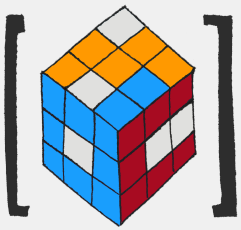
Data Science problems and complex networks optimization ... in a nutshell

Justo Puerto

Universidad de Sevilla

(joint work with B. Blanco, A. Japón and A.M. Rodríguez-Chía)



Desafíos de la Matemática Combinatoria
FQM-5849

GDO 2019. Cluj-Napoca.

Contents

ON ℓ_p-SUPPORT VECTOR MACHINES AND MULTIDIMENSIONAL KERNELS

VÍCTOR BLANCO†, JUSTO PUERTO‡, AND ANTONIO M. RODRÍGUEZ-CHÍA⋆

†Dpt. Quantitative Methods for Economics & Business, Universidad de Granada
E-mail address: vblanco@ugr.es

‡Dpt. Statistics & OR, Universidad de Sevilla
E-mail address: puerto@us.es

⋆Dpt. Statistics & OR, Universidad de Cádiz
E-mail address: antonio.rodriguezchia@uca.es

Abstract. In this paper, we extend the methodology developed for Support Vector Machines (SVM) using ℓ_2-norm (ℓ_2-SVM) to the more general case of ℓ_p-norms with $p \geq 1$ (ℓ_p-SVM). The resulting primal and dual problems are formulated as mathematical programming problems; namely, in the primal case, as a second order cone optimization problem and in the dual case, as a polynomial optimization problem involving homogeneous polynomials. Scalability of the primal problem is obtained via general transformations based on the expansion of functionals in Schauder spaces. The concept of Kernel function, widely applied in ℓ_2-SVM, is extended to the more general case by defining a new operator called multidimensional Kernel. This object gives rise to reformulations of dual problems, in a transformed space of the original data, which are solved by a moment-sdp based approach. The results of some computational experiments on real-world datasets are presented showing rather good behavior in terms of standard indicators such a *accuracy index* and its ability to classify new data.

1. Introduction

In supervised classification, given a finite set of objects partitioned into classes, the goal is to build a mechanism, based on current available information, for classifying new objects into these classes. Due to their successful applications in the last decades, as for instance in writing recognition [1], insurance companies (to determine whether an applicant is a high insurance risk or not) [19], banks (to decide whether an applicant is a good credit risk or not) [15], medicine (to determine whether a tumor is benigne or maligne) [32, 26], etc; support vector machines (SVMs) have become a popular methodology for supervised classification [4].

Support vector machine (SVM) is a mathematical programming tool, originally developed by Vapnik [35, 36] and Cortes and Vapnik [11], which consists in finding a hyperplane to separate a set of data into two classes, so that the distance from the hyperplane to the nearest point of each class is maximized. In order to do that, the standard SVM solves an optimization problem that accounts for both

2010 *Mathematics Subject Classification.* 62H30, 90C26, 53A45, 15A60.

Optimal arrangements of hyperplanes for multiclass classification

Víctor Blanco[†], Alberto Japón[‡] and Justo Puerto[‡]

[†]IEMath-GR, Universidad de Granada
[‡]IMUS, Universidad de Sevilla

Abstract. In this paper, we present a novel approach to construct multiclass clasifiers by means of arrangements of hyperplanes. We propose different mixed integer non linear programming formulations for the problem by using extensions of widely used measures for misclassifying observations. We prove that kernel tools can be extended to these models. Some strategies are detailed that help solving the associated mathematical programming problems more efficiently. An extensive battery of experiments has been run which reveal the powerfulness of our proposal in contrast to other previously proposed methods.

1. Introduction

Support Vector Machine (SVM) is a widely-used methodology in supervised binary classification, firstly proposed by Cortes and Vapnik [6]. Given a set of data together with a label, the general idea under the SVM methodologies is to find a partition of the feature space and an assignment rule from data to each of the cells in the partition that maximizes the separation between the classes of a training sample and that minimizes certain measure for the misclassifying errors. At that point, convex optimization tools come into scene and the shape of the obtained dual problem allows one to project the data out onto a higher dimensional space where the separation of the classes can be more adequately performed, but whose problem can be solved with the same computational effort that the original one. This fact is the so-called *kernel trick*, and has motivated the use of this tool with success in a wide range of applications [2, 14, 10, 20, 25].

Most of the SVM proposals and extensions concern instances with only two different classes. Some extensions have been proposed for this case by means of choosing different measures for the separation between classes [12, 13, 5], incorporating feature selection tasks [19], regularization strategies [18], etc. However, the analysis of SVM-based methods for instances with more than two classes has been, from our point of view, only partially investigated. To construct a k-label classification rule for $k \geq 2$, one is provided with a *training sample* of observations $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ and labels for each of the observations in such a sample, $(y_1, \ldots, y_n) \in \{1, \ldots, k\}$. The goal is to find a decision rule which is

Computers and Operations Research 95 (2018) 172–193

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

Locating hyperplanes to fitting set of points: A general framework

Víctor Blanco[a,*], Justo Puerto[b], Román Salmerón[c]

[a] Dpt. Quant. Methods for Economics & Business and IEMath-GR, Universidad de Granada, Spain
[b] Dpt. Statistics & OR and IMUS, Universidad de Sevilla, Spain
[c] Dpt. Quant. Methods for Economics & Business, Universidad de Granada, Spain

ARTICLE INFO

Article history:
Received 17 November 2016
Revised 21 March 2018
Accepted 22 March 2018
Available online 23 March 2018

2010 MSC:
90B85
90C26
52C35
65D10

Keywords:
Fitting hyperplanes
Mathematical programming
Location of structures
Robust fitting
Linear regression

ABSTRACT

This paper presents a family of methods for locating/fitting hyperplanes with respect to a given set of points. We introduce a general framework for a family of aggregation criteria, based on ordered weighted operators, of different distance-based errors. The most popular methods found in the specialized literature, namely least sum of squares, least absolute deviation, least quantile of squares or least trimmed sum of squares among many others, can be cast within this family as particular choices of the errors and the aggregation criteria. Unified mathematical programming formulations for these methods are provided and some interesting cases are analyzed. The most general setting give rise to mixed integer nonlinear programming problems. For those situations we present inner and outer linear approximations to assess tractable solution procedures. It is also proposed a new goodness of fitting index which extends the classical coefficient of determination and allows one to compare different fitting hyperplanes. A series of illustrative examples and extensive computational experiments implemented in R are provided to show the applicability of the proposed methods.

1. Introduction

The problem of locating hyperplanes with respect to a given set of point is well-known in Location Analysis (LA) Schöbel (1999). This problem is closely related to another common question in Data Analysis (DA): to study the behavior of a given set of data $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)$ that minimizes some measure of the deviation of the data with respect to the hyperplane it induces, $\mathcal{H}(\hat{\beta}) = \{z \in \mathbb{R}^d : \hat{\beta}_0 + \sum_{k=1}^{d} \hat{\beta}_k z_k = 0\}$. For a given point $x \in \mathbb{R}^d$, we define the *residual* with respect to a generic x as a mapping $\varepsilon_x : \mathbb{R}^{d+1} \to \mathbb{R}_+$, that maps any set of coefficients $\beta = (\beta_0, \ldots, \beta_d) \in \mathbb{R}^{d+1}$, into a measure $\varepsilon_x(\beta)$ that represents the deviation of the given point x

European Journal of Operational Research 000 (2017) 1–11

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Discrete Optimization

Clustering data that are graph connected

Stefano Benati[a], Justo Puerto[b], Antonio M. Rodríguez-Chía[c,*]

[a] Dipartimento di Sociologia e Ricerca Sociale, Università di Trento, Via Verdi 26, 38122 Trento. Italy
[b] IMUS. Universidad de Sevilla, Avda. Reina Mercedes, s/n, 41012 Sevilla. Spain
[c] Faculty of Sciences, Universidad de Cádiz, Avda. República Saharaui, 11510 Puerto Real (Cádiz). Spain

ARTICLE INFO

Article history:
Received 18 June 2016
Accepted 7 February 2017
Available online xxx

Keywords:
Combinatorial optimization
Clustering
Clique partitioning
Integer programming

ABSTRACT

A new combinatorial model for clustering is proposed for all applications in which individual and relational data are available. Individual data refer to the intrinsic features of units, they are stored in a matrix D, and are the typical input of all clustering algorithms proposed so far. Relational data refer to the observed links between units, representing social ties such as friendship, joint participation to social events, and so on. Relational data are stored in the graph $G = (V, E)$, and the data available for clustering are the triplet $G = (V, E, D)$, called attributed graph. Known clustering algorithms can take advantage of the relational structure of G to redefine and refine the units membership. For example, uncertain membership of units to groups can be resolved using the sociological principle that ties are more likely to form between similar units. The model proposed here shows how to take into account the graph information, combining the clique partitioning objective function (a known clustering methodology) with connectivity as the structural constraint of the resulting clusters. The model can be formulated and solved using Integer Linear Programming and a new family of cutting planes. Moderate size problems are solved, and heuristic procedures are developed for instances in which the optimal solution can only be approximated. Finally, tests conducted on simulated data show that the clusters quality is greatly improved through this methodology.

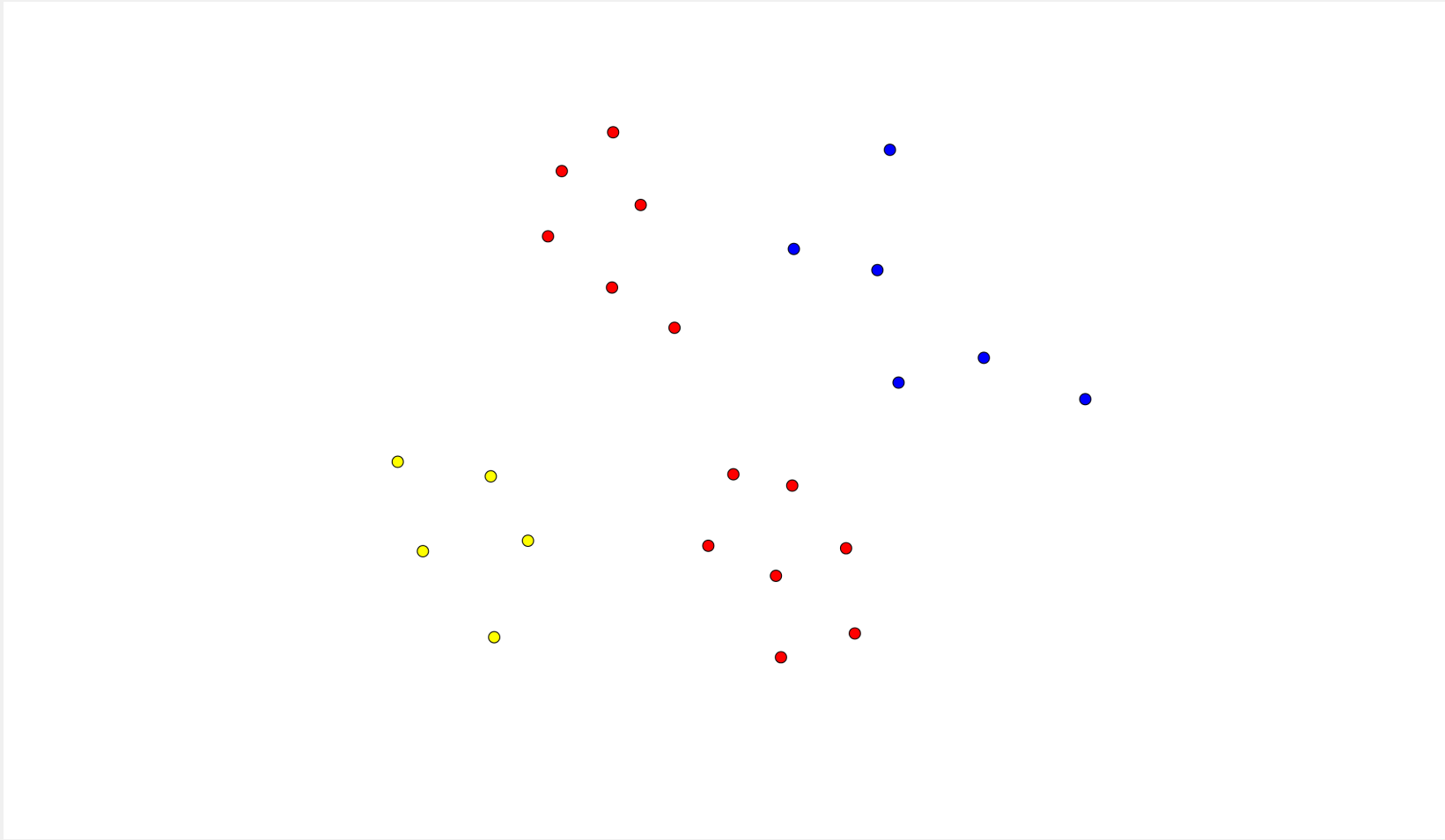
Data

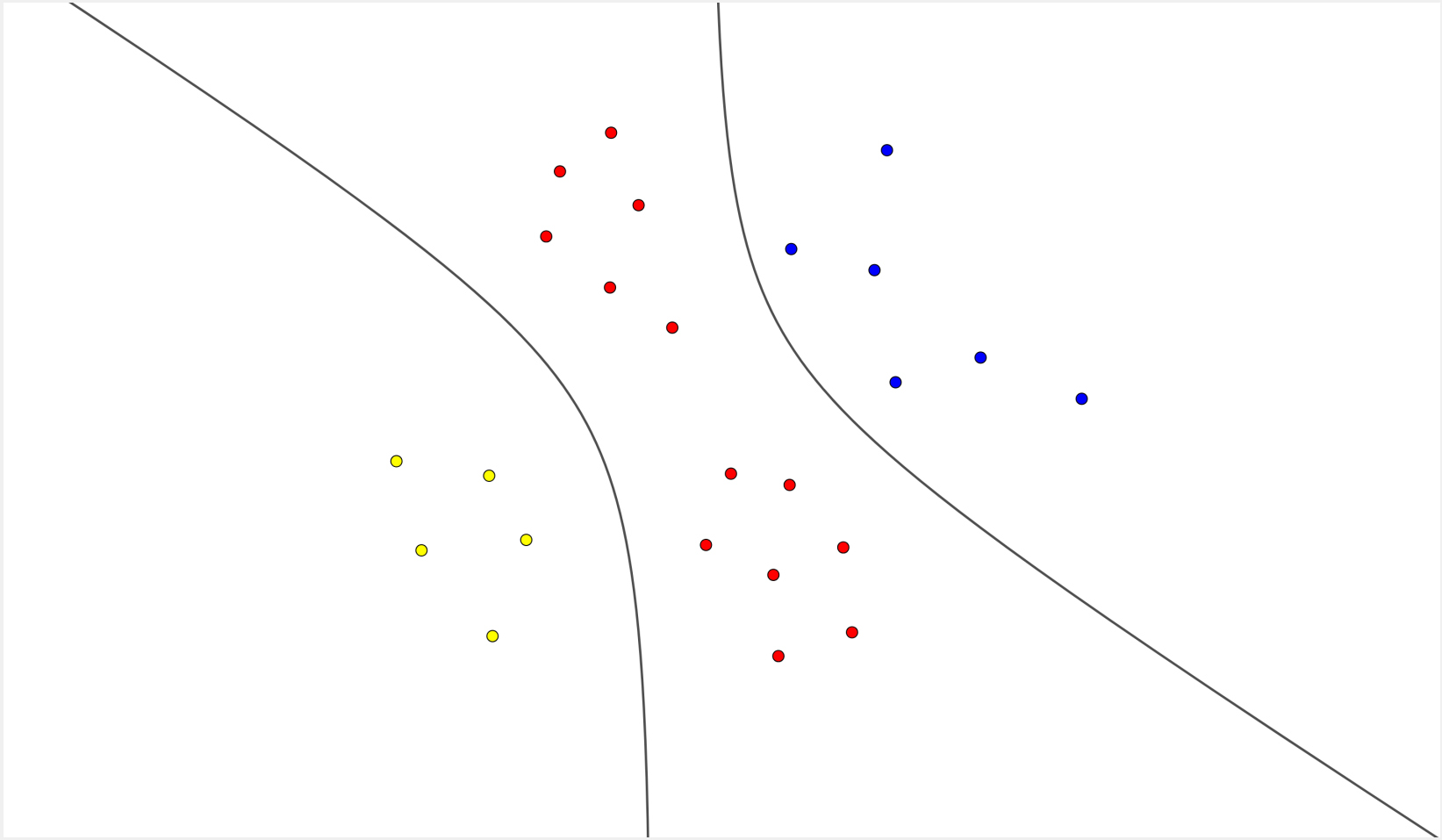
- Random sample of size n
- d predictor variables

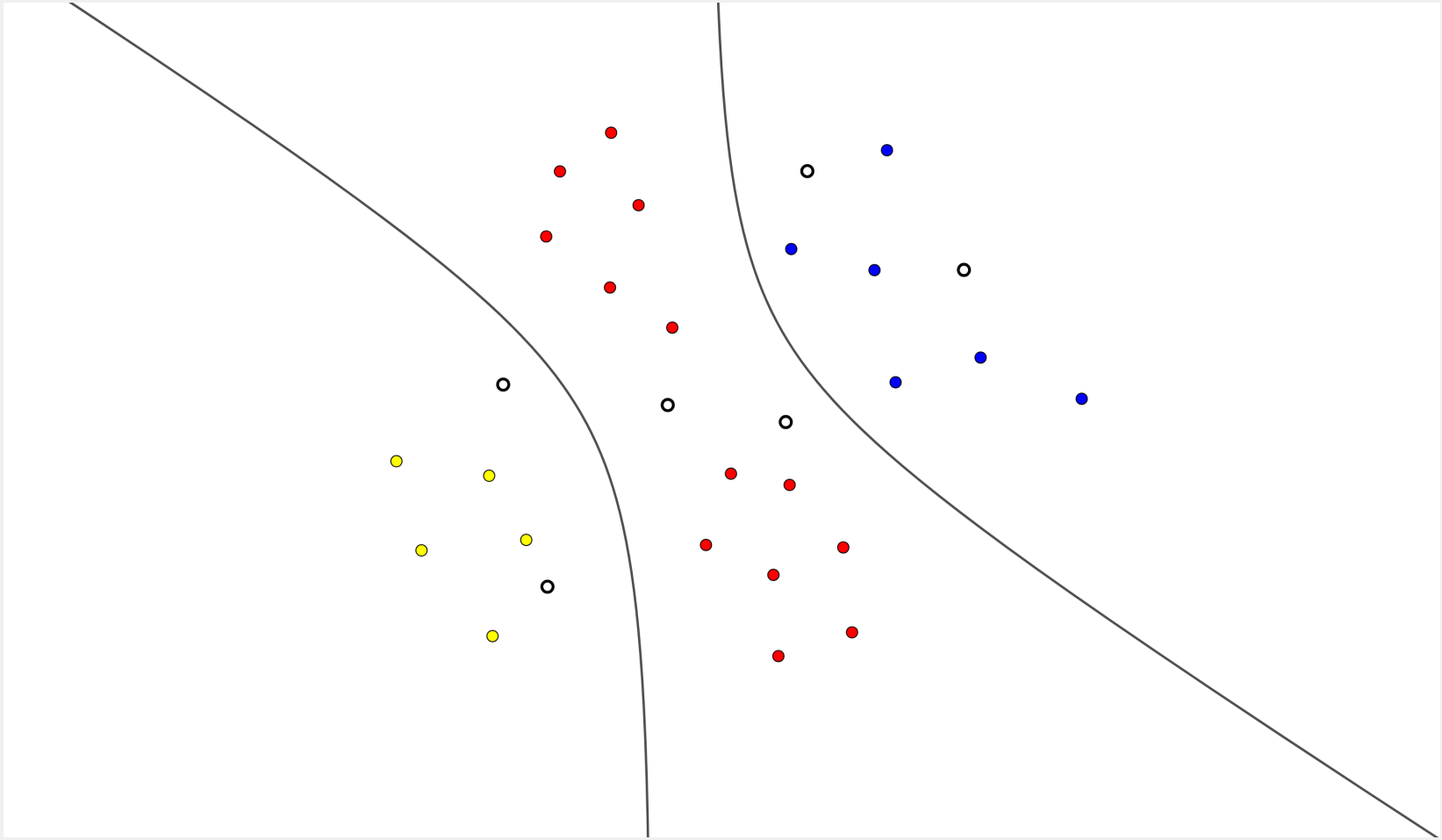
$$x_i \in \mathbb{R}^d, \;\; i = 1, \ldots, n$$

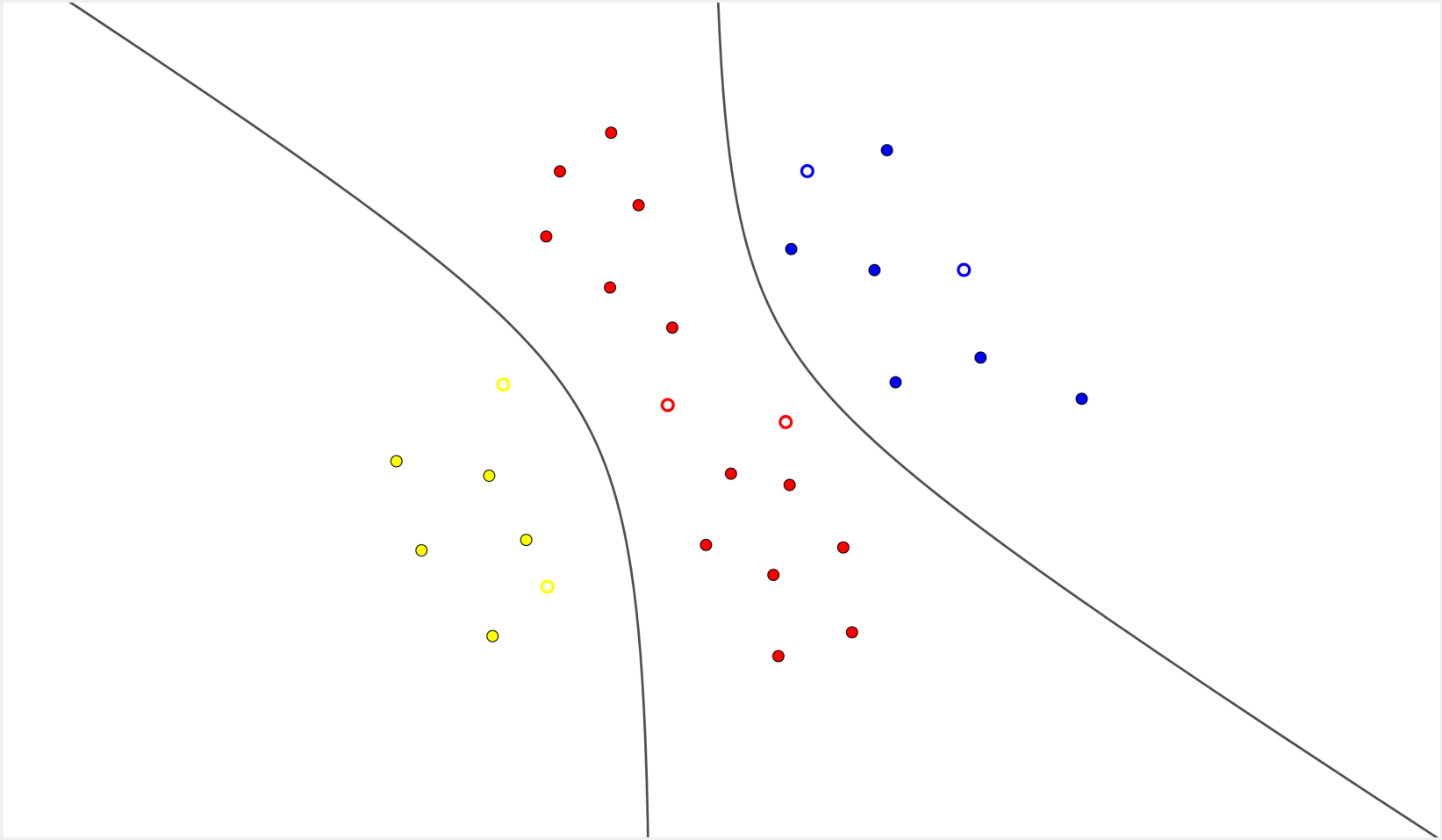
- One target variable with k classes

$$y_i \in \{y_{i1}, \ldots, y_{ik}\}, \;\; i = 1, \ldots, n$$

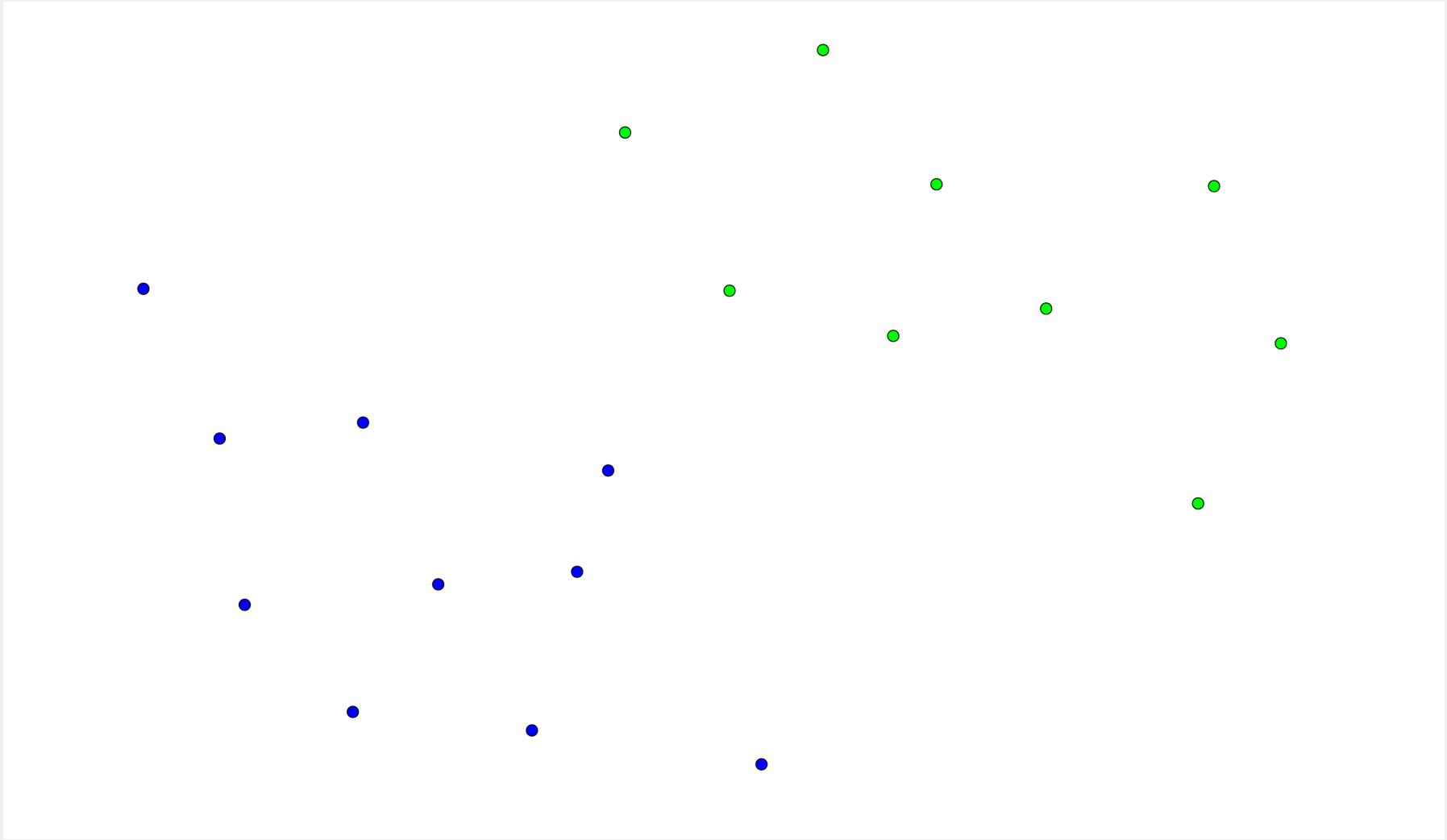




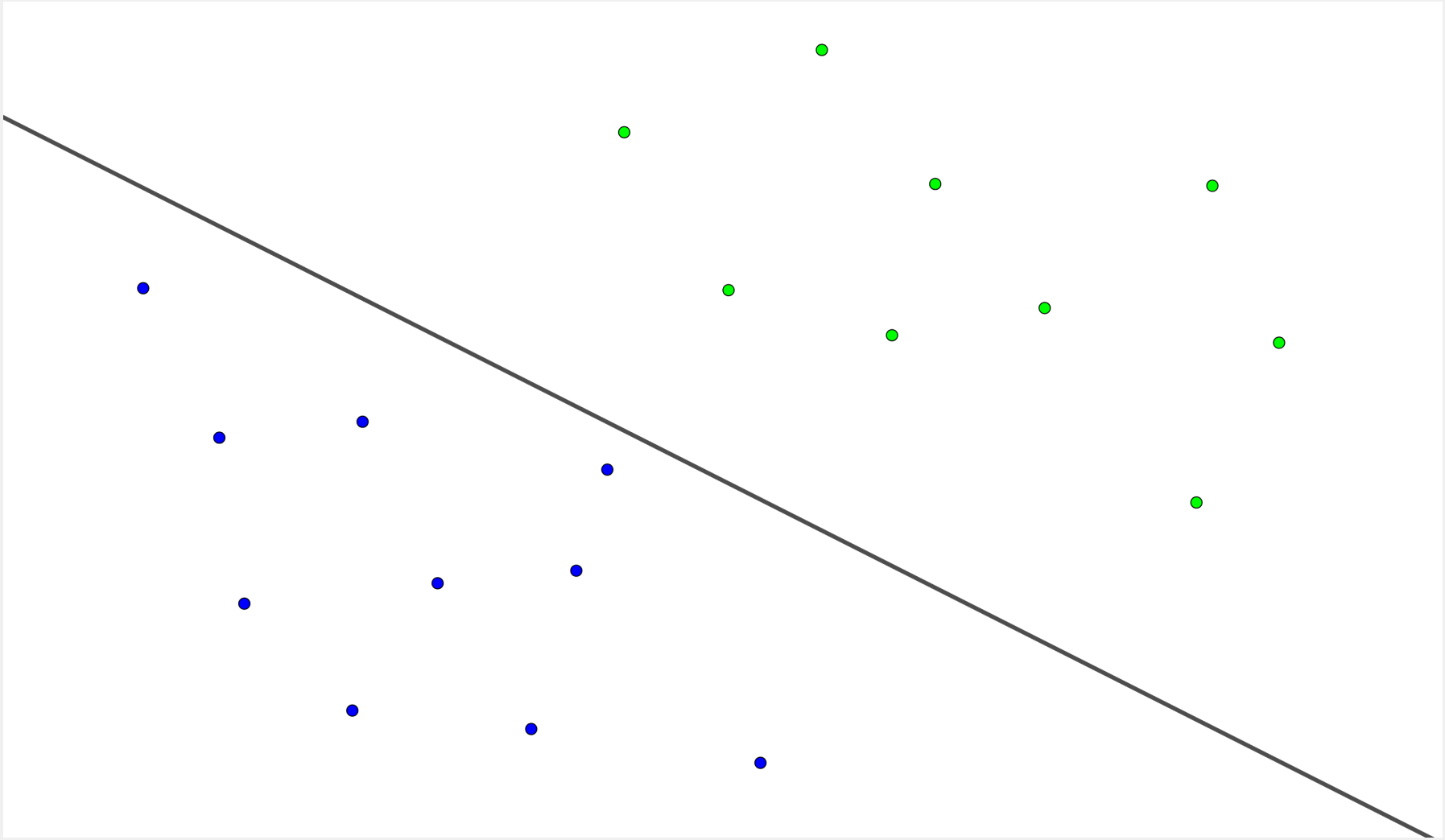




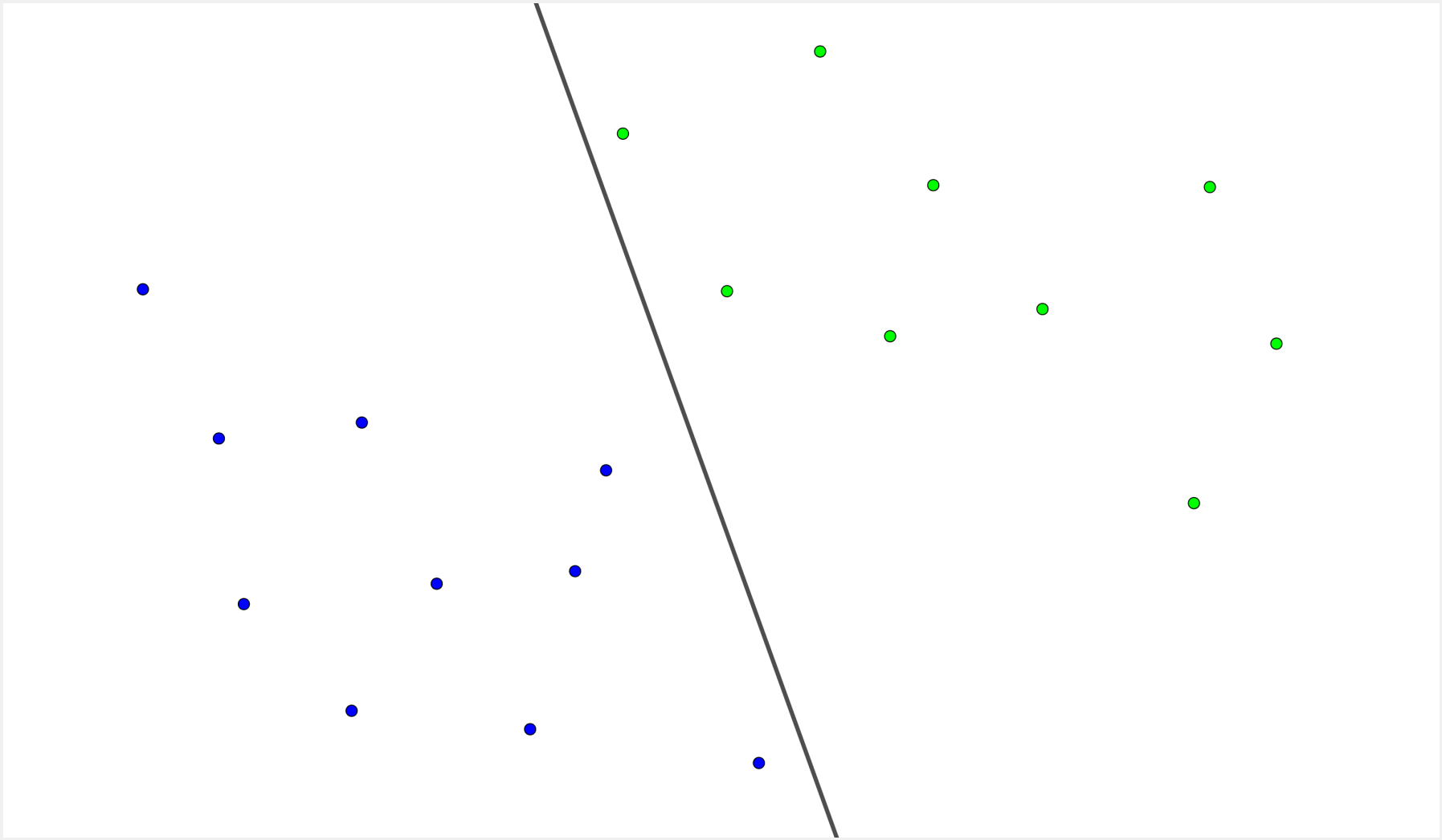
Support Vector Machine (SVM)



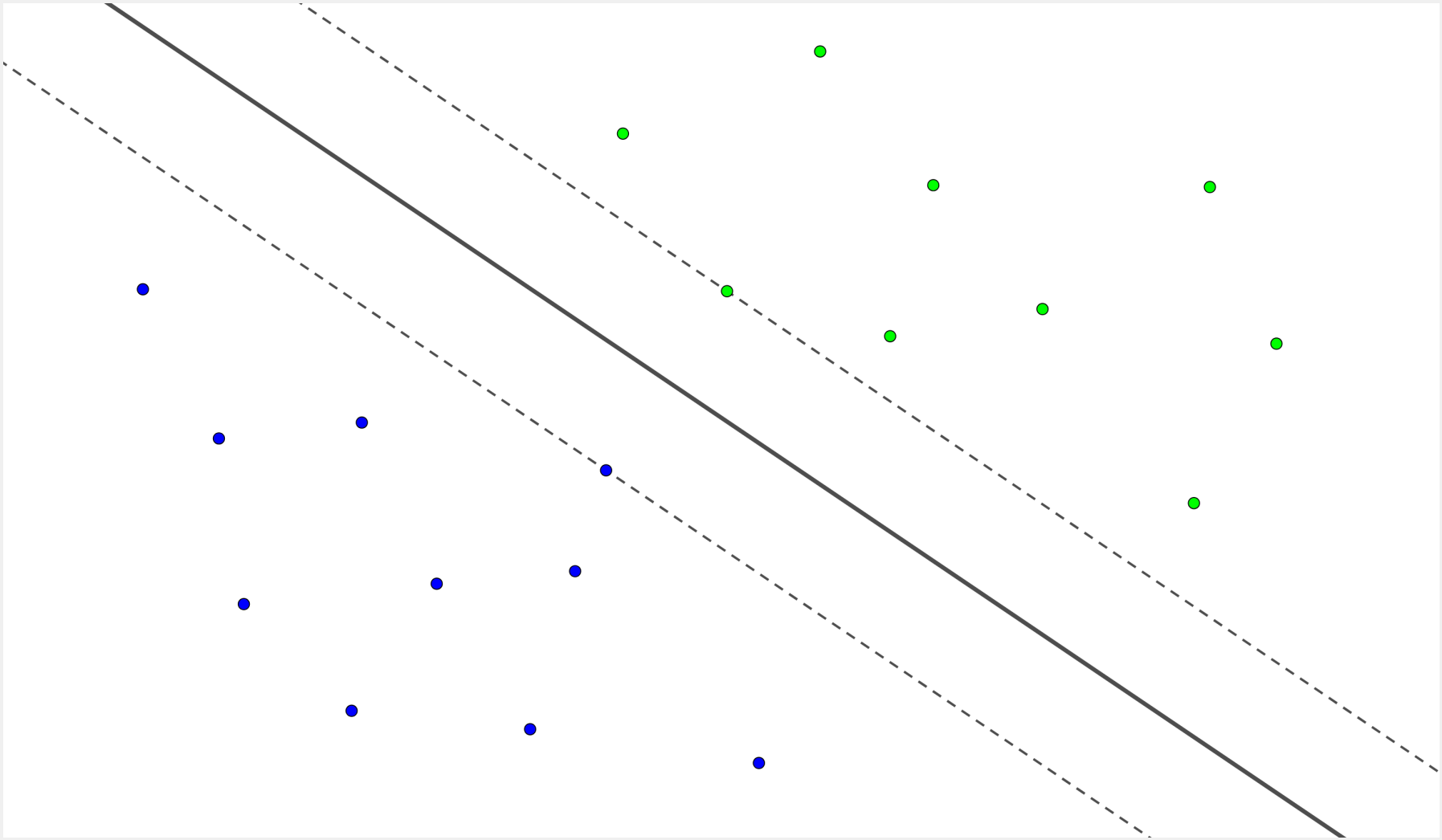
SVM



SVM



SVM



Support vector machines

Given a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, each of them labeled with a class $y_i \in \{-1, +1\}$, find an **hyperplane** in $\mathbb{R}^d$ that **separates** both classes.

Support vector machines

Given a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, each of them labeled with a class $y_i \in \{-1, +1\}$, find an **hyperplane** in $\mathbb{R}^d$ that **separates** both classes.

Find $\mathcal{H} = \{z \in \mathbb{R}^d : \omega^t z + b = 0\}$ such that:

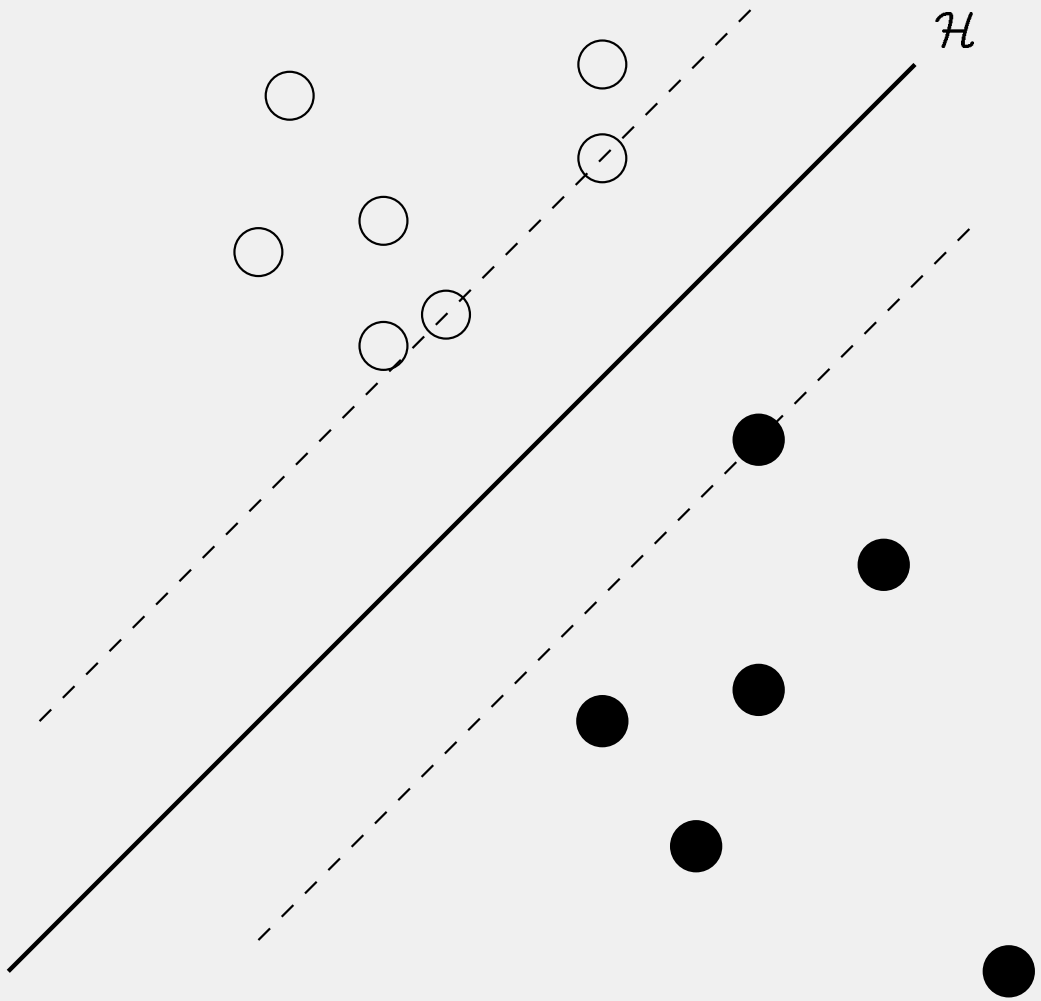
- **-1 Class** belongs to $\{z : \omega^t z + b < 0\}$,
- **+1 Class** belongs to $\{z : \omega^t z + b > 0\}$,

Support Vector Machines

SVM (Vapnik & Chervonenkis '63; Vapnik & Cortes '95): Hyperplane such that the distance between the classes through $\mathcal{H}$ is **maximized**:

Support Vector Machines

SVM (Vapnik & Chervonenkis '63; Vapnik & Cortes '95): Hyperplane such that the distance between the classes through $\mathcal{H}$ is **maximized**:.



Support Vector Machines

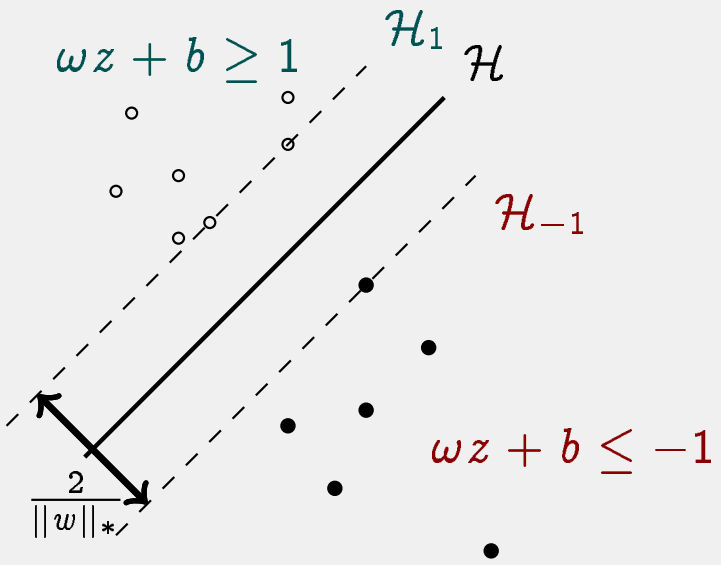
- Consider $\mathcal{H}$ and shifted hyperplanes $\mathcal{H}_1 = \{z : \omega^t x + b = 1\}$ and $\mathcal{H}_{-1} = \{z : \omega^t x + b = -1\}$.
- Each observation should verify $y_i(\omega^t x_i + b) \geq 1$ (Separation).
- Choose a norm $||\cdot||$ to measure the distances between both hyperplanes, then (Mangasarian, 99):

$$D(\mathcal{H}_1, \mathcal{H}_{-1}) = \frac{2}{||\omega||_*}$$

(where $||\cdot||_*$ is the dual norm of $||\cdot||$).

- Solve

$$\max_{y_i(\omega^t x_i + b) \geq 1} \frac{2}{||\omega||_*} \equiv \min_{y_i(\omega^t x_i + b) \geq 1} \frac{1}{2}||\omega||_*.$$



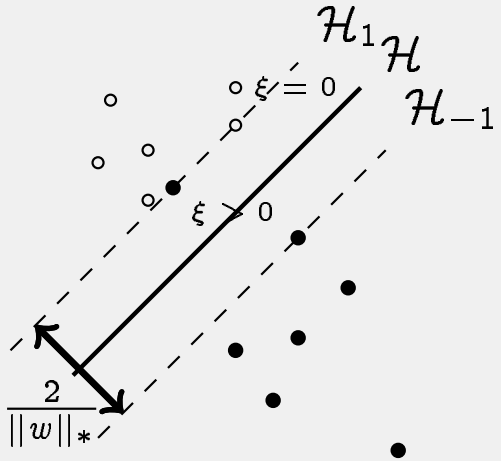
Support Vector Machines

If points are non-linearly separable case: **soft margin constraints**:

Support Vector Machines

If points are non-linearly separable case: **soft margin constraints**:
$\xi_i = \max\{0, 1 - y_i(\omega^t x_i + b)\}$ (Hinge Loss)

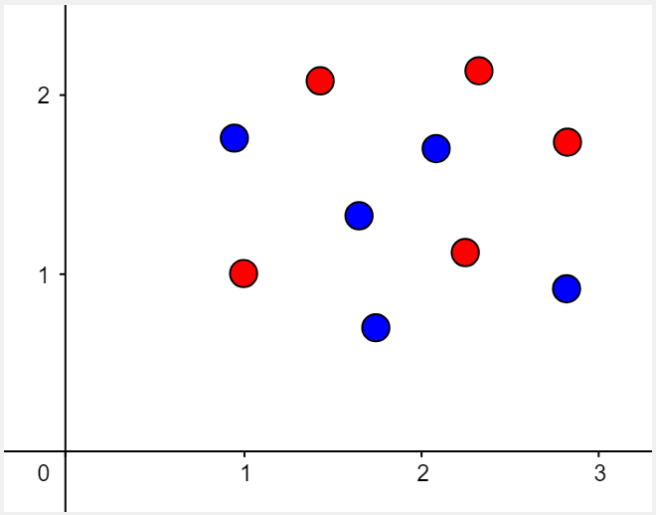
$$\begin{aligned} \min \ & ||w||_* + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \ & y_i(\omega^t x_i + b) \geq 1 - \xi_i, \forall i = 1, \ldots, n, \\ & \xi_i \geq 0, \forall i = 1, \ldots, n, \\ & \omega \in \mathbb{R}^d, b \in \mathbb{R}. \end{aligned}$$

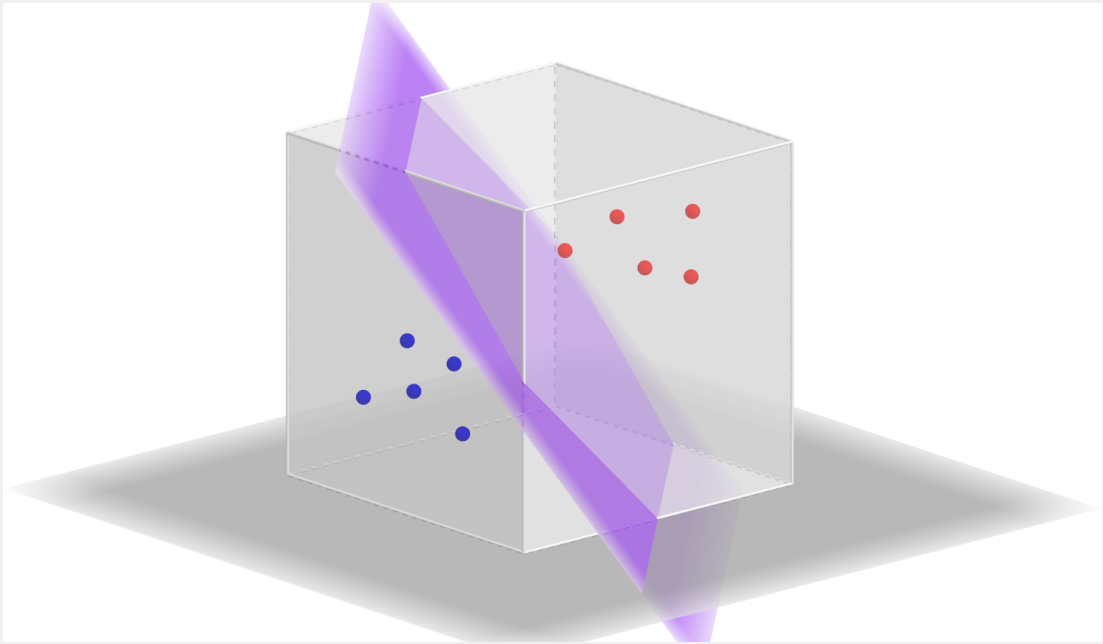


Minimization of the risk incurred applying SVM to outsample data and the one of classifying the insample data.

Kernel trick

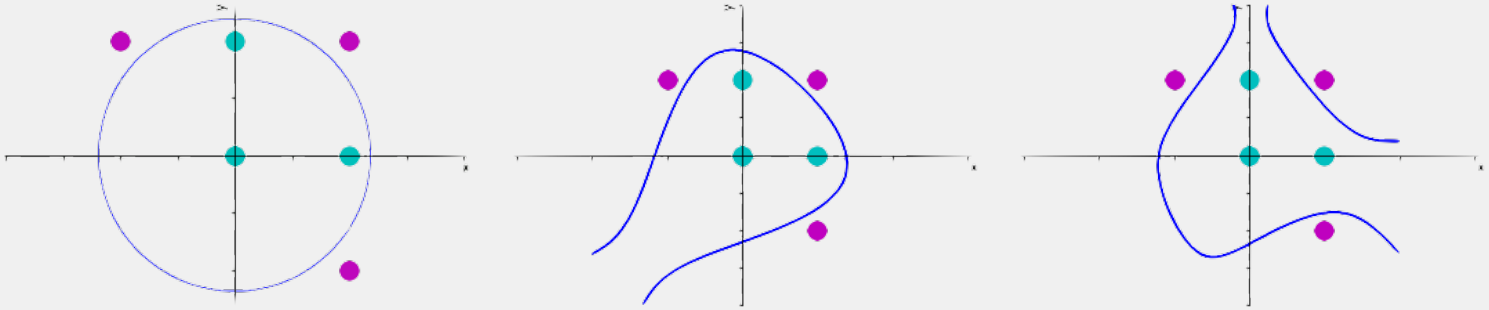
$$\Phi : X \longrightarrow F$$





Kernels

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^D$. Can we manage the dual problem without the explicit knowledge of Φ?



Kernels

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^D$. Can we manage the dual problem without the explicit knowledge of Φ? For the Euclidean case, **YES**:

$$\max \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k y_i y_k \Phi(\mathbf{x}_{i\cdot})^t \cdot \Phi(\mathbf{x}_{k\cdot}) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \; \sum_{i=1}^{n}\alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, n.$$

Only the products $K_{ij} := \Phi(\mathbf{x}_{i\cdot})^t \cdot \Phi(\mathbf{x}_{k\cdot})$ are needed! (Kernel trick)

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $\Big(K(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot})\Big)_{i,j} \succ 0$. Then, there exists $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ with $K(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot}) = \Phi(\mathbf{x}_{i\cdot})^t \cdot \Phi(\mathbf{x}_{k\cdot})$.
(Mercer, 1909)

Also, the optimal ℓ_2-SVM is $\sum_{i=1}^{n}\alpha_i^* y_i K(\mathbf{x}_{i\cdot}, z) + b^* = 0$, $\forall z \in \mathbb{R}^d$.

NO NEED TO KNOW Φ NOT EVEN D.

ℓ_p-SVM

- Standard SVM = ℓ_2-SVM: Successfully applied to classify data of different nature (Finance, Medicine, Biology, etc).
- ℓ_1 and ℓ_∞ explored (Bradley & Mangasarian, 1998; Pedroso & Murata, 2001, Bennet and Bredensteiner 2000).
- Geometry under ℓ_p-SVMs (Ikeda & Murata; 2005; Liu et. al, 2007).
- Different norms for different classes (ℓ_p-SVM-ℓ_q).

ℓ_p-SVM

- Standard SVM = ℓ_2-SVM: Successfully applied to classify data of different nature (Finance, Medicine, Biology, etc).
- ℓ_1 and ℓ_∞ explored (Bradley & Mangasarian, 1998; Pedroso & Murata, 2001, Bennet and Bredensteiner 2000).
- Geometry under ℓ_p-SVMs (Ikeda & Murata; 2005; Liu et. al, 2007).
- Different norms for different classes (ℓ_p-SVM-ℓ_q).

Our Contirbution:

- SOCP Formulations for the **primal** and **dual** problems, for ℓ_p-SVMs ($p \geq 1$)..
- Extend the theory under the *Kernel Trick* through **Multidimensional Kernels**.
- **Apply** ℓ_p-SVM to real standard benchmarking problems.

ℓ_p-SVMs

Let $q = \frac{r}{s} > 1$, with $r, s \in \mathbb{Z}_+$ and $\gcd(r, s) = 1$.
We are given a set of n points in $\mathbb{R}^d$, $\mathbf{x}$, and their classes $\mathbf{y} \in \{-1, 1\}^n$.
Let p such that $\frac{1}{p} + \frac{1}{q} = 1$: $\|\cdot\|_{q^*} = \|\cdot\|_p$.

$$\begin{aligned} \rho^* = \min \ & \|\omega\|_p^p + C \sum_{i=1}^{n} \xi_i \\ \text{s.t. } & y_i(\omega^t \mathbf{x}_{i\cdot} + b) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n, \qquad (\ell_p - \text{SVM}) \\ & \xi \geq 0, \omega \in \mathbb{R}^d, b \in \mathbb{R} \end{aligned}$$

ℓ_p-SVMs

$$\begin{aligned}
\min\ & t + C\sum_{i=1}^{n}\xi_i \\
\text{s.t. } & y_i(\omega^t \mathbf{x}_{i\cdot} + b) \geq 1 - \xi_i, && \forall i = 1,\dots,n, \\
& \mathbf{t \geq \|\omega\|_p^p}, \\
& \xi \geq 0, \omega \in \mathbb{R}^d, b \in \mathbb{R}
\end{aligned}$$

$$t \geq \|\omega\|_p^p \quad \overset{p=\frac{r}{r-s}\geq 1}{\equiv} \quad \begin{cases} -v_j \leq \omega_j \leq v_j, & \forall j = 1,\dots,d, \\ t \geq \sum_{j=1}^{d} u_j, & \\ \mathbf{u_j^{r-s} \geq v_j^r}, & \forall j = 1,\dots,d, \\ u_j, v_j \geq 0, & \forall j = 1,\dots,d, \end{cases}$$

Polynomial constraints in the form $u_j^{r-s} \geq v_j^r$ can be **explicitly** and **efficiently** rewritten as SOC-constraints (B., Puerto, ElHaj, 2014).

The Dual Problem

$$\min ||\omega||_p^p + C \sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} \text{s.t. } & y_i(\omega^t \mathbf{x}_{i\cdot} + b) \geq 1 - \xi_i, && \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)} \\ & \xi_i \geq 0, && \forall i = 1, \ldots, n. \end{aligned}$$

The Dual Problem

$$\min ||\omega||_p^p + C \sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} \text{s.t. } & y_i(\omega^t \mathbf{x}_{i\cdot} + b) \geq 1 - \xi_i, && \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)} \\ & \xi_i \geq 0, && \forall i = 1, \ldots, n. \end{aligned}$$

Conic Dual is also SOCP.

$$\max_{\alpha, u, \delta} \quad \left(\frac{1}{p^q} - \frac{1}{p^{q-1}} \right) \sum_{j=1}^{d} u_j + \sum_{i=1}^{n} \alpha_i \qquad (\text{SOC}_{\text{LD}})$$

$$\begin{aligned} s.t. \; & -\delta_j \leq \sum_{i=1}^{n} \alpha_i y_i x_{ij} \leq \delta_j, && \forall j = 1, \ldots, d, \\ & u_j^s \geq \delta_j^r, && \forall j = 1, \ldots, d, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, && \forall i = 1, \ldots, n. \end{aligned}$$

The Dual Problem

$$\min \|\omega\|_p^p + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y_i(\omega^t \mathbf{x}_{i\cdot} + b) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)}$$

$$\xi_i \geq 0, \qquad \forall i = 1, \ldots, n.$$

Reformulation the Lagrangean Dual is convenient:

$$\max \left(\frac{1}{p^q} - \frac{1}{p^{q-1}} \right) \sum_{j=1}^{d} \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right|^q + \sum_{i=1}^{n} \alpha_i \qquad \text{(LAG-DUAL)}$$

$$s.t. \sum_{i=1}^{n} \alpha_i y_i = 0,$$

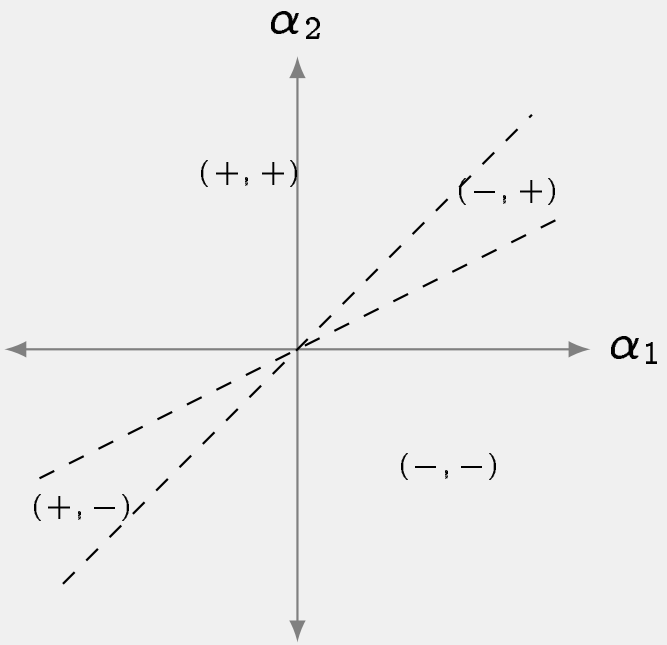
$$0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, n.$$

Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: s_j, for $j = 1, \ldots, d$.

Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: s_j, for $j = 1, \ldots, d$.



Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: s_j, for $j = 1, \ldots, d$.

For each α in a cell:

$$\sum_{j=1}^{d} \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right|^r = \sum_{j=1}^{d} \mathcal{S}_{\alpha,j}^r \left(\sum_{i=1}^{n} \alpha_i y_i x_{ij} \right)^r = \sum_{\gamma \in \mathbb{N}_r^n} c_\gamma \alpha^\gamma y^\gamma \sum_{j=1}^{d} \mathrm{s}_j^r \mathbf{x}_{\cdot j}^\gamma$$

where $c_\gamma = \binom{(\sum_{i=1}^{n} \gamma_i)}{\gamma_1, \ldots, \gamma_n} = \frac{(\sum_{i=1}^{n} \gamma_i)!}{\gamma_1! \cdots \gamma_n!}$, $\mathbb{N}_a^n := \{\gamma \in \mathbb{N}^n : \sum_{i=1}^{n} \gamma_i = a\}$.

$\mathcal{S}_{\alpha,j}^r = \mathrm{sg} \left(\sum_{i=1}^{n} \alpha_i y_i x_{ij} \right)^r$.

NOTE: For **even** r a single cell is enough.

Alternative Dual Formulation

For $p \in \mathbb{N}$, and sign-patterns of the cell, s:

$$\max f_{\mathbf{s}}(\alpha) := \left(\frac{1}{p^q} - \frac{1}{p^{q-1}}\right) \sum_{\gamma \in \mathbb{N}_r^n} c_\gamma \alpha^\gamma y^\gamma \sum_{j=1}^{d} \mathbf{s}_j^r \mathbf{x}_{\cdot j}^\gamma + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad (\mathrm{SOC}'_{\mathrm{LD}}(\mathbf{s}))$$

$$\mathbf{s}_j \sum_{i=1}^{n} \alpha_i y_i x_{ij} \geq 0, \qquad \forall j = 1, \ldots, d,$$

$$0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, n.$$

Optimal hyperplane from duals

Let $\bar{\alpha}$ optimal for a subdivision:
For i_0 such that $0 < \bar{\alpha}_{i_0} < C$:

$$b = y_{i_0} - \frac{1}{p^{q-1}} \sum_{j=1}^{d} \mathcal{S}^r_{\bar{\alpha},j} \left(\sum_{i=1}^{n} \bar{\alpha}_i y_i x_{ij} \right)^{r-1} x_{i_0 j}.$$

and the induced hyperplane is:

$$\frac{1}{p^{r-1}} \sum_{\gamma \in \mathbb{N}^n_{r-1}} c_\gamma \bar{\alpha}^\gamma y^\gamma \sum_{j=1}^{d} \mathcal{S}^r_{\bar{\alpha},j} \mathbf{x}^\gamma_{\cdot j} z_j + b = 0.$$

for all $z \in \mathbb{R}^d$.

Kernels

- Let $\Phi : \mathbb{R}^d \to \mathbb{R}^D$. Can we manage the dual problem without the explicit knowledge of Φ?
- Is there some extension of Mercer's theorem to the case of ℓ_p-norm?

ℓ_p-Kernels

We are given a data set $[\mathbf{x}] = (x_{1\cdot}, \ldots, x_{n\cdot})$ together with their classification patterns $\mathbf{y} = (y_1, \ldots, y_n)$ and $r \in \mathbb{N}$.

$$\mathrm{H}_{\mathbf{y}} = \{\alpha \in [0, C]^n : \sum_{i=1}^{n} \alpha_i y_i = 0\}$$

$$S_{\Phi}(R) := \left\{ \mathrm{s} = (\mathrm{s}_1, \ldots, \mathrm{s}_D) \in \{-1, 1\}^D : \mathrm{s}_j = \mathrm{sg}(\sum_{i=1}^{n} \alpha_i y_i \Phi_j(\mathbf{x}_{i\cdot}))^r, \alpha \in R, \forall j \right\}.$$

Definition

Given a transformation function, $\Phi : \mathbb{R}^d \to \mathbb{R}^D$, a subdivision of $\mathrm{H}_{\mathbf{y}}$, $\{R_k\}_{k \in \mathcal{K}}$, is said a ***suitable Φ-subdivision*** of $\mathrm{H}_{\mathbf{y}}$ if

$$\mathrm{S}_{\Phi}(R_k) = \{\mathrm{s}_{R_k}\} \text{ for some } \mathrm{s}_{R_k} \in \{-1, 1\}^D \text{ and for all } k \in \mathcal{K}.$$

Definition

Given a suitable Φ-subdivision, $\{R_k\}_{k\in\mathcal{K}} \subseteq 2^{\mathbf{H}_\mathbf{y}}$, and $(\gamma, \lambda) \in \mathbb{N}_r^{n+1}$, $\lambda \in \{0, 1\}$, the operator

$$\mathrm{K}[\mathbf{x}]_{R_k,\gamma,\lambda}(z) := \sum_{j=1}^{D} \mathbf{s}_{R_k,j}^r \Phi_j(\mathbf{x})^\gamma \Phi_j(z)^\lambda, \forall z \in \mathbb{R}^d, \forall k \in \mathcal{K}, \tag{1}$$

is called a r-order Kernel function of Φ. For $k \in \mathcal{K}$, $\mathrm{K}[\mathbf{x}]_{R_k,\gamma,\lambda}(z)$ is called the k-th slice of the kernel function.

Proposition

The separating hyperplane and the objective function can be rewritten for the Φ-transformed data using the Kernel function.

From Kernels to Tensors...

Given $z \in \mathbb{R}^d$, and for any $k \in \mathcal{K}$, the kernel operator $\mathrm{K}[\mathbf{x}]_{R_k,\gamma,\lambda}$ induces a symmetric r-order $(n+1)$-dimensional real tensor namely $\mathbb{K}^k = (\mathbb{K}^k_{i_1 \dots i_r})^n_{i_1,\dots,i_r=1}$ with $\mathbb{K}^k_{i_1 \dots i_r} \in \mathbb{R}$ such that

$$\mathbb{K}^k_{i_1 \dots i_r} = \begin{cases} \mathrm{K}[\mathbf{x}]_{R_k,\gamma_0,0}(z) & \text{if } i_1, \dots, i_r < n+1, \\ \mathrm{K}[\mathbf{x}]_{R_k,\gamma_1,1}(z) & \text{if there exists } s \in \{1, \dots, r\} \text{ such that } i_s = n+1. \end{cases}$$

being $(\gamma_0, \lambda) = \sum_{l=1}^r \mathrm{e}_{i_l}$ with $\lambda = 0$ and $(\gamma_1, \lambda) = \sum_{l=1}^r \mathrm{e}_{i_l}$ with $\lambda = 1$.

... and from Tensors to Kernels...

Theorem

Let $\{R_k\}_{k\in\mathcal{K}}$ *be a subdivision of* $\mathrm{H}_{\mathbf{y}}$ *and* $\mathbb{K}^k$*, for* $k \in \mathcal{K}$*, be a* r*-order* $(n+1)$*-dimensional symmetric tensor such that each* $\mathbb{K}^k$ *can be decomposed as:*

$$\mathbb{K}^k = \sum_{j=1}^{\widehat{D}} \psi_{kj}\, v_j \otimes \overset{r}{\cdots} \otimes v_j,\ \forall k \in \mathcal{K},$$

and satisfying, either

1. r *is even and* $\psi_j := \psi_{kj} \geq 0$*, or*
2. r *is odd and* $\psi_j := |\psi_{kj}|$ *and for all* $k \in \mathcal{K}$*:*

$$\mathrm{sg}(\psi_{kj}) = \mathrm{sg}\left(\sum_{i=1}^{n} \alpha_i\, y_i \sqrt[r]{\psi_j}\, v_{ji}\right),\ \textit{for all } \alpha \in \mathbb{R}_k.$$

Then, there exists a transformation Φ*, such that* $\{R_k\}_{k\in\mathcal{K}}$ *is a* Φ*-suitable subdivision of* $\mathrm{H}_{\mathbf{y}}$ *and* $\{\mathbb{K}^k\}_{k\in\mathcal{K}}$ *induces a* r*-order kernel function of* Φ*.*

... and from Tensors to Kernels...

Theorem

Let $\{R_k\}_{k\in\mathcal{K}}$ *be a subdivision of* $\mathrm{H_y}$ *and* $\mathbb{K}^k$*, for* $k \in \mathcal{K}$*, be a* r*-order* $(n+1)$*-dimensional symmetric tensor such that each* $\mathbb{K}^k$ *can be decomposed as:*

$$\mathbb{K}^k = \sum_{j=1}^{\widehat{D}} \psi_{kj}\, v_j \otimes \overset{r}{\cdots} \otimes v_j,\ \forall k \in \mathcal{K},$$

and satisfying, either

1. r *is even and* $\psi_j := \psi_{kj} \geq 0$*, or*
2. r *is odd and* $\psi_j := |\psi_{kj}|$ *and for all* $k \in \mathcal{K}$*:*

$$\mathrm{sg}(\psi_{kj}) = \mathrm{sg}\left(\sum_{i=1}^{n} \alpha_i y_i \sqrt[r]{\psi_j}\, v_{ji} \right),\ \textit{for all } \alpha \in \mathbb{R}_k.$$

Then, there exists a transformation Φ*, such that* $\{R_k\}_{k\in\mathcal{K}}$ *is a* Φ*-suitable subdivision of* $\mathrm{H_y}$ *and* $\{\mathbb{K}^k\}_{k\in\mathcal{K}}$ *induces a* r*-order kernel function of* Φ*.*

For even r: P tensors, B tensors, B_0 tensors, diagonally dominated tensors, positive Cauchy tensors, SOS tensors, ... verify the hyphoteses.

Avoiding kernels: Schauder Bases

Theorem (Lindenstrauss & Tzafriri '77)

The Banach space of continuous functions $C_{\mathbb{R}^D}(T)$ from a compact set $T \subseteq \mathbb{R}^d$ admits a Schauder basis.

For instance, $\mathcal{B} = \{\mathbf{z}^\gamma : \gamma \in \mathbb{N}^d\}$, the standard basis of multidimensional monomials is a Schauder basis for this space. (also Bernstein or trigonometric polynomials are Schauder bases).
Any continuous function $\Phi : T \longmapsto \mathbb{R}^D$

$$\Phi(\mathbf{z}) = \sum_{j=1}^{\infty} \tau_j \mathbf{z}_j$$

with $\tau_j \in \mathbb{R}$ and $\mathbf{z}_j \in \mathcal{B}$ for any $j = 1, \ldots, \infty$.

Avoiding kernels: Schauder Bases

Theorem (Lindenstrauss & Tzafriri '77)

The Banach space of continuous functions $C_{\mathbb{R}^D}(T)$ from a compact set $T \subseteq \mathbb{R}^d$ admits a Schauder basis.

For instance, $\mathcal{B} = \{\mathbf{z}^\gamma : \gamma \in \mathbb{N}^d\}$, the standard basis of multidimensional monomials is a Schauder basis for this space. (also Bernstein or trigonometric polynomials are Schauder bases).
Any continuous function $\Phi : T \longmapsto \mathbb{R}^D$

$$\Phi(\mathbf{z}) = \sum_{j=1}^{\infty} \tau_j \mathbf{z}_j$$

with $\tau_j \in \mathbb{R}$ and $\mathbf{z}_j \in \mathcal{B}$ for any $j = 1, \ldots, \infty$.

Strategy: Fix a truncation degree η and find the best polynomial with degree up to η that separates the classes.

Experiments

Datasets (UCI repository):

- `cleveland`: heart disease (303 obs., 13 features).
- `housing`: prices of Boston houses (303 obs., 13 features).
- `gc`: loan defaulters (1000 obs., 21 features)
- `colon`: cancerous colon tissues (62 obs., 2002 features)

Models were coded in `Python` 3.6, and solved using `Gurobi` 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported.

- $\Phi[\eta] : \mathbb{R}^d \to \mathbb{R}^{\mathbb{N}^d_\eta}$. Its components, $\Phi[\eta]_\gamma(\mathbf{z}) = z^\gamma$ for $\gamma \in \mathbb{N}^d_\eta$, are the monomials (in d variables) up to degree η.
- $\widetilde{\Phi}[\eta] : \mathbb{R}^d \to \mathbb{R}^{\mathbb{N}^d_\eta}$, with $\widetilde{\Phi}[\eta]_\gamma(\mathbf{z}) = \exp(-\sigma||z||_2^2)\frac{\sqrt[r]{2\sigma}\mathbf{z}^\gamma}{\sqrt[r]{\gamma_1! \cdots \gamma_d!}}$, for $\mathbf{z} \in \mathbb{R}^d$, for $\gamma \in \mathbb{N}^d_\eta$ and $\sigma > 0$.

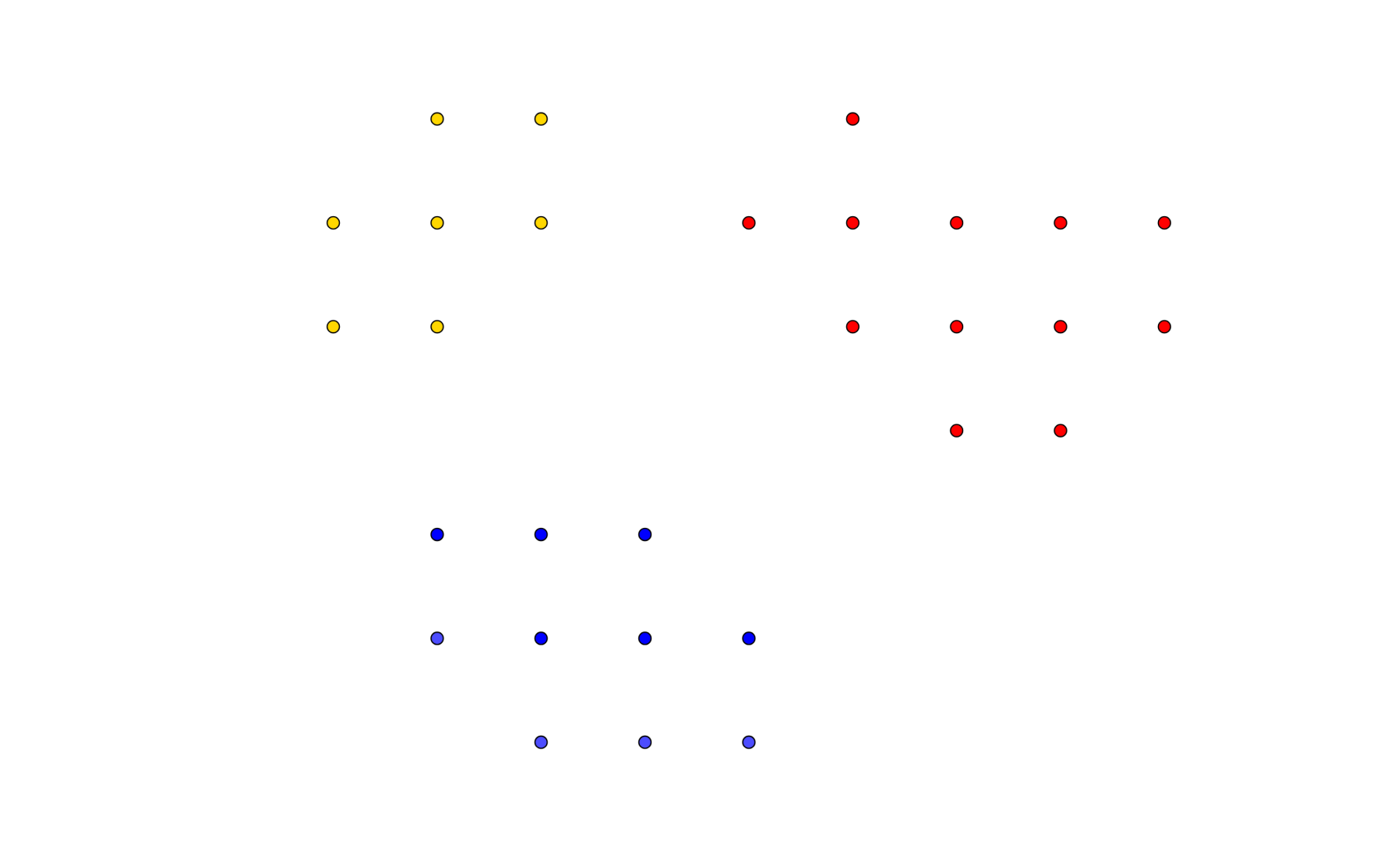
Experiments ($\Phi[\eta]$)

	$\ell_{\frac{4}{3}}$			$\ell_{\frac{3}{2}}$			ℓ_2			ℓ_3		
η	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time
	cleveland dataset											
1	85.11%	82.84%	0.01	85.11%	83.16%	0.01	85.15%	**83.48%**	0.01	85.33%	83.15%	0.01
2	94.02%	**82.57%**	0.44	93.58%	81.57%	0.40	93.33%	81.58%	0.04	93.35%	79.61%	0.41
3	99.34%	74.93%	5.49	99.41%	75.60%	2.87	99.67%	78.53%	0.14	99.67%	**80.23%**	2.65
4	99.67%	76.56%	28	99.67%	76.92%	22.5	99.74%	**79.21%**	0.47	100%	78.60%	17.56
	housing dataset											
1	88.56%	**85.36%**	0.01	88.25%	85.16%	0.02	88.10%	84.36%	0.02	87.92%	83.35%	0.04
2	94.93%	78.85%	0.22	94.14%	80.03%	0.42	92.31%	80.02%	0.14	91.15%	**81.38%**	0.39
3	98.60%	**80.95%**	9.57	98.24%	80.00%	6.13	97.34%	79.81%	0.51	96.07%	78.84%	5.86
4	99.23%	**79.99%**	45.09	98.90%	77.78%	31.69	98.37%	78.63%	1.59	97.98%	78.43%	27.42
	german credit dataset											
1	78.53%	**76.20%**	0.02	78.53%	**76.20%**	0.04	78.53%	**76.20%**	0.05	78.54%	**76.20%**	0.04
2	93.03%	67.50%	0.92	93.04%	67.60%	2.50	92.98%	67.40%	0.50	93.00%	**67.70%**	3.32
3	100%	**71.90%**	85.86	100%	70.50%	94.12	100%	70.20%	3.14	100%	68.90%	98.58
	colon dataset											
1	100%	**82.14%**	20.3	100%	80.48%	15.73	100%	80.48%	0.05	100%	80.48%	14.61

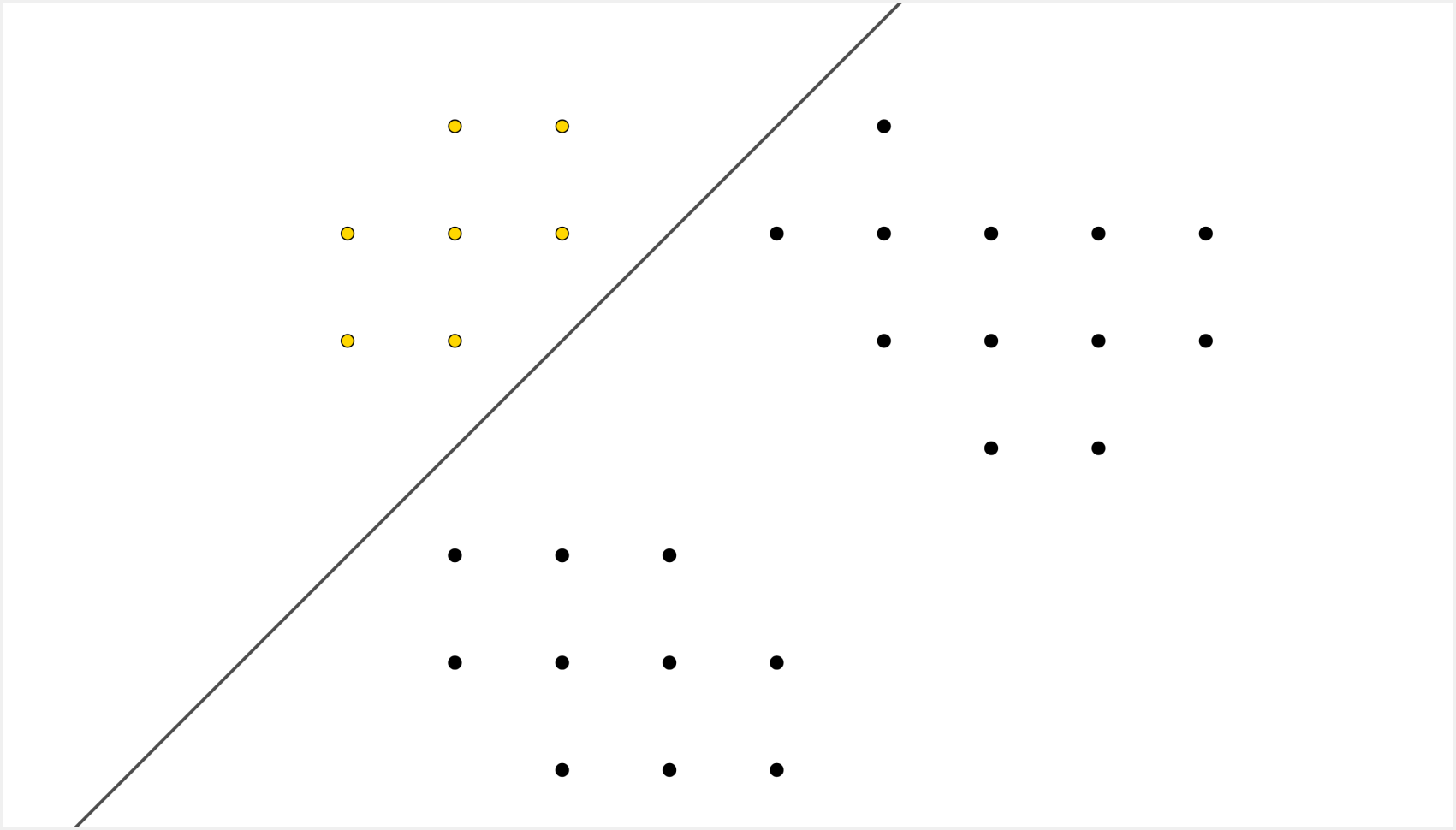
Experiments ($\widetilde{\Phi}[\eta]$)

	$\ell_{\frac{4}{3}}$			$\ell_{\frac{3}{2}}$			ℓ_2			ℓ_3		
η	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time	ACC^{Tr}	ACC^{Test}	Time
	cleveland dataset											
1	85.15%	83.16%	0.01	85.11%	83.16%	0.01	85.33%	**83.48%**	0.01	85.22%	**83.48%**	0.01
2	88.30%	**84.19%**	0.24	88.05%	82.55%	0.28	86.72%	80.58%	0.04	84.01%	77.26%	0.24
3	92.15%	80.87%	4.91	92.12%	81.54%	2.77	92.41%	**81.55%**	0.13	92.59%	81.20%	2.54
4	84.38%	83.47%	19.57	84.41%	83.46%	12.83	84.71%	83.46%	0.19	85.18%	**83.48%**	15.51
	housing dataset											
1	88.56%	**85.36%**	0.01	88.25%	85.16%	0.02	88.10%	84.36%	0.02	87.53%	84.71%	0.04
2	89.53%	**83.53%**	0.25	88.84%	82.95%	0.48	87.42%	82.94%	0.11	86.72%	82.46%	0.66
3	94.01%	80.03%	4.47	93.30%	79.82%	4.29	91.50%	**80.21%**	0.25	90.36%	79.95%	3.05
4	90.80%	82.37%	14.43	90.58%	**83.36%**	20.98	88.95%	81.59%	0.17	86.69%	82.95%	12.2
	german credit dataset											
1	78.35%	**79.00%**	0.02	78.33%	78.88%	0.04	78.25%	78.63%	0.05	78.26%	78.75%	0.04
2	77.29%	74.38%	2.96	77.83%	75.00%	2.37	79.23%	74.44%	0.45	81.15%	**75.22%**	2.13
3	76.72%	76.75%	57.01	92.78%	**79.00%**	63.64	96.36%	77.88%	2.75	98.24%	76.57%	48.4
	colon dataset ($C = 1$)											
1	100%	**82.14%**	20.3	100%	80.48%	15.73	100%	80.48%	0.05	100%	80.48%	14.61

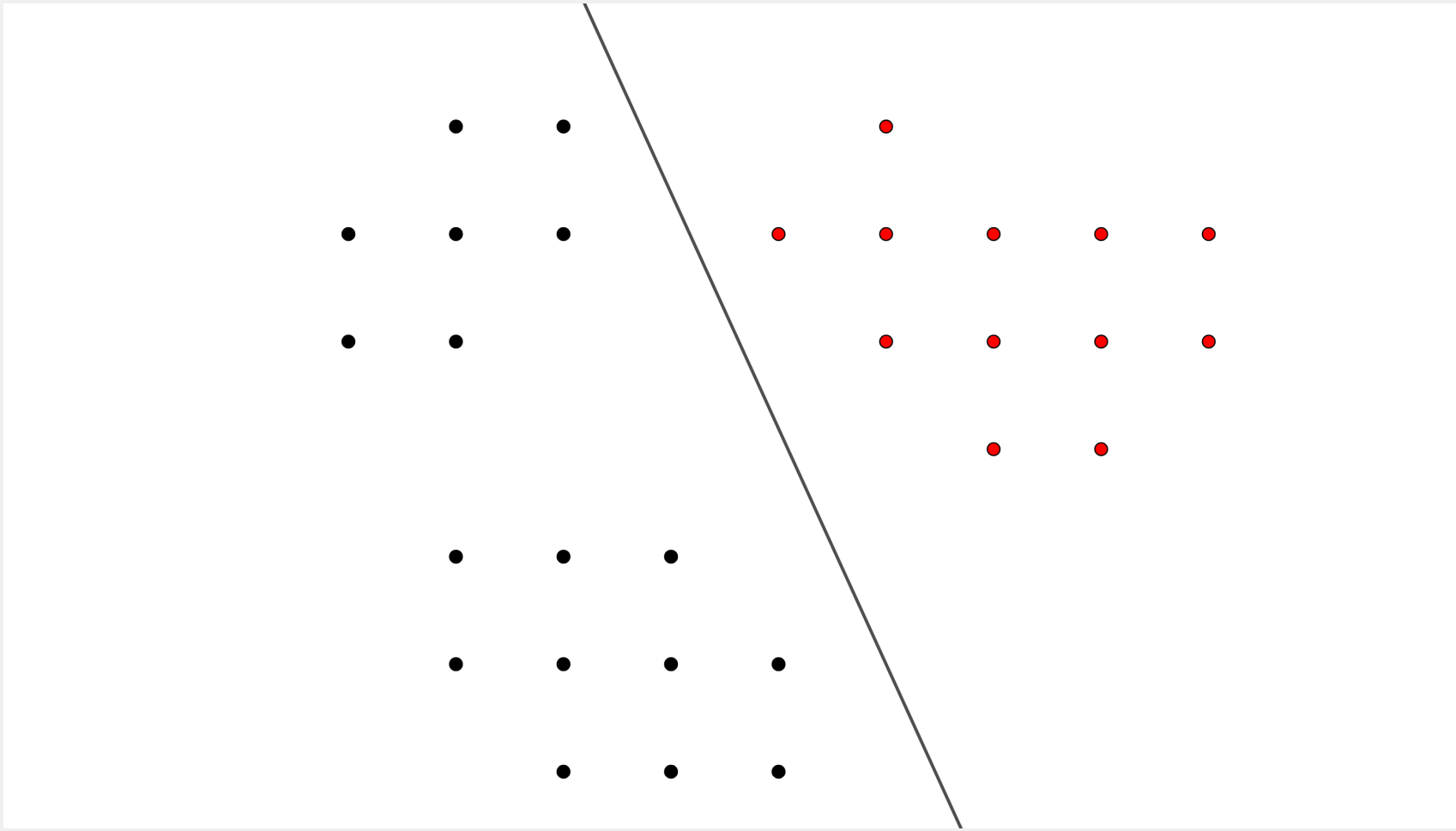
Multiclass problem



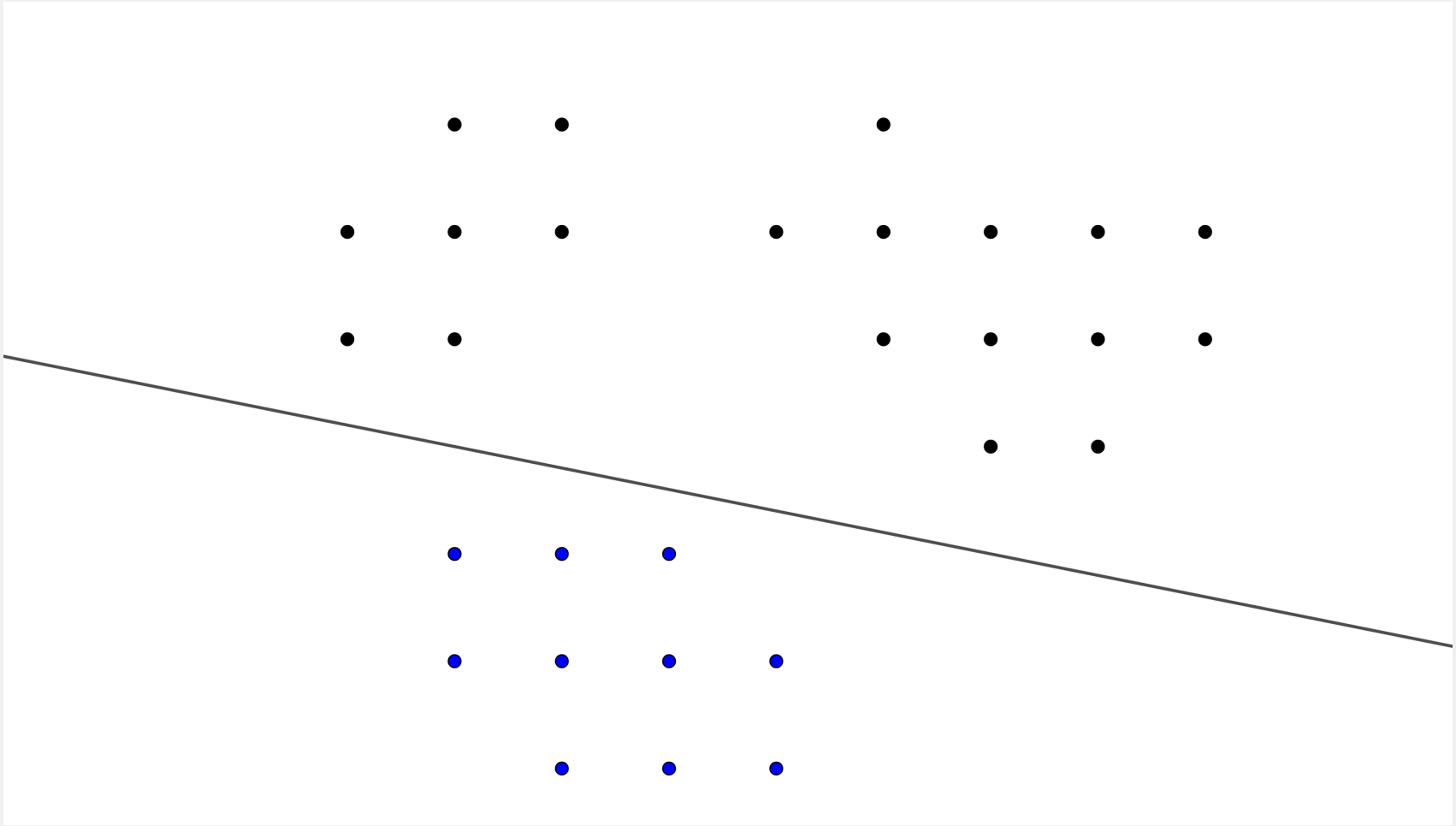
One Vs All



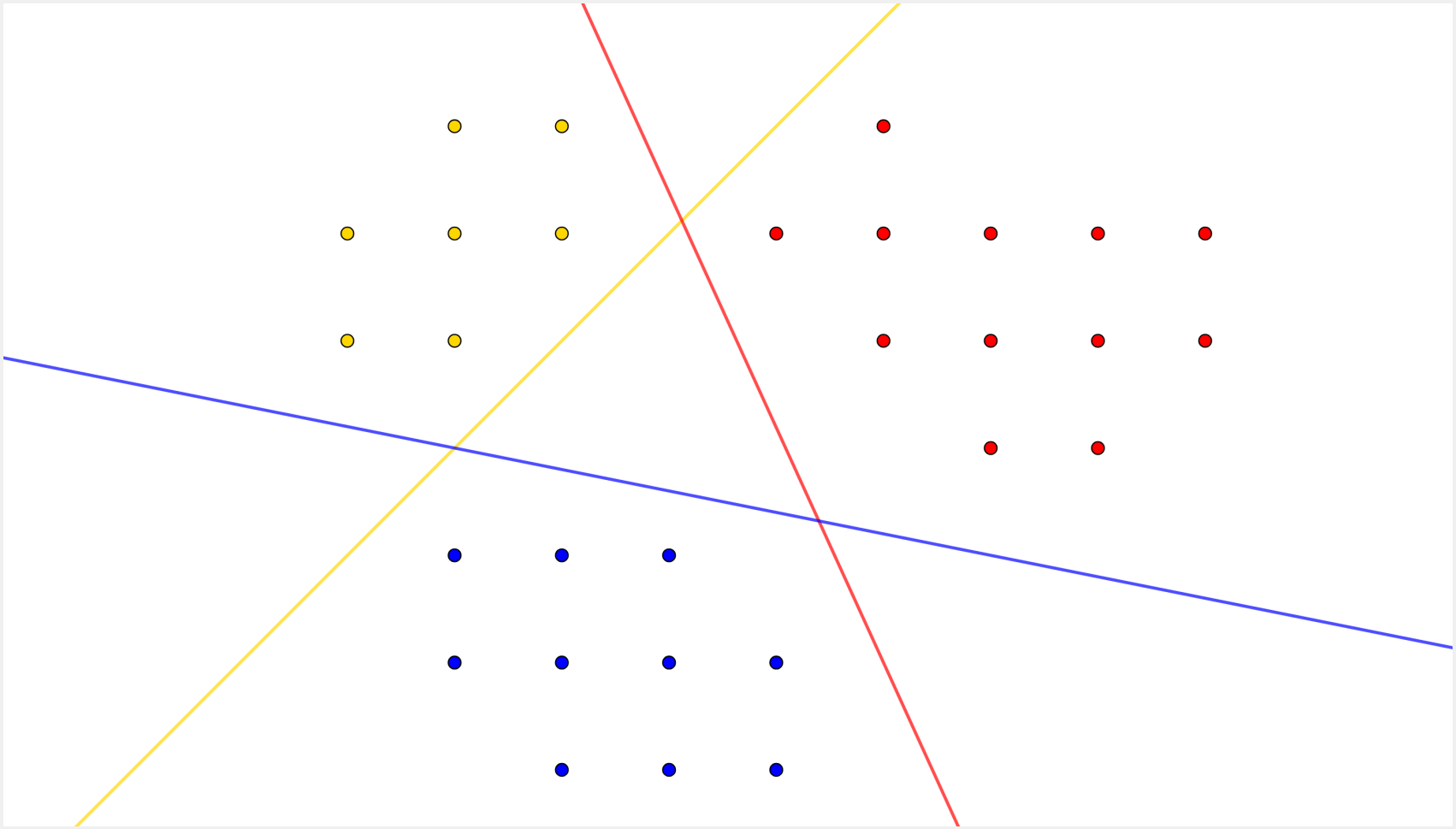
One Vs All



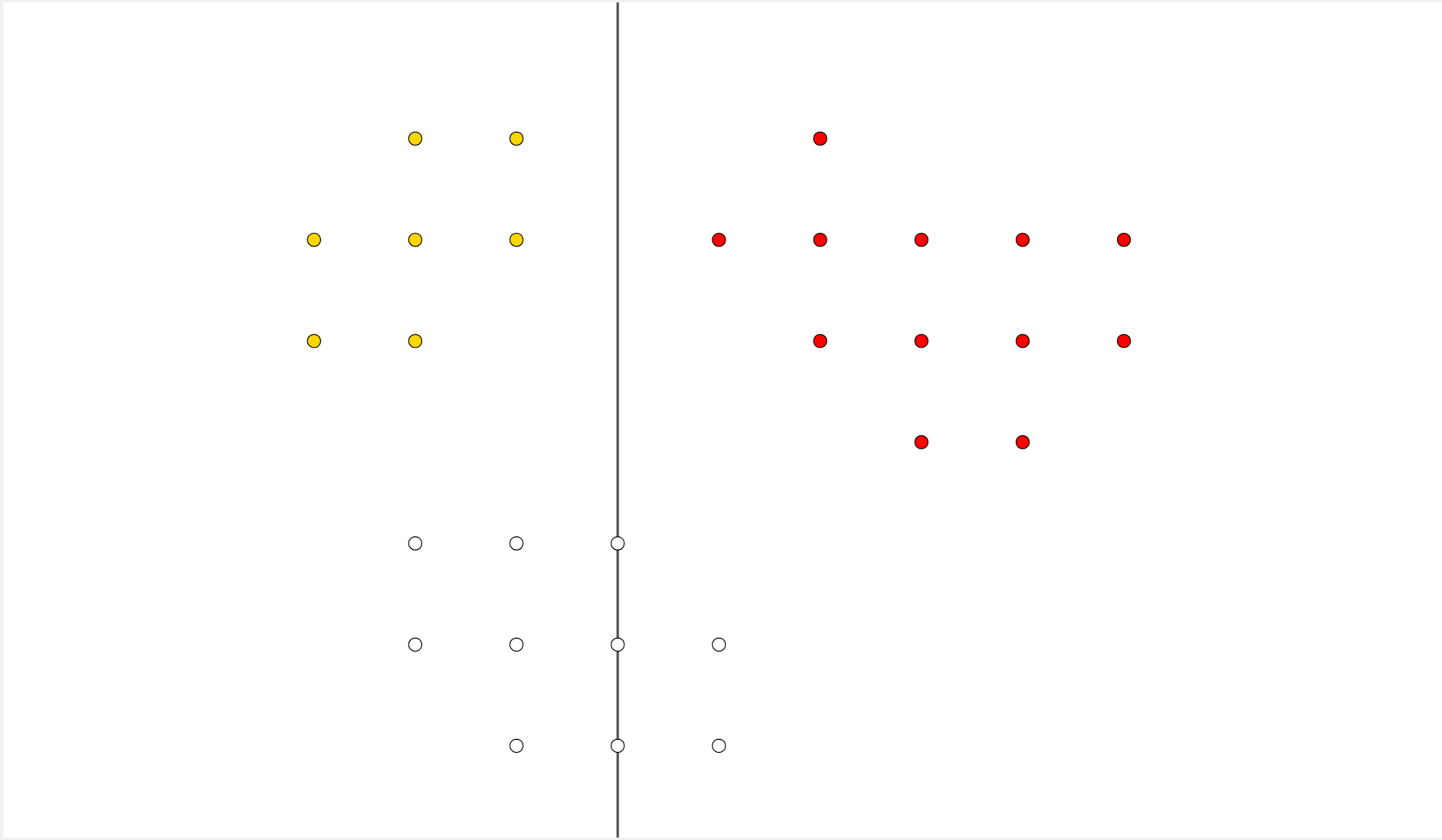
One Vs All



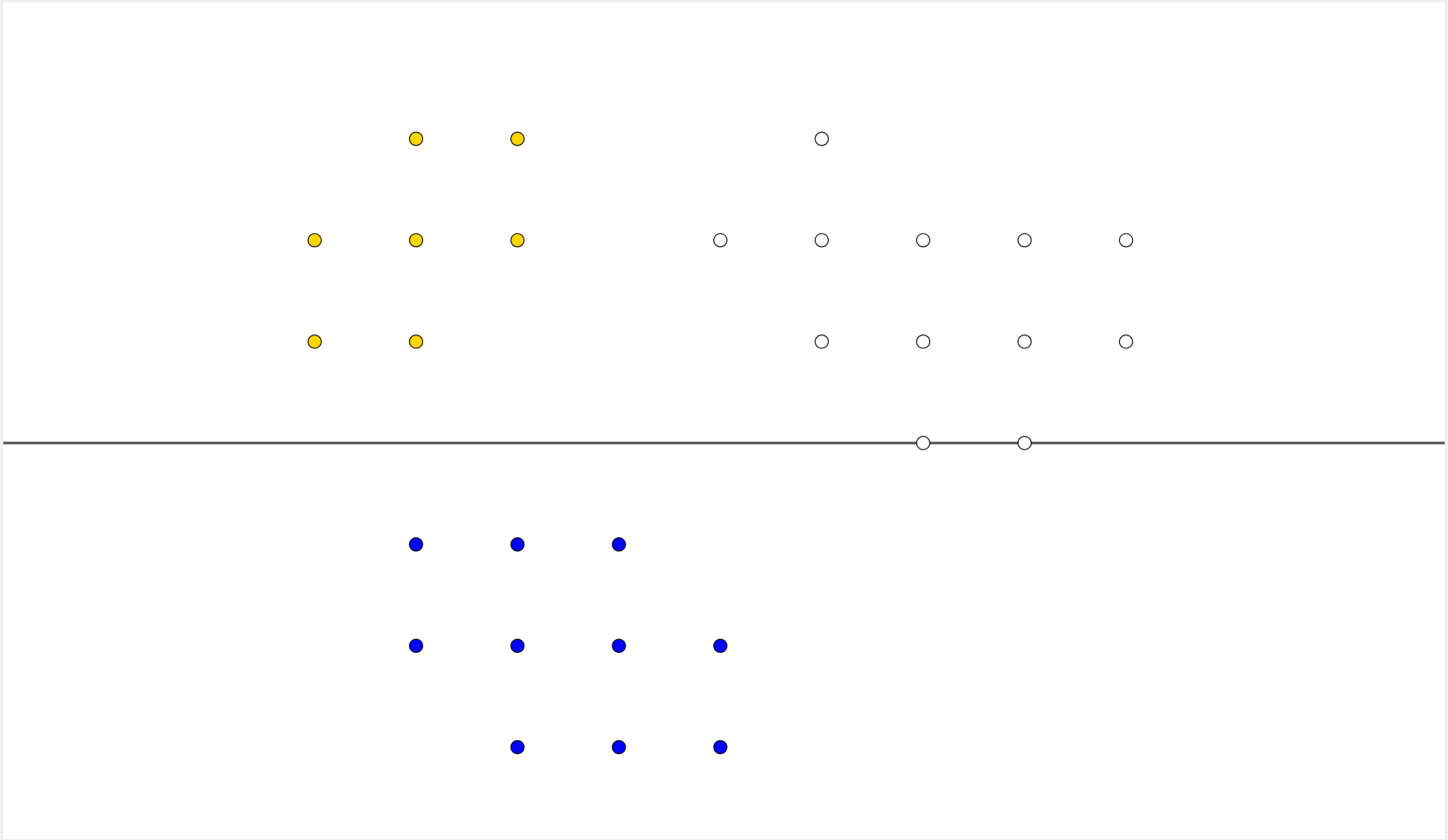
One Vs All



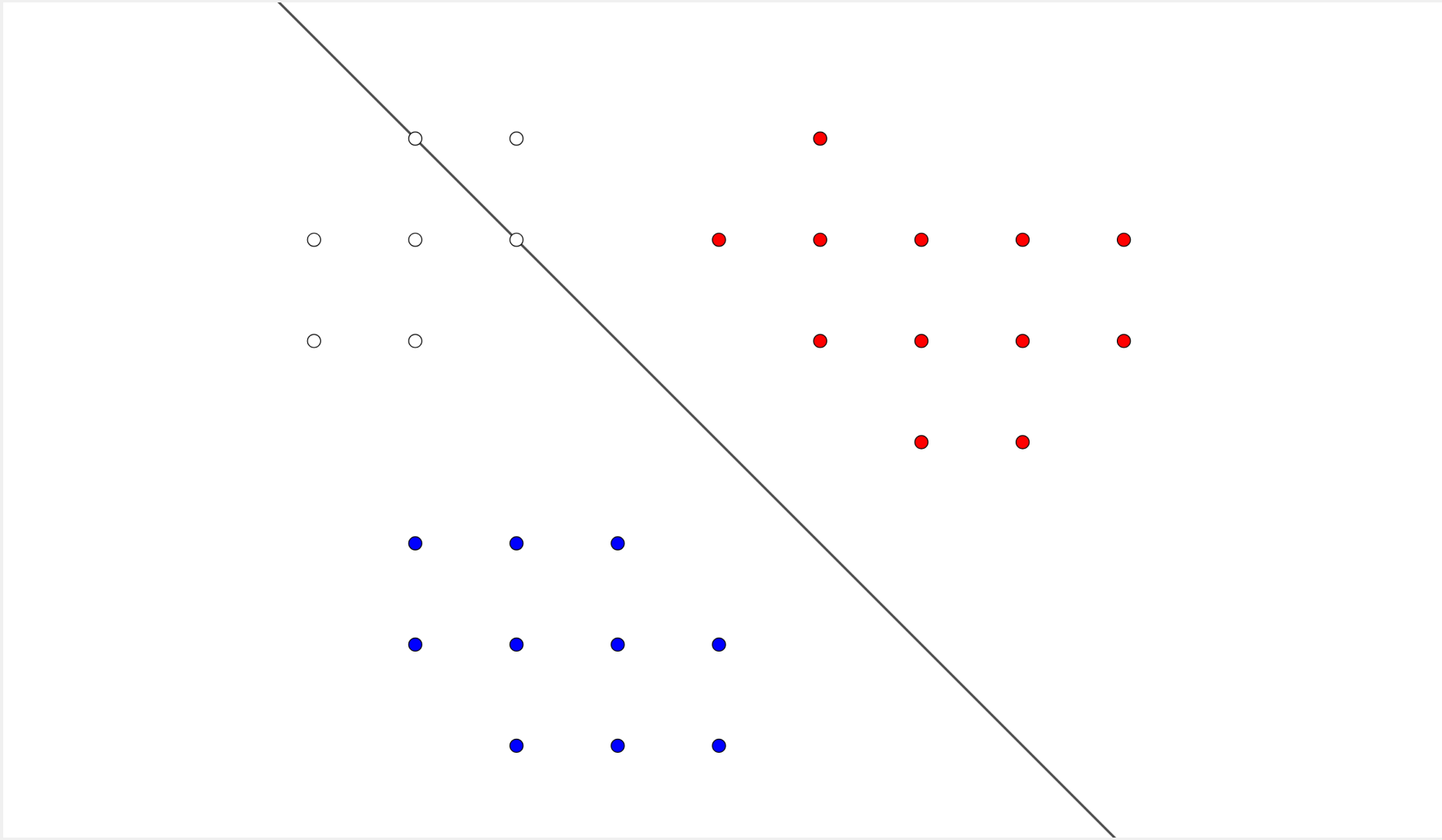
One Vs One



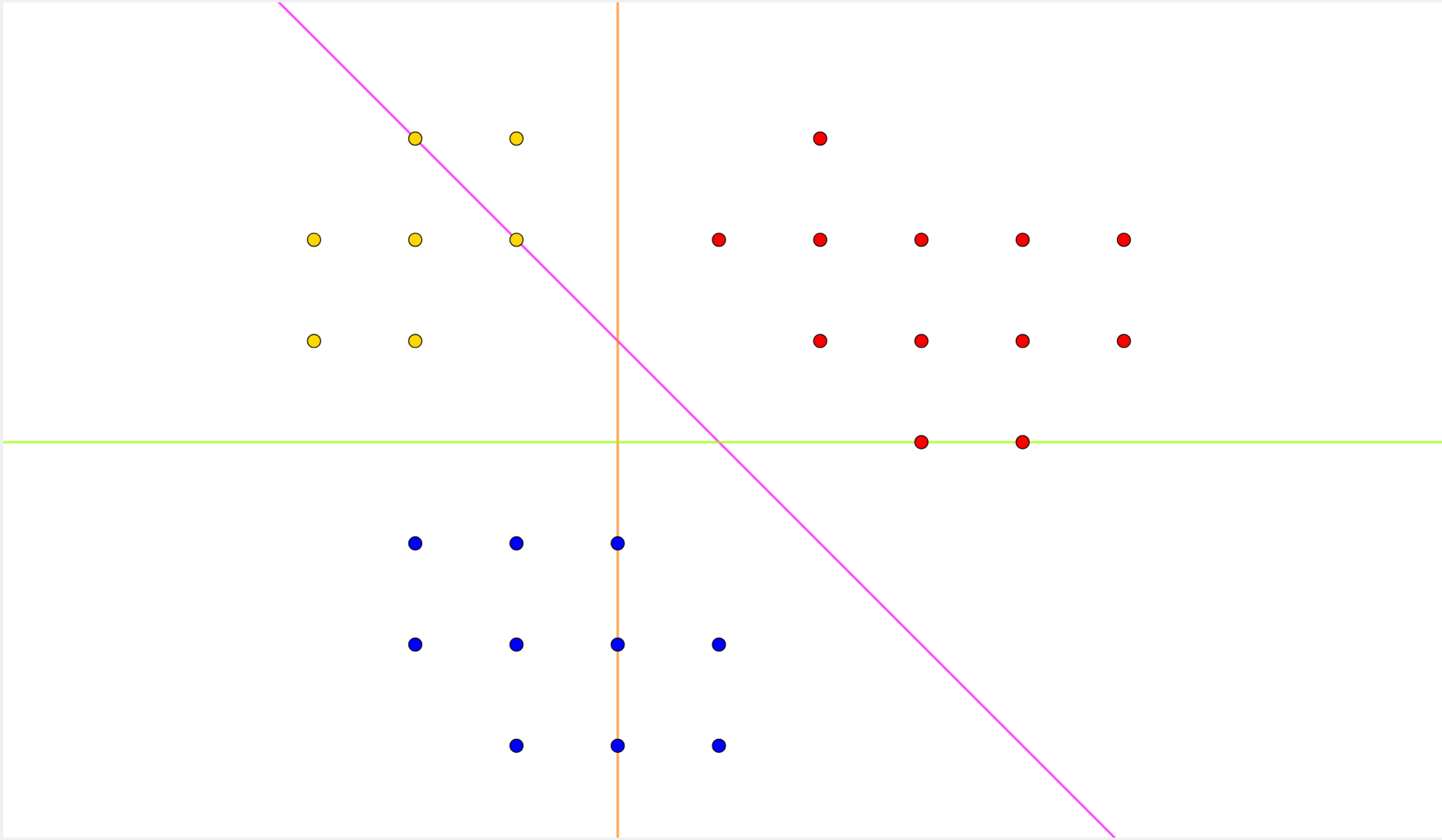
One Vs One



One Vs One



One Vs One



Global: Weston-Watkins

$$\min \left\{ \sum_{r=1}^{k} \frac{\omega_r' \omega_r}{2} + C \sum_{i=1}^{n} \sum_{j \neq y_i} \xi_i^j \right\}$$

$$\begin{aligned}
\text{s.t:} \quad & \omega_{y_i}' x_i + \omega_{y_i 0} \geq \omega_j' x_i + \omega_{j0} + 2 - \xi_i^j && \forall i = 1, \ldots, n, \; j \in \{1, \ldots, k\} \setminus y_i \\
& \xi_i^j \geq 0 && \forall i = 1, \ldots, n, \; j \in \{1, \ldots, k\} \setminus y_i \\
& \omega_r \in \mathbb{R}^p, \omega_{r0} \in \mathbb{R} && \forall r = 1, \ldots, k
\end{aligned}$$

Global: Weston-Watkins

$$\min \left\{ \sum_{r=1}^{k} \frac{\omega_r' \omega_r}{2} + C \sum_{i=1}^{n} \sum_{j \neq y_i} \xi_i^j \right\} \left(\textbf{rather than} \left(\sum_{r=1}^{k} \frac{2}{\omega_r' \omega_r} \right)^{-1} + C \sum_{i=1}^{n} \sum_{j \neq y_i} \xi_i^j \right)$$

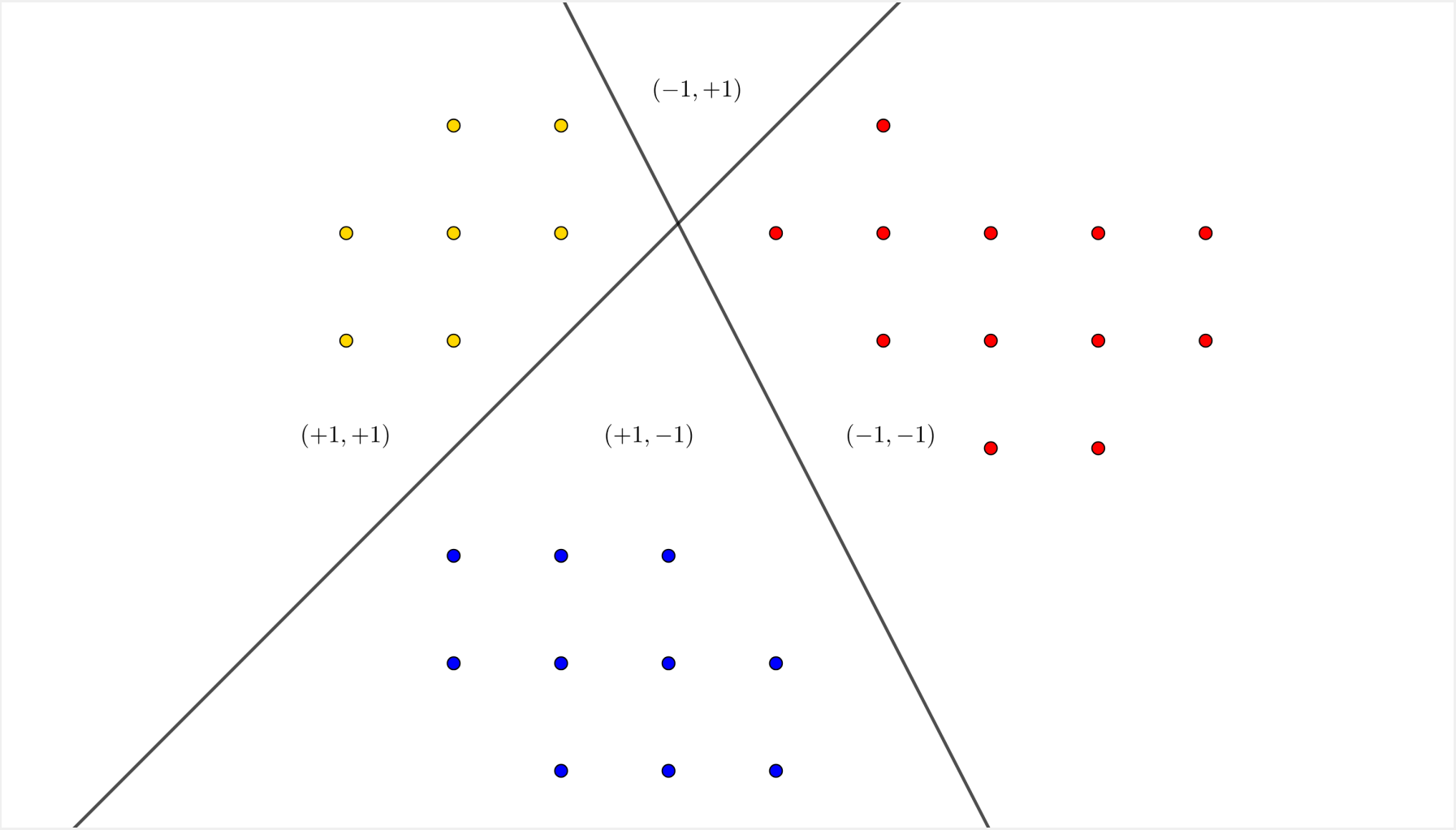
$$\begin{aligned}
\text{s.t:} \quad & \omega_{y_i}' x_i + \omega_{y_i 0} \geq \omega_j' x_i + \omega_{j0} + 2 - \xi_i^j && \forall i = 1, \ldots, n, \; j \in \{1, \ldots, k\} \setminus y_i \\
& \xi_i^j \geq 0 && \forall i = 1, \ldots, n, \; j \in \{1, \ldots, k\} \setminus y_i \\
& \omega_r \in \mathbb{R}^p, \omega_{r0} \in \mathbb{R} && \forall r = 1, \ldots, k
\end{aligned}$$

Global: Crammer-Singer

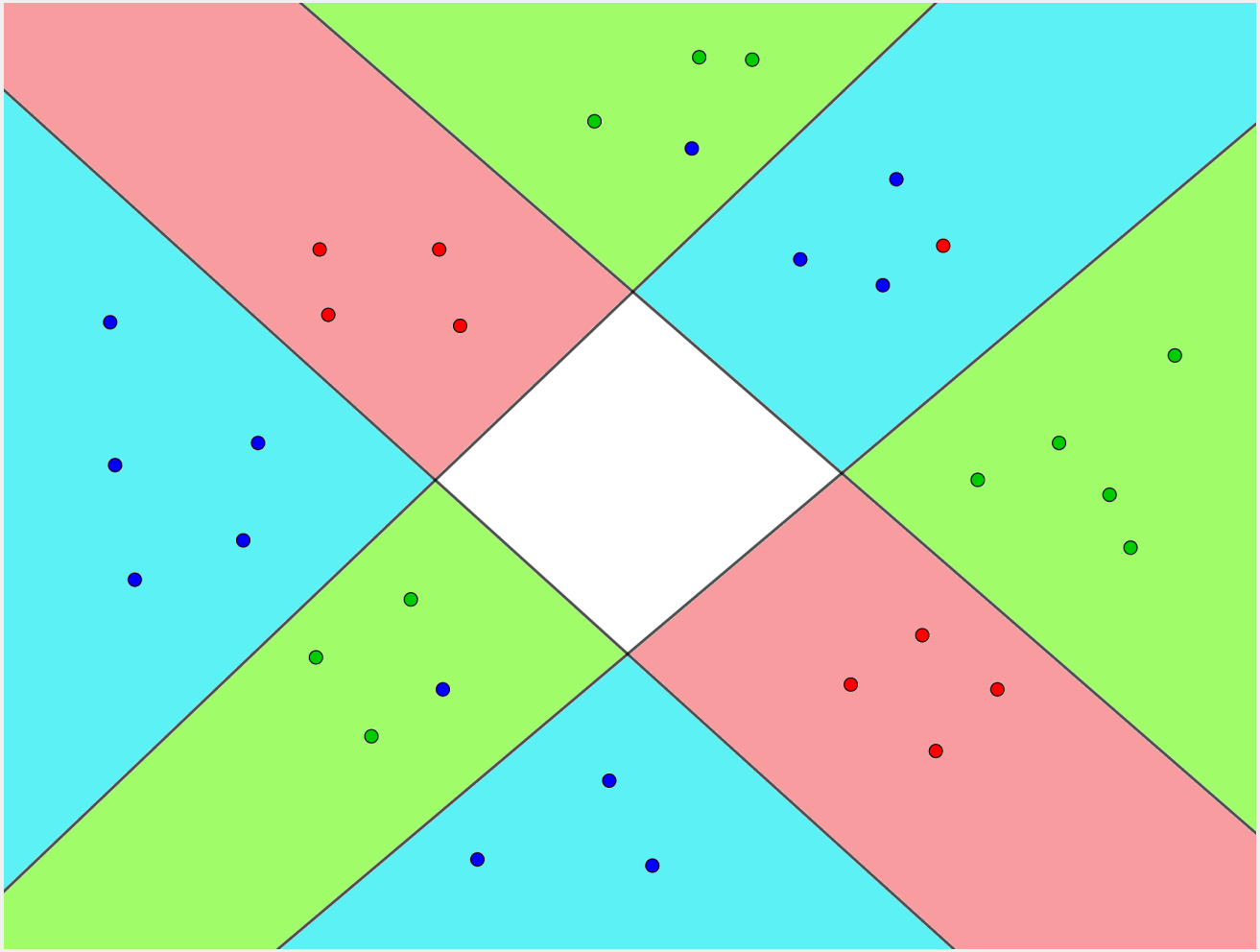
$$\min \left\{ \sum_{r=1}^{k} \frac{\omega_r' \omega_r}{2} + C \sum_{i=1}^{n} \xi_i \right\}$$

$$\begin{aligned}
\text{s.t:} \quad & \omega_{y_i}' x_i + \delta_{y_i j} - \omega_j' x_i \geq 1 - \xi_i && \forall i = 1, \ldots, n, \; j \in \{1, \ldots, k\} \\
& \xi_i \geq 0 && \forall i = 1, \ldots, n \\
& \omega_r \in \mathbb{R}^p, \omega_{r0} \in \mathbb{R} && \forall r = 1, \ldots, k \\
& \delta_{y_i j} \in \{0, 1\} && \forall y_i, j \in \{1, \ldots, k\}
\end{aligned}$$

MCSVM: The model



Class assignment through cells



Separation between classes

$$\max \min \left\{ \frac{2}{||\omega_1||_2}, \ldots, \frac{2}{||\omega_m||_2} \right\}$$

Separation between classes

$$\max \min \left\{ \frac{2}{||\omega_1||_2}, \ldots, \frac{2}{||\omega_m||_2} \right\}$$

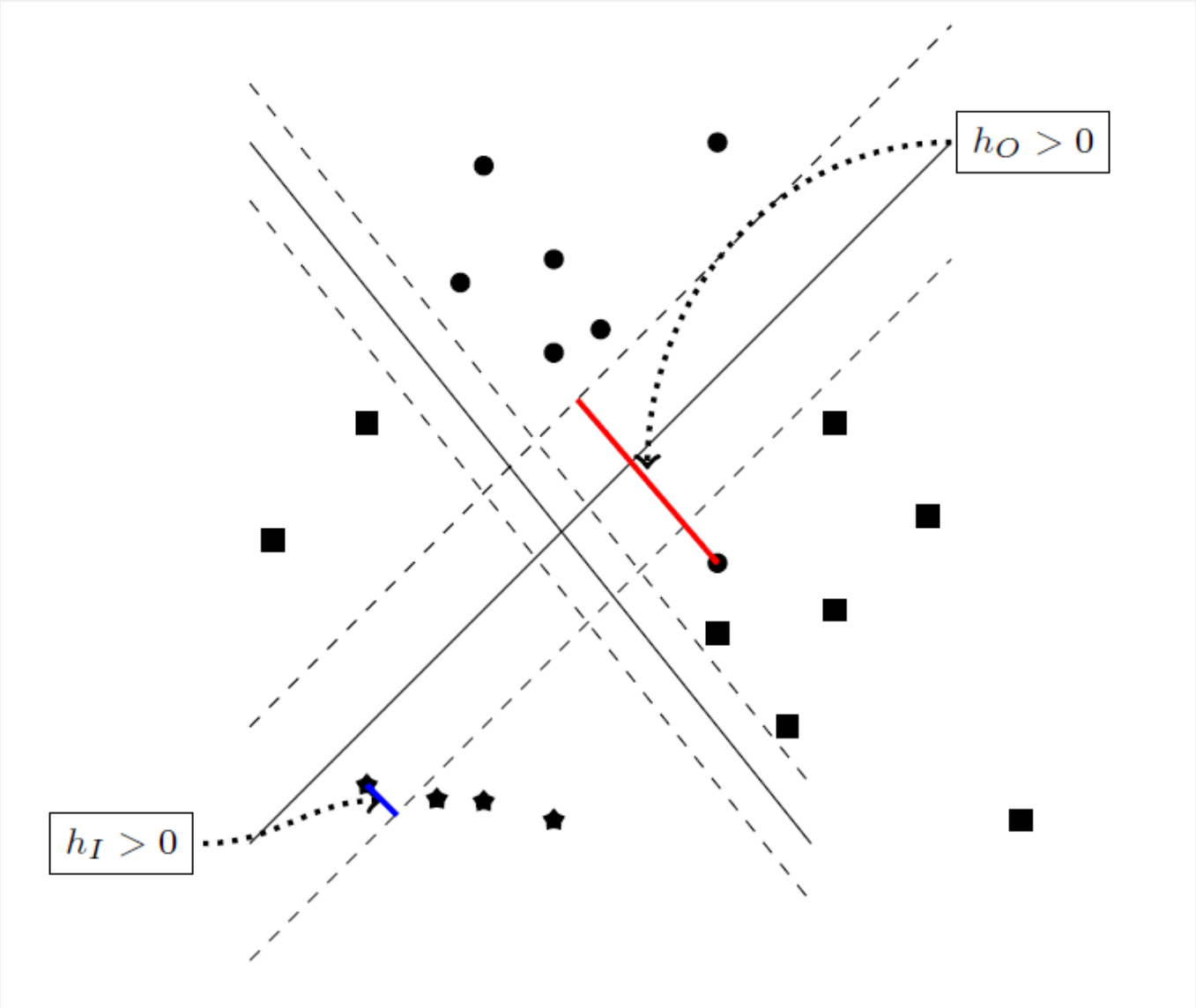
$$\min \max \left\{ \frac{1}{2}||\omega_1||_2^2, \ldots, \frac{1}{2}||\omega_m||_2^2 \right\}$$

Error functions

$$h_I\left(x, y, \mathcal{H}\right) = \begin{cases} \max\{0, \min\{1, 1 - \mathrm{s}_r(x)(\omega_r^t x + \omega_{r0})\} & \text{if } x \text{ is well classified with respect to } \mathcal{H} \\ 0 & \text{otherwise} \end{cases}$$

$$h_O\left(x, y, \mathcal{H}\right) = \begin{cases} 1 - \mathrm{s}(x)_r(\omega_r^t x + \omega_{r0}) & \text{if } x \text{ is wrong classified with respect to } \mathcal{H} \\ 0 & \text{otherwise} \end{cases}$$

Error functions



Continuous variables

$$\omega_r \in \mathbb{R}^p,\ \omega_{r0} \in \mathbb{R}, \qquad r = 1, \dots m$$

$$e_{ir} \geq 0, \qquad i = 1, \dots, n,\ \ r = 1, \dots, m$$

$$d_{ir} \geq 0, \qquad i = 1, \dots, n,\ \ r = 1, \dots, m$$

Binary variables

- $t_{ir} = \begin{cases} 1 & \text{if } \omega_r^t x_i + \omega_{r0} \geq 0 \\ 0 & \text{otherwise} \end{cases}$, $i = 1, \ldots, n,\ r = 1, \ldots, m$
- $z_{is} = \begin{cases} 1 & \text{if } i \text{ is assigned to class } s \\ 0 & \text{otherwise} \end{cases}$, $i = 1, \ldots, n,\ s = 1, \ldots, k$

Binary variables

- $\xi_i = \begin{cases} 0 & \text{if the class assigned to } i \\ & \text{coincides with } y_i \\ 1 & \text{otherwise} \end{cases}, \; i = 1, \ldots, n$

- $h_{ij} = \begin{cases} 1 & \text{if } x_j \text{ is well classified and} \\ & \text{is the representative of } x_i \\ 0 & \text{otherwise} \end{cases}, \; i, j = 1, \ldots, n \; (y_i = y_j)$

MCSVM formulation

$$\min \ \frac{1}{2}||\omega_1||_2^2 + C_1 \sum_{i=1}^{n}\sum_{r=1}^{m} e_{ir} + C_2 \sum_{i=1}^{n}\sum_{r=1}^{m} d_{ir}$$

$$\text{s.a:} \quad \frac{1}{2}||\omega_1||_2^2 \geq \frac{1}{2}||\omega_r||_2^2 \qquad \forall r = 2, \ldots, m \qquad (2)$$

$$\omega_{ir}^t x_i + w_{r_0} \geq -T(1 - t_{ir}) \qquad \forall i \in N, r \in M \qquad (3)$$

$$\omega_{ir}^t x_i + w_{r_0} \leq T t_{ir} \qquad \forall i \in N, r \in M \qquad (4)$$

$$\sum_{s=1}^{k} z_{is} = 1 \qquad \forall i \in N \qquad (5)$$

MCSVM formulation

$$||z_i - z_j||_1 \leq 2||t_i - t_j||_1 \qquad \forall i, j \in N \qquad (6)$$

$$\xi_i = \frac{1}{2}||z_i - \delta_i||_1 \qquad \forall i \in N \qquad (7)$$

$$\sum_{\substack{j \in N: \\ y_i = y_j}} h_{ij} = 1 \qquad \forall i \in N \qquad (8)$$

$$\xi_j + h_{ij} \leq 1 \qquad \forall i, j \in N (y_i = y_j) \qquad (9)$$

$$h_{ii} = 1 - \xi_i \qquad \forall i \in N \qquad (10)$$

MCSVM formulation

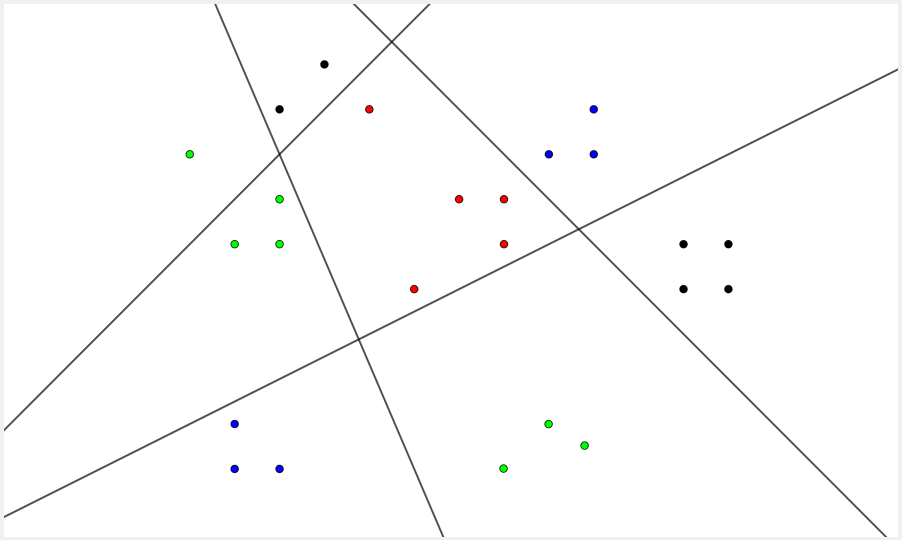
$$\omega_r^t x_i + \omega_{r0} \geq 1 - e_{ir} - T\,(3 - t_{ir} - t_{jr} - h_{ij}) \qquad \forall i, j \in N, r \in M \qquad (11)$$

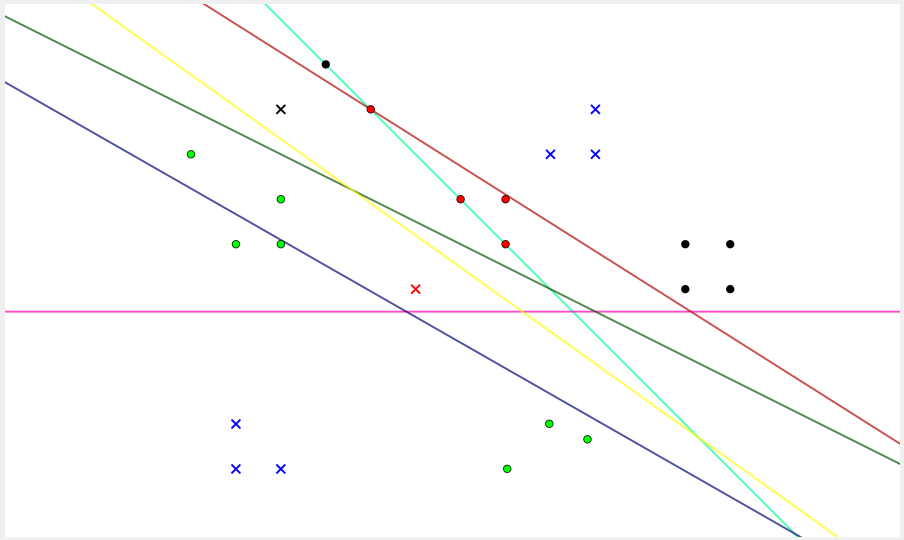
$$\omega_r^t x_i + \omega_{r0} \leq -1 + e_{ir} + T\,(1 + t_{ir} + t_{jr} - h_{ij}) \qquad \forall i, j \in N, r \in M \qquad (12)$$

$$d_{ir} \geq 1 - \omega_r^t x_i - \omega_{r0} - T\,(2 + t_{ir} - t_{jr} - h_{ij}) \qquad \forall i, j \in N, r \in M \qquad (13)$$

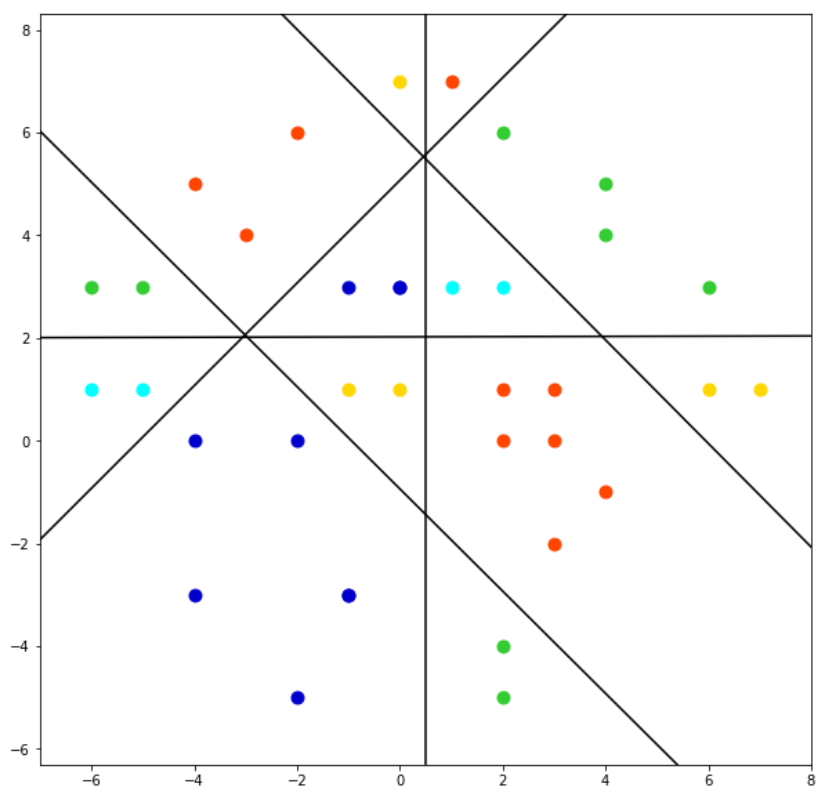
$$d_{ir} \geq 1 + \omega_r^t x_i + \omega_{r0} - T\,(2 - t_{ir} + t_{jr} - h_{ij}) \qquad \forall i, j \in N, r \in M \qquad (14)$$

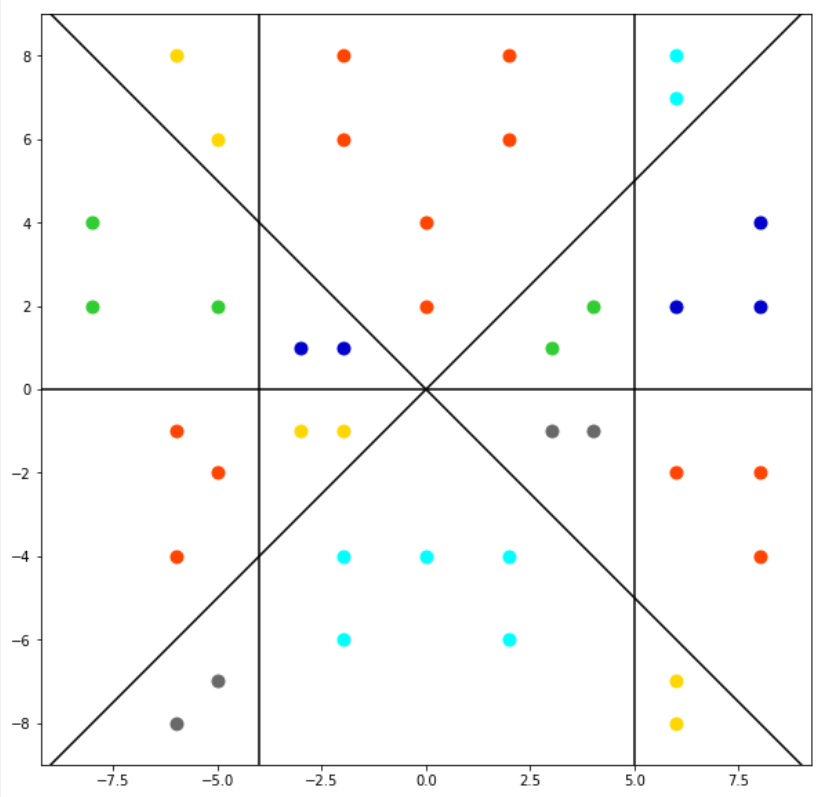
MCSVM (left) and OVO (right)





Some MCSVM examples





Kernel trick

Theorem

Let $\varphi : \mathbb{R}^p \to \mathbb{R}^d$ be a transformation of the feature space. Then, one can obtain a multiclass classifier which only depends on the original data by means of the inner products $\varphi(x_i)^t \varphi(x_j)$, for $i, j = 1, \ldots, n$.

Classification rule

- $s(x)$ the sign-pattern of x with respect to the optimal arrangement of hyperplanes.
- $J = \{j \in \{1, \ldots, n\} : \xi_j^* = 0\}$ (here ξ^* stand for the optimal vector obtained by solving the model above.

Among all the well-classified observations in the training sample, J, we assign to x the class of the one whose cell is the *closest* (less separated from x).

$$\min \sum_{j \in J} \sum_{\substack{r=1 \\ s(x_j)_r + s(x)_r = 0,}}^{m} \gamma_j \left| (\omega_r^*)^t x + \omega_{r0}^* \right|$$

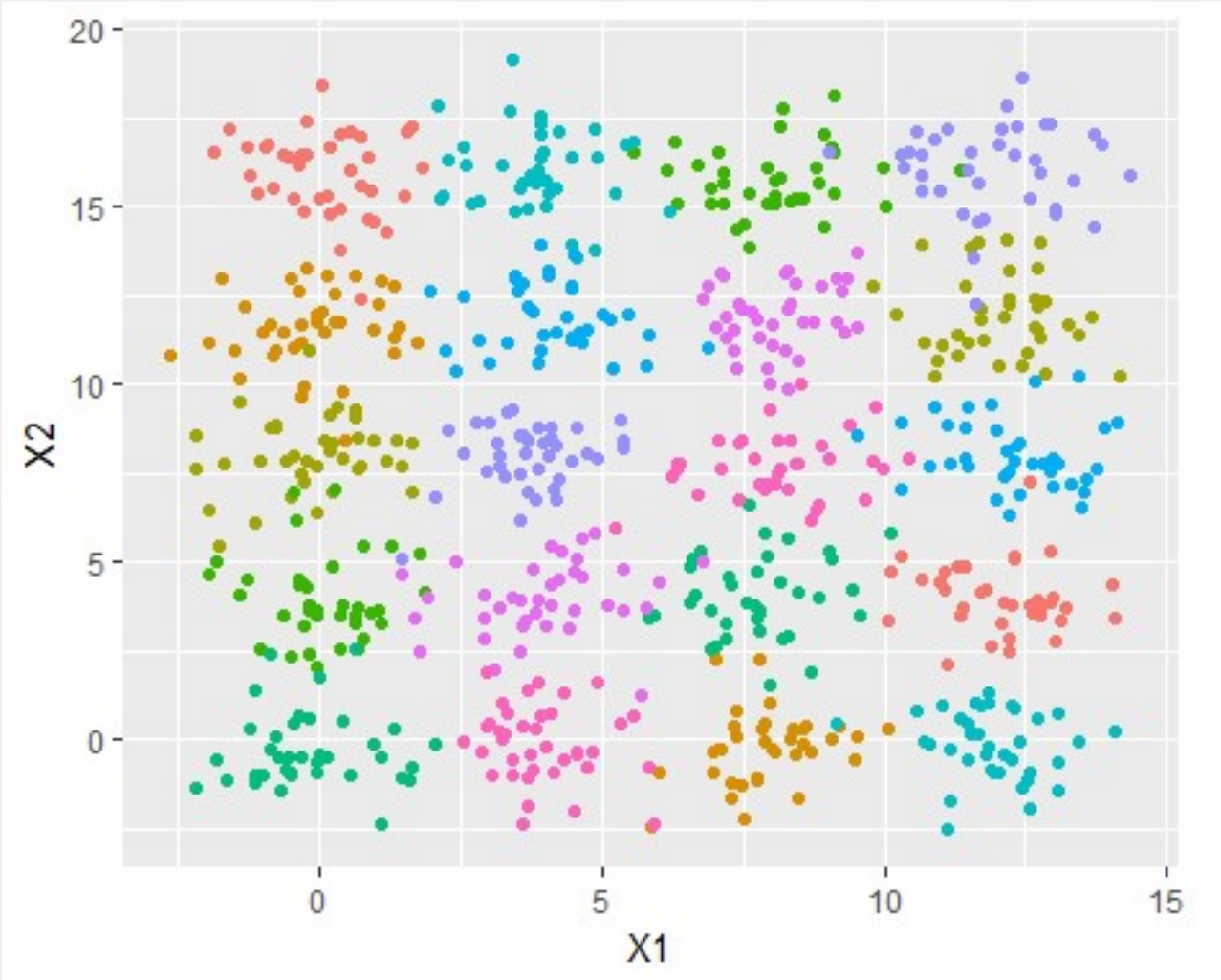
$$\text{s.t.} \sum_{j \in J}^{n} \gamma_j = 1,$$

$$\gamma_j \in \{0, 1\}, \forall j \in J$$

where $\gamma_j = \begin{cases} 1 & \text{if } x \text{ is assigned to the same cell as } x_j, \\ 0 & \text{otherwise.} \end{cases}$

Computational experiments

Synthetic data: Example of data



Computational experiments

Synthetic data: Dataset description

Dataset	n	p	k	m	π
Data 1	750	10	2	2	4
Data 2	750	10	3	3	6
Data 3	750	10	4	4	8
Data 4	750	10	4	4	8
Data 5	750	10	7	6	15
Data 6	750	10	10	8	20

Computational experiments

Synthetic data: Average accuracy results

Dataset	MCSVM	OVO	WW	CS
Data 1	**96.64**	61.85 (SVM)		
Data 2	**85.67**	41.94	43.18	39.21
Data 3	**87.46**	36.53	32.09	28.33
Data 4	**92.85**	48.88	34.07	35.852
Data 5	**90.42**	25.92	20.53	19.11
Data 6	**86.65**	29.92	16.88	15.85

Computational experiments

Real data: Dataset description

Dataset	n_{Tr}	n_{Te}	p	k	m	m_{OVO}
Forest	75	448	28	4	3	6
Glass	75	139	10	6	6	15
Iris	75	75	4	3	2	3
Seeds	75	135	7	3	2	3
Wine	75	103	13	3	2	3
Zoo	75	26	17	7	4	21

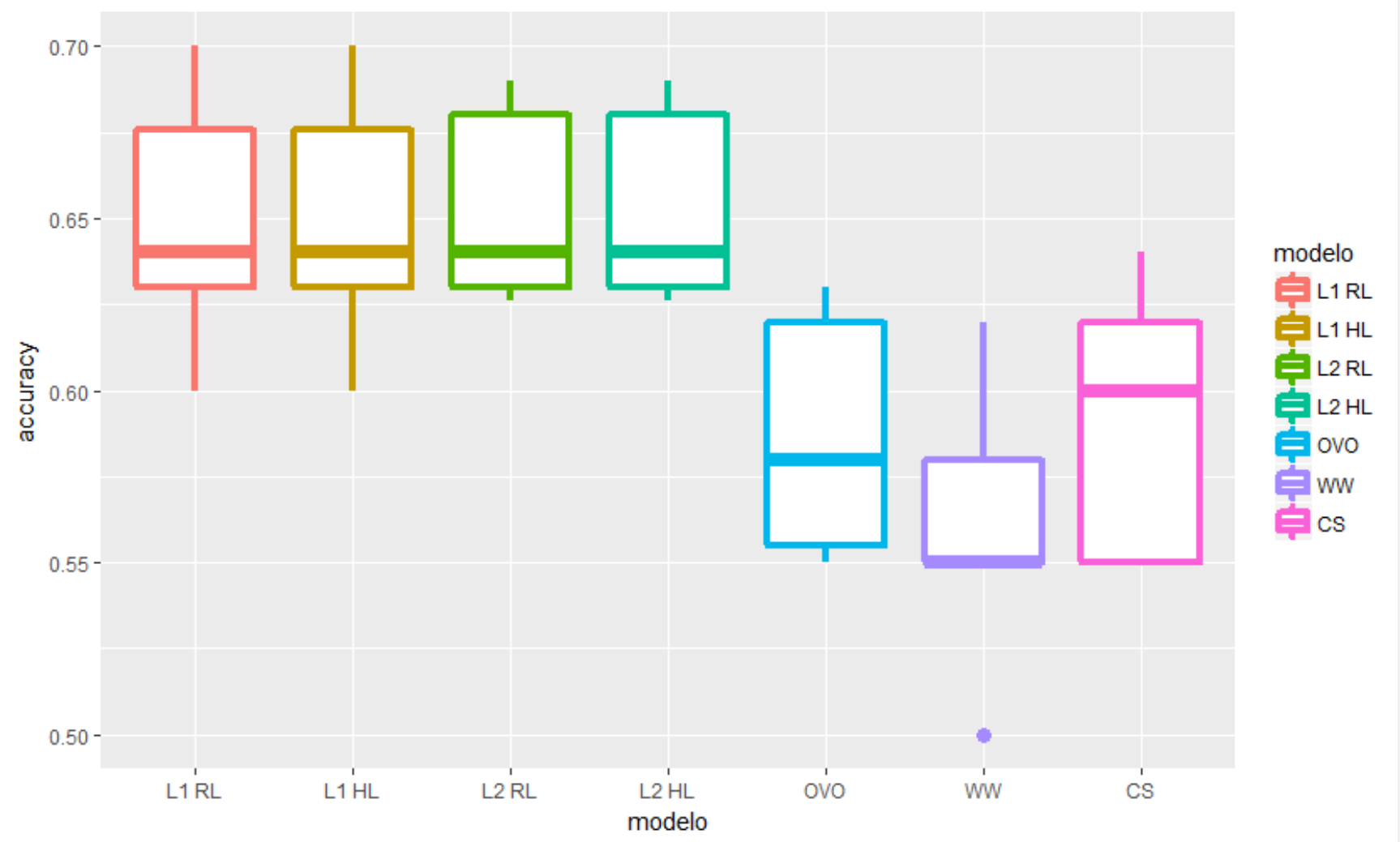
Computational experiments

Real data: Average accuracy results

Dataset	ℓ_1 RL	ℓ_1 HL	ℓ_2 RL	ℓ_2 HL	OVO	WW	CS
Forest	80.66	80.12	**82.30**	81.62	82.10	78.40	78.60
Glass	64.92	64.92	**65.32**	**65.32**	58.76	56.25	59.26
Iris	95.08	95.40	96.44	**96.66**	93.80	96.44	96.44
Seeds	**93.66**	**93.66**	93.52	93.52	91.02	93.52	93.52
Wine	95.20	95.20	**96.82**	**96.82**	96.34	96.09	96.17
Zoo	**89.75**	**89.75**	**89.75**	**89.75**	87.44	87.68	87.68

Computational experiments

Glass detailed experiment



Thank you for your attention

puerto@us.es