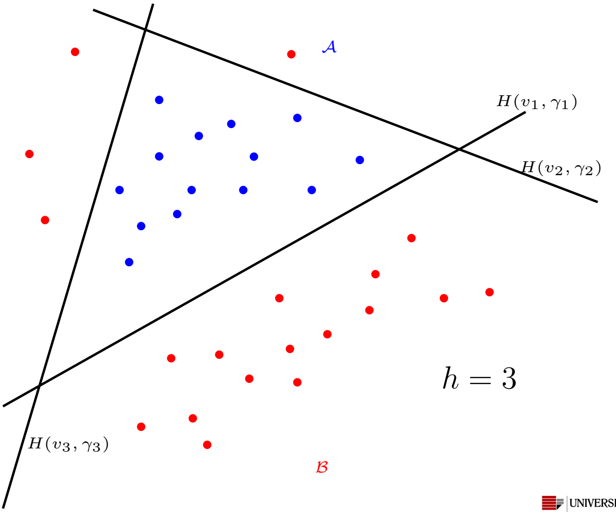


Polyhedral separation: first example

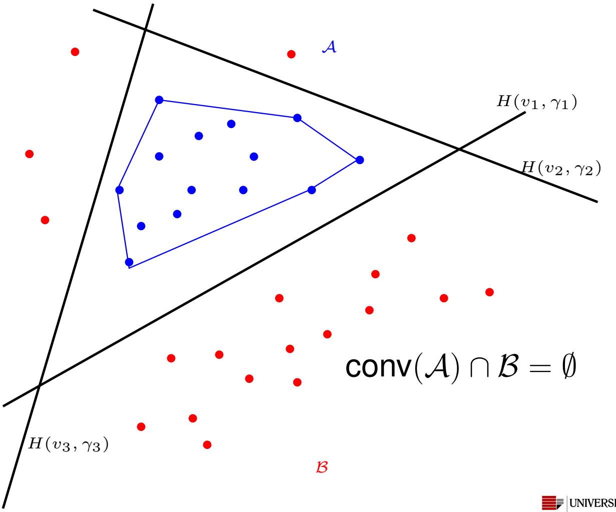


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Polyhedral separation: first example



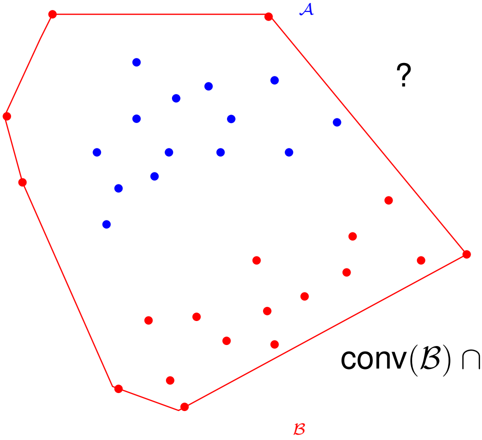
Polyhedral separation: first example



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

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

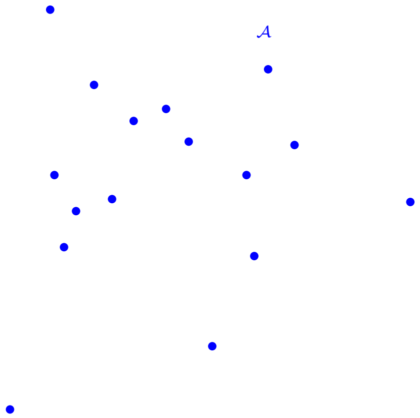
Polyhedral separation: first example



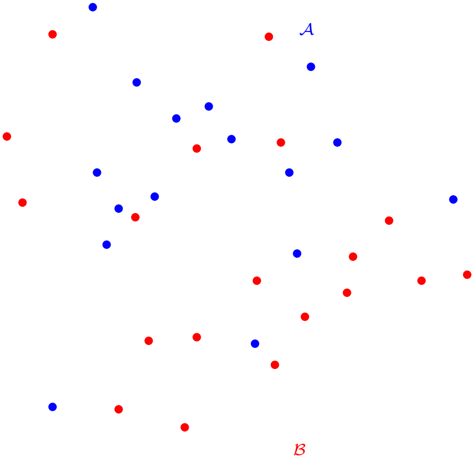
$\text{conv}(\mathcal{B}) \cap \mathcal{A} \neq \emptyset$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

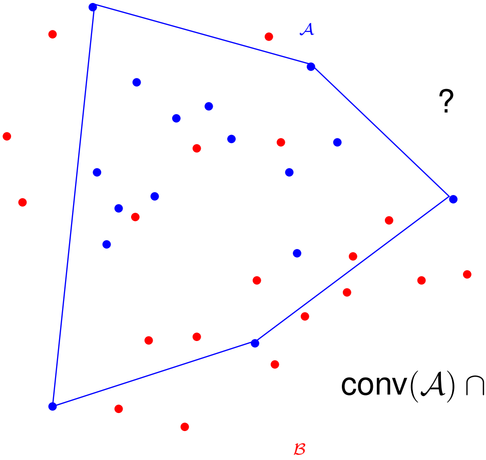
Polyhedral separation: second example



Polyhedral separation: second example

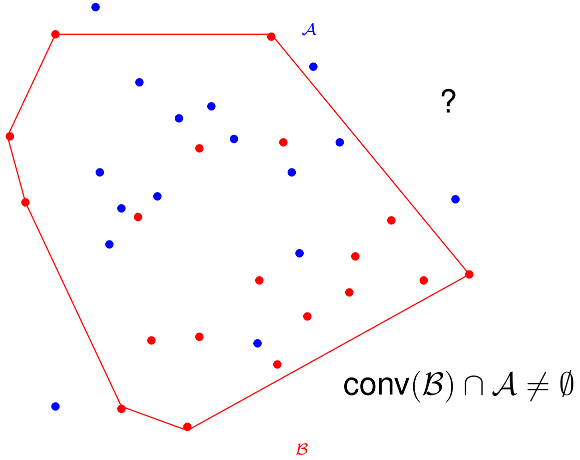


Polyhedral separation: second example



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

Polyhedral separation: second example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Polyhedral separation: error function

What can we do if $\mathcal{A}$ is not polyhedrally separable from $\mathcal{B}$?

- A point $a_i \in \mathcal{A}$ is correctly classified if

$$v_j^T a_i - \gamma_j + 1 \leq 0, \quad \text{for all } j = 1, \ldots, h,$$

  i.e. if

$$\max_{j=1,\ldots,h} v_j^T a_i - \gamma_j + 1 \leq 0.$$

- As a consequence, a point $a_i \in \mathcal{A}$ is misclassified if

$$\max_{j=1,\ldots,h} v_j^T a_i - \gamma_j + 1 > 0.$$

- Then the classification error, in correspondence to a point $a_i \in \mathcal{A}$, is
$$\max\{0, \max_{j=1,\ldots,h} v_j^T a_i - \gamma_j + 1\} = \max_{j=1,\ldots,h}\{0, v_j^T a_i - \gamma_j + 1\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Polyhedral separation: error function

- A point $b_l \in \mathcal{B}$ is correctly classified if there exists an index $j \in \{1, \dots, h\}$ such that

$$v_j^T b_l \geq \gamma_j + 1, \text{ i.e. } -v_j^T b_l + \gamma_j + 1 \leq 0.$$

- As a consequence, a point $b_l \in \mathcal{B}$ is misclassified if

$$\text{for all } j = 1, \dots, h, \quad -v_j^T b_l + \gamma_j + 1 > 0,$$

- i.e. if

$$\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1 > 0.$$

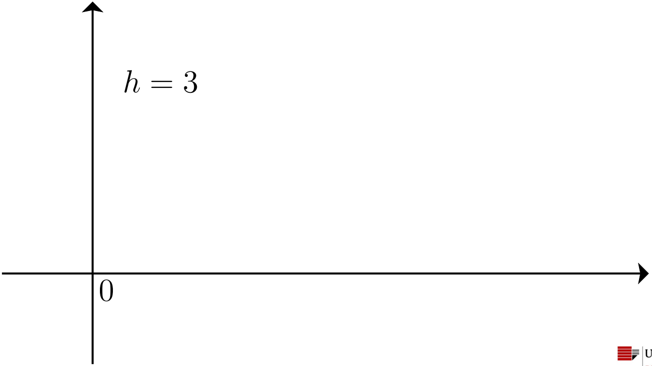
- Then the classification error, in correspondence to a point $b_l \in \mathcal{B}$, is

$$\max\{0, \min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

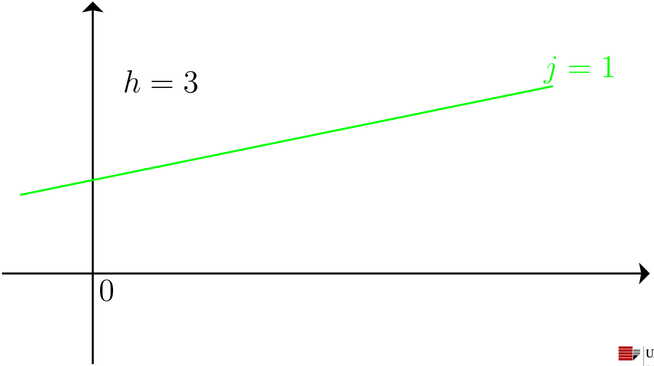
Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}}_{f_l}.$$



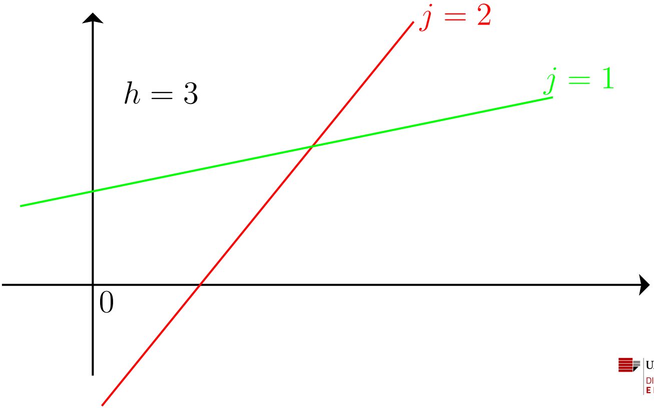
Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}}_{f_l}.$$



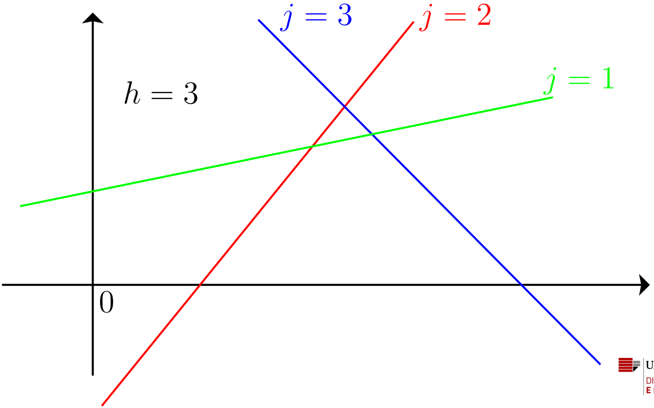
Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}}_{f_l}.$$



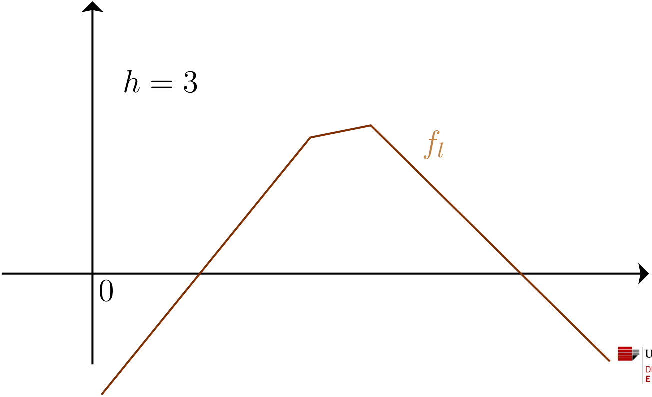
Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}}_{f_l}.$$



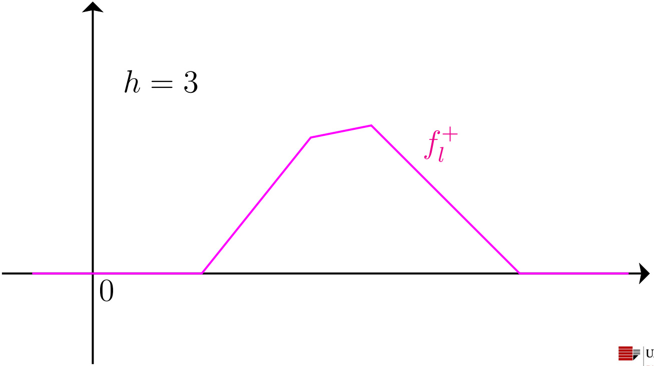
Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1\}}_{f_l}.$$



Polyhedral separation: error function

$$f_l^+ = \max\{0, \underbrace{\min_{j=1,\dots,h} -v_j^T b_l + \gamma_j + 1}_{f_l}\}.$$



Polyhedral separation: error function

We obtain the following classification error function:

$$\begin{aligned} f(v_1,\dots,v_h;\gamma_1,\dots,\gamma_h) \quad \triangleq \quad & \frac{1}{m}\sum_{i=1}^{m}\max_{1\le j\le h}\{0, v_j^T a_i - \gamma_j + 1\} + \\ & + \frac{1}{k}\sum_{l=1}^{k}\max\{0, \min_{1\le j\le h} -v_j^T b_l + \gamma_j + 1\}. \end{aligned}$$

- Function f is nonsmooth and nonconvex.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation

Spherical separation - (Tax and Duin, 1999 [TD99])

- The set $\mathcal{A}$ is spherically separable from the set $\mathcal{B}$ if there exists a sphere

$$S(x_0, R) \triangleq \{x \in \mathbb{R}^n \mid \|x - x_0\|^2 = R^2\},$$

with $x_0 \in \mathbb{R}^n$ and $R \in \mathbb{R}$,

such that

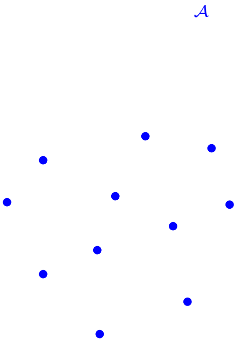
  - $$\|a_i - x_0\|^2 \leq R^2, \quad i = 1, \dots, m$$

    and
  - $$\|b_l - x_0\|^2 \geq R^2, \quad l = 1, \dots . k$$

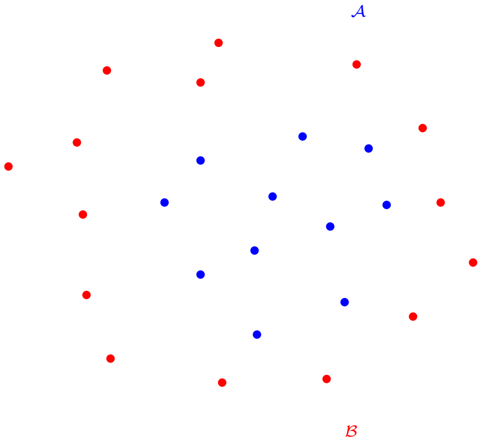
- **NOTE 1**: The role played by $\mathcal{A}$ and $\mathcal{B}$ is not symmetric.
- **NOTE 2**: $\mathcal{A}$ is spherically separable from $\mathcal{B} \Rightarrow$ conv$(\mathcal{A}) \cap \mathcal{B} = \emptyset$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation: first example

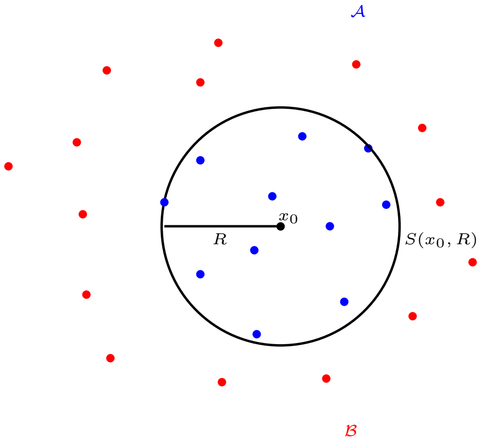


Spherical separation: first example



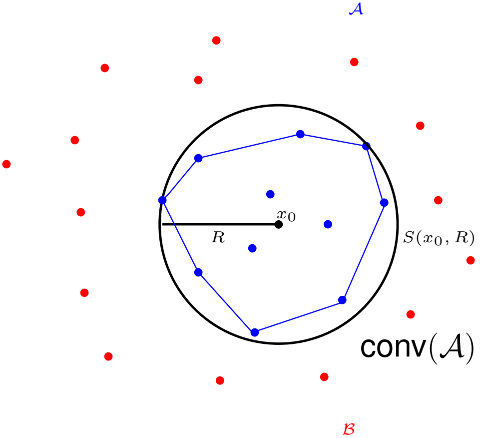
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation: first example



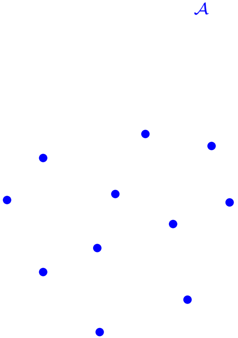
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation: first example

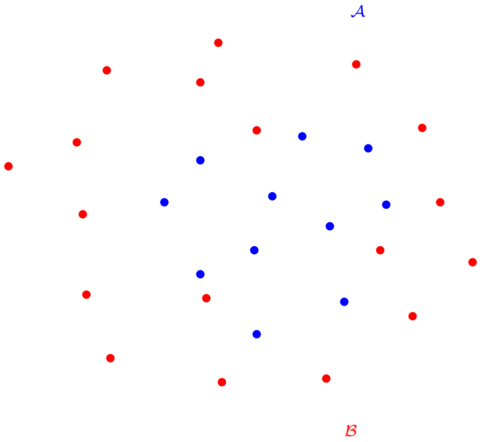


$$\text{conv}(\mathcal{A}) \cap \mathcal{B} = \emptyset$$

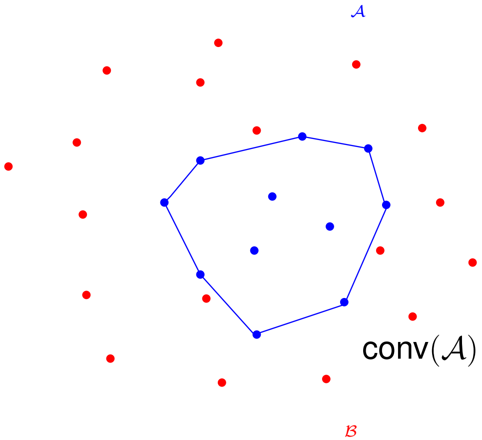
Spherical separation: second example



Spherical separation: second example



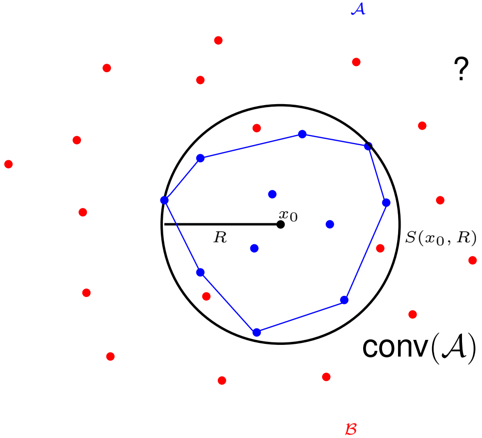
Spherical separation: second example



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

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation: second example



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

Spherical separation: error function

What can we do if $\mathcal{A}$ is not spherically separable from $\mathcal{B}$?

- A point $a_i \in \mathcal{A}$ is correctly classified if

$$\|a_i - x_0\|^2 - R^2 \leq 0.$$

- As a consequence, a point $a_i \in \mathcal{A}$ is misclassified if

$$\|a_i - x_0\|^2 - R^2 > 0.$$

- Then the classification error, in correspondence to a point $a_i \in \mathcal{A}$, is

$$\max\{0, \|a_i - x_0\|^2 - R^2\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation: error function

- A point $b_l \in \mathcal{B}$ is correctly classified if

$$R^2 - \|b_l - x_0\|^2 \leq 0.$$

- As a consequence, a point $b_l \in \mathcal{B}$ is misclassified if

$$R^2 - \|b_l - x_0\|^2 > 0.$$

- Then the classification error, in correspondence to a point $b_l \in \mathcal{B}$, is

$$\max\{0, R^2 - \|b_l - x_0\|^2\}.$$

Spherical separation: error function

- We obtain the following classification error function:

$$
\begin{aligned}
f(x_0, R) \stackrel{\triangle}{=} R^2 \quad & + C \sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - R^2\} + \\
& + C \sum_{l=1}^{k} \max\{0, R^2 - \|b_l - x_0\|^2\},
\end{aligned}
$$

  with $C > 0$, tuning the trade-off between the minimization of the volume of the sphere and the minimization of the misclassification error.
- Function f is nonsmooth and nonconvex.

Spherical separation: fixing the center (Astorino and Gaudioso, 2009 [AG09])

$$\begin{aligned} f(x_0,R) \overset{\triangle}{=} R^2 \quad & +C\sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - R^2\} + \\ & +C\sum_{l=1}^{k} \max\{0, R^2 - \|b_l - x_0\|^2\}, \end{aligned}$$

NOTE: If x_0 is fixed, setting $z \overset{\triangle}{=} R^2 \geq 0$, then function f is convex in z.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

Spherical separation: fixing the center

$$f(z) \stackrel{\triangle}{=} z + C\sum_{i=1}^{m} \overbrace{\max\{0, \|a_i - x_0\|^2 - z\}}^{\xi_i} + C\sum_{l=1}^{k} \overbrace{\max\{0, z - \|b_l - x_0\|^2\}}^{\psi_l}.$$

In this case, minimization of f corresponds to solve the following linear program:

$$\begin{cases} \displaystyle\min_{z,\xi,\psi} & z + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l & \\ & \xi_i \geq \|a_i - x_0\|^2 - z & i = 1,\ldots,m \\ & \psi_l \geq z - \|b_l - x_0\|^2 & l = 1,\ldots,k \\ & \xi_i \geq 0 & i = 1,\ldots,m \\ & \psi_l \geq 0 & l = 1,\ldots,k. \\ & z \geq 0 & \end{cases}$$

Support Vector Machine

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Support Vector Machine (SVM) (Vapnik, 1995 [Vap95])

- Motivation: To maximize the generalization capability of the classifier, i.e. to maximize the probability that a new point is correctly classified.
- This minimizes also possible overfitting phenomena.

OVERFITTING

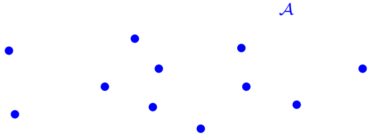
$\Downarrow$

There is overfitting, when the classifier fits too much the training set.

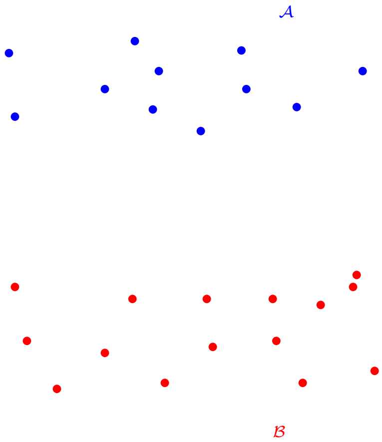
$\Downarrow$

Bad performance on the classification of new points.

SVM: an example

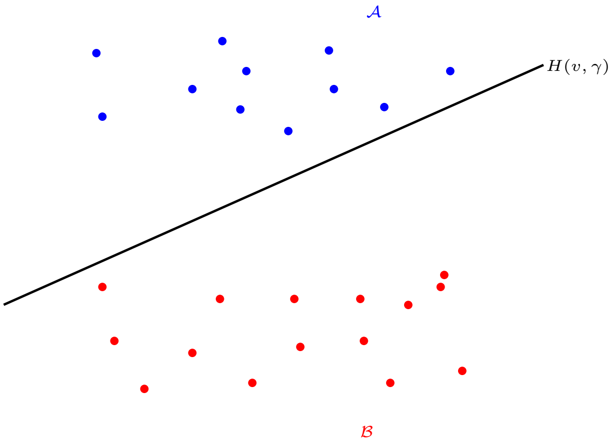


SVM: an example

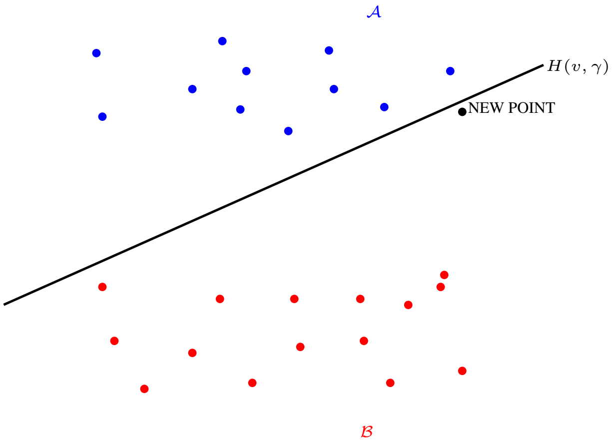


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

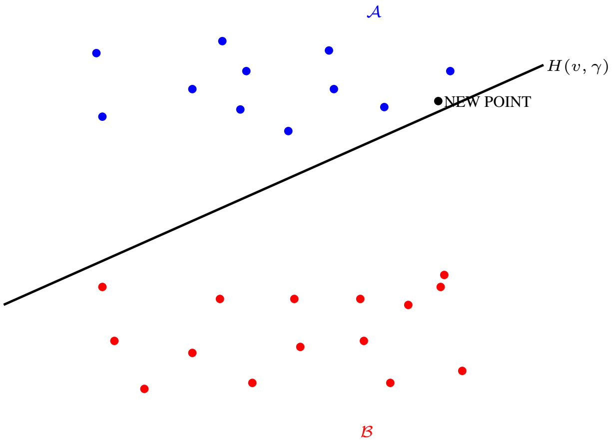
SVM: an example



SVM: an example

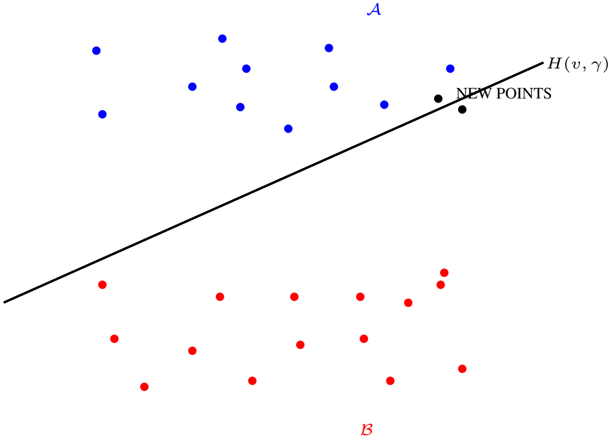


SVM: an example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM: an example

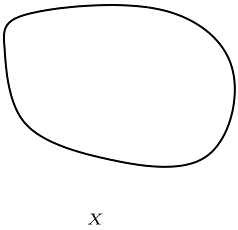


Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.

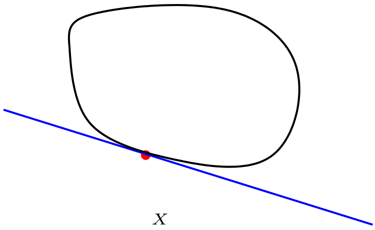


Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.



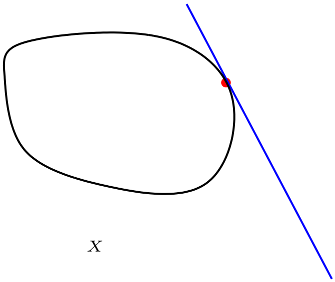
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.

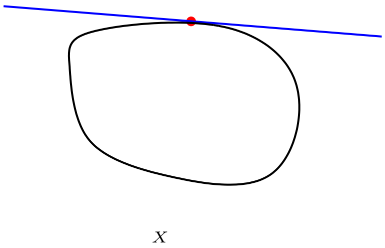


Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.

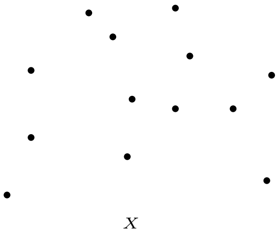


Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.

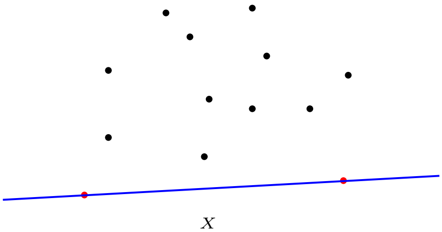


Support Vector Machine (SVM)

Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.



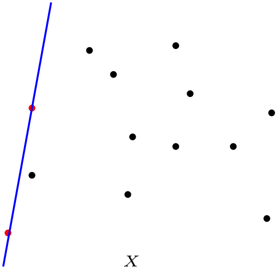
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Support Vector Machine (SVM)

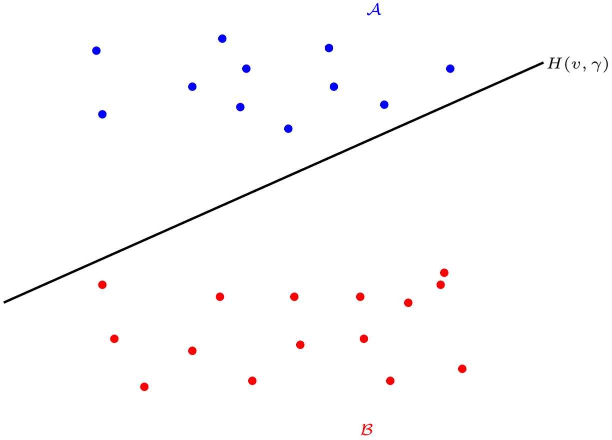
Definition (Supporting hyperplane)

Let $X \subset R^n$. A supporting hyperplane of X is a hyperplane such that:

- X is entirely contained in one of the two half-spaces generated by the hyperplane;
- at least one boundary point of X is on the hyperplane.

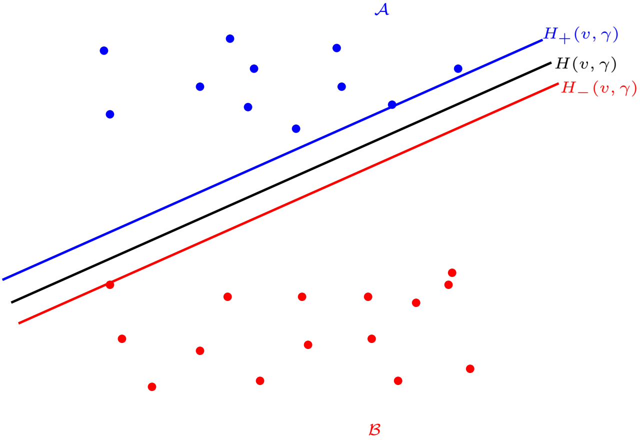


SVM: an example



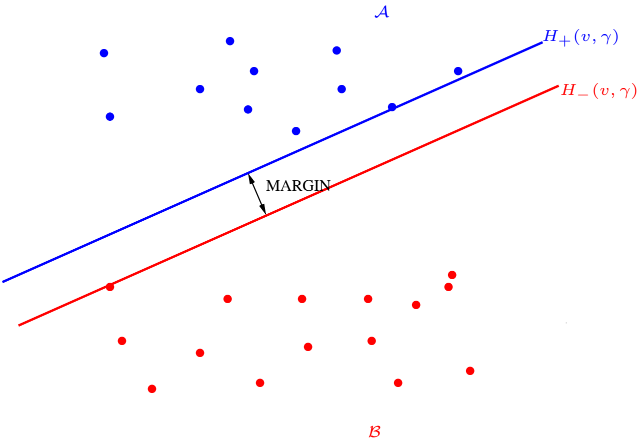
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM: an example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM: an example



SVM

- The margin is the area between the two parallel hyperplanes

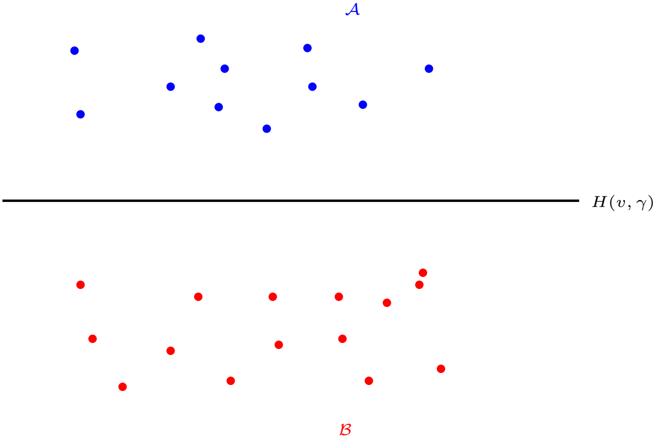
$$H_+(v,\gamma) \triangleq \{x \in \mathbb{R}^n | v^T x = \gamma + 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

and

$$H_-(v,\gamma) \triangleq \{x \in \mathbb{R}^n | v^T x = \gamma - 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

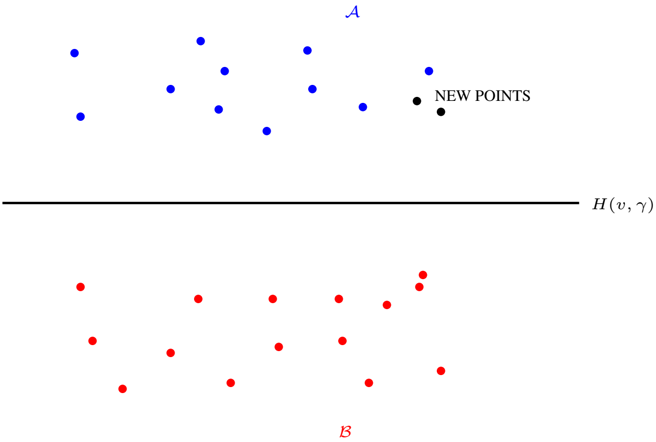
which are the supporting hyperplanes of $\mathcal{A}$ and $\mathcal{B}$, respectively.

SVM: an example

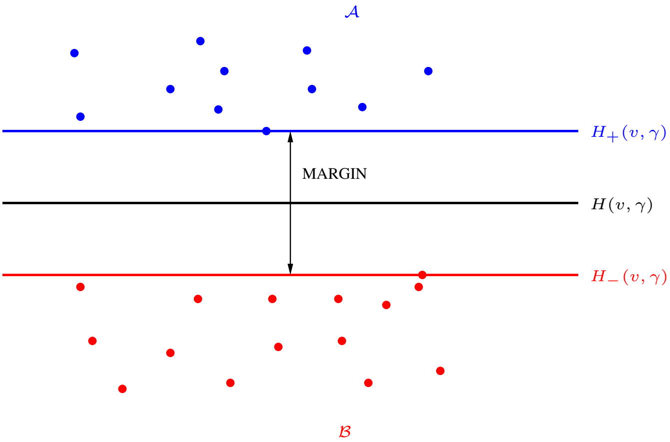


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM: an example



SVM: an example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

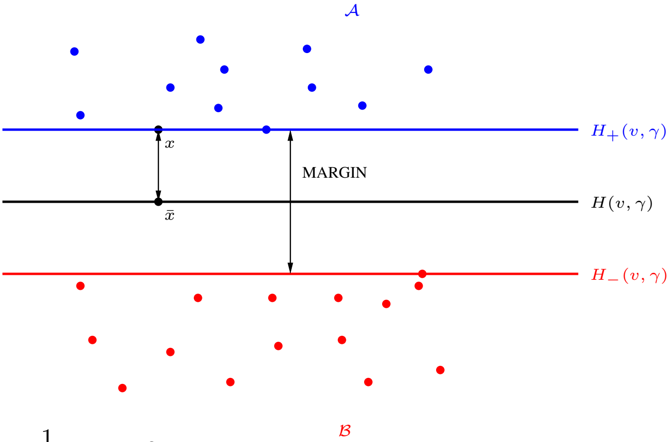
SVM: the margin

How to compute the margin? We solve the following problem:

$$P \begin{cases} \min\limits_{x} & \dfrac{1}{2}\|x-\bar{x}\|^2 \\ & v^T x = \gamma + 1 \end{cases},$$

with $\bar{x}$, such that $v^T \bar{x} = \gamma$.

SVM: an example



$$P \begin{cases} \min\limits_{x} & \frac{1}{2}\|x-\bar{x}\|^2 \\ & v^T x = \gamma + 1 \end{cases}$$

with $\bar{x}$ such that $v^T\bar{x} = \gamma$.

SVM: the margin

KKT conditions

$$\mathcal{L}(x,\lambda) = \frac{1}{2}\|x-\bar{x}\|^2 - \lambda(v^T x - \gamma - 1), \text{ with } \lambda \in \mathbb{R}.$$

$$\Downarrow$$

$$\nabla_x \mathcal{L}(x,\lambda) = \frac{1}{2}2(x-\bar{x}) - \lambda v = 0, \text{ i.e.}$$

$$x = \bar{x} + \lambda v.$$

$$\Downarrow$$

$$\underbrace{v^T x}_{\gamma+1} = \underbrace{v^T \bar{x}}_{\gamma} + \lambda\|v\|^2 \Rightarrow \lambda = \frac{1}{\|v\|^2}$$

SVM: the margin

We have obtained:

$$\begin{cases} x - \bar{x} = \lambda v \\ \\ \lambda = \dfrac{1}{\|v\|^2} \end{cases}$$

$$\Downarrow$$

$$\|x - \bar{x}\| = |\lambda|\|v\| = \lambda\|v\| = \frac{1}{\|v\|^2}\|v\| = \frac{1}{\|v\|.}$$

$$\Downarrow$$

$$\text{MARGIN} = \frac{2}{\|v\|}$$

$$\Downarrow$$

$$\max\ \text{MARGIN} \Leftrightarrow \min_v \|v\| \Leftrightarrow \min_v \frac{1}{2}\|v\|^2.$$

Support Vector Machine (SVM) (Vapnik, 1995 [Vap95])

To summarize:

- The sets

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

  and

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

  are given.

- We compute a separating hyperplane

$$H(v, \gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | v^T x = \gamma\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

  called the support vector machine, which is furthest from the closest points in the two sets.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Support Vector Machine (SVM)

The separation hyperplane (the support vector machine) is constructed by minimizing the following nonsmooth error function:

$$
\begin{aligned}
f(v,\gamma) \stackrel{\triangle}{=} \quad & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\max\{0, -v^T a_i + \gamma + 1\} + \\
& + C\sum_{l=1}^{k}\max\{0, v^T b_l - \gamma + 1\}.
\end{aligned}
$$

- The first term maximizes the margin.
- By minimizing the last two terms we minimize the misclassification measure of the points of the two sets $\mathcal{A}$ and $\mathcal{B}$, respectively.
- Parameter $C > 0$ tunes the weight of the two objectives.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Smoothing...

$$f(v,\gamma) \triangleq \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m} \overbrace{\max\{0, -v^T a_i + \gamma + 1\}}^{\xi_i} + C\sum_{l=1}^{k} \overbrace{\max\{0, v^T b_l - \gamma + 1\}}^{\psi_l}$$

Minimization of f corresponds to solve the following quadratic program:

$$\begin{cases} \displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{i=1}^{k}\psi_i & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\ldots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\ldots,k \\ & \xi_i \geq 0 & i = 1,\ldots,m \\ & \psi_l \geq 0 & l = 1,\ldots,k. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

PRIMAL

$$
P \left\{
\begin{array}{lll}
\displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l & \\
 & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\ldots,m \\
 & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\ldots,k \\
 & \xi_i \geq 0 & i = 1,\ldots,m \\
 & \psi_l \geq 0 & l = 1,\ldots,k
\end{array}
\right.
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The Wolfe dual

The Wolfe dual is defined as follows:

$$D \left\{ \begin{array}{ll} \max\limits_{x,\lambda} & \mathcal{L}(x,\lambda) \\ & \nabla_x \mathcal{L}(x,\lambda) = 0 \\ & \lambda_i \geq 0, \quad i \in I \end{array} \right.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

PRIMAL

$$
P \begin{cases}
\displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l & & \\
& \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m & \lambda_i \\
& \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k & \mu_l \\
& \xi_i \geq 0 & i = 1,\dots,m & \alpha_i \\
& \psi_l \geq 0 & l = 1,\dots,k & \beta_l
\end{cases}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

Objective function (max)

$$
\begin{aligned}
\mathcal{L}(v,\xi,\psi,\lambda,\mu,\alpha,\beta) = \quad & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l \\
& -\sum_{i=1}^{m}\lambda_i(\xi_i + v^T a_i - \gamma - 1) \\
& -\sum_{l=1}^{k}\mu_l(\psi_l - v^T b_l + \gamma - 1) \\
& -\sum_{i=1}^{m}\alpha_i\xi_i - \sum_{l=1}^{k}\beta_l\psi_l
\end{aligned}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

Objective function (max)

$$
\begin{aligned}
\mathcal{L}(v,\xi,\psi,\lambda,\mu,\alpha,\beta) \quad &= \frac{1}{2}\|v\|^2 - v^T\left(\sum_{i=1}^{m}\lambda_i a_i - \sum_{l=1}^{k}\mu_l b_l\right) \\
&+\gamma\left(\sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l\right) \\
&+\sum_{i=1}^{m}\xi_i(C-\lambda_i-\alpha_i) + \sum_{l=1}^{k}\psi_l(C-\mu_l-\beta_l) \\
&+\sum_{i=1}^{m}\lambda_i + \sum_{l=1}^{k}\mu_l
\end{aligned}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

Constraints

$$\nabla_v \mathcal{L} = v + \sum_{l=1}^{k} \mu_l b_l - \sum_{i=1}^{m} \lambda_i a_i = 0 \Leftrightarrow v = \sum_{i=1}^{m} \lambda_i a_i - \sum_{l=1}^{k} \mu_l b_l$$

$$\nabla_\gamma \mathcal{L} = \sum_{i=1}^{m} \lambda_i - \sum_{l=1}^{k} \mu_l = 0$$

$$\nabla_{\xi_i} \mathcal{L} = C - \lambda_i - \alpha_i = 0 \Leftrightarrow \lambda_i = C - \alpha_i \Leftrightarrow \lambda_i \leq C \quad i = 1, \ldots, m$$

$$\nabla_{\psi_l} \mathcal{L} = C - \mu_l - \beta_l = 0 \Leftrightarrow \mu_l = C - \beta_l \Leftrightarrow \mu_l \leq C \quad l = 1, \ldots, k$$

$$\lambda, \mu, \alpha, \beta \geq 0$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

Objective function (max)

$$
\begin{aligned}
\mathcal{L}(v,\xi,\psi,\lambda,\mu,\alpha,\beta) = \quad & \frac{1}{2}\|v\|^2 - v^T \overbrace{\left(\sum_{i=1}^{m}\lambda_i a_i - \sum_{l=1}^{k}\mu_l b_l\right)}^{v} \\
& +\gamma \overbrace{\left(\sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l\right)}^{0} \\
& +\sum_{i=1}^{m}\xi_i \overbrace{(C-\lambda_i-\alpha_i)}^{0} + \sum_{l=1}^{k}\psi_l \overbrace{(C-\mu_l-\beta_l)}^{0} \\
& +\sum_{i=1}^{m}\lambda_i + \sum_{l=1}^{k}\mu_l
\end{aligned}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

Objective function (max)

$$\mathcal{L}(v,\lambda,\mu) = \frac{1}{2}\|v\|^2 - \overbrace{v^T v}^{\|v\|^2} + \sum_{i=1}^{m}\lambda_i + \sum_{l=1}^{k}\mu_l = -\frac{1}{2}\|v\|^2 + \sum_{i=1}^{m}\lambda_i + \sum_{l=1}^{k}\mu_l$$

Constraints

$$v = \sum_{i=1}^{m}\lambda_i a_i - \sum_{l=1}^{k}\mu_l b_l$$

$$\sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l = 0$$

$$0 \le \lambda_i \le C \quad i = 1,\dots,m$$

$$0 \le \mu_l \le C \quad l = 1,\dots .k$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The SVM Wolfe dual

$$
D \begin{cases}
\max\limits_{\lambda,\mu} & -\frac{1}{2}\|\sum\limits_{i=1}^{m}\lambda_i a_i - \sum\limits_{l=1}^{k}\mu_l b_l\|^2 + \sum\limits_{i=1}^{m}\lambda_i + \sum\limits_{l=1}^{k}\mu_l \\
& \sum\limits_{i=1}^{m}\lambda_i - \sum\limits_{l=1}^{k}\mu_l = 0 \\
& 0 \leq \lambda_i \leq C \quad i = 1,\ldots,m \\
& 0 \leq \mu_l \leq C \quad l = 1,\ldots.k
\end{cases}
$$

The SVM Wolfe dual

$$\max \overbrace{\Rightarrow}^{-} \min$$

$$D \begin{cases} \min\limits_{\lambda,\mu} & \frac{1}{2}\|\sum_{i=1}^{m}\lambda_i a_i - \sum_{l=1}^{k}\mu_l b_l\|^2 - \sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l \\ & \sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l = 0 \\ & 0 \le \lambda_i \le C \quad i = 1, \dots, m \\ & 0 \le \mu_l \le C \quad l = 1, \dots . k \end{cases}$$

NOTE: Quadratic program with one constraint and $m + k$ box constraints.

The SVM Wolfe dual

$$(\lambda^*, \mu^*)$$

$$\Downarrow$$

$$v^* = \sum_{i=1}^{m} \lambda_i^* a_i - \sum_{l=1}^{k} \mu_l^* b_l$$

$$\gamma^* = {v^*}^T a_i - 1, \text{ with } i \text{ such that } 0 < \lambda_i^* < C$$

or

$$\gamma^* = {v^*}^T b_l + 1, \text{ with } l \text{ such that } 0 < \mu_l^* < C$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The kernel trick

Motivation: To separate $\mathcal{A}$ and $\mathcal{B}$ by means of a nonlinear surface, using the SVM approach.

- We indicate by $X_I \subseteq \mathbb{R}^n$ the so-called input space, such that $\mathcal{A}, \mathcal{B} \subset X_I$.
- We define the so-called feature space $X_F \subseteq \mathbb{R}^N$, with generally $N > n$.
- Given a map

  $$\phi : X_I \mapsto X_F,$$

  the kernel function is defined as:

  $$K : X_I \times X_I \mapsto \mathbb{R}$$

  such that

  $$K(x_1, x_2) = \phi(x_1)^T \phi(x_2).$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The kernel trick

Some kernel functions

- Linear:

$$K(x_1, x_2) = x_1^T x_2$$

- RBF (Radial Basis Function) or Gaussian:

$$K(x_1, x_2) = \exp(-\|x_1 - x_2\|^2 / 2\sigma), \text{ for some value of } \sigma$$

- Hyperbolic tangent:

$$K(x_1, x_2) = \tanh(\beta x_1^T x_2 + \gamma), \text{ for some values of } \beta \text{ and } \gamma.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The kernel trick

The linear kernel

$$
D \left\{
\begin{array}{l}
\displaystyle \min_{\lambda,\mu} \quad \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j \overbrace{a_i^T a_j}^{K(a_i,a_j)} + \sum_{l=1}^{k}\sum_{j=1}^{k}\mu_l\mu_j \overbrace{b_l^T b_j}^{K(b_l,b_j)} - 2\sum_{i=1}^{m}\sum_{l=1}^{k}\lambda_i\mu_l \overbrace{a_i^T b_l}^{K(a_i,b_l)}\right) \\
\qquad\quad \displaystyle -\sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l \\
\\
\qquad\quad \displaystyle \sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l = 0 \\
\qquad\quad 0 \le \lambda_i \le C \quad i = 1,\ldots,m \\
\qquad\quad 0 \le \mu_l \le C \quad l = 1,\ldots.k
\end{array}
\right.
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

The kernel trick

The general case

$$
D \begin{cases}
\min\limits_{\lambda,\mu} & \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j \overbrace{\phi(a_i)^T\phi(a_j)}^{K(a_i,a_j)} + \sum_{l=1}^{k}\sum_{j=1}^{k}\mu_l\mu_j \overbrace{\phi(b_l)^T\phi(b_j)}^{K(b_l,b_j)} \right. \\
& \left. -2\sum_{i=1}^{m}\sum_{l=1}^{k}\lambda_i\mu_l \overbrace{\phi(a_i)^T\phi(b_l)}^{K(a_i,b_l)}\right) - \sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l \\
& \sum_{i=1}^{m}\lambda_i - \sum_{l=1}^{k}\mu_l = 0 \\
& 0 \le \lambda_i \le C \quad i = 1,\dots,m \\
& 0 \le \mu_l \le C \quad l = 1,\dots .k
\end{cases}
$$

NOTE: There is no need to know explicitly the map ϕ.

The kernel trick

$$(\lambda^*, \mu^*)$$

$$\Downarrow$$

$$v^* = \sum_{i=1}^{m} \lambda_i^* \phi(a_i) - \sum_{l=1}^{k} \mu_l^* \phi(b_l) \quad \text{not explicitly needed}$$

$$\gamma^* = {v^*}^T \phi(a_i) - 1, \text{ with } i \text{ such that } 0 < \lambda_i^* < C$$

or

$$\gamma^* = {v^*}^T \phi(b_l) + 1, \text{ with } l \text{ such that } 0 < \mu_l^* < C$$

NOTE: Substituting v^* in the expression of γ^*, γ^* is expressed in terms of the kernel function K.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The kernel trick

The decision function

In correspondence to any new point $\bar{x}$, we compute:

$$
\begin{aligned}
\overbrace{v^{*T}\phi(\bar{x})}^{\text{linear in} X_F} - \gamma^* &= \left(\sum_{i=1}^{m} \lambda_i^* \phi(a_i) - \sum_{l=1}^{k} \mu_l^* \phi(b_l)\right)^T \phi(\bar{x}) - \gamma^* \\
&= \sum_{i=1}^{m} \lambda_i^* \phi(a_i)^T \phi(\bar{x}) - \sum_{l=1}^{k} \mu_l^* \phi(b_l)^T \phi(\bar{x}) - \gamma^* \\
&= \underbrace{\sum_{i=1}^{m} \lambda_i^* K(a_i, \bar{x}) - \sum_{l=1}^{k} \mu_l^* K(b_l, \bar{x}) - \gamma^*}_{\text{nonlinear in} X_I, \text{ if } K \text{ is nonlinear}}.
\end{aligned}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Fixed-center spherical separation with kernel

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Fixed-center spherical separation with kernel (Astorino and Gaudioso, 2009 [AG09])

$$
\begin{aligned}
f(x_0, R) \triangleq R^2 \quad & + C \sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - R^2\} + \\
& + C \sum_{l=1}^{k} \max\{0, R^2 - \|b_l - x_0\|^2\},
\end{aligned}
$$

NOTE: If x_0 is fixed, setting $z \triangleq R^2 \geq 0$, then function f is convex in z.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Fixed-center spherical separation with kernel

$$f(z) \stackrel{\triangle}{=} z + C\sum_{i=1}^{m}\overbrace{\max\{0, \|a_i - x_0\|^2 - z\}}^{\xi_i} + C\sum_{l=1}^{k}\overbrace{\max\{0, z - \|b_l - x_0\|^2\}}^{\psi_l}.$$

In this case, minimization of f corresponds to solve the following linear program:

$$\begin{cases} \min\limits_{z,\xi,\psi} & z + C\sum\limits_{i=1}^{m}\xi_i + C\sum\limits_{l=1}^{k}\psi_l & \\ & \xi_i \geq \|a_i - x_0\|^2 - z & i = 1,\dots,m \\ & \psi_l \geq z - \|b_l - x_0\|^2 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \\ & z \geq 0 & \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Fixed-center spherical separation with kernel

NOTE

$$\|a_i - x_0\|^2 = \|a_i\|^2 + \|x_0\|^2 - 2a_i^T x_0,$$

i.e.

$$\|a_i - x_0\|^2 = \underbrace{a_i^T a_i}_{K(a_i,a_i)} + \underbrace{x_0^T x_0}_{K(x_0,x_0)} - 2 \underbrace{a_i^T x_0}_{K(a_i,x_0)} .$$

Moreover:

$$\|b_l - x_0\|^2 = \|b_l\|^2 + \|x_0\|^2 - 2b_l^T x_0,$$

i.e.

$$\|b_l - x_0\|^2 = \underbrace{b_l^T b_l}_{K(b_l,b_l)} + \underbrace{x_0^T x_0}_{K(x_0,x_0)} - 2 \underbrace{b_l^T x_0}_{K(b_l,x_0)} .$$

Fixed-center spherical separation with kernel

$$f(z) \stackrel{\triangle}{=} z \quad + C \sum_{i=1}^{m} \overbrace{\max\{0, K(a_i, a_i) + K(x_0, x_0) - 2K(a_i, x_0) - z\}}^{\xi_i}$$
$$+ C \sum_{l=1}^{k} \underbrace{\max\{0, z - K(b_l, b_l) - K(x_0, x_0) + 2K(b_l, x_0)\}}_{\psi_l}.$$

Minimization of f corresponds to solve the following linear program:

$$\begin{cases} \displaystyle\min_{z,\xi,\psi} & z + C \displaystyle\sum_{i=1}^{m} \xi_i + C \sum_{l=1}^{k} \psi_l \\ & \xi_i \geq K(a_i, a_i) + K(x_0, x_0) - 2K(a_i, x_0) - z \quad & i = 1, \ldots, m \\ & \psi_l \geq z - K(b_l, b_l) - K(x_0, x_0) + 2K(b_l, x_0) \quad & l = 1, \ldots, k \\ & \xi_i \geq 0 & i = 1, \ldots, m \\ & \psi_l \geq 0 & l = 1, \ldots, k. \\ & z \geq 0 & \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The kernel trick

The decision function

In correspondence to any new point $\bar{x}$, given z^*, we compute:

$$\|\phi(\bar{x}) - \phi(x_0)\|^2 = \|\phi(\bar{x})\|^2 + \|\phi(x_0)\|^2 - 2\phi(\bar{x})^T\phi(x_0),$$

i.e.

$$\|\phi(\bar{x}) - \phi(x_0)\|^2 = \underbrace{\phi(\bar{x})^T\phi(\bar{x})}_{K(\bar{x},\bar{x})} + \underbrace{\phi(x_0)^T\phi(x_0)}_{K(x_0,x_0)} - 2\underbrace{\phi(\bar{x})^T\phi(x_0)}_{K(\bar{x},x_0)}.$$

- if $K(\bar{x},\bar{x}) + K(x_0,x_0) - 2K(\bar{x},x_0) \leq z^*$ then $\bar{x}$ is classified as a point of $\mathcal{A}$;
- if $K(\bar{x},\bar{x}) + K(x_0,x_0) - 2K(\bar{x},x_0) > z^*$ then $\bar{x}$ is classified as a point of to $\mathcal{B}$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Proximal Support Vector Machine

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Proximal Support Vector Machine (PSVM) (Fung and Mangasarian, 2001 [FM01])

Given $\mathcal{A}$ and $\mathcal{B}$, the standard SVM model is:

$$\begin{cases} \min\limits_{v,\gamma,\xi,\psi} & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \end{cases}$$

$$\Downarrow$$

The two hyperplanes

$$H_+(v,\gamma) \triangleq \{x \in \mathbb{R}^n | v^T x = \gamma + 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

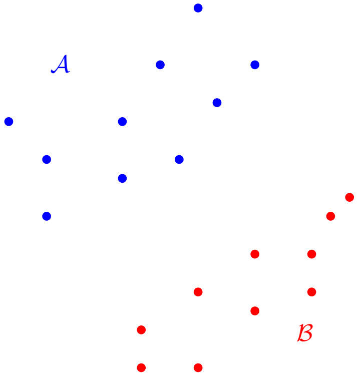
and

$$H_-(v,\gamma) \triangleq \{x \in \mathbb{R}^n | v^T x = \gamma - 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R}.$$

are the supporting hyperplanes.

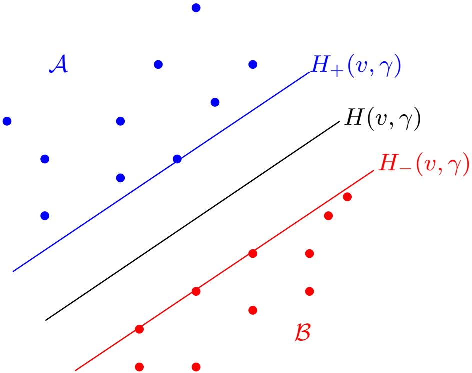
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

SVM example



H_+ and H_- are supporting hyperplanes

Proximal Support Vector Machine (PSVM)

Given the sets $\mathcal{A}$ and $\mathcal{B}$, the standard SVM model is:

$$\begin{cases} \min\limits_{v,\gamma,\xi,\psi} & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \end{cases}$$

i.e.

$$\begin{cases} \min\limits_{v,\gamma,\xi,\psi} & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}|\xi_i| + C\sum_{l=1}^{k}|\psi_l| & \\ & & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Proximal Support Vector Machine (PSVM)

Instead of

$$\begin{cases} \displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}|\xi_i| + C\sum_{l=1}^{k}|\psi_l| & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots,k \\ & \xi_i \geq 0 & i = 1,\dots,m \\ & \psi_l \geq 0 & l = 1,\dots,k. \end{cases}$$

consider

$$\begin{cases} \displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\left\|\begin{array}{c} v \\ \gamma \end{array}\right\|^2 + \frac{C}{2}\|\xi\|^2 + \frac{C}{2}\|\psi\|^2 & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1,\dots,m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1,\dots .k \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Proximal Support Vector Machine (PSVM)

Instead of

$$\begin{cases} \displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\left\| \begin{array}{c} v \\ \gamma \end{array} \right\|^2 + \frac{C}{2}\|\xi\|^2 + \frac{C}{2}\|\psi\|^2 & \\ & \xi_i \geq -v^T a_i + \gamma + 1 & i = 1, \ldots, m \\ & \psi_l \geq v^T b_l - \gamma + 1 & l = 1, \ldots. k \end{cases}$$

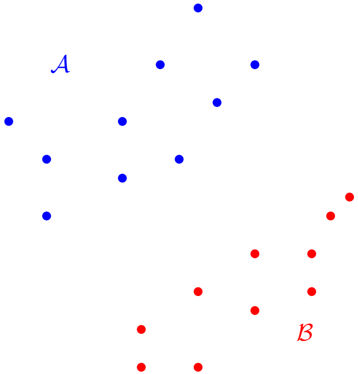
consider

$$\begin{cases} \displaystyle\min_{v,\gamma,\xi,\psi} & \displaystyle\frac{1}{2}\left\| \begin{array}{c} v \\ \gamma \end{array} \right\|^2 + \frac{C}{2}\|\xi\|^2 + \frac{C}{2}\|\psi\|^2 & \\ & \xi_i = -v^T a_i + \gamma + 1 & i = 1, \ldots, m \\ & \psi_l = v^T b_l - \gamma + 1 & l = 1, \ldots. k \end{cases}$$

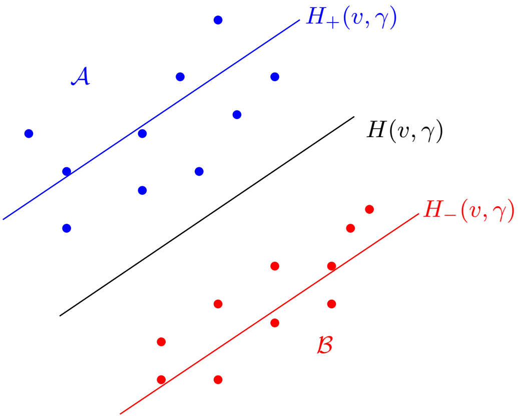
which can be solved very quickly in a closed form.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

PSVM example



PSVM example



H_+ and H_- are proximal hyperplanes

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

Spherical separation with margin (Astorino et al., 2012 [AFG12])

Motivation: To extend the concept of margin to spherical separation.
We recall that:
The set $\mathcal{A}$ is spherically separable from the set $\mathcal{B}$ if there exists a sphere

$$S(x_0, R) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | \|x - x_0\|^2 = R^2\},$$

with $x_0 \in \mathbb{R}^n$ and $R \in \mathbb{R}$,

such that

$$\|a_i - x_0\|^2 \leq R^2, \quad i = 1, \ldots, m$$

and

$$\|b_l - x_0\|^2 \geq R^2, \quad l = 1, \ldots, k.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

The set $\mathcal{A}$ is strictly spherically separable from the set $\mathcal{B}$ if there exists a sphere

$$S(x_0, R) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid \|x - x_0\|^2 = R^2\},$$

with $x_0 \in \mathbb{R}^n$ and $R \in \mathbb{R}$,

such that

$$\|a_i - x_0\|^2 \le (R - M)^2, \quad i = 1, \dots, m$$

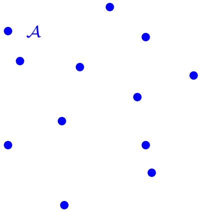
and

$$\|b_l - x_0\|^2 \ge (R + M)^2, \quad l = 1, \dots, k$$

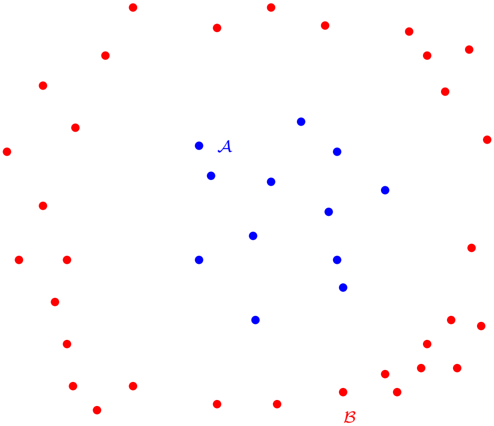
for some M with $0 < M \le R$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

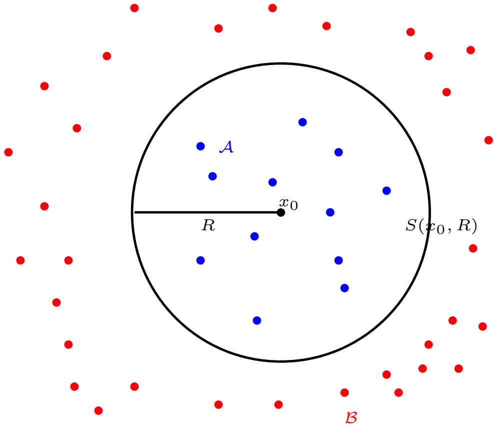


Spherical separation with margin

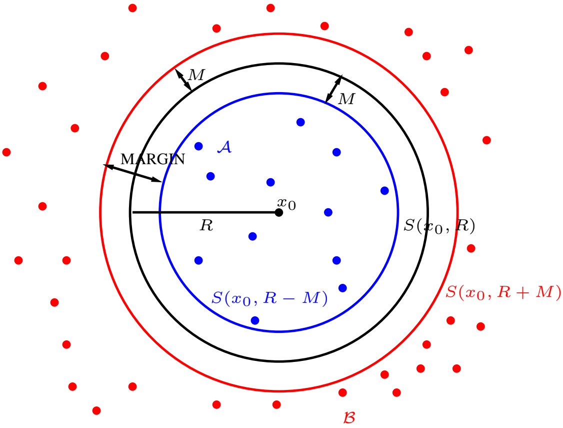


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin



Spherical separation with margin



Margin = $2M$.

Spherical separation with margin

Error function

- A point $a_i \in \mathcal{A}$ is correctly classified if

$$\|a_i - x_0\|^2 \leq (R - M)^2.$$

- As a consequence, a point $a_i \in \mathcal{A}$ is misclassified if

$$\|a_i - x_0\|^2 - (R - M)^2 > 0.$$

- Then the classification error, in correspondence to a point $a_i \in \mathcal{A}$, is

$$\max\{0, \|a_i - x_0\|^2 - (R - M)^2\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

Error function

- A point $b_l \in \mathcal{B}$ is correctly classified if

$$\|b_l - x_0\|^2 \geq (R + M)^2.$$

- As a consequence, a point $b_l \in \mathcal{B}$ is misclassified if

$$(R + M)^2 - \|b_l - x_0\|^2 > 0.$$

- Then the classification error, in correspondence to a point $b_l \in \mathcal{B}$, is

$$\max\{0, (R + M)^2 - \|b_l - x_0\|^2\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

Error function

$$
\begin{aligned}
e(x_0, R, M) \quad & \triangleq \sum_{i=1}^{m} \max\left\{0, \|a_i - x_0\|^2 - (R-M)^2\right\} \\
& + \sum_{l=1}^{k} \max\left\{0, (R+M)^2 - \|b_l - x_0\|^2\right\}.
\end{aligned}
$$

Setting $z \triangleq R^2 + M^2 \geq 0$ and $q \triangleq 2RM \geq 0$, we have:

$$
\left.
\begin{aligned}
e(x_0, z, q) \quad & = \sum_{i=1}^{m} \max\left\{0, q - z + \|a_i - x_0\|^2\right\} \\
& + \sum_{l=1}^{k} \max\left\{0, q + z - \|b_l - x_0\|^2\right\}
\end{aligned}
\right\} \text{nonsmooth and nonconvex}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Spherical separation with margin

Then we solve the following nonsmooth nonconvex optimization problem:

$$P \begin{cases} \min\limits_{x_0,z,q} & f(x_0,z,q) \\ & 0 \le q \le z, \end{cases}$$

where

$$f(x_0,z,q) \stackrel{\triangle}{=} Ce(x_0,z,q) - q$$

with $C > 0$.

NOTE 1: Minimizing $-q$ corresponds to maximize the margin.
NOTE 2: The parameter $C > 0$ tunes the weight of the two objectives.

Spherical separation with margin: fixing the center

NOTE: If x_0 is fixed, function f is convex in z and q.

$$f(z,q) \triangleq -q + C\sum_{i=1}^{m} \overbrace{\max\{0, q - z + \|a_i - x_0\|^2\}}^{\xi_i} + C\sum_{l=1}^{k} \overbrace{\max\{0, q + z - \|b_l - x_0\|^2\}}^{\psi_l}.$$

In this case, minimization of f corresponds to solve the following linear program:

$$\begin{cases} \min\limits_{z,q,\xi,\psi} & -q + C\sum_{i=1}^{m}\xi_i + C\sum_{l=1}^{k}\psi_l \\ & \xi_i \geq q - z + \|a_i - x_0\|^2 \qquad i = 1,\dots,m \\ & \psi_l \geq q + z - \|b_l - x_0\|^2 \qquad l = 1,\dots,k \\ & \xi_i \geq 0 \qquad i = 1,\dots,m \\ & \psi_l \geq 0 \qquad l = 1,\dots,k. \\ & 0 \leq q \leq z. \end{cases}$$

PART VI

UNSUPERVISED CLASSIFICATION

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Unsupervised classification

Unsupervised classification: we have only unlabelled objects, that we would like to cluster on the basis of their similarities.

The clustering problem

- A set

$$\mathcal{X} = \{x_1, \ldots, x_q\}$$

  of q unlabelled points is given.
- The objective is to group the points into h clusters, with $h \leq q$, on the basis of their similarities.
- Criterion: for each cluster j, $j = 1 \ldots, h$, compute the center x_{0_j} of the cluster such that each point x_p, $p = 1, \ldots, q$, is assigned to the cluster with the closest center.

A mixed integer model

A mixed integer model (Bagirov and Yearwood, 2006 [BY06])

A constrained optimization model is:

$$\begin{cases} \min\limits_{x_0, w} & \displaystyle\frac{1}{q}\sum_{p=1}^{q}\sum_{j=1}^{h} w_{pj}\|x_p - x_{0_j}\|^2 \\ & \displaystyle\sum_{j=1}^{h} w_{pj} = 1 \quad p = 1, \dots, q \\ & w_{pj} \in \{0, 1\} \quad p = 1, \dots, q \quad j = 1 \dots, h \end{cases}$$

where x_{0_j} is the center of the cluster j, for $j = 1, \dots, h$ and

$$w_{pj} = \begin{cases} 1 & \text{if the point } x_p \text{ is assigned to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

NOTE: It is a mixed integer nonlinear nonconvex program.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A mixed integer model

KKT conditions

$$L(x_0, w, \lambda) = \frac{1}{q}\sum_{p=1}^{q}\sum_{j=1}^{h} w_{pj}\|x_p - x_{0_j}\|^2 - \sum_{p=1}^{q}\lambda_p\left(\sum_{j=1}^{h} w_{pj} - 1\right)$$

$$\nabla L_{x_{0_j}} = \frac{1}{q}\sum_{p=1}^{q} 2w_{pj}(x_{0_j} - x_p) = 0, \quad j = 1\ldots,h$$

$$\sum_{p=1}^{q} w_{pj}x_{0_j} = \sum_{p=1}^{q} w_{pj}x_p, \quad j = 1\ldots,h$$

barycenter $\rightarrow x_{0_j} = \dfrac{\displaystyle\sum_{p=1}^{q} w_{pj}x_p}{\displaystyle\sum_{p=1}^{q} w_{pj}}, \quad j = 1\ldots,h.$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

An integer model

The model becomes an integer program:

$$\begin{cases} \min\limits_{w} & \dfrac{1}{q}\displaystyle\sum_{p=1}^{q}\sum_{j=1}^{h} w_{pj}\|x_p - \sum_{r=1}^{q} w_{rj}x_r / \sum_{r=1}^{q} w_{rj}\|^2 \\ & \displaystyle\sum_{j=1}^{h} w_{pj} = 1 \quad p = 1,\dots,q \\ & w_{pj} \in \{0,1\} \quad p = 1,\dots,q \quad j = 1\dots,h \end{cases}$$

where

$$w_{pj} = \begin{cases} 1 & \text{if the point } x_p \text{ is assigned to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

A nonsmooth model

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A nonsmooth model (Bagirov and Yearwood, 2006 [BY06])

An unconstrained optimization model is:

$$\min_{x_0} \frac{1}{q} \sum_{p=1}^{q} \min_{1 \le j \le h} \|x_p - x_{0_j}\|^2,$$

where x_{0_j} is the center of the cluster j, for $j = 1, \ldots, h$.

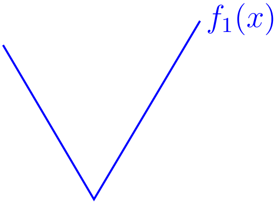
NOTE: If $h > 1$, the objective function is nonconvex and nonsmooth.

Unsupervised classification

Example: $f(x) = \min_{1 \le j \le 4} f_j(x),$ with $f_j(x)$ convex, for $j = 1, \ldots, 4$.

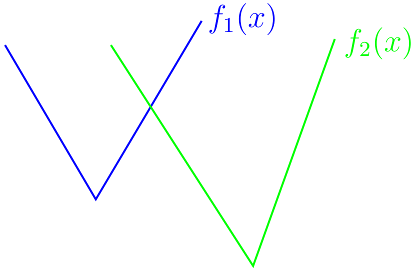
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Unsupervised classification



Example: $f(x) = \min_{1 \le j \le 4} f_j(x)$, with $f_j(x)$ convex, for $j = 1, \ldots, 4$.

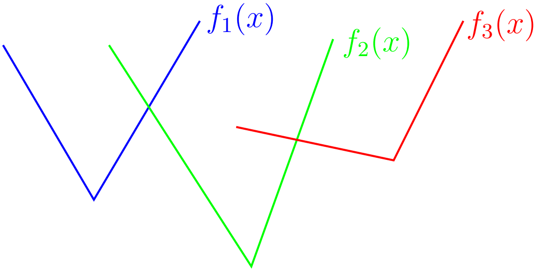
Unsupervised classification



Example: $f(x) = \min_{1 \le j \le 4} f_j(x),$ with $f_j(x)$ convex, for $j = 1, \ldots, 4.$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

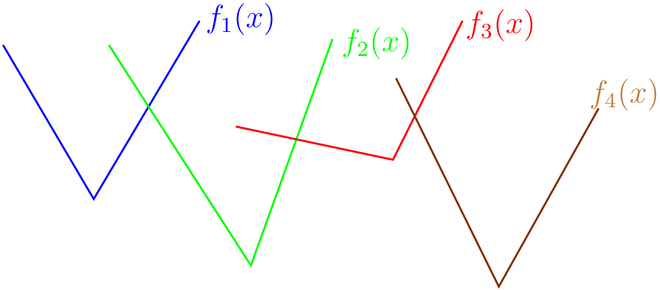
Unsupervised classification



Example: $f(x) = \min\limits_{1 \leq j \leq 4} f_j(x),$ with $f_j(x)$ convex, for $j = 1, \ldots, 4.$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

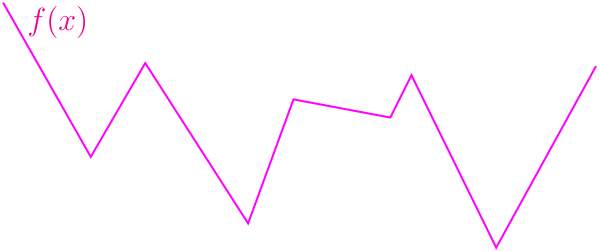
Unsupervised classification



Example: $f(x) = \min_{1 \le j \le 4} f_j(x)$, with $f_j(x)$ convex, for $j = 1, \ldots, 4$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Unsupervised classification



Example: $f(x) = \min_{1 \le j \le 4} f_j(x),$ with $f_j(x)$ convex, for $j = 1, \ldots, 4.$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A spherical model for unsupervised classification

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The clustering problem

- A set
$$\mathcal{X} = \{x_1, \ldots, x_q\}$$
of q unlabelled points is given.
- The objective is to group the points into h clusters
$$\mathcal{X}_1, \ldots, \mathcal{X}_h,$$
on the basis of their similarities, with $h \leq q$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The idea (Astorino et al., 2023 [AAC+23])

Proposed criterion: given h fixed-center spheres

$$S_j(x_{0_j}, R_j) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \,|\, \|x - x_{0_j}\|^2 = R_j^2\}, \quad j = 1, \ldots, h,$$

assign each point to the closest sphere. In other words a point $x_p \in \mathcal{X}$ is assigned to the cluster $\mathcal{X}_j$ if

$$||x_p - x_{0_j}||^2 - R_j^2 \leq ||x_p - x_{0_r}||^2 - R_r^2 \qquad r = 1, \ldots, h,$$

or, equivalently, if

$$||x_p - x_{0_j}||^2 - R_j^2 \leq \min_{1 \leq r \leq h} \{||x_p - x_{0_r}||^2 - R_r^2\}.$$

The error function

Letting

$$d_{pj} \stackrel{\triangle}{=} \|x_p - x_{0_j}\|^2,$$

where the center x_{0_j} of the sphere S_j is fixed, the error function e_{pj} of x_p with respect to cluster $\mathcal{X}_j$ is defined as follows:

$$\begin{aligned}
e_{pj}(R_1, \ldots, R_h) &\stackrel{\triangle}{=} \max \left\{ 0, d_{pj} - R_j^2 - \min_{1 \le r \le h} \{d_{pr} - R_r^2\} \right\} \\
&= \max \left\{ 0, d_{pj} - R_j^2 + \max_{1 \le r \le h} \{R_r^2 - d_{pr}\} \right\} \\
&= \max_{1 \le r \le h} \left\{ 0, (R_r^2 - d_{pr}) - (R_j^2 - d_{pj}) \right\} \\
&= \max_{1 \le r \le h} \{R_r^2 - d_{pr}\} - R_j^2 + d_{pj}.
\end{aligned}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The model

Two objectives

- Minimizing the overall error function;
- minimizing the volume of each clustering sphere, in order to increase the “generalization capability”.

$$\Downarrow$$

$$\min_{R_1,\ldots,R_h} \sum_{p=1}^{q} \sum_{j=1}^{h} e_{pj}(R_1,\ldots,R_h) + C \sum_{j=1}^{h} R_j^2,$$

with $C > 0$.

NOTE: Nonsmooth and nonconvex objective function.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Obtaining a linear model

Letting $z_j \stackrel{\triangle}{=} R_j^2 \geq 0$, for $j = 1, \ldots, h$, the previous problem reduces to a linear program. In fact:

$$\begin{cases} \min\limits_{z} & \displaystyle\sum_{p=1}^{q}\sum_{j=1}^{h} \max_{1 \leq r \leq h} \{z_r - d_{pr}\} + \sum_{p=1}^{q}\sum_{j=1}^{h}(d_{pj} - z_j) + C\sum_{j=1}^{h} z_j \\ & z_j \geq 0 \quad j = 1, \ldots, h. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Obtaining a linear model

$$\begin{cases} \displaystyle\min_{z} \quad \sum_{p=1}^{q}\sum_{j=1}^{h} d_{pj} + (C-m)\sum_{j=1}^{h} z_j + h\sum_{p=1}^{q}\max_{1\le j\le h}\{z_j - d_{pj}\} \\ \\ z_j \ge 0 \quad j=1,\dots,h. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Obtaining a linear model

Neglecting the constant $\sum_{p=1}^{q} \sum_{j=1}^{h} d_{pj}$, we obtain the following linear program:

$$
LP \begin{cases}
\min\limits_{z,\xi} & (C-q)\sum_{j=1}^{h} z_j + h\sum_{p=1}^{q} \xi_p \\[2ex]
& \xi_p - z_j \geq -d_{pj} \qquad\qquad p=1,\dots,q \quad j=1,\dots,h \\[2ex]
& z_j \geq 0 \qquad\qquad j=1,\dots,h.
\end{cases}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Computing the dual

Computing the dual of problem LP, we obtain:

$$
TP \begin{cases}
\min\limits_{\alpha} & \displaystyle\sum_{p=1}^{q}\sum_{j=1}^{h} d_{pj}\alpha_{pj} \\
& \displaystyle\sum_{j=1}^{h} \alpha_{pj} = h \qquad p = 1,\dots,q \\
& \displaystyle\sum_{p=1}^{q} \alpha_{pj} \ge q - C \quad j = 1,\dots,h \\
& \alpha_{pj} \ge 0 \qquad p = 1,\dots,q \quad j = 1\dots,h.
\end{cases}
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The transportation problem

- Problem TP is a transportation problem;
- the pth source, for $p = 1, \ldots, q$, is represented by the point x_p;
- the jth destination, for $j = 1, \ldots, h$, is represented by the cluster $\mathcal{X}_j$;
- the supply of each source is equal to h and the demand of each destination is equal to $q - C$.
- problem TP admits an optimal solution: it is feasible since $C > 0$ and it cannot be unbounded, since $\alpha_{pj} \leq h$.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Comparison with the K-Means criterion

NOTE: In case in problem LP $z_j^* = 0$ for any $j = 1, \dots, h$, the clustering criterion, based on assigning each point of $\mathcal{X}$ to the closest sphere, reduces to the standard criterion used by the K-Means algorithm.

Theoretical results

Theorem (sufficient condition)

If $C > q$, any optimal solution to problem LP is such that $z_j^* = 0$ for $j = 1, \ldots, h$.

Proof.

If $C > q$, then $q - C < 0$, which is the right hand side of the demand constraints of the transportation problem TP. Because of the nonnegativity of the dual variables α_{pj}, all the demand constraints of problem TP are strictly satisfied in correspondence to any optimal solution. The proof follows from the complementary slackness relationships. □

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Theoretical results

Theorem (necessary condition)

Under an appropriate assumption concerning problem LP, if an optimal solution of LP is such that $z_j^* = 0$ for any $j = 1, \dots, h$, then

$$C \geq q - h \min_{1 \leq j \leq h} |D_j|,$$

where

$$D_j \triangleq \left\{ p \in \{1, \dots, q\} \mid d_{pj} = \min_{1 \leq r \leq h} d_{pr} \right\} \quad \textit{for } j = 1, \dots, h.$$

NOTE 1: If $C < q - h \min_{1 \leq j \leq h} |D_j|$, then any optimal solution of LP provides at least a value $z_j^* > 0$.

NOTE 2: The r.h.s. of the above condition is nonnegative since

$$\min_{1 \leq j \leq h} |D_j| \leq \frac{q}{h}.$$

PART VII

BINARY SEMISUPERVISED CLASSIFICATION

Semisupervised classification

Semisupervised classification: on the basis of the labelled and unlabelled objects, we would like to predict the class of the unlabelled objects.

Transductive Support Vector Machine

Transductive Support Vector Machine (TSVM) (Chapelle and Zien, 2005 [CZ05])

- The TSVM (Transductive Support Vector Machine) technique is the semisupervised version of the SVM approach.
- We compute the best support vector machine, on the basis of the labelled points (i.e. the sets $\mathcal{A}$ and $\mathcal{B}$) and some unlabelled points.
- The objective is to classify the unlabelled points.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Transductive Support Vector Machine (TSVM)

- The sets

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

  and

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

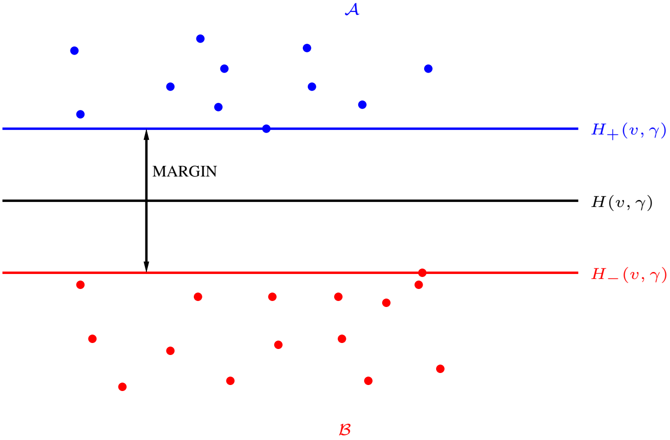
  are given.
- Another set

$$\mathcal{X} = \{x_1, \dots, x_q\}$$

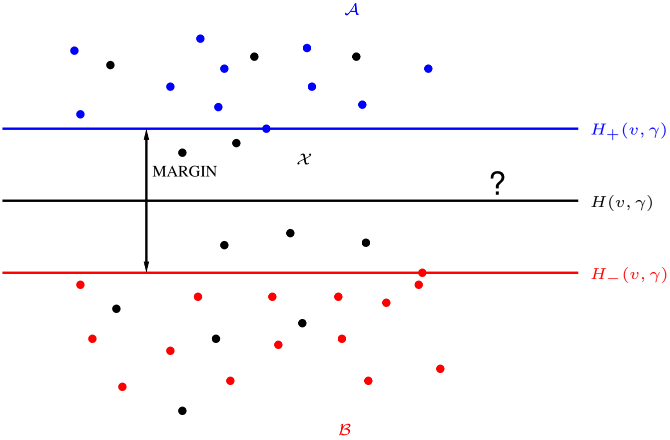
  of q unlabelled points is given.
- The objective is to obtain the best SVM having as few unlabelled points as possible in the margin.
- **NOTE**: Number q in the practical cases is very large.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

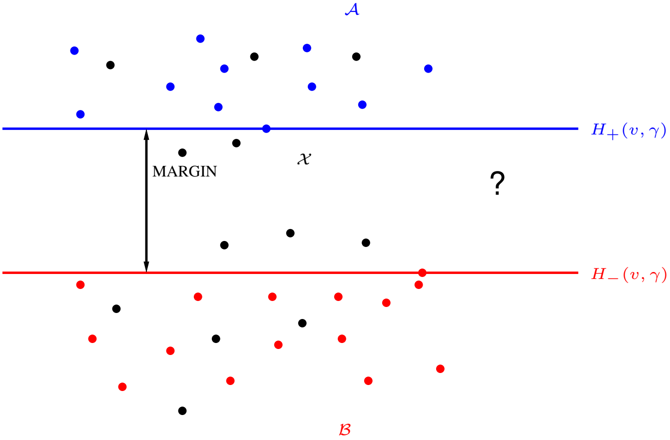
SVM: an example



TSVM: an example

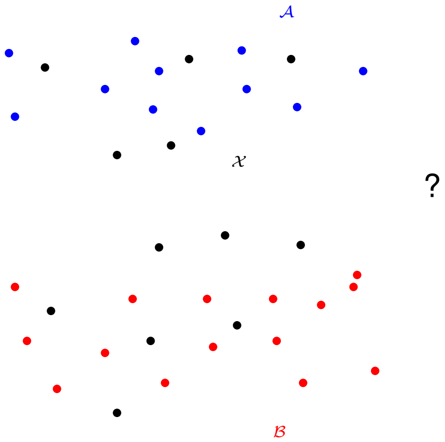


TSVM: an example

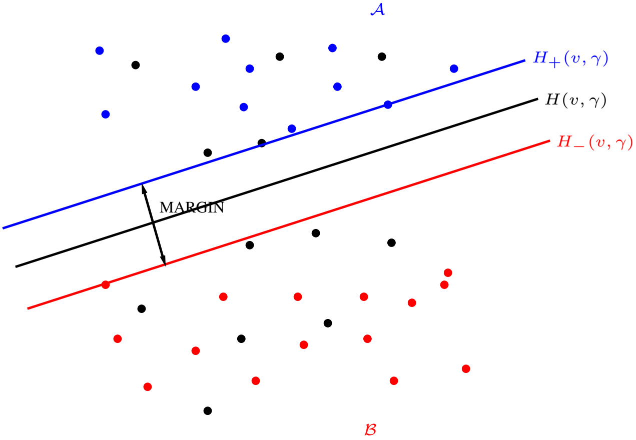


UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

TSVM: an example



TSVM: an example



TSVM: the error function

Question: How to minimize the number of unlabelled points in the margin?

TSVM: the error function

The margin is the area between the two supporting hyperplanes

$$H_+(v,\gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | v^T x = \gamma + 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R},$$

and

$$H_-(v,\gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n | v^T x = \gamma - 1\}, \text{ with } v \in \mathbb{R}^n \text{ and } \gamma \in \mathbb{R}.$$

Then, a point $x \in \mathcal{X}$ belongs to the margin if

$$v^T x < \gamma + 1 \text{ and } v^T x > \gamma - 1,$$

i.e. if

$$-1 < v^T x - \gamma < 1,$$

i.e. if

$$|v^T x - \gamma| < 1,$$

i.e. if

$$1 - |v^T x - \gamma| > 0.$$

TSVM: the error function

We want to find a separating hyperplane $H(v, \gamma)$, with $v \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, by minimizing the following function:

$$
\begin{aligned}
f(v,\gamma) \triangleq \;& \frac{1}{2}\|v\|^2 + \\
& + C_1 \left[\sum_{i=1}^{m} \max\{0, -v^T a_i + \gamma + 1\} + \sum_{l=1}^{k} \max\{0, v^T b_l - \gamma + 1\} \right] \\
& + C_2 \sum_{p=1}^{q} \max\{0, 1 - |v^T x_p - \gamma|\}.
\end{aligned}
$$

- f is nonsmooth;
- f is nonconvex, due to the last term involving the unlabelled points;
- $C_1, C_2 > 0$ tune the weights of the three objectives (generally $C_2 \leq C_1$).

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised polyhedral separation

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised polyhedral separation (Astorino and Fuduli, 2015 [AF15b])

- The sets

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

  and

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

  are given.
- Another set

$$\mathcal{X} = \{x_1, \ldots, x_q\}$$

  of q unlabelled points is given.
- The objective is to obtain the best polyhedral separation having as few unlabelled points as possible in the margin.

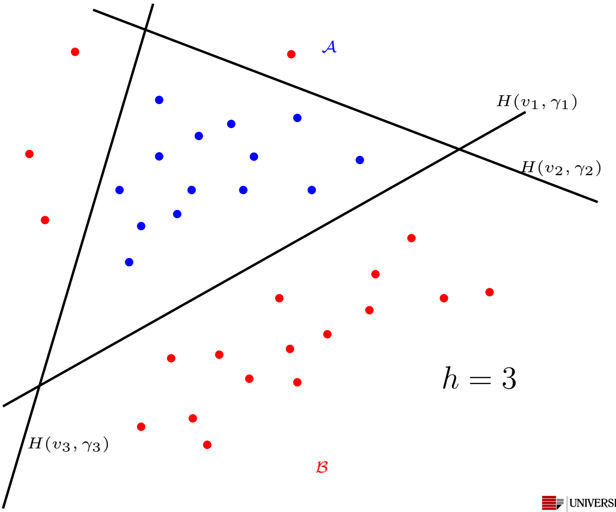
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised polyhedral separation

In the standard polyhedral separation we minimize the following function:

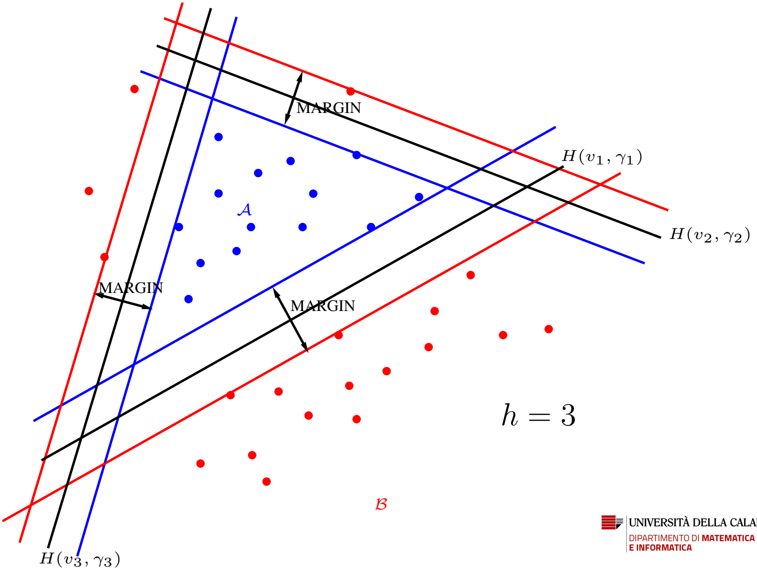
$$
\begin{aligned}
f(v_1, \ldots, v_h; \gamma_1, \ldots, \gamma_h) \quad \stackrel{\triangle}{=} \quad & \frac{1}{m} \sum_{i=1}^{m} \max_{1 \le j \le h} \{0, v_j^T a_i - \gamma_j + 1\} + \\
& + \frac{1}{k} \sum_{l=1}^{k} \max\{0, \min_{1 \le j \le h} -v_j^T b_l + \gamma_j + 1\}.
\end{aligned}
$$

Semisupervised polyhedral separation



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised polyhedral separation



Semisupervised polyhedral separation

Then, combining the TSVM approach and the polyhedral separation, we obtain the following error function:

$$
\begin{aligned}
f(v_1,\dots,v_h;\gamma_1,\dots,\gamma_h) \quad \stackrel{\triangle}{=} \quad & \frac{1}{2}\sum_{j=1}^{h}\|v_j\|^2 + C_1\sum_{i=1}^{m}\max_{1\le j\le h}\{0, v_j^T a_i - \gamma_j + 1\} + \\
& + C_1\sum_{l=1}^{k}\max\{0, \min_{1\le j\le h} -v_j^T b_l + \gamma_j + 1\} \\
& + C_2\sum_{j=1}^{h}\sum_{p=1}^{q}\max\{0, 1 - |v_j^T x_p - \gamma_j|\}.
\end{aligned}
$$

- f is nonsmooth and nonconvex;
- $C_1, C_2 > 0$ tune the weights of the three objectives (generally $C_2 \le C_1$).

Semisupervised spherical separation

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised spherical separation (Astorino and Fuduli, 2015 [AF15a])

- In the semisupervised spherical separation approach, we compute a separating sphere, on the basis of the labelled points (i.e. the sets $\mathcal{A}$ and $\mathcal{B}$) and some unlabelled points.
- The objective is to classify the unlabelled points.

Semisupervised spherical separation

- The sets

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

and

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

are given.

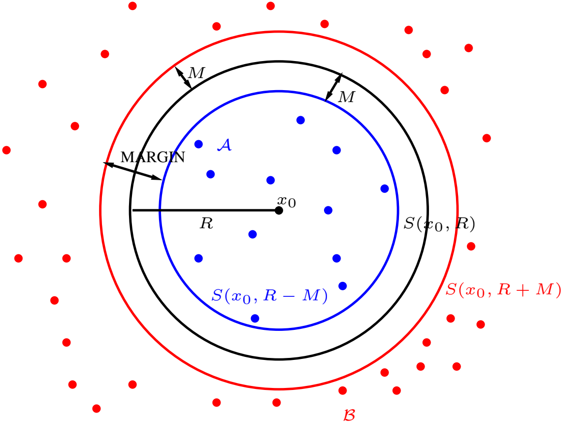
- Another set

$$\mathcal{X} = \{x_1, \dots, x_q\}$$

of q unlabelled points is given.

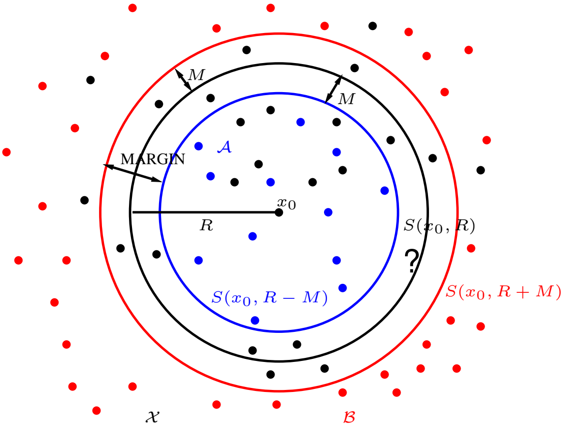
- The objective is to obtain a separating sphere having as few unlabelled points as possible in the margin.

Semisupervised spherical separation: an example



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Semisupervised spherical separation: an example



The error function

Question: How to minimize the number of unlabelled points in the margin?

The error function

A point $x \in \mathcal{X}$ belongs to the margin if

$$\|x - x_0\|^2 < (R+M)^2 \text{ and } \|x - x_0\|^2 > (R-M)^2,$$

i.e. if

$$(R+M)^2 - \|x - x_0\|^2 > 0 \text{ and } \|x - x_0\|^2 - (R-M)^2 > 0,$$

i.e. if

$$\min\{(R+M)^2 - \|x - x_0\|^2, \|x - x_0\|^2 - (R-M)^2\} > 0.$$

Setting $z \stackrel{\triangle}{=} R^2 + M^2 \geq 0$ and $q \stackrel{\triangle}{=} 2RM \geq 0$, we have that x belongs to the margin if:

$$\min\{q + z - \|x - x_0\|^2, \|x - x_0\|^2 - z + q\} > 0.$$

The error function

Then we minimize the following function:

$$
\begin{aligned}
f(x_0, z, q) &= -q \\
&+C_1 \sum_{i=1}^{m} \max\left\{0, q - z + \|a_i - x_0\|^2\right\} \\
&+C_1 \sum_{l=1}^{k} \max\left\{0, q + z - \|b_l - x_0\|^2\right\} \\
&+C_2 \sum_{p=1}^{q} \max\{0, \min[q + z - \|x_p - x_0\|^2, \|x_p - x_0\|^2 - z + q]\}
\end{aligned}
$$

such that $0 \leq q \leq z$.

- f is nonsmooth and nonconvex;
- $C_1, C_2 > 0$ tune the weights of the three objectives (generally $C_2 \leq C_1$).

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

PART VIII

MULTIPLE INSTANCE LEARNING

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Introduction to Multiple Instance Learning

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Multiple instance learning (MIL)

- Supervised learning: the objective is to categorize points into different classes, on the basis of labelled points.
- Multiple instance learning (MIL): the objective is to classify bags of points, each point being an instance.
- **NOTE**: In the learning phase of a MIL approach, we know the label of each bag, but the label of each instance inside the bags is unknown.

The first MIL problem (Dietterich et al., 1997 [DLLP97])

- Drug design problem: we want to discriminate between active and non-active drug molecules;
- a drug molecule is active if it is able to bind to a particular target site (typically a larger protein molecule);
- each molecule can assume different conformations;
- ...but indeed it is not known which conformation makes a molecule active;
- in the MIL perspective, each molecule is a bag and the conformations of the molecules are the instances.

MIL: the binary case

- Binary case: we would like to discriminate between two classes of bags (positive and negative) and to predict the class label of new bags.
- **NOTE**: Even in the binary case, we can have more than two classes of instances.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Example n. 1

- We have some images and we would like to discriminate between beach and non-beach;
- each image is a bag containing some “subregions” (instances): sea, countryside, cities, cars, offices, sky, sand, trees, mountains, etc.;
- an image is positive (i.e. a beach) if it contains both sea and sand;
- an image is negative if it does not contain both sea and sand.

Example n. 2

- Objective: to discriminate between non-healthy and healthy patients on the basis of their medical scan (bag);
- a patient is positive if he/she presents at least an abnormal subregion (instance) in his/her medical scan;
- a patient is negative if all the subregions (instances) in his/her medical scan are healthy.

Multiple instance learning (MIL)

NOTE: In both previous examples, only some portions of the image (or medical scan) make the image positive.

$$\Downarrow$$

The MIL approach can be interpreted as a weakly supervised approach.

NOTE: In the binary case, a crucial issue is to specify what a positive bag is.

Possible applications of MIL

- Classification of images;
- drug discovery;
- classification of text documents;
- bankruptcy prediction;
- speaker identification.

Classification of the MIL approaches

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The binary case: bag-space learning

We have two classes of bags: positive and negative.

- In the bag-space learning we separate directly the positive bags from the negative ones, considering each bag as a whole entity.
- This approach is necessary when there is no class of instances appearing only in positive bags.

The binary case: instance-space learning

We have two classes of bags: positive and negative.

- In the instance-space learning we separate the instances belonging to the positive bags from the instances belonging to the negative ones.
- Then the class label information of a bag is obtained as aggregation of the instance-space responses.
- This approach is possible when some classes of instances appear only in positive bags.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

The binary case: embedding-space learning

We have two classes of bags: positive and negative.

- In the embedding-space learning we map each bag to a single feature vector (typically the most representative instance belonging to the bag), resulting in a classical supervised classification problem to be solved in the instance space.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

MIL surveys

- J. Foulds and E. Frank, A review of multi-instance learning assumptions, Knowledge Engineering Review, 25 (2010), pp. 1–25.
- J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artificial Intelligence, 201 (2013), pp. 81–105.
- M. Carbonneau, V. Cheplygina, E. Granger and G. Gagnon, Multiple instance learning: a survey of problem characteristics and applications, Pattern Recognition, 77 (2018), pp. 329 – 353.
- G. Quellec, G. Cazuguel, B. Cochener and M. Lamard, Multiple-Instance Learning for Medical Image and Video Analysis, IEEE Reviews in Biomedical Engineering, 10 (2017), pp. 213–234.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Binary MIL problem: assumptions

MIL STANDARD ASSUMPTION

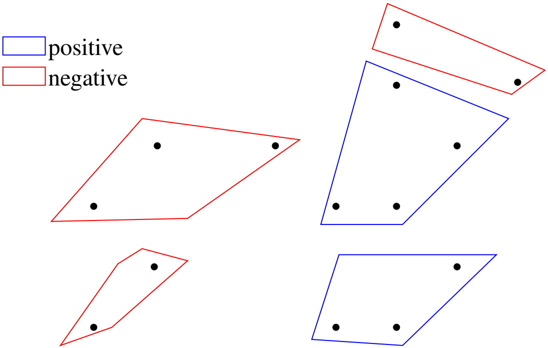
- Two classes of bags: positive and negative;
- two classes of instances: positive and negative.

$$\Downarrow$$

- A bag is positive if it contains at least a positive instance;
- a bag is negative if all its instances are negative.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

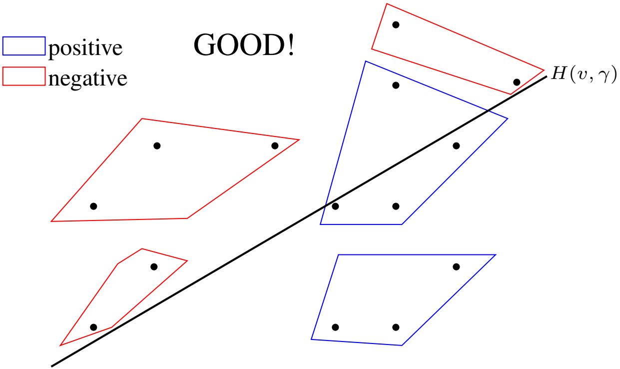
Standard MIL assumption: an example



- A bag is classified positive if at least one of its instances is classified positive.
- A bag is classified negative if all its instances are classified negative.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

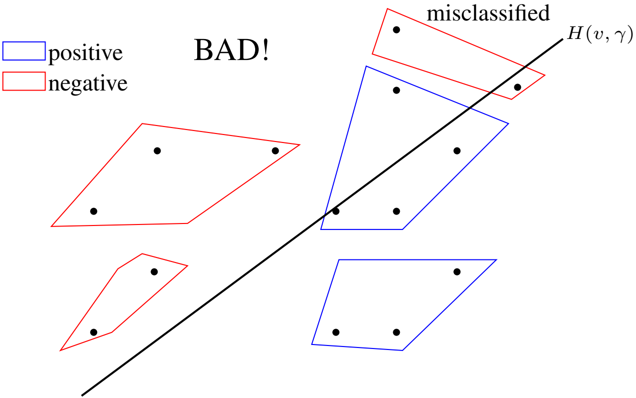
Standard MIL assumption: an example



- A bag is classified positive if at least one of its instances is classified positive.
- A bag is classified negative if all its instances are classified negative.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Standard MIL assumption: an example



- A bag is classified positive if at least one of its instances is classified positive.
- A bag is classified negative if all its instances are classified negative.

Support Vector Machine for Multiple Instance Learning

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model (Andrews et al., 2003 [ATH03])

NOTATION

- $\mathcal{A}_1, \ldots, \mathcal{A}_m$: m positive bags;
- $\mathcal{B}_1, \ldots, \mathcal{B}_k$: k negative bags;
- J_i^+: index set corresponding to $\mathcal{A}_i$, $i = 1, \ldots, m$;
- J_l^-: index set corresponding to $\mathcal{B}_l$, $l = 1, \ldots, k$;
- x_j: the jth instance;
- $y_j \in \{1, -1\}$: the class label of the instance x_j, when x_j belongs to a positive bag.

$$\Downarrow$$

$$H(v, \gamma) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid v^T x = \gamma\}.$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model

Minimize $f(v, \gamma, y)$, where

$$f(v,\gamma,y) \triangleq \frac{1}{2}\|v\|^2 \quad +C\sum_{i=1}^{m}\sum_{j\in J_i^+}\max\{0, 1+y_j(-v^Tx_j+\gamma)\}$$
$$+C\sum_{l=1}^{k}\sum_{j\in J_l^-}\max\{0, 1+(v^Tx_j-\gamma)\},$$

such that:

$$\sum_{j\in J_i^+}\frac{y_j+1}{2} \geq 1, \quad i=1,\ldots,m$$

and

$$y_j \in \{-1, 1\}, \quad j\in J_i^+, \quad i = \ \ldots, m.$$

A MIL SVM model

$$\begin{cases} f^* = \min\limits_{v,\gamma,y} & f(v,\gamma,y) \\ & \displaystyle\sum_{j \in J_i^+} \frac{y_j + 1}{2} \geq 1 \quad i = 1,\dots,m \\ & y_j \in \{-1,1\}, \quad j \in J_i^+, \quad i = 1,\dots,m. \end{cases}$$

NOTE: Constrained, nonlinear, nonconvex, mixed integer problem.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model

$$
MIL-SVM \left\{
\begin{array}{ll}
\displaystyle\min_{v,\gamma,y,\xi,\psi} & \displaystyle\frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\sum_{j\in J_i^+}\xi_j + C\sum_{l=1}^{k}\sum_{j\in J_l^-}\psi_j \\[2ex]
& \xi_j \geq 1 + y_j(-v^T x_j + \gamma) \quad j \in J_i^+,\ i = 1,\dots,m \\[2ex]
& \psi_j \geq 1 + (v^T x_j - \gamma) \quad j \in J_l^-,\ l = 1,\dots,k \\[2ex]
& \displaystyle\sum_{j\in J_i^+} \frac{y_j + 1}{2} \geq 1 \quad i = 1,\dots,m \\[2ex]
& y_j \in \{-1,+1\} \quad j \in J_i^+, \quad i = 1,\dots,m \\[2ex]
& \xi_j \geq 0 \quad j \in J_i^+,\ i = 1,\dots,m \\[2ex]
& \psi_j \geq 0 \quad j \in J_l^-,\ l = 1,\dots,k.
\end{array}
\right.
$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: the BCD approach

- BCD = Block Coordinate Descent method.
- Once y_j is fixed, solve the SVM problem to compute v and γ.
- Once v and γ are fixed, compute y_j by inspection.

A MIL SVM model: the BCD approach

COMPUTING y_j BY INSPECTION

$$z_j \stackrel{\triangle}{=} \max\{0, y_j L_j + 1\}, \text{ where } L_j \stackrel{\triangle}{=} -v^T x_j + \gamma$$

$$\begin{cases} \text{if } L_j > 0 \Rightarrow y_j^* = -1 \\ \text{if } L_j \leq 0 \Rightarrow y_j^* = +1 \end{cases}$$

NOTE 1: $L_j > 0 \Rightarrow -v^T x_j + \gamma > 0 \Rightarrow v^T x_j - \gamma > 0 \Rightarrow v^T x_j < \gamma$

NOTE 2: $L_j \leq 0 \Rightarrow -v^T x_j + \gamma \leq 0 \Rightarrow v^T x_j - \gamma \geq 0 \Rightarrow v^T x_j \geq \gamma$

A MIL SVM model: the BCD approach

1. Set $\bar{y}_j := +1$, for any $j \in J_i^+$, $i = 1, \dots, m$.
2. Solve $MIL - SVM$ with $y = \bar{y}$, to compute $\bar{v}$ and $\bar{\gamma}$.
3. If $\bar{v}^T x_j \geq \bar{\gamma}$ set $\bar{y}_j := +1$, else set $\bar{y}_j := -1$.
4. For any $i \in \{1, 2, \dots, m\}$ such that

$$\sum_{j \in J_i^+} \frac{\bar{y}_j + 1}{2} = 0,$$

   compute k_i such that

$$\bar{v}^T x_{k_i} - \bar{\gamma} = \max_{j \in J_i^+} \{\bar{v}^T x_j - \gamma\}$$

   and set $\bar{y}_{k_i} := +1$.
5. If $\bar{y}$ has changed go to Step 2, else STOP.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

$$LR(\lambda) \left\{ \begin{array}{ll} z^*_{LR}(\lambda) = \min\limits_{v,\gamma,y} & \mathcal{L}(v,\gamma,y,\lambda) \\ & y_j \in \{-1,1\}, \quad j \in J_i^+, \quad i = \ldots, m, \end{array} \right.$$

where

$$\begin{aligned} \mathcal{L}(v,\gamma,y,\lambda) \stackrel{\triangle}{=} \quad & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\sum_{j\in J_i^+} \max\{0, 1 + y_j(-v^T x_j + \gamma)\} \\ & + C\sum_{l=1}^{k}\sum_{j\in J_l^-} \max\{0, 1 + (v^T x_j - \gamma)\} \\ & - \sum_{i=1}^{m} \lambda_i \left(\sum_{j\in J_i^+} \frac{y_j+1}{2} - 1 \right). \end{aligned}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

BCD APPROACH FOR SOLVING $LR(\lambda)$, when $\lambda \geq 0$ is fixed

$$\begin{aligned}\mathcal{L}(v,\gamma,y,\lambda) \stackrel{\triangle}{=} \quad & \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{m}\sum_{j\in J_i^+}\max\{0, 1+y_j(-v^Tx_j+\gamma)\} \\ & + C\sum_{l=1}^{k}\sum_{j\in J_l^-}\max\{0, 1+(v^Tx_j-\gamma)\} \\ & + \sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\sum_{j\in J_i^+}\lambda_i\frac{y_j+1}{2}\end{aligned}$$

- Once y_j is fixed, solve the SVM problem to compute v and γ.
- Once v and γ are fixed, compute y_j by inspection.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

COMPUTING y_j BY INSPECTION

$$z_j \stackrel{\triangle}{=} \max\{0, y_j L_j + 1\} - \lambda_i \frac{y_j + 1}{2},$$

where $L_j \stackrel{\triangle}{=} -v^T x_j + \gamma$ and λ_i is the Lagrangian multiplier such that $j \in J_i^+$.

$$\Downarrow$$

3 cases, on the basis of the value of L_j.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

CASE 1 $(L_j \leq -1)$

- $y_j = +1 \Rightarrow y_j L_j + 1 = L_j + 1 \leq 0 \Rightarrow z_j = -\lambda_i \leq 0.$
- $y_j = -1 \Rightarrow y_j L_j + 1 = \underbrace{-L_j}_{\geq 1} + 1 \geq 2 \Rightarrow z_j = C \underbrace{(-L_j + 1)}_{\geq 2} \geq 2C > 0.$

$$\Downarrow$$

If $L_j \leq 1,$ then $y_j^* = +1.$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

CASE 2 ($L_j \geq 1$)

- $y_j = +1 \Rightarrow y_j L_j + 1 = L_j + 1 \geq 2 \Rightarrow z_j = C(L_j + 1) - \lambda_i.$
- $y_j = -1 \Rightarrow y_j L_j + 1 = \underbrace{-L_j}_{\leq -1} +1 \leq 0 \Rightarrow z_j = 0.$

$$\Downarrow$$

$$\text{If } L_j \geq 1, \text{ then } \begin{cases} \text{if } C(L_j+1) - \lambda_i \leq 0, & \text{then } y_j^* = +1 \\ \text{if } C(L_j+1) - \lambda_i > 0, & \text{then } y_j^* = -1. \end{cases}$$

$$\Downarrow$$

$$\text{If } L_j \geq 1, \text{ then } \begin{cases} \text{if } \lambda_i \geq C(L_j+1), & \text{then } y_j^* = +1 \\ \text{if } \lambda_i < C(L_j+1), & \text{then } y_j^* = -1. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

A MIL SVM model: a Lagrangian relaxation approach

CASE 3 $(-1 < L_j < 1)$

- $y_j = +1 \Rightarrow y_j L_j + 1 = L_j + 1 > 0 \Rightarrow z_j = C(L_j + 1) - \lambda_i.$
- $y_j = -1 \Rightarrow y_j L_j + 1 = -L_j + 1 > 0 \Rightarrow z_j = C(-L_j + 1).$

$$\Downarrow$$

$$\text{If } -1 < L_j < 1, \text{ then } \begin{cases} \text{if } C(L_j+1) - \lambda_i \le C(-L_j+1), & \text{then } y_j^* = +1 \\ \text{if } C(L_j+1) - \lambda_i > C(-L_j+1), & \text{then } y_j^* = -1. \end{cases}$$

$$\Downarrow$$

$$\text{If } -1 < L_j < 1, \text{ then } \begin{cases} \text{if } \lambda_i \ge 2CL_j, & \text{then } y_j^* = +1 \\ \text{if } \lambda_i < 2CL_j, & \text{then } y_j^* = -1. \end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

PART IX

EVALUATION OF A CLASSIFIER

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

10-fold cross-validation

Evaluation of a classifier

Question: How to evaluate the quality of a binary classifier?

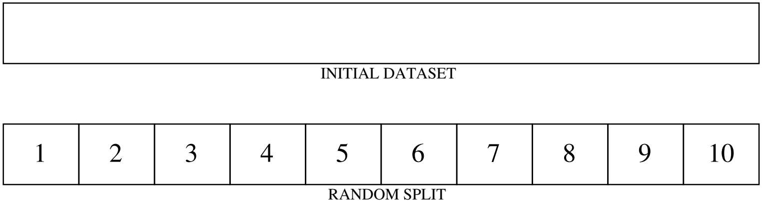
Answer: A possibility is to use a 10-fold cross-validation, which consists in randomly generating 10 folds, each of them constituted by a training set and a testing set.

1. The training set: (90% of the data) is used to construct (to learn) the classifier, i.e. the separation surface (such as a hyperplane, a sphere, and so on). It corresponds to the m positive points of $\mathcal{A}$ and to the k negative points of $\mathcal{B}$.
2. The testing set (10% of the data) simulates the unknown data to be classified.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

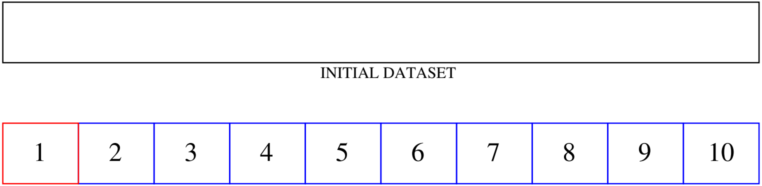
Evaluation of a classifier

10 fold cross-validation (first level)



Evaluation of a classifier

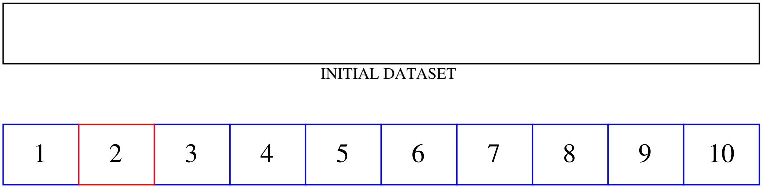
10 fold cross-validation



$$\text{FOLD 1} \begin{cases} \text{testing set:} & 1 \\ \text{training set:} & 2, 3, \ldots, 9, 10 \end{cases}$$

Evaluation of a classifier

10 fold cross-validation (first level)

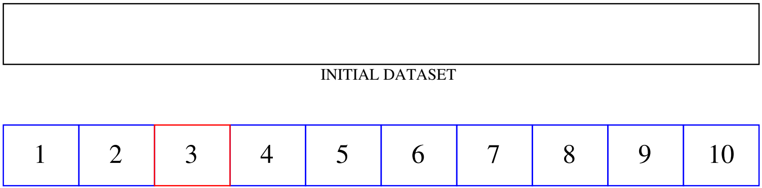


$$\text{FOLD 2}\begin{cases}\text{testing set:} & 2\\ \text{training set:} & 1, 3, \ldots, 9, 10\end{cases}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

10 fold cross-validation



FOLD 3 $\begin{cases} \text{testing set:} & 3 \\ \text{training set:} & 1, 2, 4, \ldots, 9, 10 \end{cases}$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

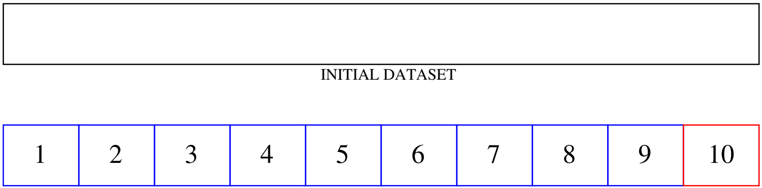
Evaluation of a classifier

...and so on...

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA
E INFORMATICA

Evaluation of a classifier

10 fold cross-validation



FOLD 10 $\begin{cases} \text{testing set:} & 10 \\ \text{training set:} & 1, 2, 3, \ldots, 9 \end{cases}$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Performance indicators

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

$\Downarrow$

For each fold, we compute the testing correctness:

$$\frac{\text{\# points correctly classified in the testing set}}{\text{\# total points in the testing set}}$$

$\Downarrow$

Average testing correctness = accuracy of the classifier.

NOTE: The average testing correctness measures the generalization capability of a classifier, i.e. the capability to correctly classify the new data.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

For each fold, we can compute also the training correctness:

$$\frac{\text{\# points correctly classified in the training set}}{\text{\# number of total points in the training set}}$$

$$\Downarrow$$

Average training correctness: measures the quality of the optimization process in the learning phase.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

OTHER INDICATORS (Testing/Training)

$$\Downarrow$$

$\mathcal{A}$: set of positive points
$\mathcal{B}$: set of negative points

$$\text{Sensitivity} = \frac{\#\text{ points of } \mathcal{A} \text{ correctly classified}}{\#\text{ points of } \mathcal{A}}$$

NOTE: The sensitivity is called also the true positive rate or recall. It measures the proportion of positive points correctly identified.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

OTHER INDICATORS (Testing/Training)

$\Downarrow$

$\mathcal{A}$: set of positive points
$\mathcal{B}$: set of negative points

$$\text{Specificity} = \frac{\#\text{ points of } \mathcal{B} \text{ correctly classified}}{\#\text{ points of } \mathcal{B}}$$

NOTE: The specificity is called also the true negative rate. It measures the proportion of negative points correctly identified.

Evaluation of a classifier

OTHER INDICATORS (Testing/Training)

$\Downarrow$

$\mathcal{A}$: set of positive points
$\mathcal{B}$: set of negative points

$$\text{Precision} = \frac{\#\text{ points of } \mathcal{A} \text{ correctly classified}}{\#\text{ points of } \mathcal{A} \text{ correctly classified} + \#\text{ points of } \mathcal{B} \text{ misclassified}}$$

$$= \frac{\#\text{ points of } \mathcal{A} \text{ correctly classified}}{\#\text{ total points classified as positive}}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Evaluation of a classifier

OTHER INDICATORS (Testing/Training)

$\Downarrow$

$\mathcal{A}$: set of positive points
$\mathcal{B}$: set of negative points

$$\text{F-score or F1-Score} = \frac{2}{\frac{1}{\text{sensitivity}} + \frac{1}{\text{precision}}}$$

$$= 2\,\frac{\text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}}$$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Leave-One-Out

Leave-One-Out

Leave-One-Out

Each time, the testing set is constituted by a single point. The remaining points of the dataset constitute the training set.

Model selection

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

Question: How to compute the suitable values of the parameters C, C_1, C_2, σ, and so on...?

Model selection

Simple case: computing C.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection - 10 fold Cross Validation

1. The data set is randomly split into ten different pieces (tenfold cross-validation - first level).
2. For ten times, each time, nine pieces form the first level training set.
3. The tenth piece forms the first level testing set, which simulates the new unknown data to be classified.
4. Then we have ten training sets and ten corresponding testing sets.
5. We fix a grid of possible values for C (for example 1, 10, 100, 1000).
6. For each first level training set, we perform a fivefold cross-validation - second level, testing each value of C.

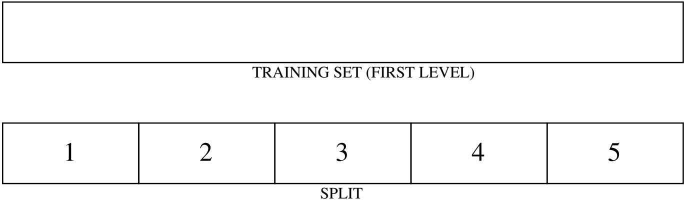
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

To compute the best value of C for the i-th first level fold, we perform a 5 fold cross-validation (second level) on the first level training set

Model selection

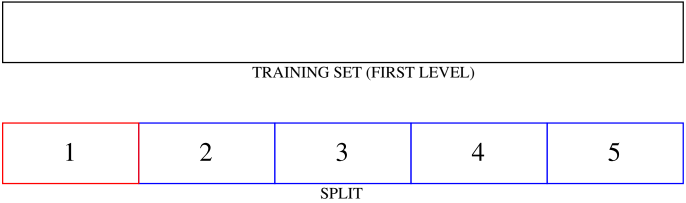
5 fold cross-validation (second level) on fold i



UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

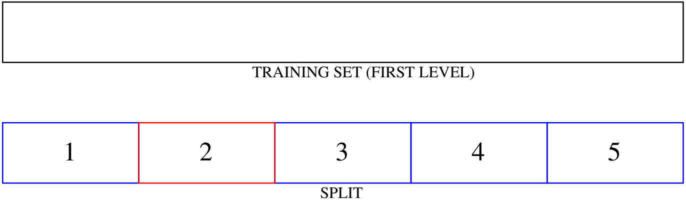
5 fold cross-validation (second level)



FOLD 1 (second level) $\begin{cases} \text{testing set (second level):} & 1 \\ \text{training set (second level):} & 2, 3, 4, 5 \end{cases}$

Model selection

5 fold cross-validation (second level)



FOLD 2 (second level) $\begin{cases} \text{testing set (second level):} & 2 \\ \text{training set (second level):} & 1, 3, 4, 5 \end{cases}$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

...and so on...

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

5 fold cross-validation (second level)

TRAINING SET (FIRST LEVEL)

1	2	3	4	5

SPLIT

FOLD 5 (second level)
- testing set (second level): 5
- training set (second level): $1, 2, 3, 4$

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Model selection

For any single first level fold:

1. For each prefixed value of C in the grid we come out with a second level average testing correctness (average of 5 values).
2. Among the values of C in the grid, we take the best value C^* such that the second level average testing correctness is maximum.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

PART X

REFERENCES

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

References I

- A. Astorino, M. Avolio, A. Canino, T. Crupi, and A. Fuduli, Partitional clustering via successive transportation problems, Operations Research Letters **51** (2023), no. 1, 40–46.
- A. Astorino and A. Fuduli, Semisupervised spherical separation, Applied Mathematical Modelling **39** (2015), no. 20, 6351–6358.
- ______, Support vector machine polyhedral separability in semisupervised learning, Journal of Optimization Theory and Applications **164** (2015), no. 3, 1039–1050.
- A. Astorino, A. Fuduli, and M. Gaudioso, Margin maximization in spherical separation, Computational Optimization and Applications **53** (2012), no. 2, 301–322.

References II

- A. Astorino and M. Gaudioso, A fixed-center spherical separation algorithm with kernel transformations for classification problems, Computational Management Science **6** (2009), no. 3, 357–372.
- S. Andrews, I. Tsochantaridis, and T. Hofmann, Support vector machines for multiple-instance learning, Advances in Neural Information Processing Systems (S. Becker, S. Thrun, and K. Obermayer, eds.), MIT Press, Cambridge, 2003, pp. 561–568.
- A. M. Bagirov and J. Yearwood, A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems, European Journal of Operational Research **170** (2006), no. 2, 578–596.

References III

- O. Chapelle and A. Zien, Semi-supervised classification by low density separation, AISTATS 2005 - Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005, pp. 57–64.
- T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial Intelligence **89** (1997), no. 1-2, 31–71.
- G. Fung and O. L. Mangasarian, Proximal support vector machine classifiers, Proceedings KDD-2001: Knowledge discovery and data mining (F. Provost and R. Srikant, eds.), ACM, 2001, pp. 77–86.
- O. L. Mangasarian, Linear and nonlinear separation of patterns by linear programming, Operations Research **13** (1965), 444–452.

References IV

- N. Megiddo, On the complexity of polyhedral separability, Discrete and Computational Geometry **3** (1988), 325–337.
- D. M. J. Tax and R. P. W. Duin, Data domain description using support vectors, ESANN'1999 proceedings Bruges (Belgium), April 1999, pp. 251–256.
- V. Vapnik, The nature of the statistical learning theory, Springer Verlag, New York, 1995.
- P. Wolfe, A duality theorem for non-linear programming, Quarterly of Applied Mathematics **19** (1961), no. 3, 239–244.

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA