DC models for machine learning problems

Annabella Astorino*

*Istituto di Calcolo e Reti ad Alte Prestazioni - CNR, Italy

Optimization, Game Theory and Data Analysis

University of Vienna, Austria

December 20-21, 2018

Joint work with Antonio Fuduli and Manlio Gaudioso

Outline

1. Supervised Classification Problems
   - Linear separability
   - Polyhedral Separability
   - Spherical Separability
2. Semi-supervised classification
   - Nonsmooth optimization for TSVM
   - Other semi-supervised approaches
3. Multiple Instance Learning
   - A DC formulation

Pattern classification

Pattern classification consists in categorizing samples into different classes, on the basis of their similarities.

- The samples are characterized by some features;
- for each sample, we want to express a particular feature, the **class label**, as a function of the remaining ones; this is done by constructing the so-called **prediction function**;
- by means of the prediction function we would like **to predict the class of each sample**, minimizing some classification error.

Classification problems

- **Supervised classification**: on the basis of the labelled samples, we would like to predict the class of any new future sample.
- **Unsupervised classification**: we have only unlabelled samples and we would like to cluster the data.
- **Semisupervised classification**: on the basis of the labelled and unlabelled samples, we would like to predict the class of the unlabelled samples.
- **Multiple instance learning**: we learn a classifier on the basis of a training set of bags. Each bag is made up of several samples, called instances. We know the label of each bag, but the label of each instance inside the bags is unknown and has to be assigned by the classification algorithm.

Outline

Binary supervised classification

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

- Two disjoint, finite point sets

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

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

- **Mathematical Programming Approach**: To find a separating surface minimizing some classification error.

Outline

1. Supervised Classification Problems
   - Linear separability

Linear separability

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

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

such that

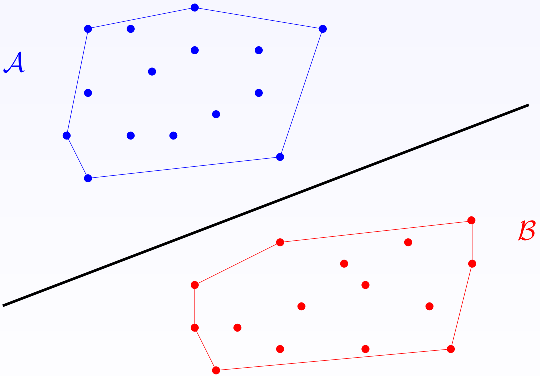
- $a_i^T w \leq \gamma - 1, \quad i = 1, \ldots, m$
- $b_l^T w \geq \gamma + 1, \quad l = 1, \ldots, k.$

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

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

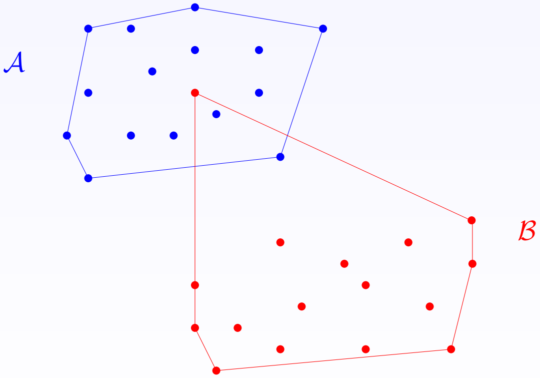
Linear separability condition 1/2

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

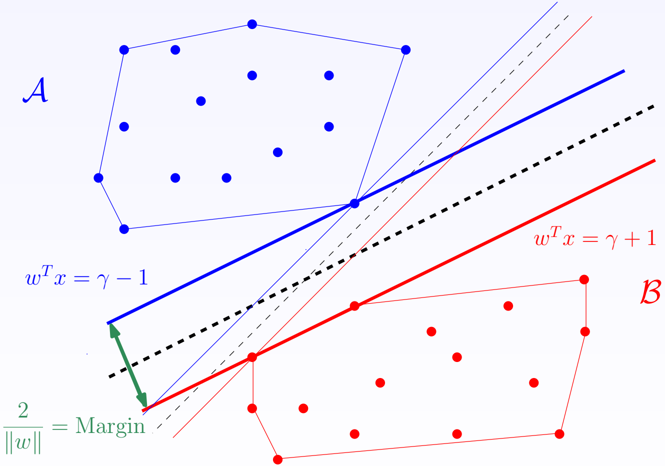


Linear separability condition 2/2

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



SVM: maximizing the margin



SVM: maximizing the margin

Idea:
The SVM technique constructs a classifier by generating a separation hyperplane far away from the objects of the two classes (the optimal separation hyperplane).

Approach:

$$\min_{w,\gamma} \ \frac{1}{2}\|w\|^2 + C\left[\sum_{i=1}^{m} \max\{0, a_i^T w - \gamma + 1\} + \sum_{l=1}^{k} \max\{0, -b_l^T w + \gamma + 1\}\right]$$

Outline

1. Supervised Classification Problems
   - Polyhedral Separability

h-Polyhedral separability definition

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

$$\{(w^{(j)}, \gamma_j)\}, \text{ with } w^{(j)} \in \mathbb{R}^n \text{ and } \gamma_j \in \mathbb{R}, \ j = 1, \ldots, h,$$

such that

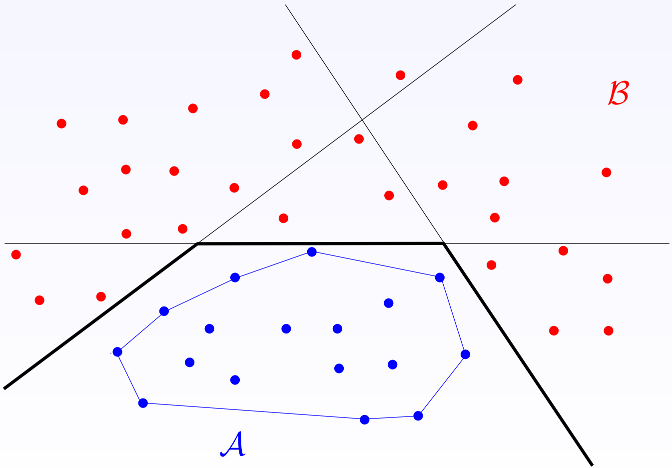
- $a_i^T w^{(j)} \leq \gamma_j - 1$, for all $i = 1, \ldots, m$ and for all $j = 1, \ldots, h$;
- for all $l = 1, \ldots, k$, there exists an index $j \in \{1, \ldots, h\}$ such that $b_l^T w^{(j)} \geq \gamma_j + 1$.

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

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

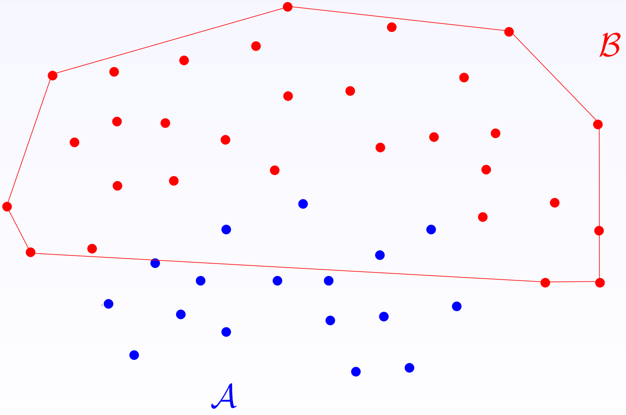
Polyhedral separability

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



Polyhedral separability

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



A mathematical programming model

Minimizing the following error function (h fixed):

Error function

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

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

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

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

DC model

$$z(\mathbf{w},\gamma) = z_1(\mathbf{w},\gamma) - z_2(\mathbf{w},\gamma),$$

Convex $z_1(\mathbf{w},\gamma) :=$

$$\frac{1}{m}\sum_{i=1}^{m}\left\{\max_{1\le j\le h}[\max(0, a_i^T w^{(j)} - \gamma_j + 1)]\right\} +$$
$$\frac{1}{k}\sum_{l=1}^{k}\left\{\max[0, \max_{1\le j\le h}(b_l^T w^{(j)} - \gamma_j - 1)]\right\}$$

Convex $z_2(\mathbf{w},\gamma) :=$

$$\frac{1}{k}\sum_{l=1}^{k}\left\{\max_{1\le j\le h}(b_l^T w^{(j)} - \gamma_j - 1)\right\}$$

Some references on polyhedral separability

- A. Astorino and M. Gaudioso, *Polyhedral separability through successive LP*, Journal of Optimization Theory and Applications, 112 (2002), pp. 265–293.
- R. N. Gasimov and G. Ozturk, *Separation via polyhedral conic functions*, Optimization Methods and Software, 21 (2006), pp. 527–540.
- C. Orsenigo and C. Vercellis, *Accurately learning from few examples with a polyhedral classifier*, Computational Optimization and Applications, 38 (2007), pp. 235–247.

Outline

1. Supervised Classification Problems
   - Spherical Separability

Spherical separability definition

The set $\mathcal{A}$ is **spherically separable** (not strictly!) from set $\mathcal{B}$ if and only if there exists a sphere

$$S(x_0, R) := \{x \in \mathbb{R}^n | (x - x_0)^T (x - x_0) \leq R^2\},$$

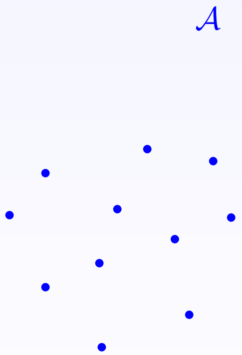
with center $x_0 \in \mathbb{R}^n$ and radius R, enclosing all points of $\mathcal{A}$ and no points of $\mathcal{B}$, i.e. such that

- $\|a_i - x_0\|^2 \leq R^2$ for all points $a_i \in \mathcal{A}$ $(i = 1, ..., m)$
- $\|b_l - x_0\|^2 \geq R^2$ for all points $b_l \in \mathcal{B}$ $(l = 1, ..., k)$.

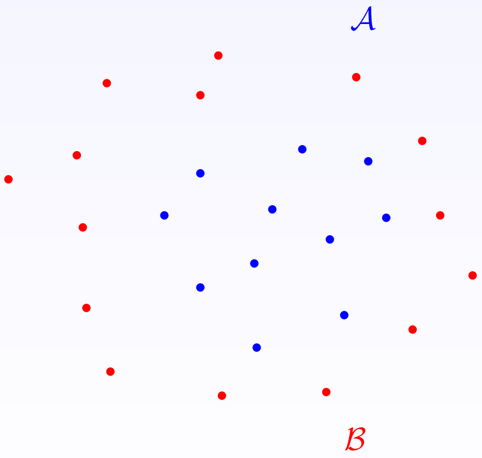
Remark:

- The role of the two sets $\mathcal{A}$ and $\mathcal{B}$ is not symmetric.
- $\mathrm{conv}(\mathcal{A}) \cap \mathcal{B} = \emptyset$ is necessary (but not sufficient) condition for the existence.

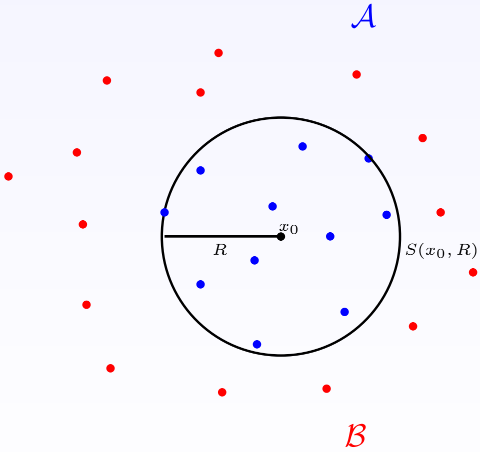
Spherical separability: first example



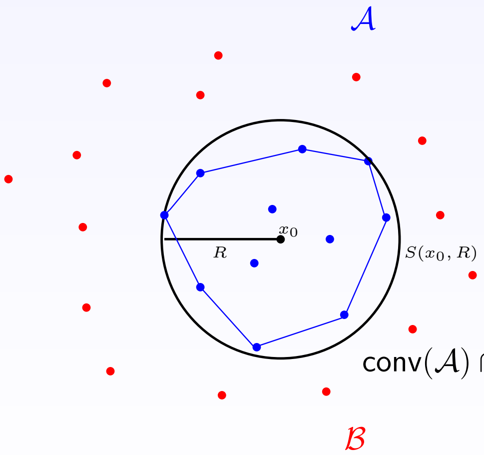
Spherical separability: first example



Spherical separability: first example

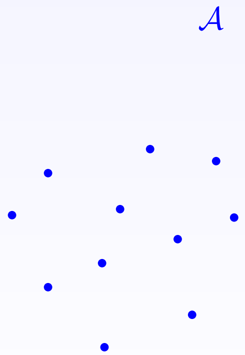


Spherical separability: first example

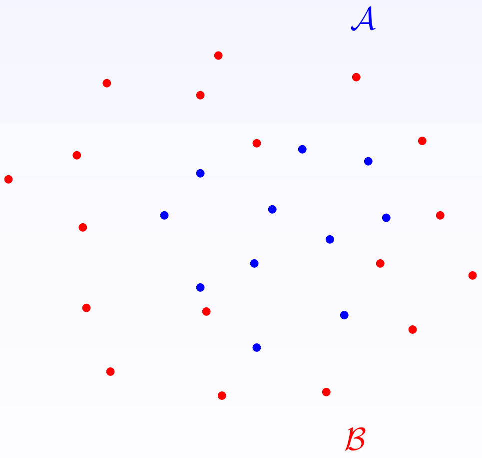


$\text{conv}(\mathcal{A}) \cap \mathcal{B} = \emptyset$

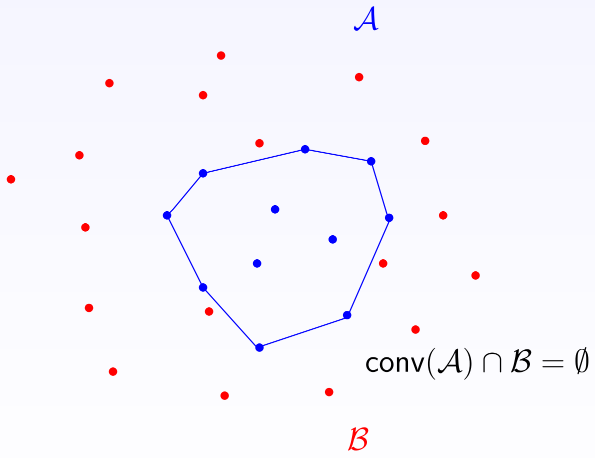
Spherical separability: second example



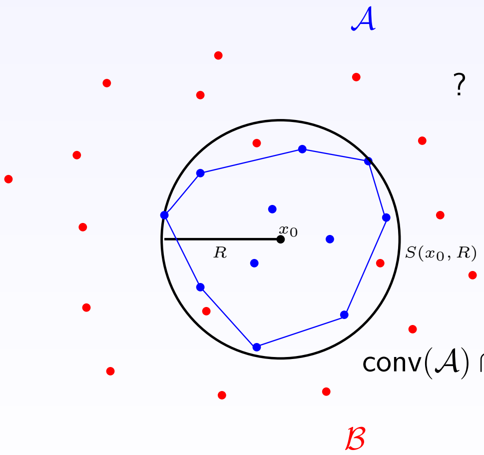
Spherical separability: second example



Spherical separability: second example



Spherical separability: second example



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

The optimization problem

Looking for a "*small*" separating sphere ⇒ Minimizing the following error function:

Error function

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

where $C > 0$ controls the tradeoff between the two objectives (the radius and the classification error).

- Nonsmooth and nonconvex optmization problem.

The fixed center case

Assume the center is fixed (take e.g. the barycentre of $\mathcal{A}$)
Define $z \triangleq R^2, \quad z \geq 0.$

The problem becomes:

$$\min_{z \geq 0} z + C \left(\sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - z\} + \sum_{l=1}^{k} \max\{0, z - \|b_l - x_0\|^2\} \right)$$

- A univariate, convex, nonsmooth optimization problem.
- Transform in a structured LP solvable in $O(p \log p)$ time ($p \triangleq \max\{m, k\}$).

The moving center case: a DC decomposition

$$h(x_0, R) = h_1(x_0, R) - h_2(x_0, R)$$

Convex $h_1(x_0, R) :=$

$$R^2 + C \sum_{i=1}^{m} \max\{R^2, \|a_i - x_0\|^2\} + C \sum_{l=1}^{k} \max\{R^2, \|b_l - x_0\|^2\}$$

Convex $h_2(x_0, R) :=$

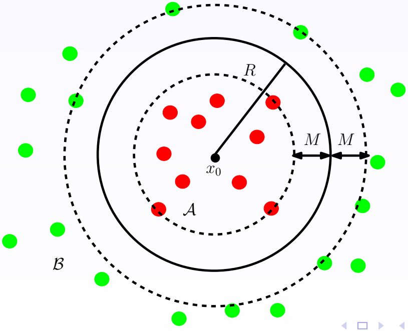
$$CmR^2 + C \sum_{l=1}^{k} \|b_l - x_0\|^2$$

Margin spherical separation

$\mathcal{A}$ and $\mathcal{B}$ are "*strictly*" spherically separated by $S(x_0, R)$ if and only if

- $\|a_i - x_0\|^2 \leq (R - M)^2, \quad i = 1, \dots, m$
- $\|b_l - x_0\|^2 \geq (R + M)^2, \quad l = 1, \dots, k$

for some "*margin*" M, $0 < M \leq R$.



The optimization problem

By setting $z \triangleq R^2 + M^2$ and $s \triangleq 2RM$, solve the following problem:

$$\min_{x_0, z, 0 \le s \le z} h(x_0, z, s)$$

Error function

$$\begin{aligned} h(x_0, z, s) := \quad & C\left(\sum_{i=1}^{m} \max\{0, s - z + \|a_i - x_0\|^2\}\right) + \\ & C\left(\sum_{l=1}^{k} \max\{0, s + z - \|b_l - x_0\|^2\}\right) - s \end{aligned}$$

where $-s$ represents the margin maximization objective and $C > 0$ is the tradeoff parameter.

- Nonsmooth and nonconvex optmization problem.

The fixed center case

Assume the center is fixed (take e.g. the barycentre of $\mathcal{A}$)

The problem becomes:

$$\min_{0 \le s \le z} C \left(\sum_{i=1}^{m} \max\{0, \|a_i - x_0\|^2 - z + s\} + \sum_{l=1}^{k} \max\{0, z + s - \|b_l - x_0\|^2\} \right) - s$$

- A convex, nonsmooth optimization problem.
- Transform in a structured LP solvable in $O(p \log p)$ ($p \triangleq \max\{m, k\}$) time.

The moving center case

$$h(x_0, z, s) = h_1(x_0, z, s) - h_2(x_0, z, s)$$

Convex $h_1(x_0, z, s) :=$

$$C \sum_{i=1}^{m} \max\{0, s - z + \|a_i - x_0\|^2\} + C \sum_{l=1}^{k} \max\{s + z, \|b_l - x_0\|^2\} - s$$

Convex $h_2(x_0, z, s) :=$

$$C \sum_{l=1}^{k} \|b_l - x_0\|^2$$

Some references on spherical separability

- A. Astorino and M. Gaudioso, *A fixed-center spherical separation algorithm with kernel transformations for classification problems*. Computational Management Science 6 (3), (2009), pp. 357–372.
- A. Astorino, A. Fuduli and M. Gaudioso, *DC models for spherical separation*. Journal of Global Optimization 48 (4), (2010), pp. 657–669.
- A. Astorino, A. Fuduli and M. Gaudioso, *Margin maximization in spherical separation*. Computational Optimization and Applications 53 (2), (2012) pp. 301–322.
- A. Astorino, I. Bomze, P. Brito and M. Gaudioso, *Two spherical separation procedures via non-smooth convex optimization*. In: Simone, V. D., Serafino, D. D., Toraldo, G. (Eds.), Recent Advances in Nonlinear Optimization and Equilibrium Problems: a Tribute to Marco D'Apuzzo. Vol. 27 of Q. di Matematica. Aracne, (2012), pp. 1–16.
- A. Astorino, I. Bomze, A. Fuduli and M. Gaudioso, *Robust spherical separation*. Optimization 66 (6), (2017), pp. 925–938.

Some references on spherical separability

A. Astorino, M. Gaudioso and W. Khalaf, *Edge detection by spherical separation*. Computational Management Science 11 (4), (2014), pp. 517–530.

W. Khalaf, A. Astorino, P. D'Alessandro and M. Gaudioso, *A DC optimization-based clustering technique for edge detection*. Optimization Letters 11 (3), (2017), pp. 627–640.

A. Astorino, A. Chiarello, M. Gaudioso and A. Piccolo, *Malicious URL detection via spherical classification*. Neural Computing and Applications 28, (2017), pp. 699–705.

Outline

Semi-supervised Classification

- There are many real–world problems where labelling often requires fairly expensive human labor, whereas unlabelled data are abundant being easier to obtain
  - Medical diagnosis, web categorization, text processing...
- **semi-supervised classification**: the prediction function is constructed on the basis of the labelled and unlabelled samples.

Transductive inference

The prediction function is derived from the information concerning all the available data, i.e. both the labelled and unlabelled samples.

This function is not aimed at predicting the class label for newly incoming samples, but only at making a decision about the currently available objects.

Outline

2. Semi-supervised classification
   - Nonsmooth optimization for TSVM

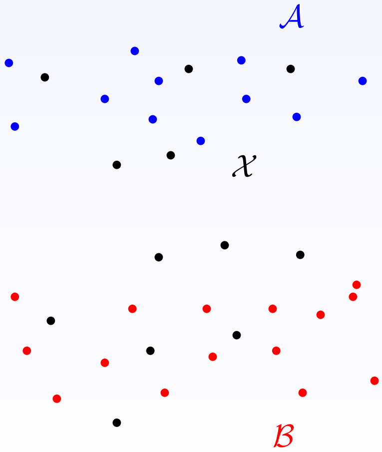
The TSVM idea

- The TSVM (Transductive Support Vector Machine) technique is the semi-supervised version of the SVM approach;
- the objective of the TSVM approach is to compute the best support vector machine, on the basis of the labelled and unlabelled points by minimizing the number of similar samples with different labels.

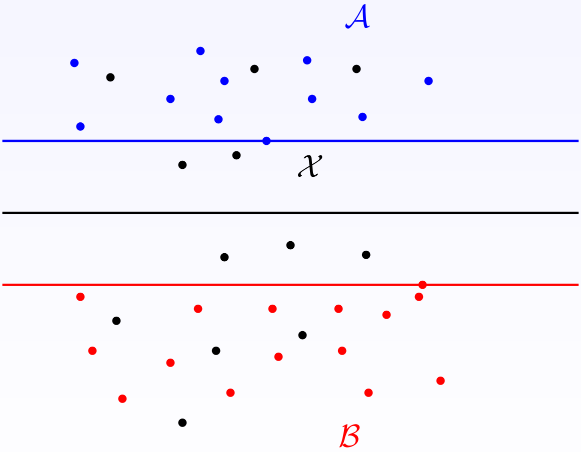
$$\Downarrow$$

Minimizing the number of unlabeled points in the margin area.

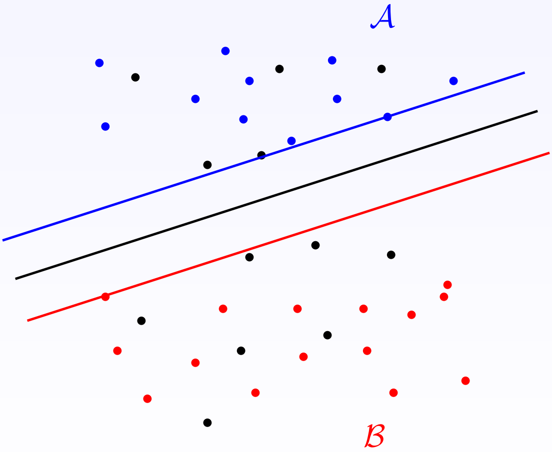
TSVM: an example



TSVM: an example



TSVM: an example



A mathematical programming TSVM model

- We have a set (training set) of $m+k$ labelled points: $a_i \in \mathcal{A}$ $(i = 1, ..., m)$ and $b_l \in \mathcal{B}$ $(l = 1, ..., k)$;
- we have a set (testing set) of q unlabelled points in $\mathbb{R}^n$: $\mathcal{X} = \{x_1, x_2, ..., x_q\}$.

TSVM approach:

$$\min_{w,\gamma} \quad \frac{1}{2}\|w\|^2 + C_1 \left[\sum_{i=1}^{m} \max\{0, a_i^T w - \gamma + 1\} + \sum_{l=1}^{k} \max\{0, -b_l^T w + \gamma + 1\} \right] + C_2 \sum_{t=1}^{q} \max\{0, 1 - |w^T x_t - \gamma|\}$$

Properties of objective function

$$\begin{aligned} h(w,\gamma) := \quad & \frac{1}{2}\|w\|^2 + \\ & C_1\left[\sum_{i=1}^{m} \max\{0, a_i^T w - \gamma + 1\} + \sum_{l=1}^{k} \max\{0, -b_l^T w + \gamma + 1\}\right] + \\ & C_2 \sum_{t=1}^{q} \max\{0, 1 - |w^T x_t - \gamma|\} \end{aligned}$$

- h is nondifferentiable;
- h is nonconvex, due to the last term involving the unlabelled points;
- h can be written as a DC function.

h as DC function

$$h(w,\gamma) = h_1(w,\gamma) - h_2(w,\gamma)$$

Convex $h_1(w,\gamma) :=$

$$\frac{1}{2}\|w\|^2 +$$
$$C_1 \left[\sum_{i=1}^{m} \max\{0, a_i^T w - \gamma + 1\} + \sum_{l=1}^{k} \max\{0, -b_l^T w + \gamma + 1\} \right] +$$
$$C_2 \sum_{t=1}^{q} \max\{1, |w^T x_t - \gamma|\}$$

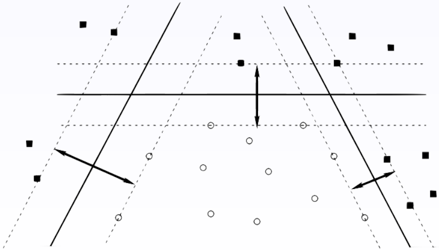
Convex $h_2(w,\gamma) :=$

$$C_2 \sum_{t=1}^{q} |w^T x_t - \gamma|$$

Outline

The TSVM polyhedral separation

$$
\begin{aligned}
z(\mathbf{w},\gamma) \triangleq \quad & \frac{1}{2}\sum_{j=1}^{h}\|w^{(j)}\|^2 + \\
& C_1 \sum_{i=1}^{m} \max_{1\le j\le h}\{\max(0, a_i^T w^{(j)} - \gamma_j + 1)\} + \\
& C_1 \sum_{l=1}^{k} \max\{0, \min_{1\le j\le h}[-b_l^T w^{(j)} + \gamma_j + 1]\} + \\
& C_2 \sum_{j=1}^{h}\sum_{t=1}^{q} \max\{0, 1 - |w^{(j)T} x_t - \gamma_j|\}.
\end{aligned}
$$



DC model

$$z(\mathbf{w},\gamma) = z_1(\mathbf{w},\gamma) - z_2(\mathbf{w},\gamma),$$

Convex $z_1(\mathbf{w},\gamma) :=$

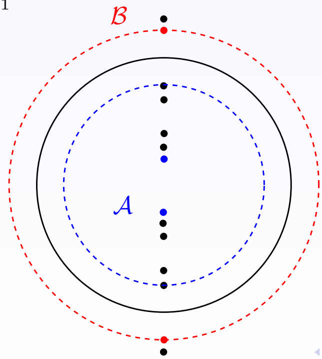
$$C_1 \sum_{i=1}^{m} \left\{ \max_{1\le j\le h} \max(0, a_i^T w^{(j)} - \gamma_j + 1) \right\} +$$
$$C_1 \sum_{l=1}^{k} \left\{ \max[0, \max_{1\le j\le h} (b_l^T w^{(j)} - \gamma_j - 1)] \right\} +$$
$$C_2 \sum_{j=1}^{h} \sum_{t=1}^{q} \max\{1, |{w^{(j)}}^T x_t - \gamma_j|\}$$

Convex $z_2(\mathbf{w},\gamma) :=$

$$C_1 \sum_{l=1}^{k} \left\{ \max_{1\le j\le h} (b_l^T w^{(j)} - \gamma_j - 1) \right\} + C_2 \sum_{j=1}^{h} \sum_{t=1}^{q} |{w^{(j)}}^T x_t - \gamma_j|$$

Semisupervised spherical separation

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



DC model

$$h(x_0, z, s) = h_1(x_0, z, s) - h_2(x_0, z, s)$$

Convex $h_1(x_0, z, s) :=$

$$C_1 \left(\sum_{i=1}^{m} \max\{0, s - z + \|a_i - x_0\|^2\} \right) +$$
$$C_1 \left(\sum_{l=1}^{k} \max\{s + z, \|b_l - x_0\|^2\} \right) +$$
$$C_2 \sum_{t=1}^{q} \max\{\|x_t - x_0\|^2 - z + s, \max[0, 2(\|x_t - x_0\|^2 - z)]\} - s$$

Convex $h_2(x_0, z, s) :=$

$$C_1 \sum_{l=1}^{k} \|b_l - x_0\|^2 + C_2 \sum_{t=1}^{q} \max\{0, 2(\|x_t - x_0\|^2 - z)\}$$

Some references on semi-supervised classification

- A. Astorino and A. Fuduli, *Nonsmooth optimization techniques for semisupervised classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, pp. 2135–2142, 2007.
- A. Astorino,E. Gorgone, M. Gaudioso and D. Pallaschke, *Data preprocessing in semi-supervised SVM classification*, Optimization, 60(1-2), pp. 143–151, 2011.
- A. Astorino and A. Fuduli, *Support Vector Machine Polyhedral Separability in Semisupervised Learning*, Journal of Optimization Theory and Applications, 164(3), pp. 1039–1050, 2015.
- A. Astorino and A. Fuduli, *Semisupervised spherical separation*, Applied Mathematical Modelling, 39(20), pp. 6351–6358, 2015.

Outline

A MIL problem

Multiple instance learning:

- We learn a classifier on the basis of a training set of bags.
- Each bag is made up of several samples, called instances.
- We know the label of each bag, but the label of each instance inside the bags is unknown and has to be assigned by the classification algorithm.

Example: Musk drug activity prediction

- A molecule (bag) has the desired drug effect (positive label) if and only if one or more of its conformations (instances) bind to the target site;
- we do not know a priori which one, so we cannot label the instances individually, and we have an overall label for the whole molecule.

Binary MIL problem: notations

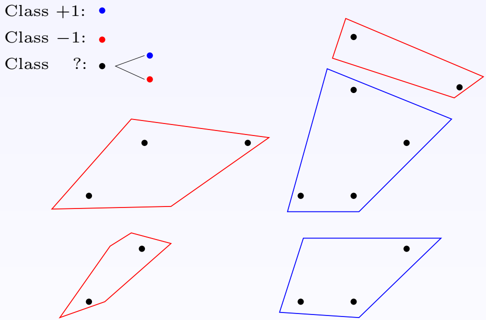
- We have p points, called "instances": $x_j \in \mathbb{R}^n$, $j = 1, \dots, p$.
- The class label of each instance $x_j \in \mathbb{R}^n$ is indicated by $y_j \in \{-1, 1\}$ and it is unknown.
- the instances are partitioned in $m + k$ bags, where m is the number of positive bags and k is the number of negative bags.
- The class label of each bag i is known and it is indicated by $Y_i \in \{-1, 1\}$, $i = 1, \dots, m + k$.
- The $i - th$ positive bag is made up by the instances identified by the index set J_i^+.
- The $i - th$ negative bag is made up by the instances identified by the index set J_i^-.

Binary MIL problem: assumptions

The objective is to find a hyperplane separating the instances according to $y_j \in \{-1, 1\}$. Correct separation is achieved whenever:

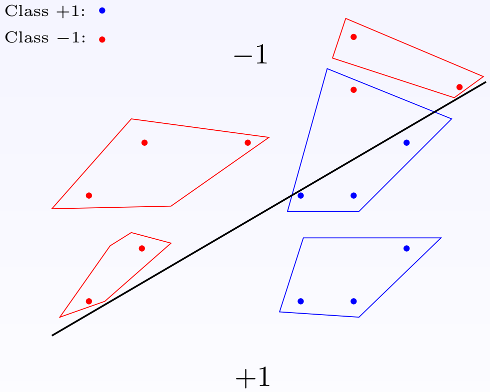
- At least one instance of each positive bag i is assigned $y_j = +1$.
- All instances of each negative bag are assigned $y_j = -1$.

MIL separation



- We want to separate the positive bags (in blue) from the negative ones (in red).
- The class label of each instance (in black) is unknown.

MIL separation



- A bag is classified positive if it contains at least an instance on the $+1$ side (in blue).
- A bag is classified negative if it contains all the instances on the -1 side (in red).

Outline

3. Multiple Instance Learning
   - A DC formulation

The optimization model (Bergeron et al., 2012)

The nonconvex nonsmooth unconstrained optimization problem

$$\min_{w,\gamma} f(w,\gamma)$$

The error function

$$f(w,\gamma) \triangleq \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}\sum_{j\in J_i^-}\max\{0, 1+(w^Tx_j-\gamma)\}$$
$$+C\sum_{i=1}^{m}\max\{0, 1-\max_{j\in J_i^+}(w^Tx_j-\gamma)\}$$

A DC decomposition

$$f(w,\gamma) = f_1(w,\gamma) - f_2(w,\gamma)$$

Convex $f_1(w,\gamma) :=$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}\sum_{j\in J_i^-}\max\{0, 1+(w^T x_j - \gamma)\}$$
$$+C\sum_{i=1}^{m}\max\{1, \max_{j\in J_i^+}(w^T x_j - \gamma)\}$$

Convex $f_2(w,\gamma) :=$

$$C\sum_{i=1}^{m}\max_{j\in J_i^+}(w^T x_j - \gamma)$$

Some references on MIL classification

A. Astorino, A. Fuduli, P. Veltri and E. Vocaturo, *On a recent algorithm for multiple instance learning. Preliminary applications in image classification*, Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, pp. 1615–1619, 2017.

A. Astorino, A. Fuduli, P. Veltri and E. Vocaturo, *A multiple instance learning algorithm for color images classification*, ACM International Conference Proceeding Series, pp. 262–266, 2018.

A. Astorino, A. Fuduli and M. Gaudioso, *A Lagrangian relaxation approach for binary Multiple Instance Classification*, IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2018.2885852.