Learn and Interpret with MINLP

Workshop on Optimization and Data Science
Pantheon-Sorbonne University, Paris
27/03/2019

Dolores Romero Morales
Copenhagen Business School
drm.eco@cbs.dk

This project has received funding from the European Union's Horizon 2020 research and Innovation programme under the Marie Skłodowska-Curie grant agreement No. 822214

Outline

Outline

Data Science

- Data Science builds mathematical models aimed to extract and represent knowledge from complex data.
- It draws expertise from different disciplines, such as, Statistics, Mathematical Optimization, Computer Science, Information Technology.
- With a business lense, it is about supporting decision making.
- It is crucial to learn and to interpret, and to interpret and to learn again.

Mathematical Optimization for Data Science

- Linear Programming,
  e.g. Mangasarian [1965]
- Quadratic Programming,
  e.g. Duarte Silva [2017]
- Global Optimization,
  e.g. Ding and Qi [2017]
- Mixed Integer Linear Programming,
  e.g. Aloise et al. [2012]
- Mixed Integer NonLinear Programming,
  e.g. Carrizosa et al. [2016]
- Multiple Criteria Decision Aiding,
  e.g. Corrente et al. [2013]
- Robust Optimization,
  e.g. Astorino et al. [2017].

For surveys

- Bottou et al. [2016]
- Carrizosa et al. [2018e]
- Carrizosa and Romero Morales [2013]
- Duarte Silva [2017]
- Olafsson et al. [2008]
- Palagi [2019]
- Piccialli and Sciandrone [2018]

Interpretability

Interpretability

- it is desirable [Freitas, 2014], e.g., in medical diagnosis [Ustun and Rudin, 2016];
- required by regulators for models aiding, e.g., credit scoring [Baesens et al., 2003] and judicial [Ridgeway, 2013] decisions;
- from 2018 onwards the EU extends this requirement by imposing the so-called *right-to-explanation* [Goodman and Flaxman, 2016].

Sparseness ... a proxy for interpretability

Classical works

- Amaldi and Kann [1998]
- Fung and Mangasarian [2004], Mangasarian [2006]
- Guyon et al. [2002], Guyon and Elisseeff [2003]
- Roth [2004]
- Weston et al. [2001, 2003]
- Zhu et al. [2004]

Sparseness continues to attract the attention of the Mathematical Optimization community.

Atamtürk and Gomez [2019], Aytug [2015], Benítez-Peña et al. [2018a], Bertolazzi et al. [2016], Bertsimas et al. [2016], Chan et al. [2007], Cotter et al. [2013], Fountoulakis and Gondzio [2016], Gaudioso et al. [2017], Ghaddar and Naoum-Sawaya [2018], Goldberg et al. [2013], Guan et al. [2009], Maldonado and Weber [2009], Maldonado et al. [2011, 2014], Rinaldi et al. [2010], Rinaldi and Sciandrone [2010]

Other ways to enhance interpretability

Knowledge Extraction

- Finding prototypes:
  Carrizosa et al. [2007], Hart [1968], Wilfong [1992]

- Building easy-to-understand structures such as rules and trees:
  Baesens et al. [2003], Martens et al. [2007], Martens and Provost [2014], Orsenigo and Vercellis [2003, 2004]

- Enhancing the interpretability of black-box methods such as SVM:
  Carrizosa et al. [2010, 2011], Chevaleyre et al. [2013], Golea and Marchand [1993], Ustun and Rudin [2016]

- Giving meaning to latent structures by means of interpretability variables:
  Carrizosa et al. [2018f], Taeb and Chandrasekaran [2018]

Knowledge Representation

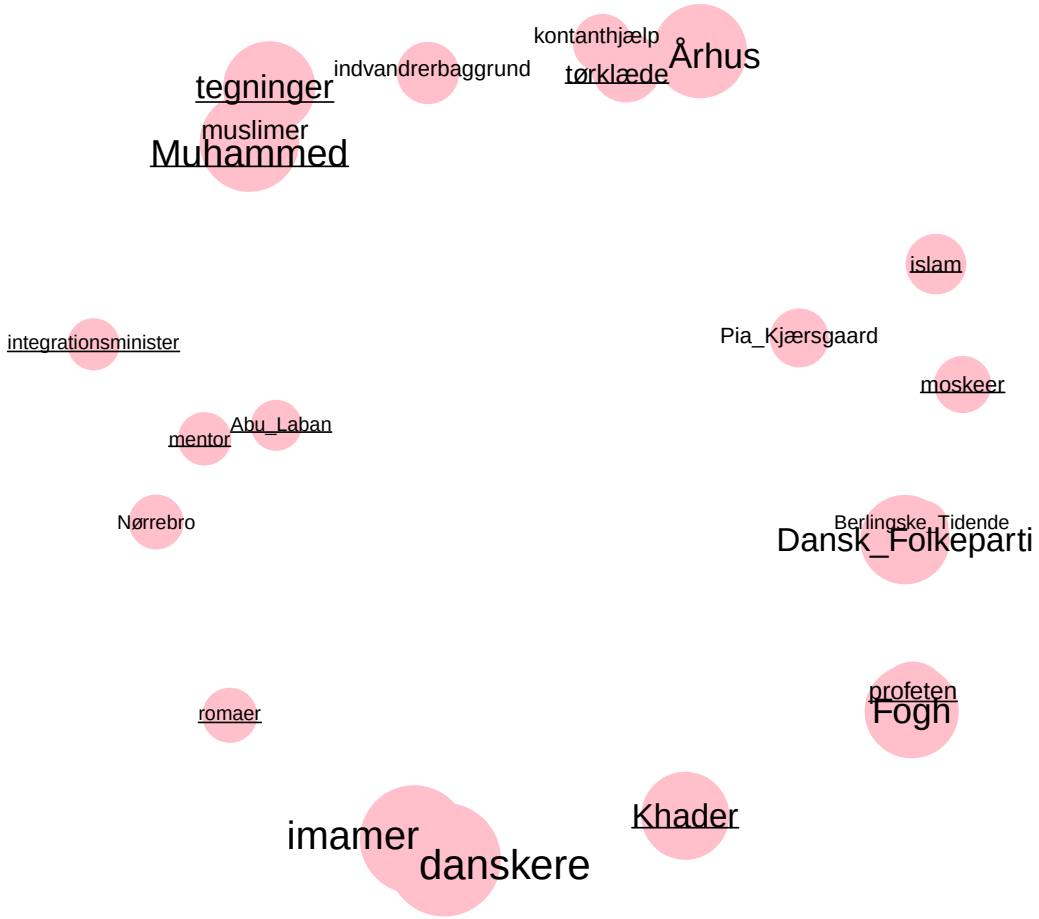
- Information Visualization:
  Carrizosa et al. [2017a, 2018b,c,d], Heer et al. [2010], Liu et al. [2014], Marron and Alonso [2014], Thomas and Wong [2004].

Knowledge Representation

We study ...

... News VIZ [Carrizosa et al., 2018a]

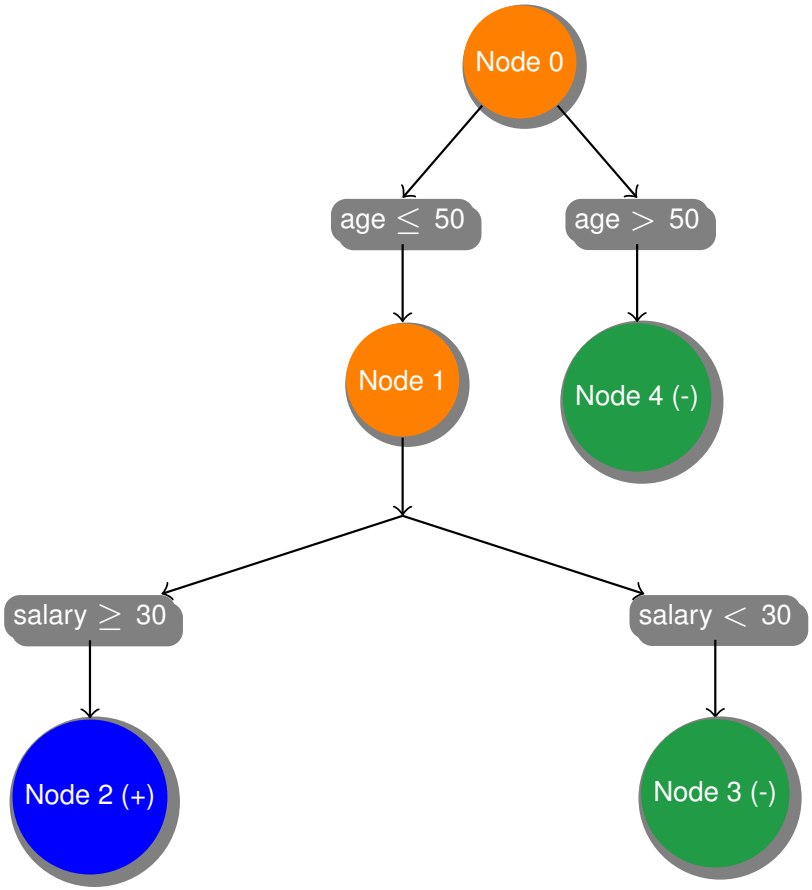
	Dansk Folkeparti	Nørrebro	tegninger	...
Dansk Folkeparti	...	...	...	
Nørrebro	...	...	...	
tegninger	...	...	...	
...	...	...	...	



Outline

Outline

Classification trees



- We have three leaf nodes
  - Node 2: age ≤ 50 and salary ≥ 30
  - Node 3: age ≤ 50 and salary < 30
  - Node 4: age > 50
- Each leaf node has an assigned class
  - Node 2: positive class
  - Node 3: negative class
  - Node 4: negative class
- New observation with age 43 and salary 34 is assigned to positive class

Optimizing Classification Trees

Optimizing Classification Trees

- Bennett [1992] optimizes the oblique cuts defining new branching nodes
- Bennett [1994] optimizes an existing tree
- Bertsimas and Dunn [2017], Günlük et al. [2018] have developed integer models to build the classification tree of a given depth.

Optimal Randomized Classification Trees (ORCT)

We propose ...

... Optimal Randomized Classification Trees [Blanquero et al., 2018a]:

- We model soft (as opposed to hard) oblique cuts

Without compromising:

- Accuracy

With the pursue of:

- Easy-to-interpret classifier
  - Small classification tree.

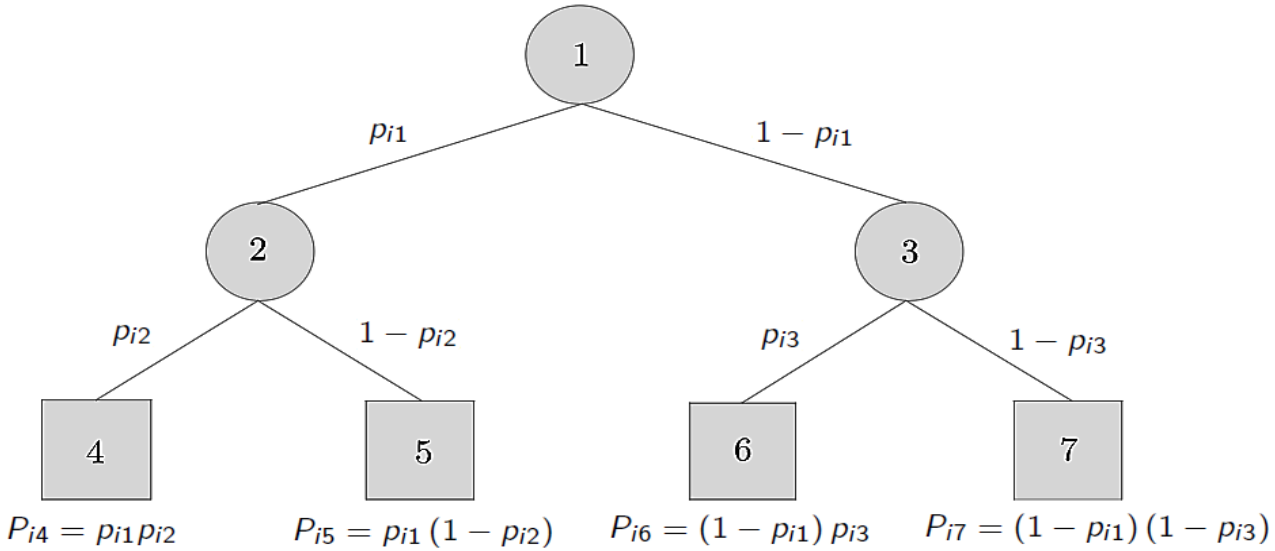
ORCT

We have at hand

- A training sample of N individuals, $I = \{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq N}$
- p predictor variables, $\boldsymbol{x}_i \in \mathbb{R}^p$
  - Wlog, $\boldsymbol{x}_i \in [0,1]^p$
- K classes, $y_i \in \{1, \ldots, K\}$
- $I_k \subset I$ the set of individuals in I in class k
- $W_{y_i k}$ the misclassification cost for classifying an individual i in class k

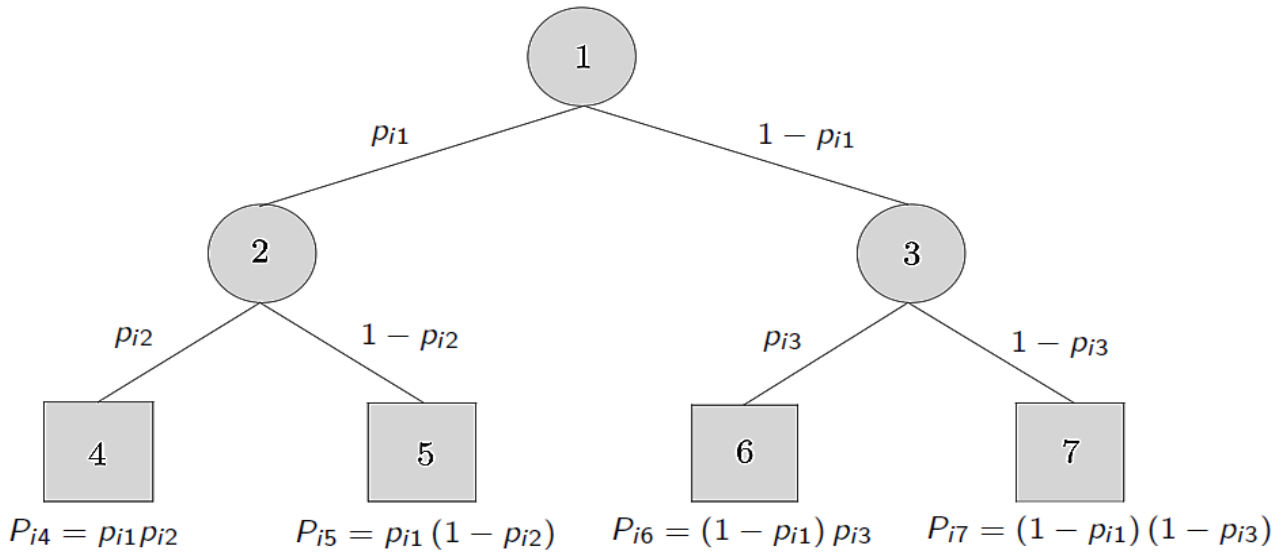
ORCT

A maximal binary tree of depth D.
Nodes: Branch $t \in \tau_B$, Leaf $t \in \tau_L$.



ORCT

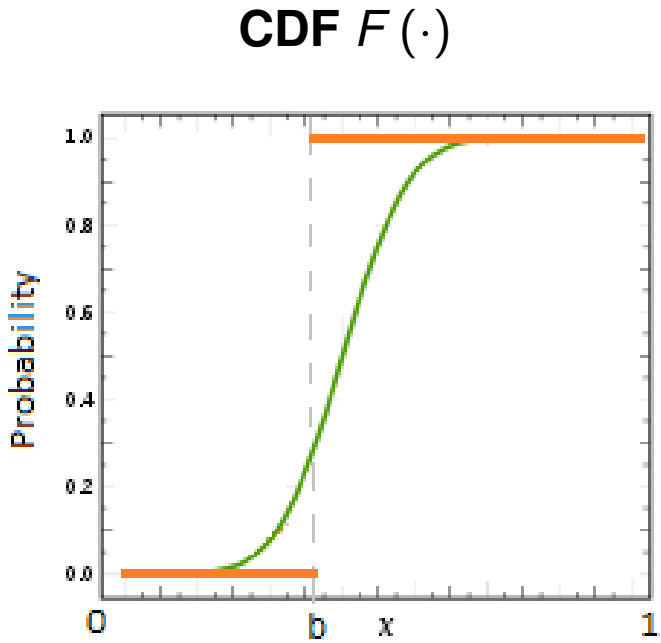
A maximal binary tree of depth D.
Nodes: Branch $t \in \tau_B$, Leaf $t \in \tau_L$.



- Oblique splits:
  - $a_{jt} \in [-1, 1]$ coefficient of predictor variable j in the oblique cut over branch node $t \in \tau_B$,
  - $\mu_t \in [-1, 1]$ location parameter at branch node $t \in \tau_B$.

ORCT

- Soft cuts



- Probabilities

$$p_{it}(\boldsymbol{a}_{\cdot t}, \mu_t) = F\left(\frac{1}{p}\sum_{j=1}^{p} a_{jt}x_{ij} - \mu_t\right), \quad i = 1, \ldots, N, \quad t \in \tau_B.$$

$$P_{it}(\boldsymbol{a}, \boldsymbol{\mu}) \equiv \mathbb{P}(\boldsymbol{x}_i \in t) = \prod_{t_l \in \mathcal{A}_L(t)} p_{it_l}(\boldsymbol{a}_{\cdot t_l}, \mu_{t_l}) \prod_{t_r \in \mathcal{A}_R(t)} (1 - p_{it_r}(\boldsymbol{a}_{\cdot t_r}, \mu_{t_r})), \quad i = 1, \ldots, N, \quad t \in \tau_L.$$

ORCT

- Each $t \in \tau_L$ is labeled with one class:

$$C_{kt} = \begin{cases} 1, & \text{node } t \text{ is labeled with class } k \\ 0, & \text{otherwise} \end{cases}, k = 1, \ldots, K, \ t \in \tau_L$$

$$\sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L.$$

- Each class $k = 1, \ldots, K$ is identified by, at least, one terminal node:

$$\sum_{t \in \tau_L} C_{kt} \geq 1, \ k = 1, \ldots, K.$$

ORCT

The formulation

$$
\begin{aligned}
\text{minimize}_{(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{C})} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\tau_L} P_{it}(\boldsymbol{a},\boldsymbol{\mu})\sum_{k=1}^{K} W_{y_i k} C_{kt} \\
\text{s.t.} \quad & \sum_{k=1}^{K} C_{kt} = 1, \;\; t \in \tau_L, \\
& \sum_{t\in\tau_L} C_{kt} \geq 1, \;\; k = 1,\ldots,K, \\
& a_{jt} \in [-1,1], \;\; j = 1,\ldots,p, \;\; t \in \tau_B, \\
& \mu_t \in [-1,1], \;\; t \in \tau_B \\
& C_{kt} \in \{0,1\}, \;\; k = 1,\ldots,K, \;\; t \in \tau_L
\end{aligned}
\qquad \text{(ORCT)}
$$

ORCT

Without loss of optimality, we can relax the integrality constraints

$$
\begin{aligned}
\text{minimize}_{(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{c})} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\tau_L} P_{it}\left(\boldsymbol{a},\boldsymbol{\mu}\right)\sum_{k=1}^{K} W_{y_i k}\, C_{kt} \\
\text{s.t.} \quad & \sum_{k=1}^{K} C_{kt} = 1, \;\; t \in \tau_L, \\
& \sum_{t\in\tau_L} C_{kt} \geq 1, \;\; k = 1,\ldots,K, \\
& a_{jt} \in [-1,1], \;\; j = 1,\ldots,p, \;\; t \in \tau_B, \\
& \mu_t \in [-1,1], \;\; t \in \tau_B, \\
& C_{kt} \in [0,1], \;\; k = 1,\ldots,K, \;\; t \in \tau_L
\end{aligned}
\qquad \text{(ORCT)}
$$

Theoretical Properties

Let $\{F_\gamma\}_{\gamma>0}$ be a family of CDFs such that

$$\lim_{\gamma\to\infty} F_\gamma(\cdot) = \begin{cases} 1, & \text{if } (\cdot) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Example: the logistic CDF family

$$F_\gamma(\cdot) = \frac{1}{1+\exp(-(\cdot)\gamma)}, \ \gamma > 0$$

Proposition. We have $\lim_{\gamma\to\infty} \text{ORCT}(\gamma) = \text{ODCT}$.

Theoretical Properties

ORCT trained with N observations can be seen as the Sample Average Approximation (SAA) problem to a true stochastic problem

- Let $\hat{\vartheta}_N$ and $\hat{\mathcal{S}}_N$ be the SAA estimators for the objective value and the optimal solution vector
- Let ϑ^* and $\mathcal{S}$ be the counterparts for the true problem

Theorem. $\hat{\vartheta}_N$ and $\hat{\mathcal{S}}_N$ are consistent estimators of ϑ^* and $\mathcal{S}$, respectively, in the sense of Shapiro et al. [2009].

ORCTs with performance constraints

ρ_k: Correct Classification Rate over the $k-$th class

$$\frac{1}{|I_k|} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it}(\boldsymbol{a}, \boldsymbol{\mu}) C_{kt} \geq \rho_k, \quad k = 1, \ldots, K,$$

ORCTs with performance constraints

ρ_k: Correct Classification Rate over the $k-$th class

$$\text{minimize}_{(\boldsymbol{a},\mu,\boldsymbol{c})} \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\tau_L} P_{it}(\boldsymbol{a},\mu)\sum_{k=1}^{K} W_{y_i k} C_{kt}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L,$$

$$\sum_{t\in\tau_L} C_{kt} \geq 1, \ k = 1,\ldots,K,$$

$$\frac{1}{|I_k|}\sum_{i\in I_k}\sum_{t\in\tau_L} P_{it}(\boldsymbol{a},\mu)\, C_{kt} \geq \rho_k, \ k = 1,\ldots,K,$$

$$a_{jt} \in [-1,1], \ j = 1,\ldots,p, \ t \in \tau_B,$$

$$\mu_t \in [-1,1], \ t \in \tau_B,$$

$$C_{kt} \in [0,1], \ k = 1,\ldots,K, \ t \in \tau_L$$

Sparse Optimal Randomized Classification Trees (S-ORCT)

We propose ...

... Sparse Optimal Randomized Classification Trees [Blanquero et al., 2018b]:

- We model soft (as opposed to hard) oblique cuts

Without compromising:

- Accuracy

With the pursue of:

- Easy-to-interpret classifier
  - Small classification tree
  - Globally sparse: few variables in the tree
  - Locally sparse: few variables in each node

S-ORCT

The formulation

$$
\begin{aligned}
\text{minimize}_{(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{C})} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\tau_L} P_{it}\left(\boldsymbol{a},\boldsymbol{\mu}\right)\sum_{k=1}^{K} W_{y_i k}\, C_{kt} \\
& + \lambda^L \sum_{j=1}^{p} \|\boldsymbol{a}_{j\cdot}\|_1 \quad \text{(local sparsity)} \\
& + \lambda^G \sum_{j=1}^{p} \|\boldsymbol{a}_{j\cdot}\|_\infty \quad \text{(global sparsity)} \\
\text{s.t.} \quad & \sum_{k=1}^{K} C_{kt} = 1, \;\; t \in \tau_L, \\
& \sum_{t\in\tau_L} C_{kt} \geq 1, \;\; k = 1,\ldots,K, \\
& a_{jt} \in [-1,1], \;\; j = 1,\ldots,p, \;\; t \in \tau_B, \\
& \mu_t \in [-1,1], \;\; t \in \tau_B, \\
& C_{kt} \in [0,1], \;\; k = 1,\ldots,K, \;\; t \in \tau_L
\end{aligned}
\qquad \text{(S-ORCT)}
$$

Theoretical Properties

Let

$$g(\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tau_L} P_{it}(\boldsymbol{a}, \boldsymbol{\mu}) \sum_{k=1}^{K} W_{y_i k} C_{kt}$$

Theorem. Let $\sigma \in [0, 1]$. For

$$\lambda^L \geq (1 - \sigma) \max_{\substack{\boldsymbol{\mu} \in [-1,1]^{|\tau_B|} \\ \boldsymbol{C} \in \{0,1\}^{K \times |\tau_L|}}} \max_{j=1,\ldots,p} \left\| \nabla_{\boldsymbol{a}_{j\cdot}} g(\boldsymbol{0}, \boldsymbol{\mu}, \boldsymbol{C}) \right\|_\infty, \text{ and}$$

$$\lambda^G \geq \sigma \max_{\substack{\boldsymbol{\mu} \in [-1,1]^{|\tau_B|} \\ \boldsymbol{C} \in \{0,1\}^{K \times |\tau_L|}}} \max_{j=1,\ldots,p} \left\| \nabla_{\boldsymbol{a}_{j\cdot}} g(\boldsymbol{0}, \boldsymbol{\mu}, \boldsymbol{C}) \right\|_1,$$

$(\boldsymbol{0}, \boldsymbol{\mu}, \boldsymbol{C})$ is a stationary point of the S-ORCT

Numerical results

- Operating System: Microsoft Windows 10, 64 bits
- Processor: Intel® Core™ i7-2600 CPU 3.40GHz
- RAM: 16 GB
- Interface: Python 3.5 (64 bits)
- Solver: IPOPT 3.11.1

Datasets

Table: Information about the data sets tested.

Data set	Abbreviation	N	p	K	Class distribution
Connectionist-bench-sonar	Sonar	208	60	2	55% - 45%
Breast-cancer-Wisconsin	Wisconsin	569	30	2	63% - 37%
Credit-approval	Creditapproval	653	37	2	55% - 45%
Pima-indians-diabetes	Pima	768	8	2	65% - 35%
Statlog-project-German-credit	Germancredit	1000	48	2	70% - 30%
Ozone-level-detection-one	Ozone	1848	72	2	97% - 3%
Spambase	Spam	4601	57	2	61% - 39%
Iris	Iris	150	4	3	33.3%-33.3%-33.3%
Wine	Wine	178	13	3	40%-33%-27%
Seeds	Seeds	210	7	3	33.3%-33.3%-33.3%
Thyroid-disease-ann-thyroid	Thyroid	3772	21	3	92.5%-5%-2.5%
Car-evaluation	Car	1728	15	4	70%-22%-4%-4%

Numerical results for ORCT

Table: Results for $D = 1$ in terms of the out-of-sample accuracy.

Data set	Average time (in secs)	Out-of-sample accuracy			
		ORCT	CART	OCT-H	RF
Connectionist-bench-sonar	12	**75.2**	70.0	70.4	83.1
Wisconsin	14	**96.4**	92.0	93.1	95.5
Credit-approval	12	83.6	85.7	**87.9**	86.7
Pima-indians-diabetes	9	**76.2**	74.2	71.6	76.3
Statlog-project-German-credit	15	**72.8**	72.1	71.6	75.2
Ozone-level-detection-one	79	96.7	95.6	**96.8**	96.4
Spambase	58	**89.8**	89.2	83.6	95.1

Numerical results for ORCT

Table: Results for $D = 2$ in terms of the out-of-sample accuracy.

Data set	Average time (in secs)	Out-of-sample accuracy			
		ORCT	CART	OCT-H	RF
Iris	7.2	**95.9**	92.7	95.1	95.4
Wine	12.4	**96.4**	88.6	91.1	98.6
Seeds	10.7	**94.0**	90.2	90.6	92.5
Thyroid-disease-ann-thyroid	122.4	92.4	**99.1**	92.5	99.1
Car-evaluation	59.2	**91.4**	88.1	87.5	88.0

Numerical results for ORCT

Table: Average number of leaf nodes per tree.

Data set	CART	RF
Connectionist-bench-sonar	6.4	18.7
Wisconsin	4.6	17.1
Credit-approval	6.8	72.2
Pima-indians-diabetes	14.9	98.1
Statlog-project-German-credit	17.2	173.0
Ozone-level-detection-one	9.3	34.8
Spambase	8.9	252.0
Iris	3	7.3
Wine	4.1	9.7
Seeds	3.7	13.2
Thyroid-disease-ann-thyroid	6.7	62.1
Car-evaluation	18.3	97.6

ORCTs with performance constraints

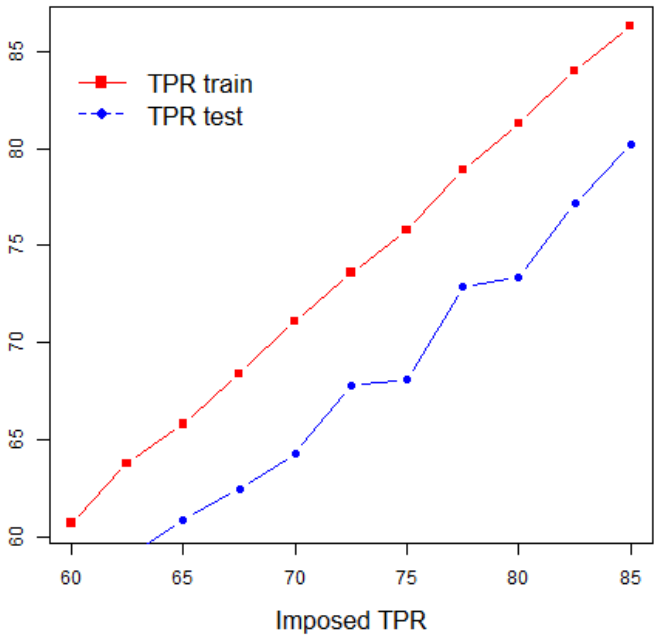
Pima-indians-diabetes dataset

- **8 characteristics** measured on **768 patients of diabetes** from two different classes: *Positive* (35%) and *Negative* (65%).
- ORCT: $\mathbf{TPR_{train}}$ 60.7 $\mathbf{TPR_{test}}$ 55.6 $\mathbf{TNR_{train}}$ 90.5 $\mathbf{TNR_{test}}$ 87.7
- ORCT with performance constraints: $\rho_{\text{Negative}} = 0$ and ρ_{Positive} varying.

ORCTs with performance constraints

Pima-indians-diabetes dataset

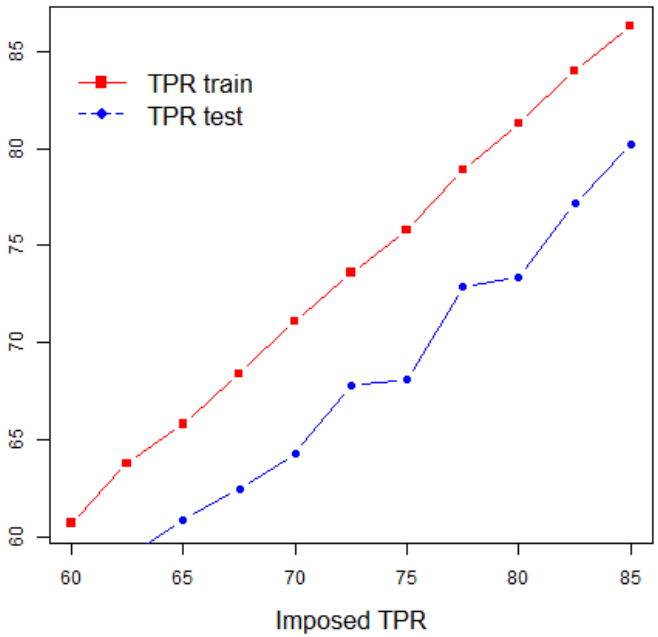
- **8 characteristics** measured on **768 patients of diabetes** from two different classes: *Positive* (35%) and *Negative* (65%).
- ORCT: $\mathbf{TPR_{train}}$ 60.7 $\mathbf{TPR_{test}}$ 55.6 $\mathbf{TNR_{train}}$ 90.5 $\mathbf{TNR_{test}}$ 87.7
- ORCT with performance constraints: $\rho_{\text{Negative}} = 0$ and ρ_{Positive} varying.

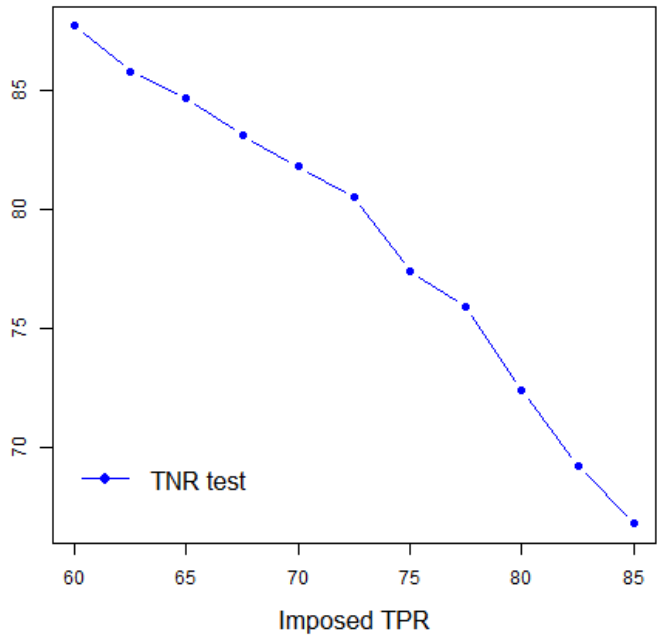


ORCTs with performance constraints

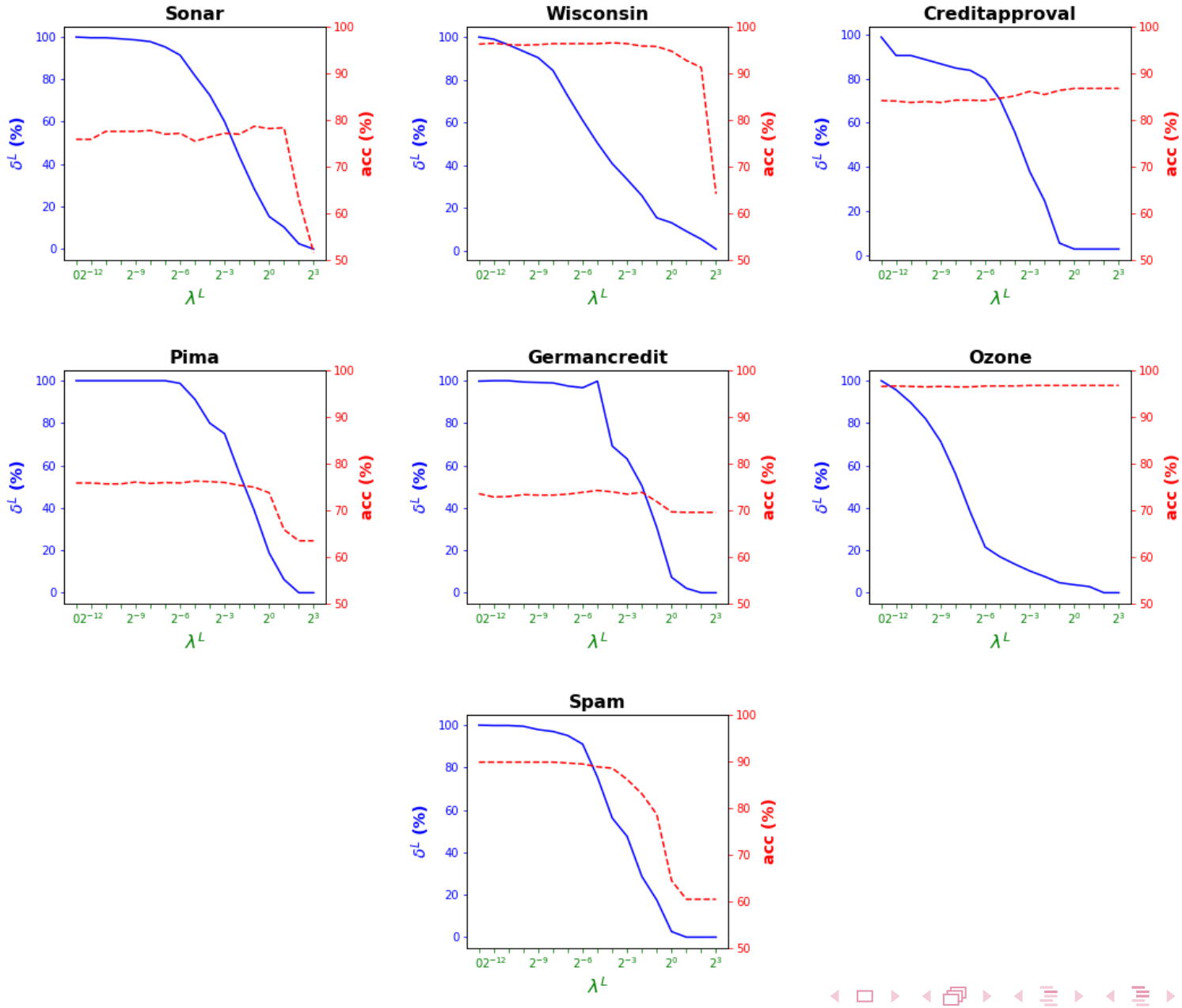
Pima-indians-diabetes dataset

- **8 characteristics** measured on **768 patients of diabetes** from two different classes: *Positive* (35%) and *Negative* (65%).
- ORCT: $\mathbf{TPR_{train}}$ 60.7 $\mathbf{TPR_{test}}$ 55.6 $\mathbf{TNR_{train}}$ 90.5 $\mathbf{TNR_{test}}$ 87.7
- ORCT with performance constraints: $\rho_{\text{Negative}} = 0$ and ρ_{Positive} varying.

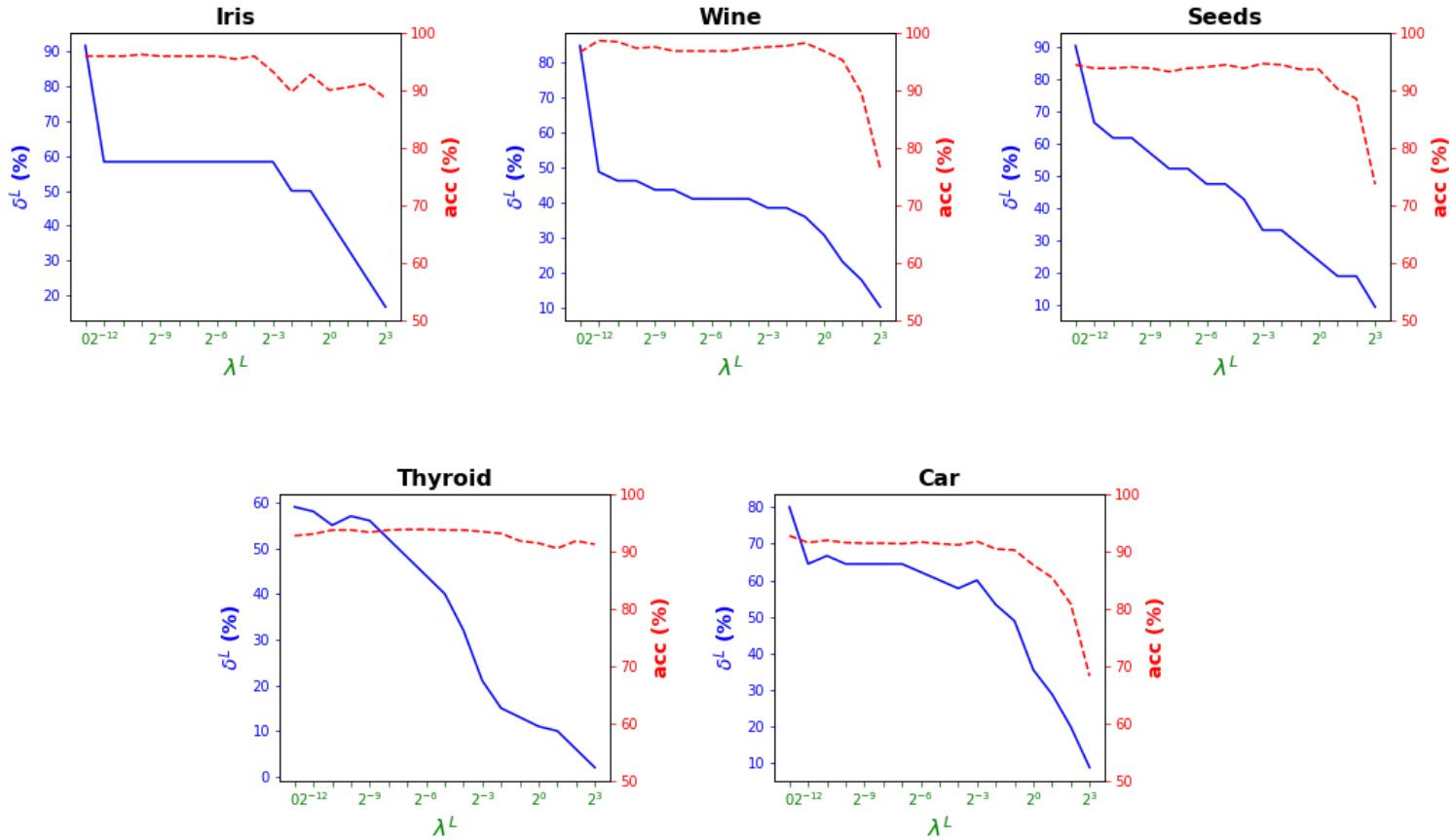




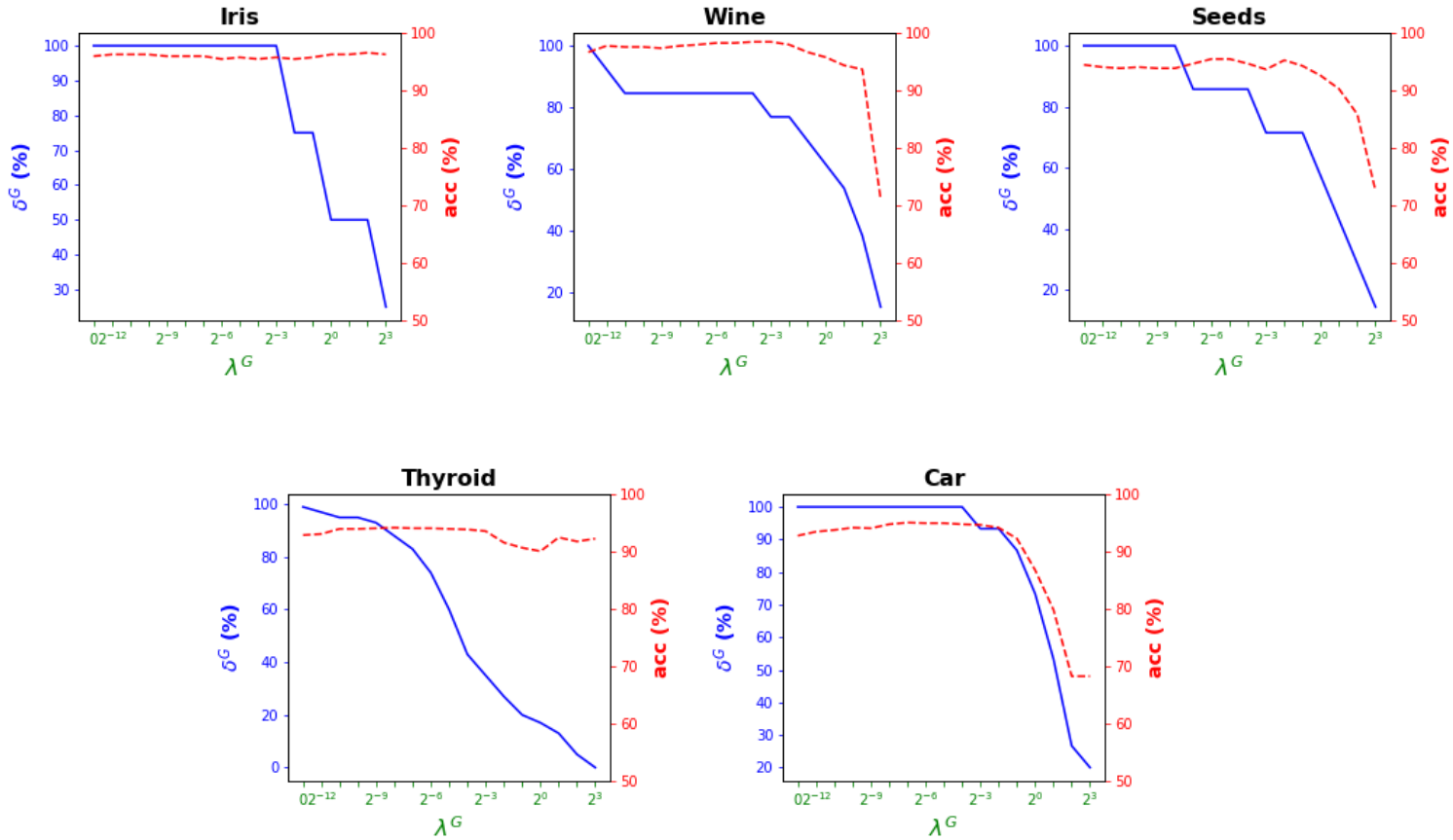
Numerical results for S-ORCT: accuracy vs local sparsity, $D = 1$



Numerical results for S-ORCT: accuracy vs local sparsity, $D = 2$

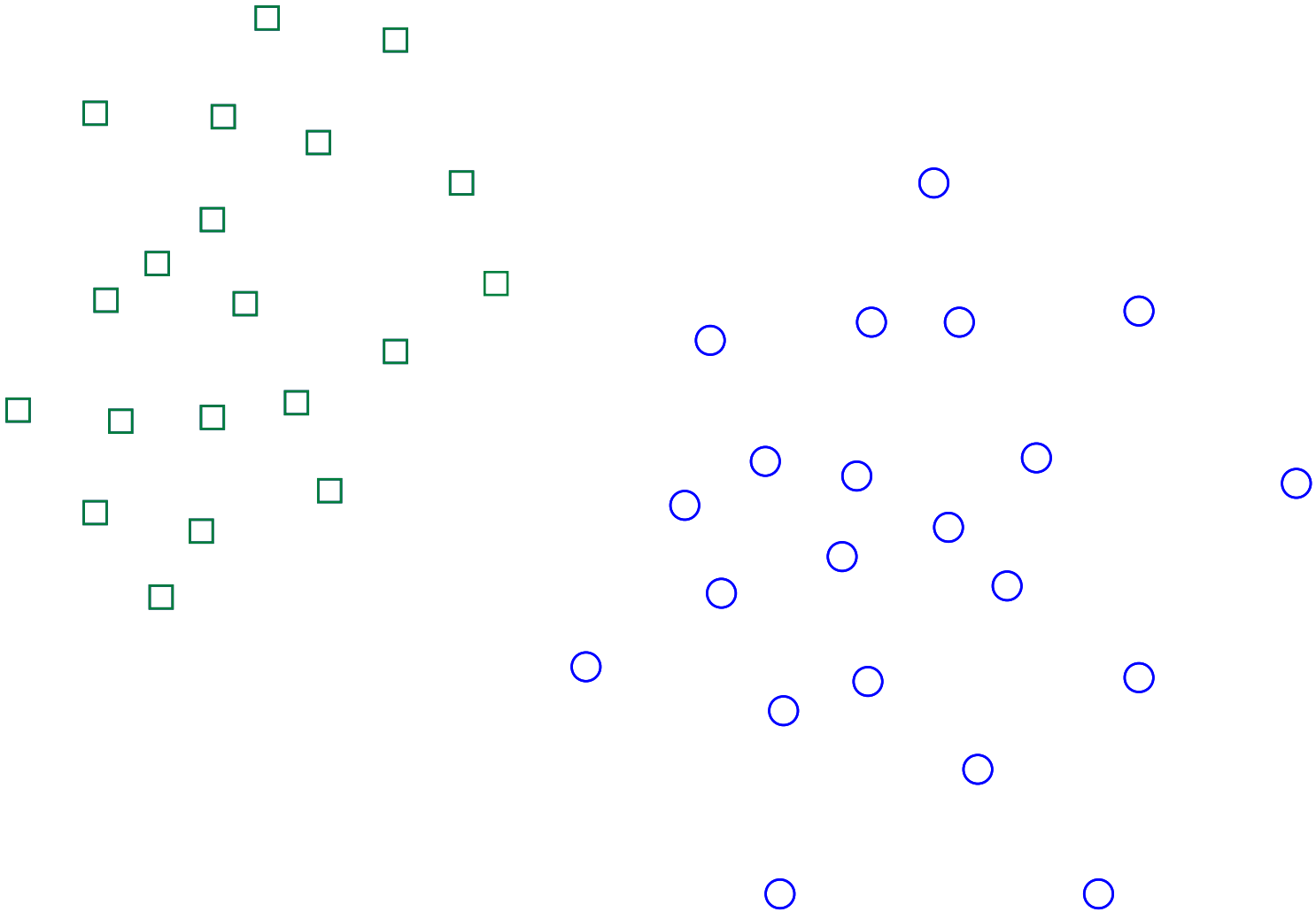


Numerical results for S-ORCT: accuracy vs global sparsity, $D = 2$

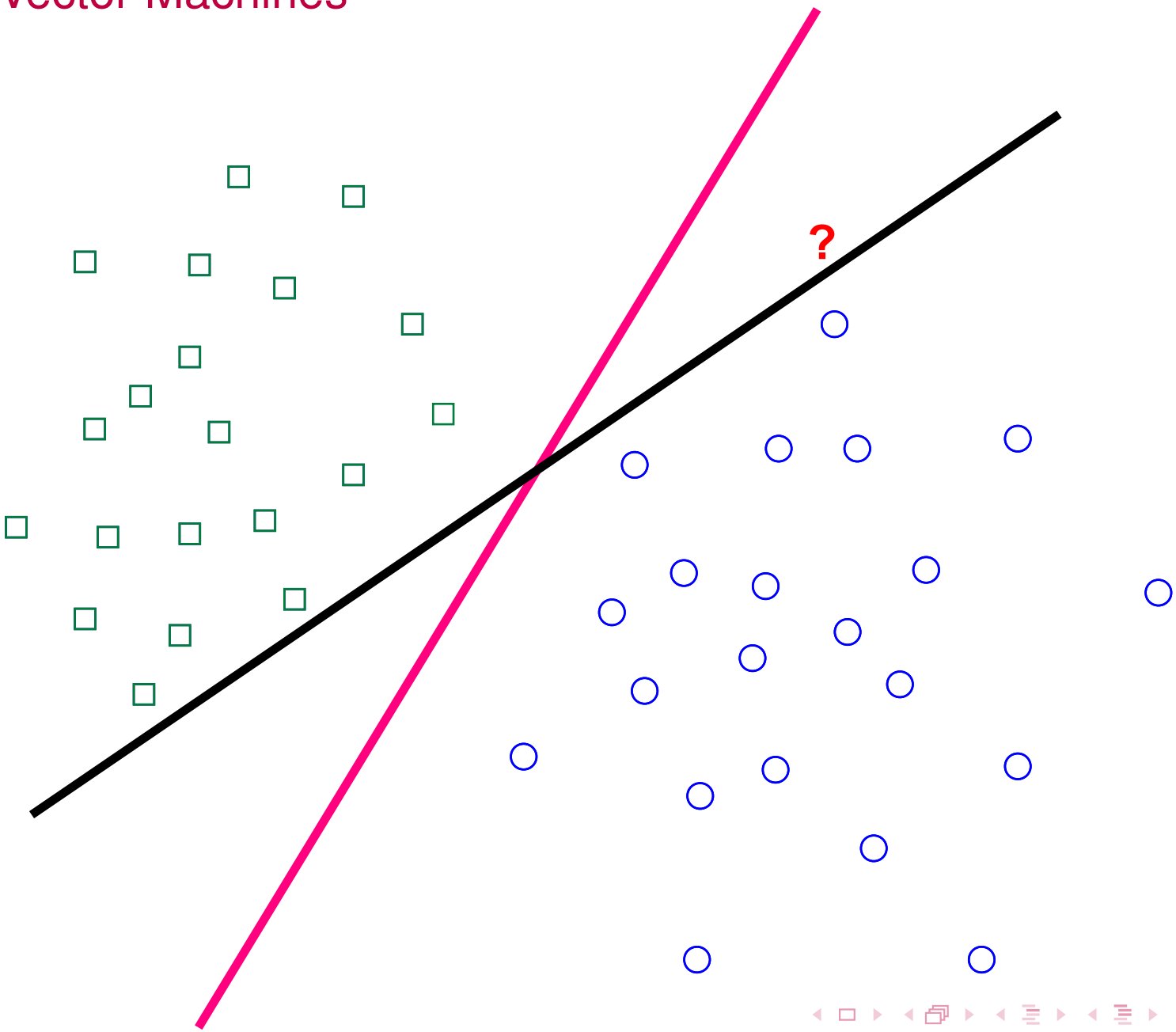


Outline

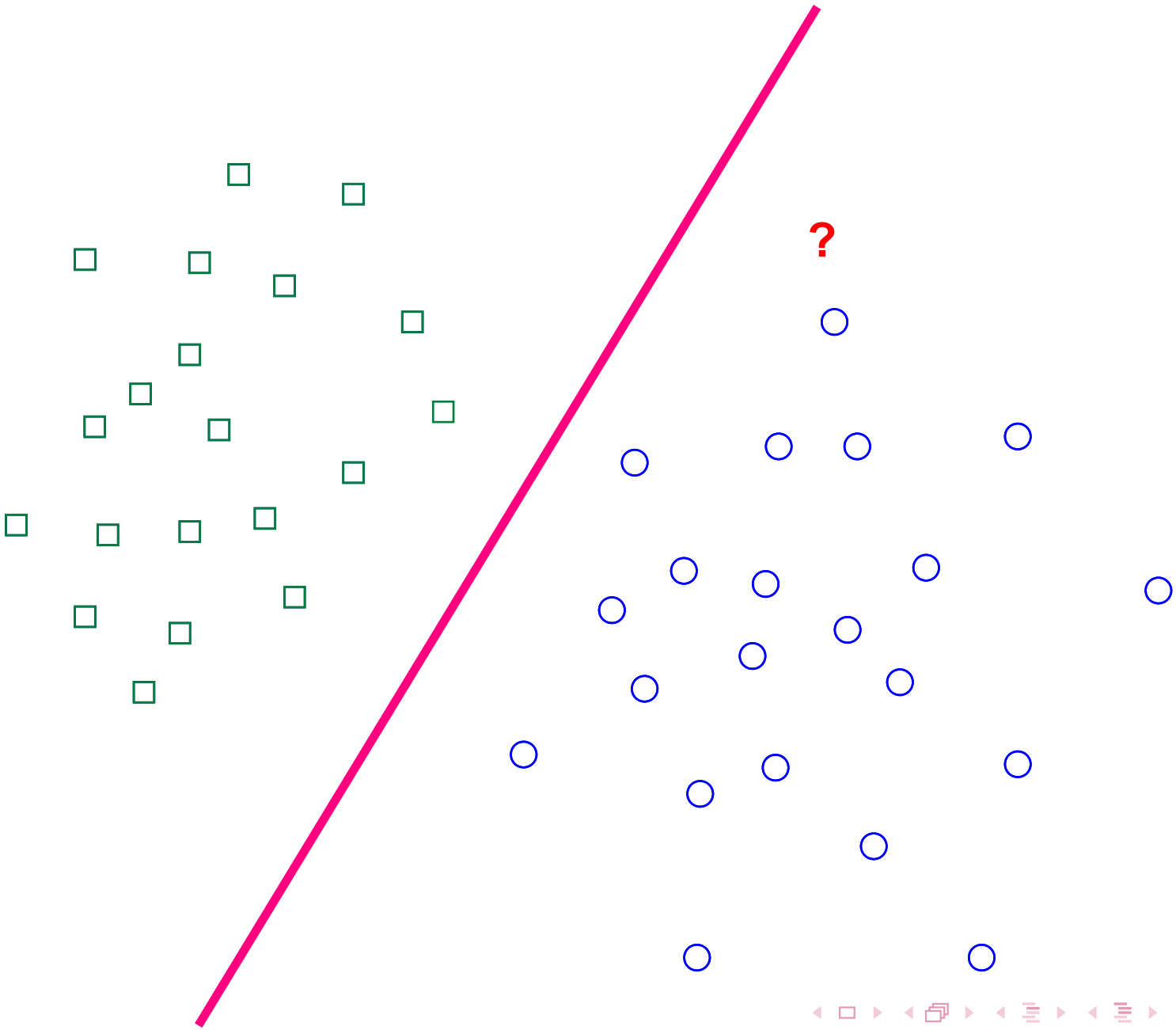
Support Vector Machines



Support Vector Machines



SVM



Support Vector Machines

Soft margin [Vapnik, 1995, 1998]

$$\text{minimize}_{\omega,\, b,\, \xi} \quad \frac{1}{2}\sum_{j=1}^{d}\omega_j^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

s.t. (SVM)

$$\begin{aligned} y_i(\omega^\top \mathbf{x}_i + b) &\geq 1 - \xi_i \qquad i = 1, \ldots, n \\ \omega &\in \mathbb{R}^k \\ b &\in \mathbb{R} \\ \xi_i &\geq 0 \qquad i = 1, \ldots, n. \end{aligned}$$

Categorical variables in SVM

With categorical variables, SVM needs to interpret many coefficients

Cluster Support Vector Machines

We propose ...

... the Cluster Support Vector Machines [Carrizosa et al., 2017b]:

- $\forall j$, the K_j categories of categorical feature j are grouped into L_j clusters

Without compromising:

- Accuracy

With the pursue of:

- Easier-to-interpret classifier
  - Sparser classifier
  - If $L_j = 2$, interpretable coefficients, since one can prove $\bar{\omega}_{j,1} \cdot \bar{\omega}_{j,2} \leq 0$

The formulation

$$\text{minimize}_{\bar{\omega},\omega',b,\xi,z} \sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \frac{(\bar{\omega}_{j,\ell})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

s.t. (CL)

$$y_i \left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \bar{\omega}_{j,\ell} \sum_{k=1}^{K_j} z_{j,k,\ell}\, x_{i,j,k} + (\omega')^\top \mathbf{x}'_i + b \right) \geq 1 - \xi_i \qquad i = 1,\dots,n$$

$$\sum_{\ell=1}^{L_j} z_{j,k,\ell} = 1 \qquad j = 1,\dots,J;\ k = 1,\dots,K_j$$

$$\xi_i \geq 0 \qquad i = 1,\dots,n$$

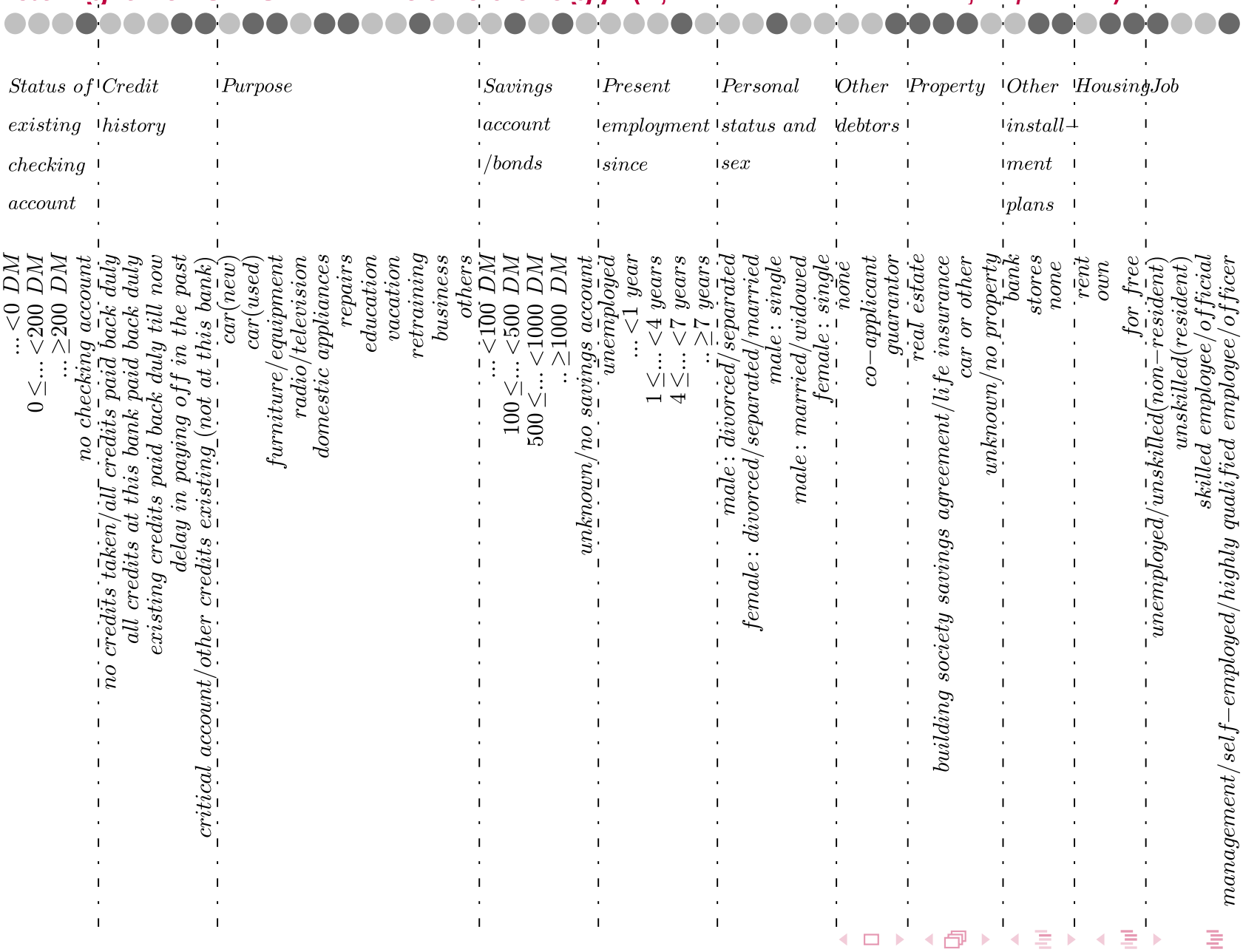
$$z \in \{0,1\}^{\sum_{j=1}^{J} L_j K_j}$$

$$\bar{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j}$$

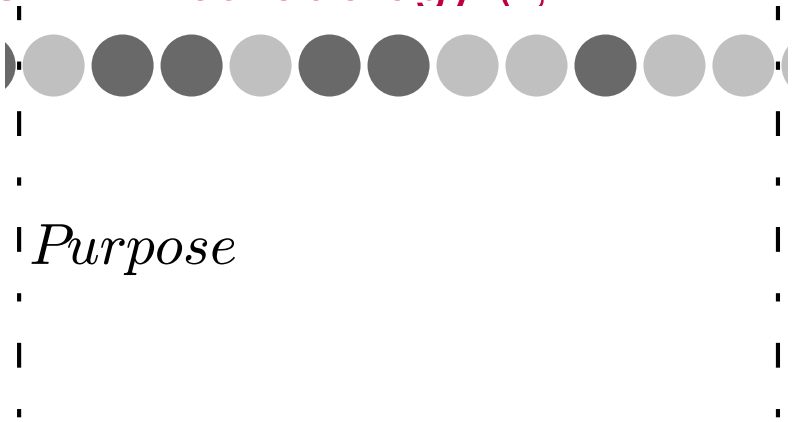
$$\omega' \in \mathbb{R}^{J'}$$

$$b \in \mathbb{R}$$

Illustrating the CLSVM methodology (`germancredit`, $L_i = 2$)



Illustrating the CLSVM methodology (germancredit, $L_j = 2$)



- Categorical feature *Purpose* has $K_j = 11$ categories
- The $K_j = 11$ categories have been clustered into $L_j = 2$ clusters
  - 1st cluster associated with *Purpose* in light grey
  - 2nd cluster associated with *Purpose* in dark grey

Clustering Categories in Logistic Regression

Logistic Regression suffers in the same way from categorical data

Clustering Categories in Logistic Regression

We study...

... transparency in the EU Public Procurement [la Cour et al., 2019]

Table: Categorical variables in the Tender Electronic Daily (TED) data set.

	country	authority	industry	open call	amount	award criteria	directive	same country	contract type	year
no. categories	30	7	12	2	4	3	3	2	3	8

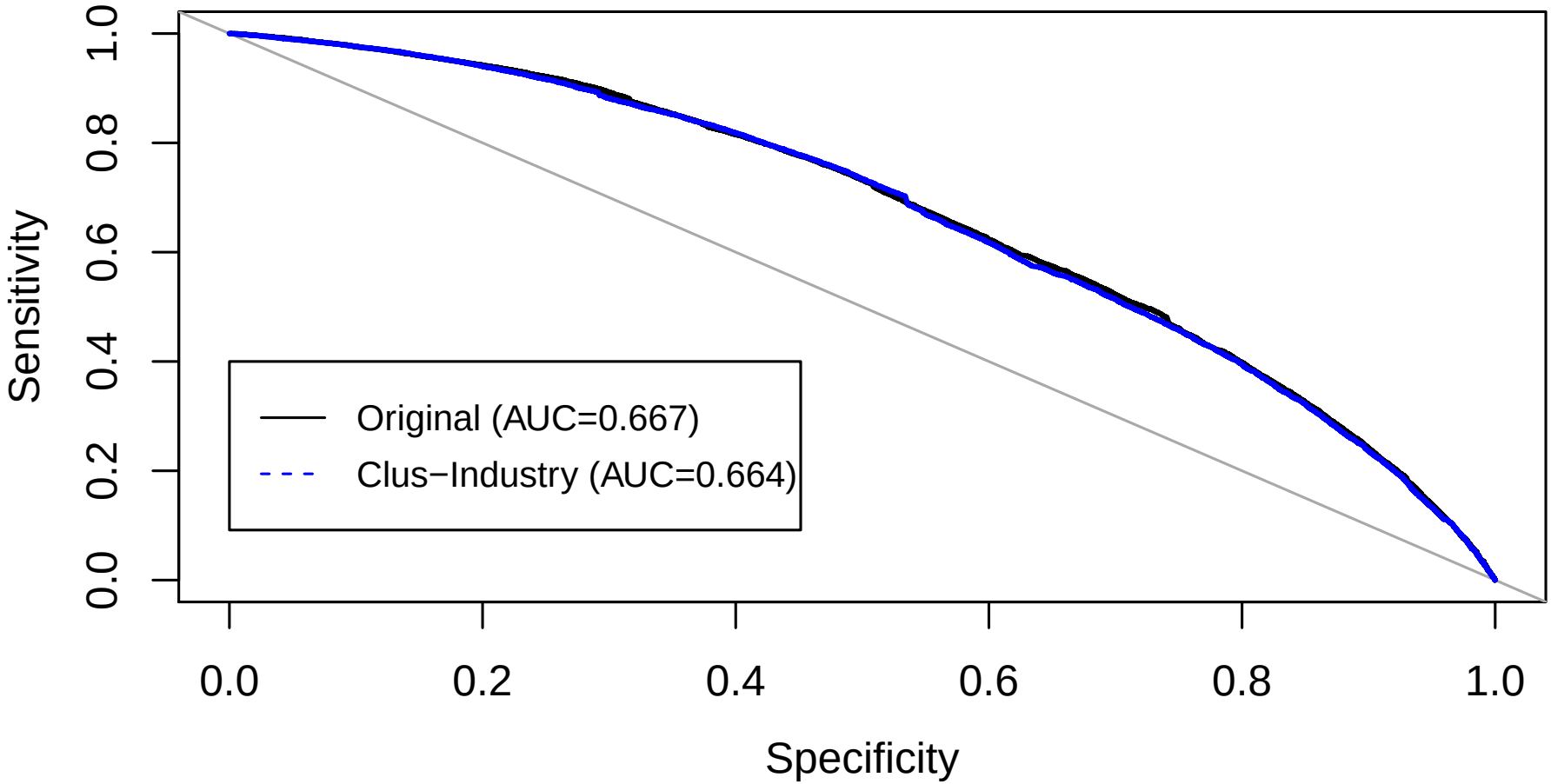
The response variable: Was there only one bid (not enough competition)?

TED Dataset

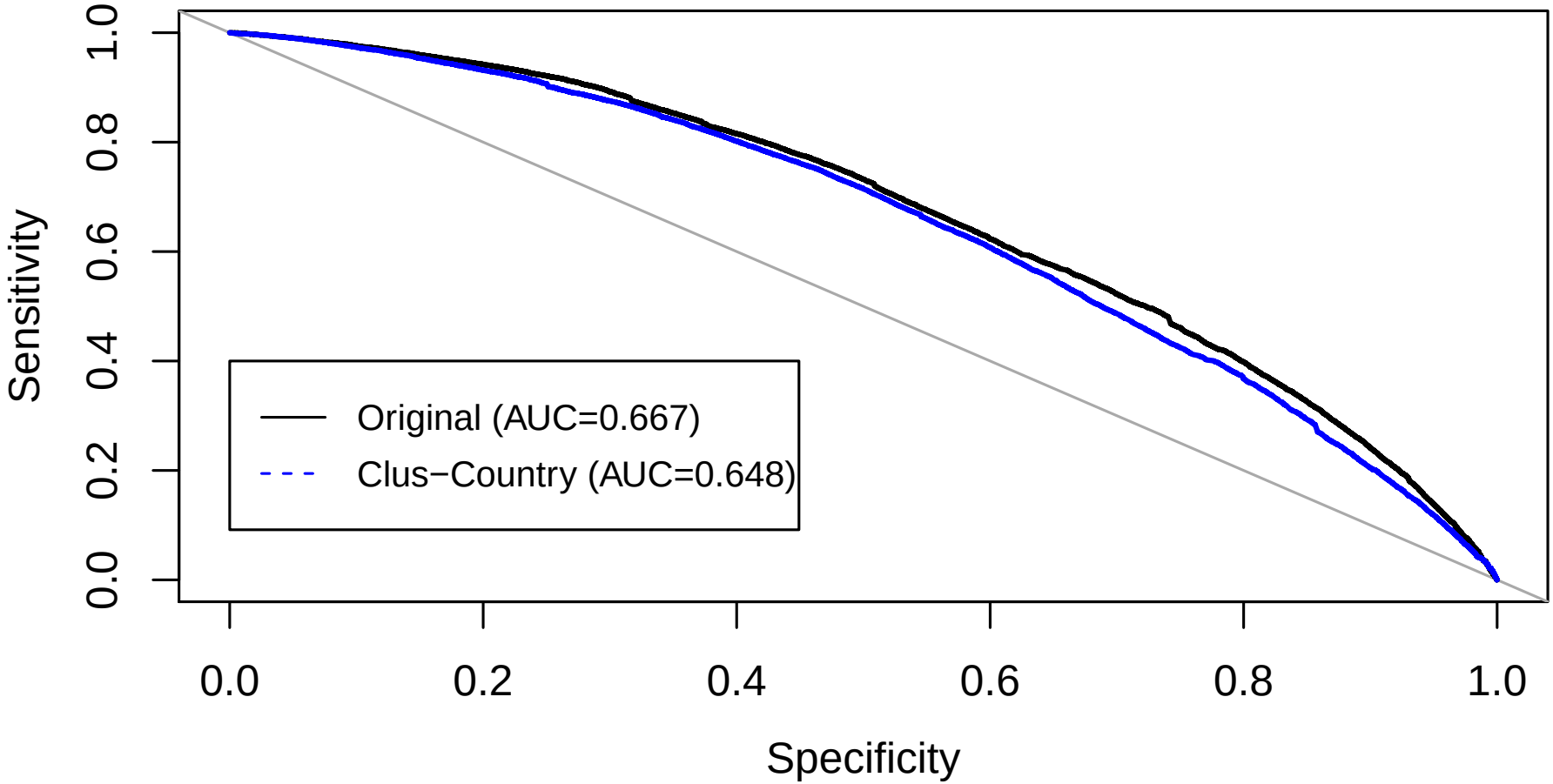
Table: Information about the TED data set.

Data set	Abbreviation	n	p	K	Class distribution
TED	TED	1.86 million	10	2	14.9% – 85.1%

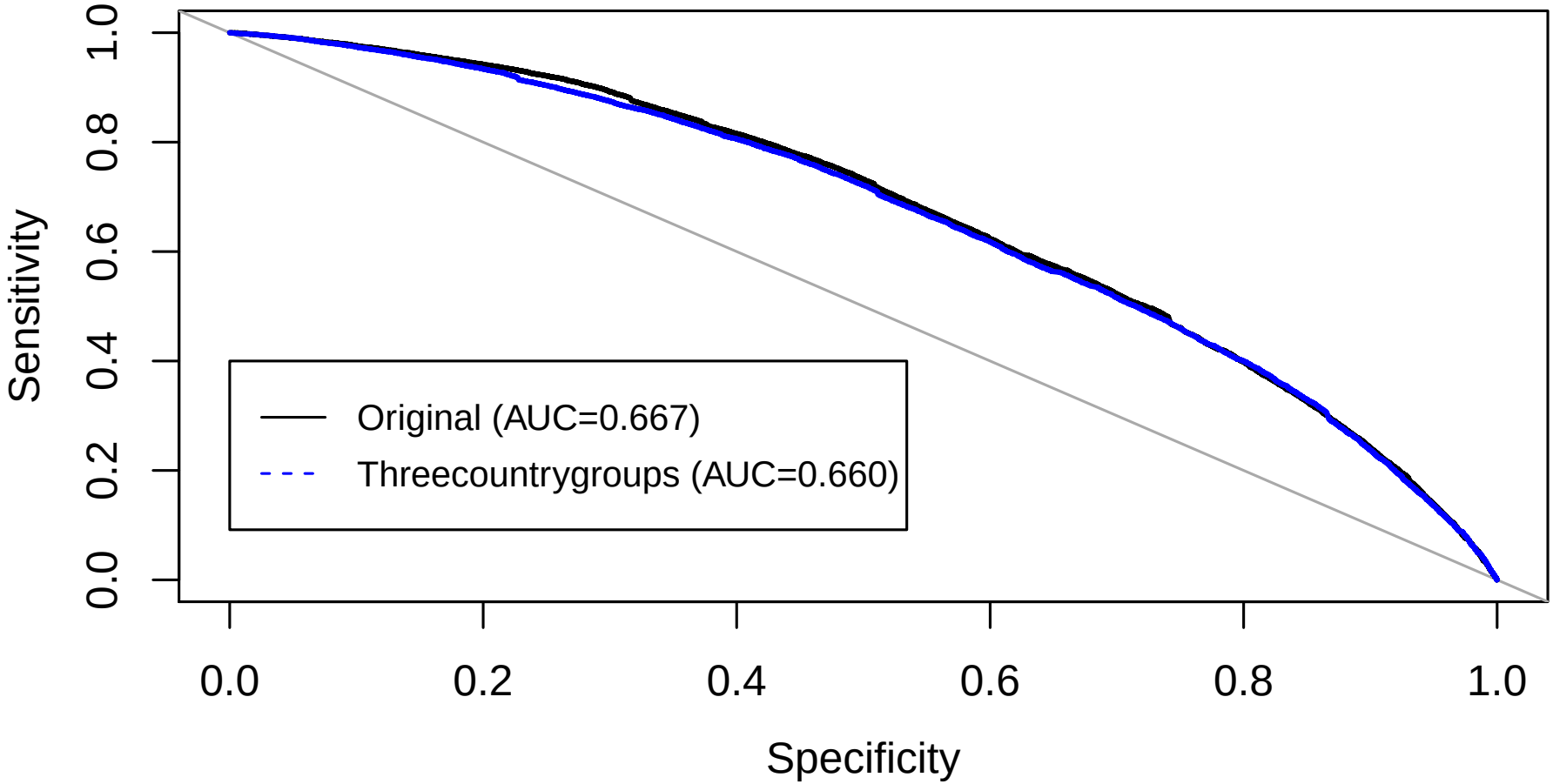
Clustering Industry: two clusters



Clustering Country: two clusters



Clustering Country: three clusters



Outline

Benchmarking

Benchmarking [Bogetoft and Otto, 2010]

- We are given a set of K firms, which use the same I inputs and produce the same O outputs.
- The performance of firms is measured in terms of outputs produced (y_o^k) in relation to inputs used (x_i^k).
- In Benchmarking, the aim is to compare the performance of firms against each other.

Applications

Benchmarking of electricity Distribution System Operators (DSOs), hospitals, universities, schools, ...

Benchmarking

Tools in Benchmarking

- Stochastic Frontier Analysis: Statistical approach.
- Data Envelopment Analysis: Mathematical Optimization approach.

Data Envelopment Analysis DEA

Data Envelopment Analysis (DEA) [Charnes et al., 1978, 2013]

- DEA is a classical tool in Benchmarking based on Mathematical Optimization.
- DEA measures performance through an efficiency score $E^k \in [0, 1]$,

$$\frac{\sum_{o=1}^{O} \beta_o^k y_o^k}{\sum_{i=1}^{I} \alpha_i^k x_i^k},$$

  - y_o^k is the amount of output o produced by firm k
  - x_i^k is the amount of input i used by firm k
  - β_o^k and α_i^k are the weights given to outputs and inputs, respectively.
- Firms with a score equal to one are deemed to be efficient.

Feature Selection in DEA

State of the Art

Lack of enough guidance in the literature [Cook et al., 2014]

- A priori rules based on Statistical Analysis
  - e.g., correlations, dimensionality reduction techniques, regression
- Information Theory
  - e.g., AIC, Shannon entropy
- Ex-post analysis
  - e.g., sensitivity of the efficient frontier to non-selected features
- Recent work on LASSO [Lee and Cai, 2018, Qin and Song, 2014]

Feature Selection in DEA [Benítez-Peña et al., 2018b]

Selecting p features

Let p be the number of outputs to be selected.

The individual and the joint models

We first model the selection for an individual firm.

We then model the joint selection for all individuals.

Feature Selection in DEA: Individual model

$$\text{maximize}_{(\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k, \mathbf{z}^k)} \sum_{o=1}^{O} \beta_o^k y_o^k$$

s.t. (OSDEA$^k(p)$)

$$\sum_{o=1}^{O} \beta_o^k y_o^j - \sum_{i=1}^{I} \alpha_i^k x_i^j \leq 0 \qquad \forall j = 1, \ldots, K$$

$$\sum_{i=1}^{I} \alpha_i^k x_i^k = 1$$

$$\beta_o^k \leq M z_o^k \qquad \forall o = 1, \ldots, O$$

$$\sum_{o=1}^{O} z_o^k = p$$

$$\boldsymbol{\alpha}^k \in \mathbb{R}_+^I$$

$$\boldsymbol{\beta}^k \in \mathbb{R}_+^O$$

$$z_o^k \in \{0, 1\} \qquad \forall o = 1, \ldots, O$$

Joint selection

Joint selection

The selection of outputs has to be the same for all firms.

Objective function

$$\phi(E^1, E^2, \ldots, E^K) = \frac{1}{K} \sum_{k=1}^{K} E^k.$$

Feature Selection in DEA: Joint model

$$\text{maximize}_{(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{z})} \ \frac{1}{K}\sum_{k=1}^{K}\sum_{o=1}^{O}\beta_o^k y_o^k$$

s.t. (OSDEA(p))

$$\sum_{o=1}^{O}\beta_o^k y_o^j - \sum_{i=1}^{I}\alpha_i^k x_i^j \leq 0 \qquad \forall j = 1,\ldots,K; \forall k = 1,\ldots,K$$

$$\sum_{i=1}^{I}\alpha_i^k x_i^k = 1 \qquad \forall k = 1,\ldots,K$$

$$\beta_o^k \leq M z_o \qquad \forall o = 1,\ldots,O; \forall k = 1,\ldots,K$$

$$\sum_{o=1}^{O} z_o = p$$

$$\boldsymbol{\alpha} \in \mathbb{R}_+^{I\cdot K}$$

$$\boldsymbol{\beta} \in \mathbb{R}_+^{O\cdot K}$$

$$z_o \in \{0,1\} \qquad \forall o = 1,\ldots,O$$

Alternative objective functions

Alternative objective functions for feature selection

$$\begin{aligned}
\phi^{\mathrm{w}}(E^1, E^2, \ldots, E^K) &= \frac{1}{K}\sum_{k=1}^{K} \omega^k E^k \\
\phi^{\mathrm{q}}(E^1, E^2, \ldots, E^K) &= -\frac{1}{K}\sum_{k=1}^{K} (1 - E^k)^2 \\
\phi^{\min}(E^1, E^2, \ldots, E^K) &= \min_{k=1,\ldots,K} E^k \\
\phi^{\pi}(E^1, E^2, \ldots, E^K) &= E^{(k(\pi))},
\end{aligned}$$

where $E^{(k(\pi))}$ represents the π percentile of the efficiencies.

In general, they can be modeled as MINLPs.

Extensions

Extensions

- Selection of inputs and outputs
- Bounds on weights
- Multicollinearity constraints
- Costs and types of outputs
- Strategic Feature Selection
- Multiple DEA Models [Benítez-Peña et al., 2019]

Numerical results

- Operating System: Microsoft Windows 10 Home, 64 bits
- Processor: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz, 3408 Mhz (4 cores)
- RAM: 16 GB
- Interface: Python 3.5 (64 bits)
- Solver: Gurobi 7.0.1

Dataset

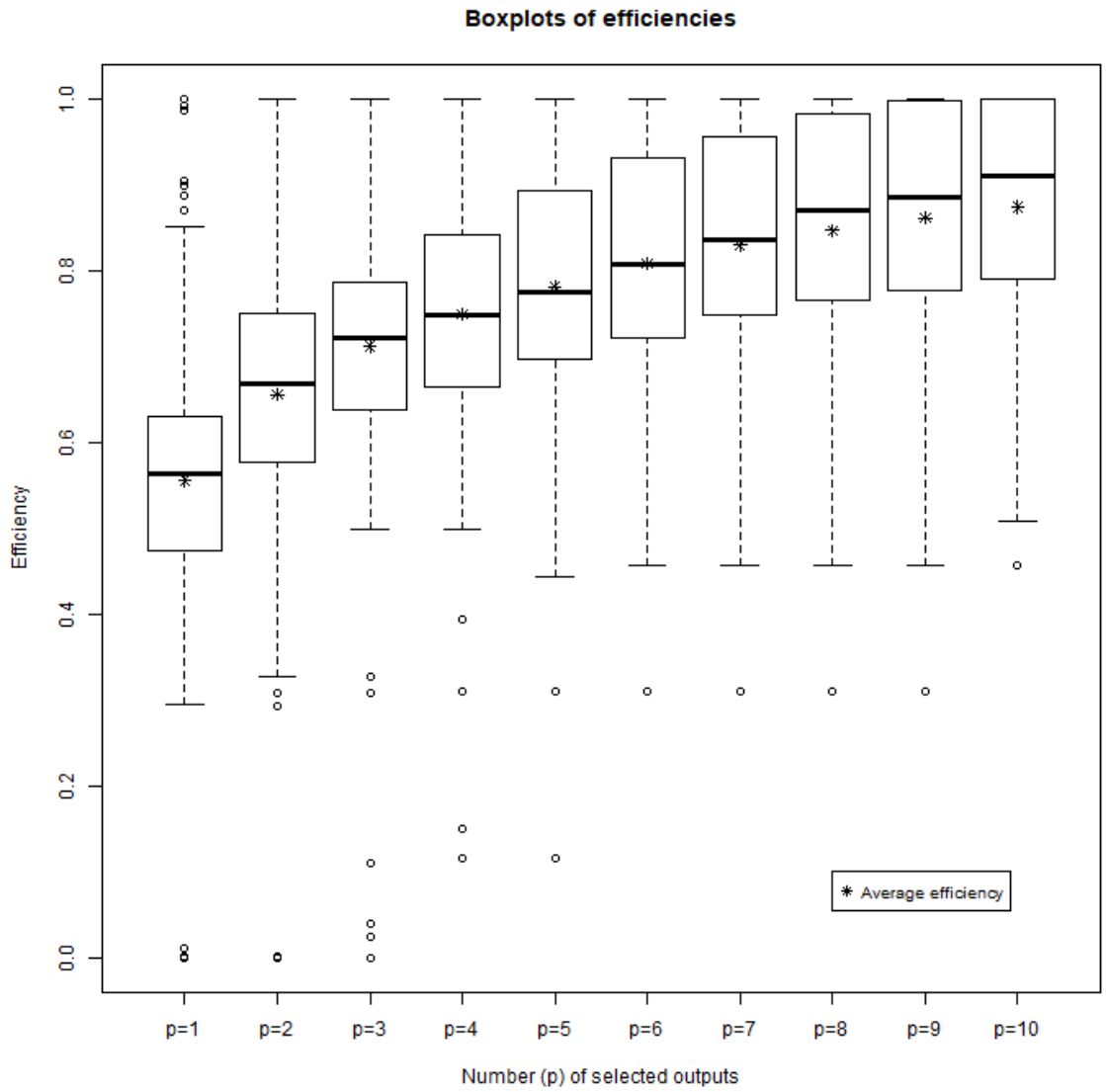
- Real-world dataset for the benchmarking of electricity Distribution System Operators (DSOs)
- $K = 182$
- $O = 100$
- $I = 1$
- Data has been normalized.

Best subset of features

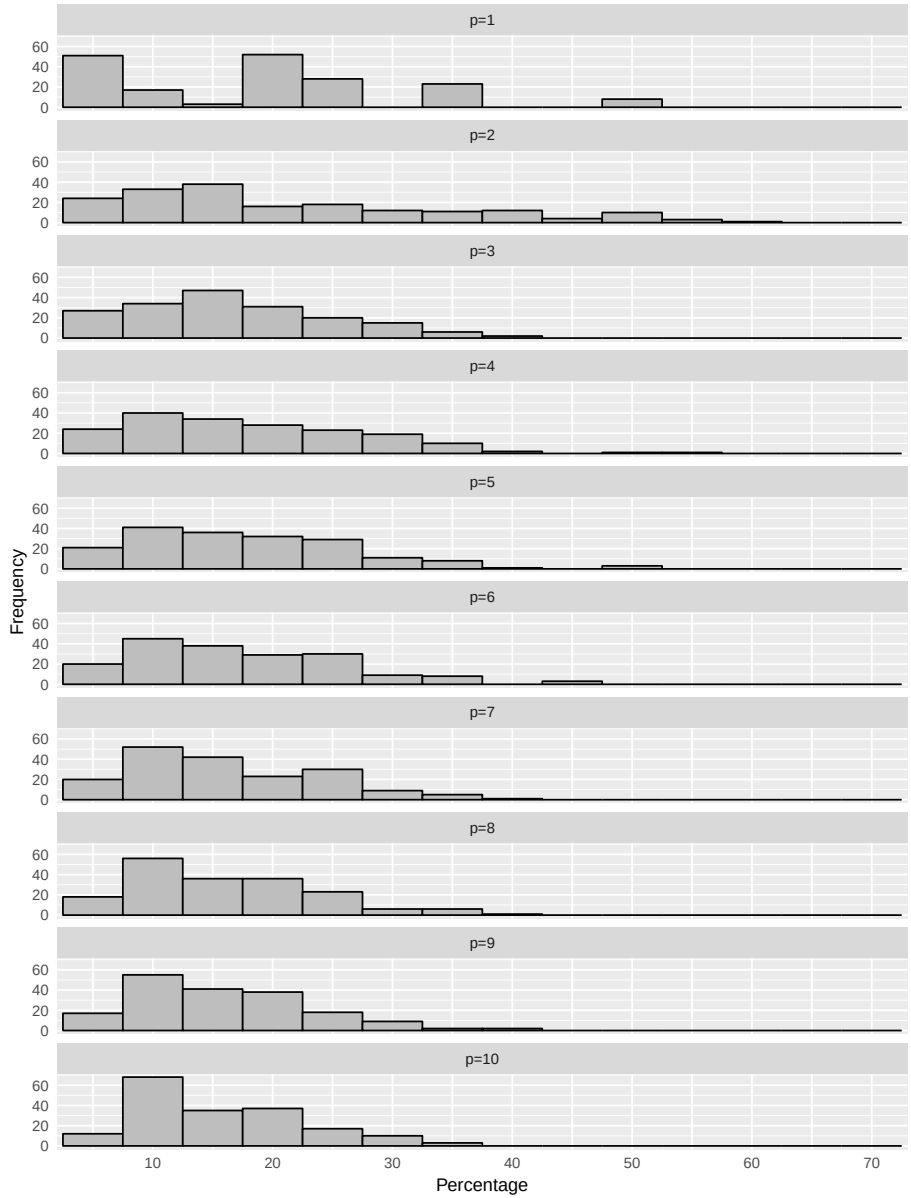
Table: Summary statistics for the distribution of efficiencies, for $p = 1, \ldots, 10$

p	min	max	mean	st. dev.	q_1	q_2	q_3	q_3-q_1	selected features
1	0.0000	1.0000	0.5555	0.1695	0.4743	0.5637	0.6300	0.1557	59
2	0.0006	1.0000	0.6553	0.1708	0.5772	0.6682	0.7482	0.1710	11 31
3	0.0009	1.0000	0.7118	0.1643	0.6391	0.7222	0.7839	0.1448	21 31 54
4	0.1161	1.0000	0.7487	0.1511	0.6645	0.7479	0.8404	0.1759	16 21 31 59
5	0.1161	1.0000	0.7812	0.1494	0.6962	0.7738	0.8911	0.1949	16 21 31 59 94
6	0.3105	1.0000	0.8082	0.1402	0.7222	0.8068	0.9305	0.2083	16 19 21 31 59 94
7	0.3105	1.0000	0.8290	0.1404	0.7474	0.8355	0.9545	0.2071	16 19 21 31 59 91 94
8	0.3105	1.0000	0.8462	0.1402	0.7658	0.8689	0.9802	0.2144	16 19 21 31 59 91 94 97
9	0.3105	1.0000	0.8610	0.1370	0.7789	0.8841	0.9972	0.2183	16 19 21 31 59 74 91 94 97
10	0.4576	1.0000	0.8732	0.1304	0.7902	0.9090	1.0000	0.2098	16 19 21 29 31 59 74 91 94 97

Best subset of features



Preference of individual strategy over joint strategy



Outline

Summary

Today

- MINLP plays an important role the enhancement of interpretability
  - It allows the optimization of classification trees
  - It can enhance the treatment of categorical variables
  - It helps with Data Driven Decision Making

There are still open problems

- quantification of categorical variables
- better VIZ of black-box models
- Data Science at the service of the Mathematical Optimization community
  e.g., Hutter et al. [2010], Khalil et al. [2016], Lodi and Zarpellon [2017]

Thank you very much!

www.riseneeds.eu

@needs_project

References I

D. Aloise, P. Hansen, and L. Liberti. An improved column generation algorithm for minimum sum-of-squares clustering. *Mathematical Programming*, 131 (1–2):195–220, 2012.

E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209: 237–260, 1998.

A. Astorino, I. Bomze, A. Fuduli, and M. Gaudioso. Robust spherical separation. *Optimization*, 66(6):925–938, 2017.

A. Atamtürk and A. Gomez. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

H. Aytug. Feature selection for support vector machines using generalized benders decomposition. *European Journal of Operational Research*, 244(1): 210–218, 2015.

B. Baesens, R. Setiono, C. Mues, and J. Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, 49(3):312–329, 2003.

S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo. Cost-sensitive feature selection for Support Vector Machines. *Forthcoming in Computers & Operations Research*, 2018a.

S. Benítez-Peña, P. Bogetoft, and D. Romero Morales. Feature selection in data envelopment analysis: A mathematical optimization approach. Technical report, Copenhagen Business School, Denmark, 2018b.

S. Benítez-Peña, P. Bogetoft, and D. Romero Morales. On clustered feature selection in data envelopment analysis. Working paper, Copenhagen Business School, Denmark, 2019.

K. Bennett. Global tree optimization: A non-greedy decision tree algorithm. In *Computing Science and Statistics*, pages 156–160, 1994.

K.P. Bennett. *Decision tree construction via linear programming*. Center for Parallel Optimization, Computer Sciences Department, University of Wisconsin, 1992.

P. Bertolazzi, G. Felici, P. Festa, G. Fiscon, and E. Weitschek. Integer programming models for feature selection: New extensions and a randomized solution algorithm. *European Journal of Operational Research*, 250(2):389–399, 2016.

D. Bertsimas and J. Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

R. Blanquero, E. Carrizosa, C. Molero-Río, and Romero Morales. Optimal randomized classification trees. Technical report, IMUS, Sevilla, Spain, 2018a.

R. Blanquero, E. Carrizosa, C. Molero-Río, and Romero Morales. Sparsity in optimal randomized classification trees. Technical report, IMUS, Sevilla, Spain, 2018b.

P. Bogetoft and L. Otto. *Benchmarking with Dea, Sfa, and R*, volume 157. Springer Science & Business Media, 2010.

L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

References II

E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers and Operations Research*, 40(1):150–165, 2013.

E. Carrizosa, B. Martín-Barragán, D. Romero Morales, and F. Plastria. On the selection of the globally optimal prototype subset for nearest-neighbor classification. *INFORMS Journal on Computing*, 19(3):470–479, 2007.

E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Binarized support vector machines. *INFORMS Journal on Computing*, 22(1):154–167, 2010.

E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Detecting relevant variables and interactions in supervised classification. *European Journal of Operational Research*, 213(1):260–269, 2011.

E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Strongly agree or strongly disagree?: Rating features in Support Vector Machines. *Information Sciences*, 329:256–273, 2016.

E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing proportions and dissimilarities by space-filling maps: a large neighborhood search approach. *Computers & Operations Research*, 78:369–380, 2017a.

E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Clustering categories in support vector machines. *Omega*, 66:28–37, 2017b.

E. Carrizosa, V. Guerrero, D. Hardt, and D. Romero Morales. On building online visualization maps for news data streams by means of mathematical optimization. *Big Data*, 6(2):139–158, 2018a.

E. Carrizosa, V. Guerrero, and D. Romero Morales. On mathematical optimization for the visualization of frequencies and adjacencies as rectangular maps. *European Journal of Operational Research*, 265:290–302, 2018b.

E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing data as objects by dc (difference of convex) optimization. *Mathematical Programming, Series B*, 169:119–140, 2018c.

E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualization of complex dynamic datasets by means of mathematical optimization. *Forthcoming in Omega*, 2018d.

E. Carrizosa, V. Guerrero, and D. Romero Morales. The role of mathematical optimization in data visualization: Optimizing distances to visualize complex data. *Invited Review for Omega*, 2018e.

E. Carrizosa, V. Guerrero, D. Romero Morales, and A. Satorra. Enhancing interpretability in factor analysis by means of mathematical optimization. Technical report, IMUS, Sevilla, Spain, 2018f.

A.B. Chan, N. Vasconcelos, and G.R.G. Lanckriet. Direct convex relaxations of sparse SVM. In *Proceedings of the 24th international conference on Machine learning*, pages 145–153, 2007.

A. Charnes, W.W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, 1978.

A. Charnes, W.W. Cooper, A.Y. Lewin, and L.M. Seiford. *Data envelopment analysis: Theory, methodology, and applications*. Springer Science & Business Media, 2013.

References III

Y. Chevaleyre, F. Koriche, and J.-D. Zucker. Rounding methods for discrete linear classification. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 651–659. JMLR Workshop and Conference Proceedings, 2013.

W.D. Cook, K. Tone, and J. Zhu. Data envelopment analysis: Prior to choosing a model. *Omega*, 44:1–4, 2014.

S. Corrente, S. Greco, M. Kadziński, and R. Słowiński. Robust ordinal regression in preference learning and ranking. *Machine Learning*, 93(2):381–422, 2013.

A. Cotter, S. Shalev-Shwartz, and N. Srebro. Learning optimally sparse support vector machines. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 266–274, 2013.

C. Ding and H.-D. Qi. Convex optimization learning of faithful euclidean distance representations in nonlinear dimensionality reduction. *Mathematical Programming*, 164(1):341–381, 2017.

A.P. Duarte Silva. Optimization approaches to supervised classification. *European Journal of Operational Research*, 261(2):772–788, 2017.

K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

A.A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

G. Fung and O.L. Mangasarian. A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications*, 28(2):185–202, 2004.

M. Gaudioso, E. Gorgone, M. Labbé, and A.M. Rodríguez-Chía. Lagrangian relaxation for SVM feature selection. *Computers & Operations Research*, 87: 137–145, 2017.

B. Ghaddar and J. Naoum-Sawaya. High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 265(3):993–1004, 2018.

N. Goldberg, S. Leyffer, and T. Munson. A new perspective on convex relaxations of sparse SVM. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 450–457, 2013.

M. Golea and M. Marchand. On learning perceptrons with binary weights. *Neural Computation*, 5(5):767–782, 1993.

B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a “right to explanation”. *arXiv preprint arXiv:1606.08813*, 2016.

W. Guan, A. Gray, and S. Leyffer. Mixed-integer support vector machine. In *NIPS Workshop on Optimization for Machine Learning*, 2009.

O. Günlük, J. Kalagnanam, M. Menickelly, and K. Scheinberg. Optimal Decision Trees for Categorical Data via Integer Programming. *arXiv preprint arXiv:1612.03225v2*, 2018.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

References IV

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

P.E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53:59–67, 2010.

F. Hutter, H.H. Hoos, and K. Leyton-Brown. Automated configuration of mixed integer programming solvers. In A. Lodi, M. Milano, and P. Toth, editors, *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, volume 6140 of *Lecture Notes in Computer Science*, pages 186–202. Springer Berlin Heidelberg, 2010.

E.B. Khalil, P. Le Bodic, L. Song, G. L Nemhauser, and B.N. Dilkina. Learning to branch in mixed integer programming. In *AAAI*, pages 724–731, 2016.

L. la Cour, M. Galvis Restrepo, D. Romero Morales, and G. Skovgaard Ølykke. Explaining low competition in public procurement procedures. Working paper, Copenhagen Business School, 2019.

C.-Y. Lee and J.-Y. Cai. LASSO variable selection in data envelopment analysis with small datasets. *Forthcoming in Omega*, 2018.

S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

A. Lodi and G. Zarpellon. On learning and branching: a survey. *TOP*, 25(2):207–236, 2017.

S. Maldonado and R. Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009.

S. Maldonado, R. Weber, and J. Basak. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1):115–128, 2011.

S. Maldonado, J. Pérez, R. Weber, and M. Labbé. Feature selection for support vector machines via mixed integer linear programming. *Information Sciences*, 279:163–175, 2014.

O.L. Mangasarian. Linear and nonlinear separation of pattern by linear programming. *Operations Research*, 31:445–453, 1965.

O.L. Mangasarian. Exact 1-norm support vector machines via unconstrained convex differentiable minimization. *Journal of Machine Learning Research*, 7: 1517–1530, 2006.

J.S. Marron and A.M. Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99, 2014.

D. Martens, B. Baesens, T.V. Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007.

S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.

C. Orsenigo and C. Vercellis. Multivariate classification trees based on minimum features discrete support vector machines. *IMA Journal of Management Mathematics*, 14(3):221–234, 2003.

References V

C. Orsenigo and C. Vercellis. Discrete support vector decision trees via tabu search. *Computational Statistics and Data Analysis*, 47(2):311–322, 2004.

L. Palagi. Global optimization issues in deep network regression: an overview. *Journal of Global Optimization*, 73(2):239–277, 2019.

V. Piccialli and M. Sciandrone. Nonlinear optimization and support vector machines. *4OR*, 16(2):111–149, 2018.

Z. Qin and I. Song. Joint Variable Selection for Data Envelopment Analysis via Group Sparsity. *ArXiv e-prints arXiv:1402.3740*, 2014.

G. Ridgeway. The pitfalls of prediction. *National Institute of Justice Journal*, 271:34–40, 2013.

F. Rinaldi and M. Sciandrone. Feature selection combining linear support vector machines and concave optimization. *Optimization Methods and Software*, 25(1):117–128, 2010.

F. Rinaldi, F. Schoen, and M. Sciandrone. Concave programming for minimizing the zero-norm over polyhedral sets. *Computational Optimization and Applications*, 46(3):467–486, 2010.

V. Roth. The generalized lasso. *IEEE Transactions on Neural Networks*, 15(1):16–28, 2004.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

A. Taeb and V. Chandrasekaran. Interpreting latent variables in factor models via convex optimization. *Mathematical Programming*, 167(1):129–154, 2018.

J. Thomas and P.C. Wong. Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, pages 668–674, Cambridge, MA, 2001. MIT Press.

J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

G. Wilfong. Nearest neighbor problems. *International Journal of Computational Geometry and Applications*, 2(4):383–416, 1992.

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16–NIPS2003*. MIT Press, 2004.