On Interpretability in Data Analytics

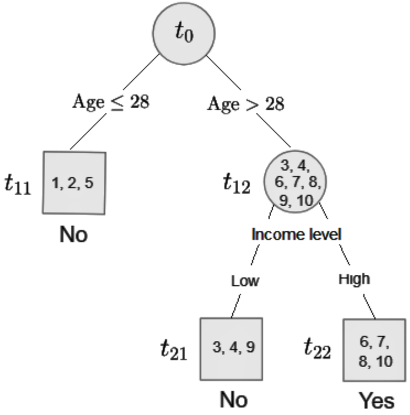
Emilio Carrizosa, IMUS - Instituto de Matemáticas de la Universidad de Sevilla

Nizhny Novgorod, 21st November 2020

Classification and regression with trees

CARTs (Breiman et al. 1984)

Applicant	Age	Income level	Loan granted
1	22	Low	No
2	26	High	No
3	30	Low	Yes
4	32	Low	No
5	20	High	No
6	45	High	Yes
7	60	High	No
8	54	High	Yes
9	50	Low	No
10	48	High	Yes

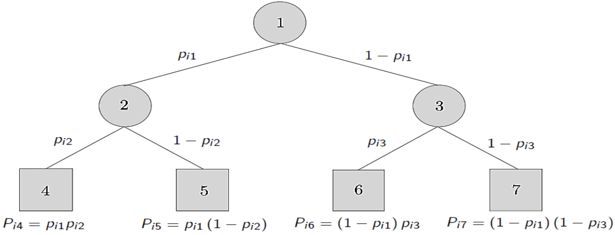


Blanquero, Carrizosa, Molero-Río, Romero Morales, EJOR, 2019

We have a sample $I = \{(\boldsymbol{x}_i, y_i)\}_{1 \le i \le n}$, where $\boldsymbol{x}_i \in [0, 1]^p$ and $y_i \in \{1, \dots, K\}$.

Blanquero, Carrizosa, Molero-Río, Romero Morales, EJOR, 2019

We have a sample $I = \{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq n}$, where $\boldsymbol{x}_i \in [0,1]^p$ and $y_i \in \{1, \ldots, K\}$.
A maximal binary tree of depth D. Nodes: Branch $t \in \tau_B$, Leaf $t \in \tau_L$.



(Mixed-Integer Non-Linear Optimization Problem)

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} \\
\text{s.t.} \quad & \sum_{j=1}^{p} a_{jt} = 1, \;\; t \in \tau_B, \\
& \sum_{k=1}^{K} C_{kt} = 1, \;\; t \in \tau_L, \\
& \sum_{t \in \tau_L} C_{kt} \geq 1, \;\; k = 1, \dots, K, \\
& a_{jt} \in \{0, 1\}, \;\; j = 1, \dots, p, \;\; t \in \tau_B, \\
& C_{kt} \in \{0, 1\}, \;\; k = 1, \dots, K, \;\; t \in \tau_L, \\
& \boldsymbol{\alpha}_t \in A, \;\; t \in \tau_B.
\end{aligned}
$$

Optimal Randomized Classification Trees

(Continuous Non-Linear Optimization Problem) **OBLIQUE** splits

$$\min \quad \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'}$$

s.t.

$$\begin{aligned}
&\sum_{k=1}^{K} C_{kt} = 1, \;\; t \in \tau_L, \\
&\sum_{t \in \tau_L} C_{kt} \geq 1, \;\; k = 1, \ldots, K, \\
&a_{jt} \in [-1, 1], \;\; j = 1, \ldots, p, \;\; t \in \tau_B, \\
&C_{kt} \in [0, 1], \;\; k = 1, \ldots, K, \;\; t \in \tau_L, \\
&\boldsymbol{\alpha}_t \in A, \;\; t \in \tau_B.
\end{aligned} \tag{ORCT}$$

ORCT's prediction

A new unlabeled observation $\boldsymbol{x}$

$$\Downarrow$$

Once the optimization problem has been solved

,

the decision variables are used for predicting its class:

$$m_n(\boldsymbol{x}) = \arg\max_k \left\{ \sum_{t \in \tau_L} \mathbb{P}\left(\boldsymbol{x} \in k | \boldsymbol{x} \in t\right) \mathbb{P}\left(\boldsymbol{x} \in t\right) \right\} = \arg\max_k \left\{ \sum_{t \in \tau_L} C_{kt} \cdot P_{\boldsymbol{x}t} \right\}$$

Computational experience

UCI Machine Learning Repository

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

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)}, \ \ \mu\in\mathbb{R}, \ \ \gamma>0.$$

$\mu_t \in [-1,1]\,, \ \ t\in\tau_L, \ \ \gamma_t=\gamma=512, \ \ t\in\tau_L.$

- Logistic CDF:
$$F\left(\cdot;\mu,\gamma\right)=\frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \ \mu\in\mathbb{R},\ \ \gamma>0.$$
$\mu_t\in\left[-1,1\right],\ \ t\in\tau_L,\ \ \gamma_t=\gamma=512,\ \ t\in\tau_L.$
- Equal misclassification weights,
$$W_{kk'}=0.5,\ \ k,k'=1,\ldots,K,\ \ k\neq k'.$$

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \ \mu\in\mathbb{R},\ \ \gamma>0.$$

  $\mu_t \in [-1,1]$, $t\in\tau_L$, $\gamma_t=\gamma=512$, $t\in\tau_L$.

- Equal misclassification weights,

$$W_{kk'} = 0.5,\ \ k,k'=1,\ldots,K,\ \ k\neq k'.$$

- 10 hold-out runs: training subset (75%) and test subset (25%).

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \ \mu\in\mathbb{R},\ \ \gamma>0.$$

  $\mu_t \in [-1,1]\,,\ \ t\in\tau_L,\ \ \gamma_t=\gamma=512,\ \ t\in\tau_L.$

- Equal misclassification weights,

$$W_{kk'} = 0.5,\ \ k,k' = 1,\ldots,K,\ \ k\neq k'.$$

- 10 hold-out runs: training subset (75%) and test subset (25%).
- Performance measure: average accuracy over the 10 test subsets.

ORCT compared with:

- **CART** (Breiman et al. 1984).
- **OCT-H** (Bertsimas and Dunn 2017).
- **RF** (Breiman 2001).

Computational experience

$D = 1$

Data set	ORCT average time (in secs)	Out-of-sample accuracy			
		ORCT	CART	OCT-H	RF
Connectionist-bench-sonar	22	**76.3**	70.0	70.4	83.1
Wisconsin	24	**96.4**	92.0	93.1	95.5
Credit-approval	22	83.7	85.7	**87.9**	86.7
Pima-indians-diabetes	21	**75.8**	74.2	71.6	76.3
Statlog-project-German-credit	28	**72.8**	72.1	71.6	75.2
Ozone-level-detection-one	94	96.7	95.6	**96.8**	96.4
Spambase	72	**89.8**	89.2	83.6	95.1

Computational experience

$D = 1$

Data set	ORCT average time (in secs)	Out-of-sample accuracy			
		ORCT	CART	OCT-H	RF
Connectionist-bench-sonar	22	**76.3**	70.0	70.4	83.1
Wisconsin	24	**96.4**	92.0	93.1	95.5
Credit-approval	22	83.7	85.7	**87.9**	86.7
Pima-indians-diabetes	21	**75.8**	74.2	71.6	76.3
Statlog-project-German-credit	28	**72.8**	72.1	71.6	75.2
Ozone-level-detection-one	94	96.7	95.6	**96.8**	96.4
Spambase	72	**89.8**	89.2	83.6	95.1

$D = 2$

Data set	ORCT average time (in secs)	Out-of-sample accuracy			
		ORCT	CART	OCT-H	RF
Iris	17	**95.9**	92.7	95.1	95.4
Wine	23	**96.6**	88.6	91.1	98.6
Seeds	20	**94.2**	90.2	90.6	92.5
Thyroid-disease-ann-thyroid	145	92.2	**99.1**	92.5	99.1
Car-evaluation	71	**90.8**	88.1	87.5	88.0

Sparsity on ORCTs at depth 1

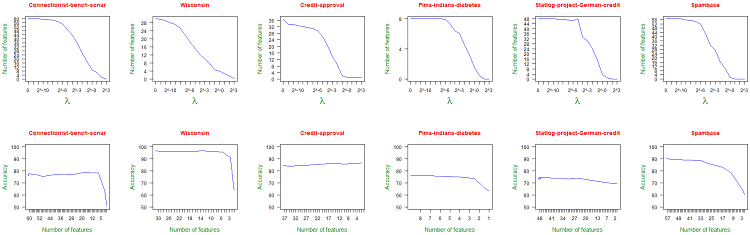
$$
\begin{aligned}
\min \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1} \in [-1, 1], \; j = 1, \ldots, p, \\
& C_{12}, \; C_{13}, \; C_{22}, \; C_{23} \in [0, 1], \\
& \mu_1 \in [-1, 1].
\end{aligned}
$$

Sparsity on ORCTs at depth 1

A lasso penalization to ORCT

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} + \lambda \quad \|\boldsymbol{a}_1\|_1 \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1} \in [-1, 1], \; j = 1, \ldots, p, \\
& C_{12}, \; C_{13}, \; C_{22}, \; C_{23} \in [0, 1], \\
& \mu_1 \in [-1, 1].
\end{aligned}
$$

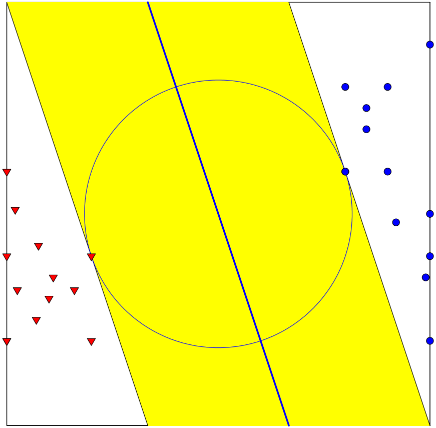
Sparsity on ORCTs at depth 1



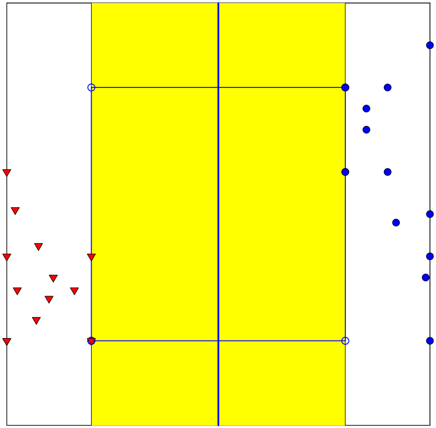
Support Vector Machines

SVM (ℓ_2)

- Roughly speaking, SVM finds the hyperplane $\omega_1 x_1 + \ldots + \omega_m x_m + \beta = 0$ separating most the sets $\{\mathrm{x}_i : i \in I, y_i = 1\}$ and $\{\mathrm{x}_i : i \in I, y_i = -1\}$



SVM (ℓ_1)



SVM (ℓ_2)

Convex quadratic optimization problem with linear constraints:

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

C., and Romero Morales, "Supervised classification and mathematical optimization", *Computers & Operations Research*, 2013.

Duarte Silva, "Optimization approaches to supervised classification", *EJOR*, 2017.

SVM (ℓ_2)

Convex quadratic optimization problem with linear constraints:

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i \in I} \xi_i & \\ \text{s.t.} & y_i \left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

C., and Romero Morales, "Supervised classification and mathematical optimization", *Computers & Operations Research*, 2013.

Duarte Silva, "Optimization approaches to supervised classification", *EJOR*, 2017.

SVM (ℓ_2)

Convex quadratic optimization problem with linear constraints:

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

$$\begin{array}{lll} \max_{\lambda} & \sum_{i\in I}\lambda_i - \frac{1}{2}\sum_{ij}\lambda_i y_i \lambda_j y_j x_i^\top x_j & \\ \text{s.t.} & \sum_{i\in I}\lambda_i y_i = 0 & \\ & 0 \leq \lambda_i \leq \frac{C}{2} & i \in I \end{array}$$

C., and Romero Morales, "Supervised classification and mathematical optimization", *Computers & Operations Research*, 2013.

Duarte Silva, "Optimization approaches to supervised classification", *EJOR*, 2017.

SVM (ℓ_2)

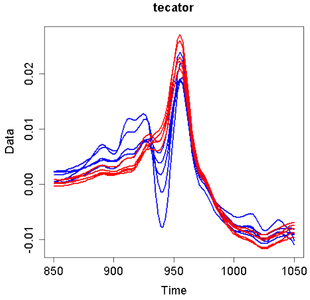
Convex quadratic optimization problem with linear constraints:

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

$$\begin{array}{lll} \max_{\lambda} & \sum_{i\in I}\lambda_i - \frac{1}{2}\sum_{ij}\lambda_i y_i \lambda_j y_j K(x_i, x_j) & \\ \text{s.t.} & \sum_{i\in I}\lambda_i y_i = 0 & \\ & 0 \leq \lambda_i \leq \frac{C}{2} & i \in I \end{array}$$

C., and Romero Morales, "Supervised classification and mathematical optimization", *Computers & Operations Research*, 2013.

Duarte Silva, "Optimization approaches to supervised classification", *EJOR*, 2017.



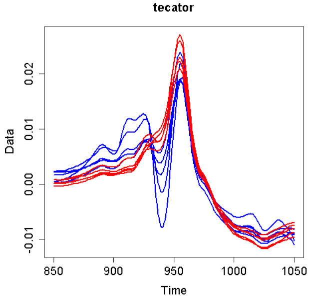
- $x \in \mathcal{C}^0([0, T])$
- $x \approx (x(t_1), \ldots, x(t_m)) \in \mathbb{R}^m$
- $(x(t_1), \ldots, x(t_m)) \approx x \in \mathcal{C}^0([0, T])$

Ferraty and Vieu. *Nonparametric functional data analysis: theory and practice*, 2006.

Ramsay and Silverman. *Functional data analysis*, 2006.

Febrero-Bande and Oviedo de la Fuente. "Statistical computing in functional data analysis: the r package fda.usc". *Journal of Statistical Software*, 2012.

Classification with functional data

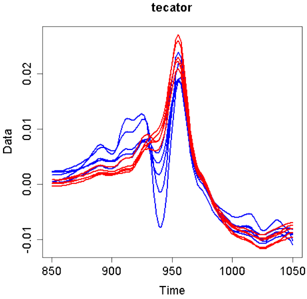


- $x \in \mathcal{C}^0([0, T])$
- $x \approx (x(t_1), \ldots, x(t_m)) \in \mathbb{R}^m$
- $(x(t_1), \ldots, x(t_m)) \approx x \in \mathcal{C}^0([0, T])$

Ferraty and Vieu. *Nonparametric functional data analysis: theory and practice*, 2006.

Ramsay and Silverman. *Functional data analysis*, 2006.

Febrero-Bande and Oviedo de la Fuente. "Statistical computing in functional data analysis: the r package fda.usc". *Journal of Statistical Software*, 2012.



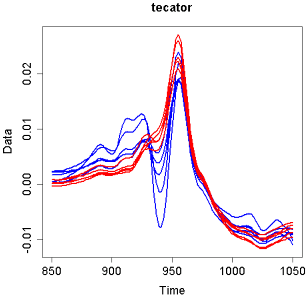
- $x \in \mathcal{C}^0([0, T])$
- $x \approx (x(t_1), \ldots, x(t_m)) \in \mathbb{R}^m$
- $(x(t_1), \ldots, x(t_m)) \approx x \in \mathcal{C}^0([0, T])$

Ferraty and Vieu. *Nonparametric functional data analysis: theory and practice*, 2006.

Ramsay and Silverman. *Functional data analysis*, 2006.

Febrero-Bande and Oviedo de la Fuente. "Statistical computing in functional data analysis: the r package fda.usc". *Journal of Statistical Software*, 2012.

Classification with functional data

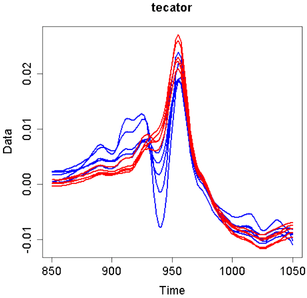


- $x \in \mathcal{C}^0([0,T])$
- $x \approx (x(t_1),\dots,x(t_m)) \in \mathbb{R}^m$
- $(x(t_1),\dots,x(t_m)) \approx x \in \mathcal{C}^0([0,T]) \approx (x(t_1),\dots,x(t_m))$

Ferraty and Vieu. *Nonparametric functional data analysis: theory and practice*, 2006.

Ramsay and Silverman. *Functional data analysis*, 2006.

Febrero-Bande and Oviedo de la Fuente. "Statistical computing in functional data analysis: the r package fda.usc". *Journal of Statistical Software*, 2012.

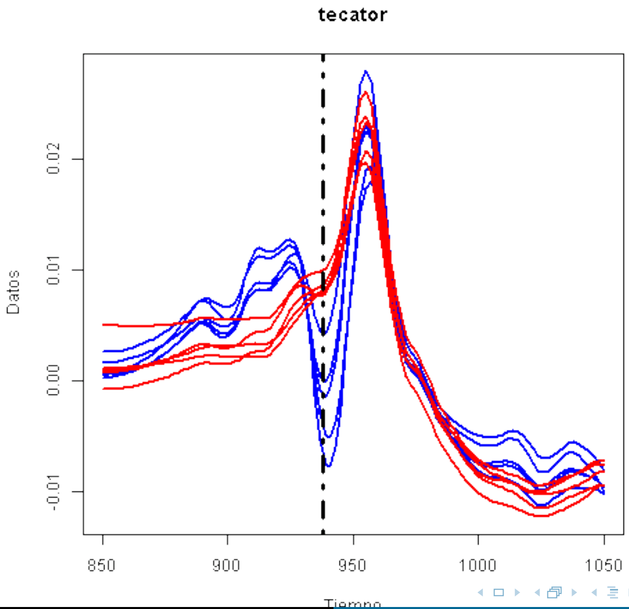


- $x \in \mathcal{C}^0([0, T])$
- $x \approx (x(t_1), \ldots, x(t_m)) \in \mathbb{R}^m$
- $(x(t_1), \ldots, x(t_m)) \approx x \in \mathcal{C}^0([0, T]) \approx (x(t_1), \ldots, x(t_m))$

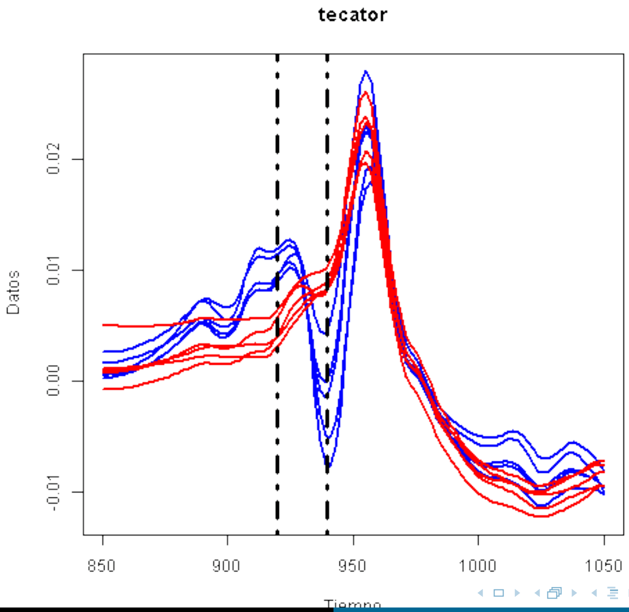
Blanquero, C., Jiménez-Cordero, Martín-Barragán. Variable Selection in Classification for Multivariate Functional Data. Researchgate, 2018.

Blanquero, C., Jiménez-Cordero, Martín-Barragán. Functional-bandwidth Kernel for Support Vector Machine with Functional Data: an Alternating Optimization Algorithm *European Journal of Operational Research* 2019

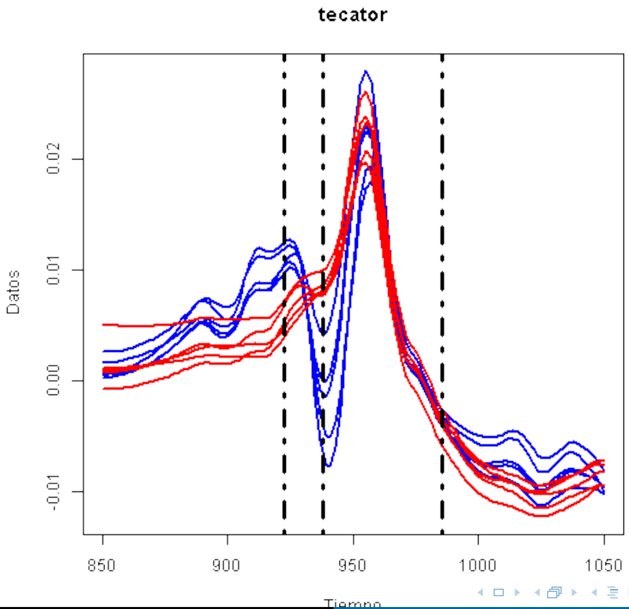
SVM with functional data



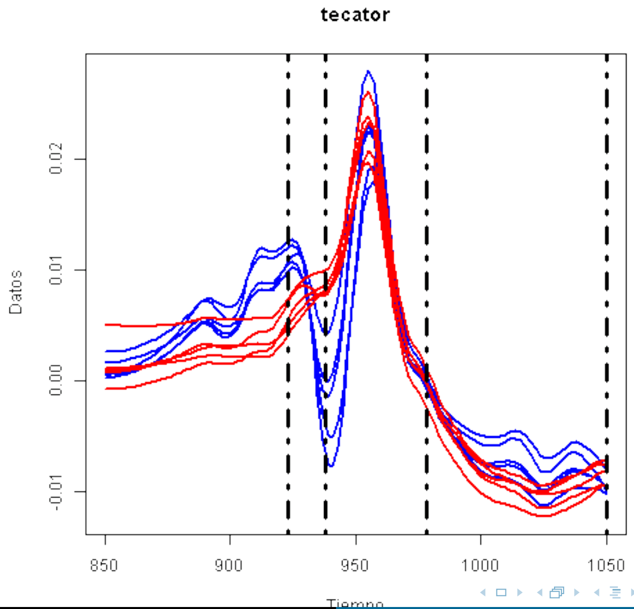
SVM with functional data



SVM with functional data



SVM with functional data



Cost-sensitive SVM

The performance measure

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I
\end{array}
$$

ω, β, C

The performance measure

- $\{(x_i, y_i) : i \in I\}$: seen as a random sample of (X, Y)
- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Distribution of (X, Y)

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

ω, β, C

The performance measure

- $\{(x_i, y_i) : i \in I\}$: seen as a random sample of (X, Y)
- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Distribution of (X, Y) : Unknown

The performance measure

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1-\xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

ω, β, C

The performance measure

- $\{(x_i, y_i) : i \in I\}$: seen as a random sample of (X, Y)
- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Distribution of (X, Y) : Unknown

The performance measure

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

$\omega, \beta,\ C$

The performance measure

- $\{(x_i, y_i) : i \in I\}$: seen as a random sample of (X, Y)
- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Distribution of (X, Y) : Unknown

(Asymmetric) costs

C., Martín-Barragán, Romero Morales. "Multi-group support vector machines with measurement costs: A biobjective approach". Discrete Applied Mathematics, 2008.

He, Ma. *Imbalanced learning: foundations, algorithms, and applications*. Wiley, 2013.

Maldonado, Pérez, Bravo. "Cost-based feature selection for support vector machines: An application in credit scoring". EJOR, 2017.

Prati, Batista, Duarte Silva. "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods". Knowledge and Information Systems. 2015.

Turney. "Types of cost in inductive concept learning". 2002.

Performance measures

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

ω, β, C

Performance measures $\pi(\omega, \beta)$:

- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Sensitivity: $TPR = P(\omega^\top X + \beta > 0 | Y = 1)$
- Specificity: $TNR = P(\omega^\top X + \beta < 0 | Y = -1)$
- Youden's index: $J = TPR + TNR - 1 = P(\omega^\top X + \beta > 0 | Y = 1) + P(\omega^\top X + \beta < 0 | Y = -1) - 1$
- Positive Predictive Value: $PPV = P(Y = 1 | \omega^\top X + \beta > 0)$
- Negative Predictive Value: $NPV = P(Y = -1 | \omega^\top X + \beta < 0)$
- ...

Performance measures

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i \in I} \xi_i & \\ \text{s.t.} & y_i \left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \end{array}$$

ω, β, C

Performance measures $\pi(\omega, \beta)$:

- Accuracy: $acc = P(Y(\omega^\top X + \beta) > 0)$
- Sensitivity: $TPR = P(\omega^\top X + \beta > 0 | Y = 1)$
- Specificity: $TNR = P(\omega^\top X + \beta < 0 | Y = -1)$
- Youden's index:
  $J = TPR + TNR - 1 = P(\omega^\top X + \beta > 0 | Y = 1) + P(\omega^\top X + \beta < 0 | Y = -1) - 1$
- Positive Predictive Value: $PPV = P(Y = 1 | \omega^\top X + \beta > 0)$
- Negative Predictive Value: $NPV = P(Y = -1 | \omega^\top X + \beta < 0)$
- ...

- Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.
- Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega,\beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\{(x_j, y_j) : j \in J\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi}_\ell(\omega,\beta;J), \ell \in L$
- Desired: $\pi_\ell(\omega,\beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega,\beta;J) \geq \gamma_\ell^*$, $\ell \in L$

- Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.
- Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega, \beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\{(x_j, y_j) : j \in J\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi}_\ell(\omega, \beta; J), \ell \in L$
- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma_\ell^*$, $\ell \in L$

Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.

Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega, \beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\{(x_j, y_j) : j \in J\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi}_\ell(\omega, \beta; J), \ell \in L$
- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma^*_\ell$, $\ell \in L$

Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.

Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega, \beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\left\{(x_j, y_j) : j \in J\right\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi}_\ell(\omega, \beta; J), \ell \in L$
- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma_\ell^*$, $\ell \in L$

Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.

Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega, \beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\{(x_j, y_j) : j \in J\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi_\ell}(\omega, \beta; J), \ell \in L$
- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma^*_\ell$, $\ell \in L$

Benítez Peña, Blanquero, C., Ramírez-Cobo. "On Support Vector Machines under a multiple-costscenario", *Advances in Data Analysis and Classification* 2019. 2018.

Benítez Peña, Blanquero, C., Ramírez-Cobo. "Cost-sensitive Feature Selection for Support Vector Machines", *Computers & OR* 2018.

- Performance measures $\pi_\ell(\omega, \beta), \ell \in L$
- Threshold values γ_ℓ for π_ℓ, $\ell \in L$
- I : training sample $\{(x_i, y_i) : i \in I\}$
- J : anchor sample $\{(x_j, y_j) : j \in J\}$ ($J \cap I = \emptyset$)
- Estimates of performance measures: $\widehat{\pi_\ell}(\omega, \beta; J), \ell \in L$
- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma^*_\ell$, $\ell \in L$

Standard approach

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I
\end{array}
$$

Constrained approach

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & \widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

Adding constraints to an SVM model

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & (\omega,\beta) \in \Omega & \end{array}$$

Ω : some (polyhedral) regions forced to be in one side of the separating hyperplane

C., and Plastria, "Optimal expected-distance separating halfspace", *Maths of OR*, 2008.

Fung, , Mangasarian, and Shavlik, "Knowledge-based support vector machine classifiers". In *Advances in NIPS*, 2002

Lauer and Bloch, "Incorporating prior knowledge in support vector machines for classification: A review", *Neurocomputing*, 2008.

Mangasarian, "Knowledge-based linear programming", *SIAM Journal on Optimization*, 2005.

Mangasarian and Wild, "Nonlinear knowledge-based classification", *IEEE Transactions on Neural Networks*, 2008.

Adding constraints to an SVM model

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & (\omega,\beta) \in \Omega & \end{array}$$

Ω : some (polyhedral) regions forced to be in one side of the separating hyperplane

- C., and Plastria, "Optimal expected-distance separating halfspace", *Maths of OR*, 2008.
- Fung, , Mangasarian, and Shavlik, "Knowledge-based support vector machine classifiers". In *Advances in NIPS*, 2002
- Lauer and Bloch, "Incorporating prior knowledge in support vector machines for classification: A review", *Neurocomputing*, 2008.
- Mangasarian, "Knowledge-based linear programming", *SIAM Journal on Optimization*, 2005.
- Mangasarian and Wild, "Nonlinear knowledge-based classification", *IEEE Transactions on Neural Networks*, 2008.

- Desired: $\pi_\ell(\omega,\beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega,\beta;J) \geq \gamma^*_\ell$, $\ell \in L$
- γ^*_ℓ : so that H_0 cannot be rejected in the test hypothesis

$$\begin{cases} H_0: & \pi_\ell(\omega,\beta) \geq \gamma_\ell \\ H_1: & \pi_\ell(\omega,\beta) < \gamma_\ell \end{cases}$$

Building γ^*_ℓ

- Hoeffding's inequality: for $Z_1,\ldots,Z_n$ $i.i.d.$, $Be(p)$, $P(\bar{Z}-p \geq c) \leq e^{-2nc^2}$.
- $100(1-\alpha)\%$ CI for p :

$$\left(\bar{Z} - \sqrt{\frac{\log\alpha}{-2n}}, 1\right)$$

- Imposing $p_0 \in CI$ means

$$\bar{Z} \geq p_0 + \sqrt{\frac{\log\alpha}{-2n}}$$

- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma_\ell^*$, $\ell \in L$
- γ_ℓ^* : so that H_0 cannot be rejected in the test hypothesis

$$\begin{cases} H_0 : & \pi_\ell(\omega, \beta) \geq \gamma_\ell \\ H_1 : & \pi_\ell(\omega, \beta) < \gamma_\ell \end{cases}$$

Building γ_ℓ^*

- Hoeffding's inequality: for $Z_1, \ldots, Z_n$ $i.i.d.$, $Be(p)$, $P(\bar{Z} - p \geq c) \leq e^{-2nc^2}$.
- $100(1-\alpha)\%$ CI for p :

$$\left(\bar{Z} - \sqrt{\frac{\log \alpha}{-2n}}, 1\right)$$

- Imposing $p_0 \in CI$ means

$$\bar{Z} \geq p_0 + \sqrt{\frac{\log \alpha}{-2n}}$$

- $\gamma_\ell^* = \gamma_\ell + \sqrt{\frac{\log \alpha}{-2|J|}}$

- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell,\ \ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma^*_\ell,\ \ell \in L$
- γ^*_ℓ : so that H_0 cannot be rejected in the test hypothesis

$$\begin{cases} H_0 : & \pi_\ell(\omega, \beta) \geq \gamma_\ell \\ H_1 : & \pi_\ell(\omega, \beta) < \gamma_\ell \end{cases}$$

Building γ^*_ℓ

- Hoeffding's inequality: for $Z_1, \ldots, Z_n$ *i.i.d.*, $Be(p)$, $P(\bar{Z} - p \geq c) \leq e^{-2nc^2}$.
- $100(1-\alpha)\%$ CI for p :

$$\left(\bar{Z} - \sqrt{\frac{\log \alpha}{-2n}}, 1\right)$$

- Imposing $p_0 \in CI$ means

$$\bar{Z} \geq p_0 + \sqrt{\frac{\log \alpha}{-2n}}$$

- $\gamma^*_\ell = \gamma_\ell + \sqrt{\frac{\log \alpha}{-2|J|}}$

- Desired: $\pi_\ell(\omega, \beta) \geq \gamma_\ell$, $\ell \in L$
- Imposed: $\widehat{\pi}_\ell(\omega, \beta; J) \geq \gamma^*_\ell$, $\ell \in L$
- γ^*_ℓ : so that H_0 cannot be rejected in the test hypothesis

$$\begin{cases} H_0 : & \pi_\ell(\omega, \beta) \geq \gamma_\ell \\ H_1 : & \pi_\ell(\omega, \beta) < \gamma_\ell \end{cases}$$

Building γ^*_ℓ

- Hoeffding's inequality: for $Z_1, \ldots, Z_n$ $i.i.d.$, $Be(p)$, $P(\bar{Z} - p \geq c) \leq e^{-2nc^2}$.
- $100(1-\alpha)\%$ CI for p :

$$\left(\bar{Z} - \sqrt{\frac{\log \alpha}{-2n}}, 1\right)$$

- Imposing $p_0 \in CI$ means

$$\bar{Z} \geq p_0 + \sqrt{\frac{\log \alpha}{-2n}}$$

- $\gamma^*_\ell = \gamma_\ell + \sqrt{\frac{\log \alpha}{-2|J|}}$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & \widehat{\pi_\ell}(\omega,\beta;J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & a_\ell^\top z \geq b_\ell & \ell \in L \\
 & z_\ell \in \{0,1\} & \ell \in L \\
 & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & y_i \left(\omega^\top x_i + \beta \right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& \widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j \left(\omega^\top x_j + \beta \right) \geq 1 \\ 0, & \text{else} \end{cases} \qquad j \in J
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & y_i \left(\omega^\top x_i + \beta \right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& a_\ell^\top z \geq b_\ell & \ell \in L \\
& z_\ell \in \{0, 1\} & \ell \in L \\
& y_j \left(\omega^\top x_j + \beta \right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& \widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{TPR}(\omega, \beta; J) \geq \gamma$

$$
\sum_{j\in J: y_j = 1} z_j \geq \gamma \# \left(\{j \in J : y_j = 1\}\right)
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& a_\ell^\top z \geq b_\ell & \ell \in L \\
& z_\ell \in \{0,1\} & \ell \in L \\
& y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & \widehat{\pi_\ell}(\omega,\beta;J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{TNR}(\omega,\beta;J) \geq \gamma$

$$
\sum_{j\in J: y_j=-1} z_j \geq \gamma\,\#\left(\{j \in J : y_j = -1\}\right)
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & a_\ell^\top z \geq b_\ell & \ell \in L \\
 & z_\ell \in \{0,1\} & \ell \in L \\
 & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & \widehat{\pi_\ell}(\omega,\beta;J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{J}(\omega,\beta;J) \geq \gamma$

$$
\sum_{j\in J} z_j \geq \gamma\,\#(J)
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & a_\ell^\top z \geq b_\ell & \ell \in L \\
 & z_\ell \in \{0,1\} & \ell \in L \\
 & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& \widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{TPR_m}(\omega, \beta; J) \geq \gamma$

$$
\sum_{j\in J: u_j = m} z_j \geq \gamma \,\#\left(\{j \in J : u_j = m\}\right)
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& a_\ell^\top z \geq b_\ell & \ell \in L \\
& z_\ell \in \{0,1\} & \ell \in L \\
& y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & \widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{PPV}(\omega, \beta; J) \geq \gamma$

$$
(1-\gamma)\, prev_+ \sum_{j\in J: y_j = 1} z_j - \gamma\,(1 - prev_+) \sum_{j\in J: y_j=-1} (1 - z_j) \geq 0
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C \sum_{i\in I} \xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \xi_i \geq 0 & i \in I \\
 & a_\ell^\top z \geq b_\ell & \ell \in L \\
 & z_\ell \in \{0,1\} & \ell \in L \\
 & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & \widehat{\pi_\ell}(\omega, \beta; J) \geq \gamma_\ell^* & \ell \in L \end{array}$$

$$z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J$$

$\widehat{NPV}(\omega, \beta; J) \geq \gamma$

$$(1-\gamma)\, prev_- \sum_{j\in J: y_j=-1} z_j - \gamma\,(1 - prev_-) \sum_{j\in J: y_j=1} (1 - z_j) \geq 0$$

$$\begin{array}{lll} \min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & a_\ell^\top z \geq b_\ell & \ell \in L \\ & z_\ell \in \{0,1\} & \ell \in L \\ & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J \end{array}$$

$$\begin{array}{lll} \min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & \widehat{\pi_\ell}(\omega,\beta;J) \geq \gamma_\ell^* & \ell \in L \end{array}$$

$$z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top x_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J$$

$\widehat{\pi}_\ell(\omega,\beta;J) \geq \gamma_\ell^*$

$$a_\ell^\top z \geq b_\ell$$

$$\begin{array}{lll} \min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\ \text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i & i \in I \\ & \xi_i \geq 0 & i \in I \\ & a_\ell^\top z \geq b_\ell & \ell \in L \\ & z_\ell \in \{0,1\} & \ell \in L \\ & y_j\left(\omega^\top x_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J \end{array}$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top \mathrm{x}_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& \widehat{\pi_\ell}(\omega,\beta;J) \geq \gamma_\ell^* & \ell \in L
\end{array}
$$

$$
z_j = \begin{cases} 1, & \text{if } y_j\left(\omega^\top \mathrm{x}_j + \beta\right) \geq 1 \\ 0, & \text{else} \end{cases} \quad j \in J
$$

$\widehat{\pi}_\ell(\omega,\beta;J) \geq \gamma_\ell^*$

$$
\mathrm{a}_\ell^\top \mathrm{z} \geq b_\ell
$$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,\mathrm{z}} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top \mathrm{x}_i + \beta\right) \geq 1 - \xi_i & i \in I \\
& \xi_i \geq 0 & i \in I \\
& \mathrm{a}_\ell^\top z \geq b_\ell & \ell \in L \\
& z_\ell \in \{0,1\} & \ell \in L \\
& y_j\left(\omega^\top \mathrm{x}_j + \beta\right) \geq 1 - M(1 - z_j) & j \in J
\end{array}
$$

- Denote $J(z) = \{j \in J : z_j = 1\}$

$$
\begin{array}{lll}
\min_{\omega,\beta,\xi,z} & \|\omega\|^2 + C\sum_{i\in I}\xi_i & \\
\text{s.t.} & y_i\left(\omega^\top x_i + \beta\right) \ge 1 - \xi_i & i \in I \\
 & \xi_i \ge 0 & i \in I \\
 & a_l^\top z \ge b_l & l \in L \\
 & z_\ell \in \{0,1\} & \ell \in L \\
 & y_j\left(\omega^\top x_j + \beta\right) \ge 1 - M(1 - z_j) & j \in J
\end{array}
$$

- Denote $J(z) = \{j \in J : \; z_j = 1\}$

$$
\begin{array}{llllll}
\min_{\mathrm{z}} & & & \min_{\omega,\beta,\xi} & \omega^\top \omega + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & z_\ell \in \{0,1\} & \ell \in L & \text{s.t.} & y_i \left(\omega^\top \mathrm{x}_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & \mathrm{a}_\ell^\top z \geq b_\ell & l \in L & & y_j \left(\omega^\top \mathrm{x}_j + \beta\right) \geq 1 & j \in J(\mathrm{z}) \\
 & & & & \xi_i \geq 0 & i \in I
\end{array}
$$

- Denote $J(z) = \{j \in J : z_j = 1\}$

$$
\begin{array}{llll}
\min_{\mathrm{z}} & & & \\
\text{s.t.} & z_\ell \in \{0,1\} & \ell \in L & \\
 & \mathrm{a}_\ell^\top z \geq b_\ell & l \in L &
\end{array}
\qquad
\begin{array}{lll}
\min_{\omega,\beta,\xi} & \omega^\top \omega + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & y_i \left(\omega^\top \mathrm{x}_i + \beta\right) \geq 1 - \xi_i & i \in I \\
 & y_j \left(\omega^\top \mathrm{x}_j + \beta\right) \geq 1 & j \in J(\mathrm{z}) \\
 & \xi_i \geq 0 & i \in I
\end{array}
$$

KKT conditions for inner problem (z fixed)

$$
\begin{array}{rcll}
\omega & = & \sum_{s \in I} \lambda_s y_s \mathrm{x}_s + \sum_{t \in J(\mathrm{z})} \mu_t y_t \mathrm{x}_t & \\
0 & = & \sum_{s \in I} \lambda_s y_s + \sum_{t \in J(\mathrm{z})} \mu_t y_t & \\
0 & \leq & \lambda_s \leq C/2 & s \in I \\
0 & \leq & \mu_t & t \in J(\mathrm{z})
\end{array}
$$

- Denote $J(z) = \{j \in J : z_j = 1\}$

$$
\begin{array}{llll}
\min_{z} & & & \min_{\omega,\beta,\xi} \quad \omega^\top\omega + C\sum_{i\in I}\xi_i \\
\text{s.t.} & z_\ell \in \{0,1\} & \ell \in L & \text{s.t.} \quad y_i\left(\omega^\top x_i + \beta\right) \geq 1 - \xi_i \quad i \in I \\
 & a_\ell^\top z \geq b_\ell & l \in L & \qquad\quad y_j\left(\omega^\top x_j + \beta\right) \geq 1 \quad j \in J(z) \\
 & & & \qquad\quad \xi_i \geq 0 \quad i \in I
\end{array}
$$

KKT conditions for inner problem (z fixed)

$$
\begin{array}{lllr}
\omega & = & \sum_{s\in I}\lambda_s y_s x_s + \sum_{t\in J}\mu_t y_t x_t & \\
0 & = & \sum_{s\in I}\lambda_s y_s + \sum_{t\in J}\mu_t y_t & \\
0 & \leq & \lambda_s \leq C/2 & s \in I \\
0 & \leq & \mu_t \leq M z_t & t \in J
\end{array}
$$

(Partial) Dual

$$
\begin{array}{lll}
\min_{\lambda,\mu,\beta,\xi,z} & \left(\sum_{s\in I} \lambda_s y_s \mathrm{x}_s + \sum_{t\in J} \mu_t y_t \mathrm{x}_t\right)^\top \left(\sum_{s\in I} \lambda_s y_s \mathrm{x}_s + \sum_{t\in J} \mu_t y_t \mathrm{x}_t\right) & \\
 & +C \sum_{i\in I} \xi_i & \\
\text{s.t.} & z_\ell \in \{0,1\} & \ell \in L \\
 & \mathrm{a}_\ell^\top z \geq b_\ell & \ell \in L \\
 & y_i \left(\left(\sum_{s\in I} \lambda_s y_s \mathrm{x}_s + \sum_{t\in J} \mu_t y_t \mathrm{x}_t\right)^\top \mathrm{x}_i + \beta \right) \geq 1 - \xi_i & i \in I \\
 & y_j \left(\left(\sum_{s\in I} \lambda_s y_s \mathrm{x}_s + \sum_{t\in J} \mu_t y_t \mathrm{x}_t\right)^\top \mathrm{x}_j + \beta \right) \geq 1 - M(1 - z_j) & j \in J \\
 & \xi_i \geq 0 & i \in I \\
 & 0 \leq \lambda_i \leq C/2 & i \in I \\
 & 0 \leq \mu_j \leq M z_j & j \in J
\end{array}
$$

(Partial) Dual: The kernel trick

$$
\begin{array}{lll}
\min & \sum_{s,s' \in I} \lambda_s y_s \lambda_{s'} y_{s'} K(\mathrm{x}_s, \mathrm{x}_{s'}) + \sum_{t,t' \in J} \mu_t y_t \mu_{t'} y_{t'} K(\mathrm{x}_t, \mathrm{x}_{t'}) & \\
 & +2 \sum_{s \in I, t \in J} \lambda_s y_s \lambda_t y_t K(\mathrm{x}_s, \mathrm{x}_t) + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & z_\ell \in \{0, 1\} & \ell \in L \\
 & \mathrm{a}_\ell^\top z \geq b_\ell & \ell \in L \\
 & y_i \left(\sum_{s \in I} \lambda_s y_s K(\mathrm{x}_s, \mathrm{x}_i) + \sum_{t \in J} \mu_t y_t K(\mathrm{x}_t, \mathrm{x}_i) + \beta \right) \geq 1 - \xi_i & i \in I \\
 & y_j \left(\sum_{s \in I} \lambda_s y_s K(\mathrm{x}_s, \mathrm{x}_j) + \sum_{t \in J} \mu_t y_t K(\mathrm{x}_t, \mathrm{x}_j) + \beta \right) \geq 1 - M(1 - z_j) & j \in J \\
 & \xi_i \geq 0 & i \in I \\
 & 0 \leq \lambda_i \leq C/2 & i \in I \\
 & 0 \leq \mu_j \leq M z_j & j \in J
\end{array}
$$

(Partial) Dual: The kernel trick

$$
\begin{array}{lll}
\min & \sum_{s,s' \in I} \lambda_s y_s \lambda_{s'} y_{s'} K(\mathrm{x}_s, \mathrm{x}_{s'}) + \sum_{t,t' \in J} \mu_t y_t \mu_{t'} y_{t'} K(\mathrm{x}_t, \mathrm{x}_{t'}) & \\
 & +2 \sum_{s \in I, t \in J} \lambda_s y_s \lambda_t y_t K(\mathrm{x}_s, \mathrm{x}_t) + C \sum_{i \in I} \xi_i & \\
\text{s.t.} & z_\ell \in \{0, 1\} & \ell \in L \\
 & \mathrm{a}_\ell^\top z \geq b_\ell & \ell \in L \\
 & y_i \left(\sum_{s \in I} \lambda_s y_s K(\mathrm{x}_s, \mathrm{x}_i) + \sum_{t \in J} \mu_t y_t K(\mathrm{x}_t, \mathrm{x}_i) + \beta \right) \geq 1 - \xi_i & i \in I \\
 & y_j \left(\sum_{s \in I} \lambda_s y_s K(\mathrm{x}_s, \mathrm{x}_j) + \sum_{t \in J} \mu_t y_t K(\mathrm{x}_t, \mathrm{x}_j) + \beta \right) \geq 1 - M(1 - z_j) & j \in J \\
 & \xi_i \geq 0 & i \in I \\
 & 0 \leq \lambda_i \leq C/2 & i \in I \\
 & 0 \leq \mu_j \leq M z_j & j \in J
\end{array}
$$

Parameters involved:

- C, to be tuned
- M, to be fixed

Experiments

- RBF kernel, parameters tuned by grid search
- Python + Gurobi
- $M = 100$, time.limit = 300 sec

Data sets

Name	$\|\Omega\|$	V	$\|\Omega_+\|$	(%)
australian	690	14	383	(55.5%)
votes	435	16	267	(61.4%)
wisconsin	567	30	357	(62.7%)
german	1000	45	700	(70%)
pageBlocks	5472	10	558	(10.2%)
biodeg	1055	41	356	(33.7%)

Experiments

- RBF kernel, parameters tuned by grid search
- Python + Gurobi
- $M = 100$, time.limit $= 300$ sec

Data sets

Name	$\|\Omega\|$	V	$\|\Omega_+\|$	(%)
australian	690	14	383	(55.5%)
votes	435	16	267	(61.4%)
wisconsin	567	30	357	(62.7%)
german	1000	45	700	(70%)
pageBlocks	5472	10	558	(10.2%)
biodeg	1055	41	356	(33.7%)

Results

Name		SVM	SVM(C_+, C_-)	Sliding β	CSVM
		Mean (Std)	Mean (Std)	Mean (Target) (Std)	Mean (Target) (Std)
australian	TPR	**0.83** (0.071)	**0.806** (0.093)	**0.821** (0.855) (0.073)	**0.903** (0.855) (0.05)
	TNR	0.863 (0.079)	0.878 (0.088)	0.855 (0.068)	0.772 (0.081)
votes	TPR	**0.963** (0.04)	**0.945** (0.042)	**0.971** (0.988) (0.037)	**0.978** (0.988) (0.026)
	TNR	0.951 (0.031)	0.941 (0.037)	0.91 (0.063)	0.922 (0.04)
wisconsin	TPR	**0.948** (0.049)	**0.962** (0.027)	**0.989** (0.973) (0.017)	**0.965** (0.973) (0.037)
	TNR	0.99 (0.017)	0.931 (0.07)	0.953 (0.045)	0.945 (0.045)
german	TPR	**0.464** (0.103)	**0.89** (0.08)	**0.043** (0.65) (0.023)	**0.671** (0.65) (0.164)
	TNR	0.847 (0.031)	0.407 (0.069)	0.996 (0.009)	0.668 (0.111)
pageBlocks	TPR	**0.807** (0.03)	**0.557** (0.361)	**0.819** (0.832) (0.981)	**0.859** (0.832) (0.045)
	TNR	0.988 (0.004)	**0.901** (0.088)	0.981 (0.006)	0.965 (0.012)
biodeg	TPR	**0.783** (0.084)	**0.793** (0.083)	**0.797** (0.808) (0.095)	**0.852** (0.808) (0.057)
	TNR	0.909 (0.032)	0.839 (0.037)	0.891 (0.037)	0.833 (0.05)

Cost-sensitive SVM Feature Selection

Feature selection

Aim

- Find a minimum-cost (e.g. minimum-cardinality) set of features
  - Attaining $\widehat{\pi}_\ell(\omega, \beta) \geq \gamma_\ell^*,\ \ell \in \boldsymbol{L}$
  - Hoping $\pi_\ell(\omega, \beta; \boldsymbol{I}) \geq \gamma_\ell,\ \ell \in \boldsymbol{L}$
- Once identified the features, solve an SVM

Feature selection. Linear kernel

$$
\begin{array}{lll}
\min_{\boldsymbol{w},\beta,z,\zeta} & \sum_{k=1}^{N} \delta_k z_k & \\
s.t. & y_i(\boldsymbol{w}^\top x_i + \beta) \geq 1 - L(1-\zeta_i), & \forall i \in I \\
& \sum_{i\in I} \zeta_i(1-y_i) \geq \lambda_{-1} \sum_{i\in I}(1-y_i) & \\
& \sum_{i\in I} \zeta_i(1+y_i) \geq \lambda_{1} \sum_{i\in I}(1+y_i) & \\
& |w_k| \leq M z_k & \forall k \in 1,\ldots,N \\
& \zeta_i \in \{0,1\} & \forall i \in I \\
& z_k \in \{0,1\} & \forall k \in 1,\ldots,N
\end{array}
$$

Results. Linear kernel

Name		SVM Mean	SVM Std	FS Mean	FS Std	Feature reduction
wisconsin	Acc	0.975	0.021	0.947	0.025	30 → 2 (0 Std)
	TPR	0.992	0.013	0.973	0.031	
	TNR	0.943	0.051	0.905	0.063	
votes	Acc	0.954	0.033	0.949	0.036	32 → 2 (0 Std)
	TPR	0.955	0.038	0.928	0.059	
	TNR	0.947	0.059	0.979	0.036	
nursery	Acc	1	0	1	0	19 → 1 (0 Std)
	TPR	1	0	1	0	
	TNR	1	0	1	0	
Australian	Acc	0.848	0.051	0.855	0.057	34 → 1 (0 Std)
	TPR	0.798	0.083	0.801	0.087	
	TNR	0.912	0.05	0.926	0.041	
careval	Acc	0.956	0.017	0.946	0.019	15 → 9 (0 Std)
	TPR	0.96	0.022	0.963	0.017	
	TNR	0.948	0.024	0.907	0.04	
leukemia	Acc	0.972	0.164	0.875	0.331	7128 → 3.139 (1.205 Std)
	TPR	0.979	0.196	0.896	0.305	
	TNR	0.96	0.144	0.833	0.373	
gastrointestinal	Acc	0.895	0.307	0.829	0.379	698 → 1 (0 Std)
	TPR	0.929	0.258	0.839	0.367	
	TNR	0.8	0.4	0.8	0.4	

Results. Radial kernel

Discrete-Level SVM

E. Carrizosa, A. Nogales-Gómez amd D. Romero Morales
Strongly Agree or Strongly Disagree?: Rating Features in Support Vector Machines, *Information Sciences*, 2015

Baesens, B., R. Setiono, C. Mues, J. Vanthienen
"Using neural network rule extraction and decision tables for credit-risk evaluation", *Management Science,* 2003.

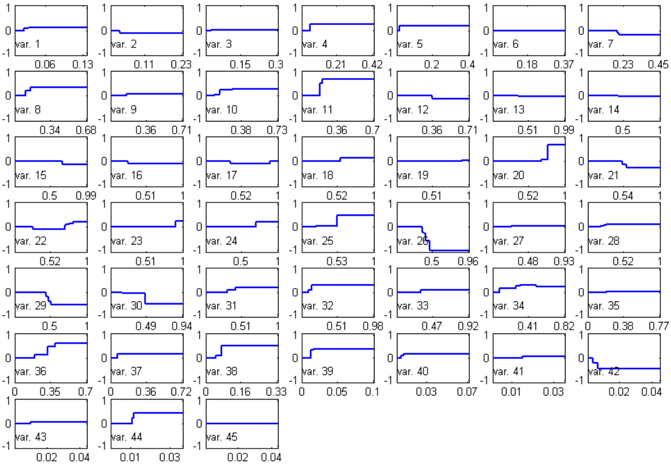
Carrizosa, E., Nogales-Gómez, A., and D. Romero Morales
"Strongly Agree or Strongly Disagree?: Rating Features in Support Vector Machines", *Information Sciences,* 2015.

Carrizosa, E., B. Martín-Barragán and D. Romero Morales
"Binarized Support Vector Machines", *IJOC,* 2010.

Carrizosa, E., B. Martín-Barragán and D. Romero Morales
"Detecting relevant variables and interactions in supervised classification", *EJOR*, 2011.

Chang, A., D. Bertsimas and C. Rudin
An Integer Optimization Approach to Associative Classification, *Advances in NIPS,* 2012.

Romero Morales, D., J. Wang.
"Forecasting cancellation rates for services booking revenue management using data mining", *EJOR,* 2010.


var. 1
var. 2
var. 3
var. 4
var. 5
var. 6
var. 7
var. 8
var. 9
var. 10
var. 11
var. 12
var. 13
var. 14
var. 15
var. 16
var. 17
var. 18
var. 19
var. 20
var. 21
var. 22
var. 23
var. 24
var. 25
var. 26
var. 27
var. 28
var. 29
var. 30
var. 31
var. 32
var. 33
var. 34
var. 35
var. 36
var. 37
var. 38
var. 39
var. 40
var. 41
var. 42
var. 43
var. 44
var. 45

Discrete-Level SVM

C., Nogales-Gómez, Romero Morales. *Information Science*, 2015

Impose $\omega_j \in \mathcal{A}$, finite, e.g., $\mathcal{A} = \{-a, 0, a\}$

The following features contribute to being a good customer:

		Strongly agree	Neutral	Strongly Disagree
Status of existing checking account	$\ldots < 0$ DM		×	
	$0 \leq \ldots < 200$ DM		×	
	$\ldots \geq 200$ DM	×		
	No checking account	×		
Credit duration				×
Credit history	no credits taken/all credits paid back duly			×
	all credits at this bank paid back duly			×
	existing credits paid back duly till now		×	
	delay in paying off in the past		×	
	critical account/other credits existing (not at this bank)	×		
Purpose	new car			×
	used car		×	
	furniture/equipement		×	
	radio/television		×	
	domestic appliances			×
	repairs		×	
	education			×
	vacation		×	
	retraining	×		
	business		×	
	others		×	
Credit amount				×
Saving accounts	$\ldots < 100$ DM		×	
	$100 \leq \ldots < 500$		×	
	$500 \leq \ldots < 1000$	×		
	$\ldots \geq 1000$	×		
	unknown/no saving accounts	×		
Duration of present employment	unemployed		×	
	$\ldots < 1$ year		×	
	$1 \leq \ldots < 4$ years		×	
	$4 \leq \ldots < 7$ years		×	
	$\ldots \geq 7$ years		×	
Installment rate in % of disposable income				×
Personal status and sex	male (divorced/separated)			×
	female (divorced/separated/married)		×	
	male (single)		×	
	male (married/widowed)		×	
	female (single)		×	
Other debtors/guarantors	none		×	
	co-applicant		×	
	guarantor	×		
Duration of present residence			×	
Property	real estate		×	
	building society savings agreement/life insurance		×	
	car/others		×	
	unknown/no property			×
Age		×		
Other installment plans	bank		×	
	stores		×	
	none		×	
Housing	rent		×	
	own		×	
	for free	×		
Number of existing credits at this bank				×
Job	unemployed/unskilled (non-resident)		×	
	unskilled (resident)		×	
	skilled employee/official		×	
	management/self-employed/highly qualified employee officer		×	
Number of people being liable to provide maintenance for			×	
Telephone	None		×	
	Yes		×	
Foreign worker	Yes		×	
	No	×		

Discrete-Level SVM

C., Nogales-Gómez, Romero Morales. *Information Science*, 2015

Impose $\omega_j \in \mathcal{A}$, finite, e.g., $\mathcal{A} = \{-b, -a, 0, a, b\}$

The following features contribute to being a good customer:

		Strongly agree	Agree	Neutral	Disagree	Strongly Disagree
Status of existing checking account	$\ldots < 0$ DM				×	
	$0 \leq \ldots < 200$ DM			×		
	$\ldots \geq 200$ DM	×				
	No checking account	×				
Credit duration						×
Credit history	no credits taken/all credits paid back duly					×
	all credits at this bank paid back duly					×
	existing credits paid back duly till now			×		
	delay in paying off in the past		×			
	critical account/other credits existing (not at this bank)	×				
Purpose	new car					×
	used car			×		
	furniture/equipment				×	
	radio/television				×	
	domestic appliances					×
	repairs				×	
	education					×
	vacation			×		
	retraining	×				
	business				×	
	others			×		
Credit amount						×
Saving accounts	$\ldots < 100$ DM		×			
	$100 \leq \ldots < 500$		×			
	$500 \leq \ldots < 1000$	×				
	$\ldots \geq 1000$	×				
	unknown/no saving accounts	×				
Duration of present employment	unemployed			×		
	$\ldots < 1$ year			×		
	$1 \leq \ldots < 4$ years			×		
	$4 \leq \ldots < 7$ years		×			
	$\ldots \geq 7$ years			×		
Installment rate in % of disposable income						×
Personal status and sex	male (divorced/separated)					×
	female (divorced/separated/married)				×	
	male (single)			×		
	male (married/widowed)			×		
	female (single)			×		
Other debtors/guarantors	none			×		
	co-applicant				×	
	guarantor	×				
Duration of present residence				×		
Property	real estate			×		
	building society savings agreement/life insurance			×		
	car/others			×		
	unknown/no property					×
Age		×				
Other installment plans	bank			×		
	stores			×		
	none		×			
Housing	rent		×			
	own		×			
	for free	×				
Number of existing credits at this bank						×
Job	unemployed/unskilled (non-resident)			×		
	unskilled (resident)			×		
	skilled employee/official			×		
	management/self-employed/highly qualified employee officer			×		
Number of people being liable to provide maintenance for				×		
Telephone	None			×		
	Yes			×		
Foreign worker	Yes			×		
	No	×				

Impose $\omega_j \in \mathcal{A}$, finite

Impose $\omega_j \in \mathcal{A}$, finite

$$\begin{array}{lll} \min & \omega^\top \omega + C \sum_{i=1}^n \xi_i & \\ & y_i(\omega^\top x_i + \beta) \geq 1 - \xi_i, & i, i = 1, 2, \ldots, n \\ & \xi_i \geq 0, & i, i = 1, 2, \ldots, n \\ & \omega \in \mathbb{R}^d,\ \beta \in \mathbb{R}. & \end{array}$$

DiLSVM

Impose $\omega_j \in \mathcal{A}$, finite

$$
\begin{array}{lll}
\min & \omega^\top \omega + C \sum_{i=1}^n \xi_i & \\
& y_i(\omega^\top x_i + \beta) \geq 1 - \xi_i, & i, i = 1, 2, \ldots, n \\
& \xi_i \geq 0, & i, i = 1, 2, \ldots, n \\
& \omega \in \mathcal{A}^d, \ \beta \in \mathbb{R}. &
\end{array}
$$

$$
\begin{array}{ll}
\min & \omega^\top \omega + C\sum_{i=1}^n \xi_i \\
& y_i(\omega^\top x_i + \beta) \geq 1 - \xi_i, \quad i, i = 1, 2, \ldots, n \\
& \xi_i \geq 0, \quad i, i = 1, 2, \ldots, n \\
& \omega \in \mathcal{A}^d, \ \beta \in \mathbb{R}.
\end{array}
$$

- Define
$$
z_{ja} = \begin{cases} 1, & \text{if } \omega_j = a \\ 0, & \text{else} \end{cases}
$$

DiLSVM problem:

$$
\begin{array}{lll}
\min & \sum_{j=1}^d \sum_{a\in\mathcal{A}} a^2 z_{ja} + C\sum_{i=1}^n \xi_i & \\
& y_i(\sum_{j=1}^d \sum_{a\in\mathcal{A}} a z_{ja} x_{ij} + \beta) \geq 1 - \xi_i & i = 1, 2, \ldots, n \\
& \sum_{a\in\mathcal{A}} z_{ja} = 1 & j = 1, \ldots, d \\
& \xi_i \geq 0 & i = 1, 2, \ldots, n \\
& z_{ja} \in \{0, 1\} & \forall j = 1, \ldots, d, \ a \in \mathcal{A} \\
& \beta \in \mathbb{R}. &
\end{array}
$$

DiLSVM

$$
\begin{array}{ll}
\min & \omega^\top \omega + C\sum_{i=1}^n \xi_i \\
& y_i(\omega^\top x_i + \beta) \geq 1 - \xi_i, \qquad i, i = 1, 2, \ldots, n \\
& \xi_i \geq 0, \qquad i, i = 1, 2, \ldots, n \\
& \omega \in \mathcal{A}^d, \ \beta \in \mathbb{R}.
\end{array}
$$

- Define
$$z_{ja} = \begin{cases} 1, & \text{if } \omega_j = a \\ 0, & \text{else} \end{cases}$$

DiLSVM problem:

$$
\begin{array}{lll}
\min & \sum_{j=1}^d \sum_{a\in\mathcal{A}} a^2 z_{ja} + C\sum_{i=1}^n \xi_i & \\
& y_i(\sum_{j=1}^d \sum_{a\in\mathcal{A}} a z_{ja} x_{ij} + \beta) \geq 1 - \xi_i & i = 1, 2, \ldots, n \\
& \sum_{a\in\mathcal{A}} z_{ja} = 1 & j = 1, \ldots, d \\
& \xi_i \geq 0 & i = 1, 2, \ldots, n \\
& z_{ja} \in \{0, 1\} & \forall j = 1, \ldots, d, \ a \in \mathcal{A} \\
& \beta \in \mathbb{R}. &
\end{array}
$$

DiLSVM

$$
\begin{array}{lll}
\min & \omega^\top \omega + C\sum_{i=1}^n \xi_i & \\
& y_i(\omega^\top x_i + \beta) \geq 1 - \xi_i, & i, i = 1, 2, \ldots, n \\
& \xi_i \geq 0, & i, i = 1, 2, \ldots, n \\
& \omega \in \mathcal{A}^d, \ \beta \in \mathbb{R}. &
\end{array}
$$

- Define

$$
z_{ja} = \begin{cases} 1, & \text{if } \omega_j = a \\ 0, & \text{else} \end{cases}
$$

DiLSVM problem:

$$
\begin{array}{lll}
\min & \sum_{j=1}^d \sum_{a\in\mathcal{A}} a^2 z_{ja} + C\sum_{i=1}^n \xi_i & \\
& y_i(\sum_{j=1}^d \sum_{a\in\mathcal{A}} a z_{ja} x_{ij} + \beta) \geq 1 - \xi_i & i = 1, 2, \ldots, n \\
& \sum_{a\in\mathcal{A}} z_{ja} = 1 & j = 1, \ldots, d \\
& \xi_i \geq 0 & i = 1, 2, \ldots, n \\
& z_{ja} \in \{0, 1\} & \forall j = 1, \ldots, d, \ a \in \mathcal{A} \\
& \beta \in \mathbb{R}. &
\end{array}
$$

Performance of the DILSVM$^{(1)}$ classifier

Name	*SVM*		*CKZ* $C = \infty, a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
adult	84.89	0.00	82.93	55.93
mushroom	99.98	0.00	99.98	49.08
german	72.96	4.12	72.52	51.43
ijcnn1	91.37	0.00	90.18	81.82
careval	95.05	0.00	86.92	30.00
gamma	79.32	0.00	70.37	31.00
abalone	79.52	0.00	74.28	15.00
shuttle	98.17	0.00	80.83	15.55
cod-rna	93.86	0.00	66.56	87.50
calhous	83.59	0.00	63.57	0.00

Name	*MILP* $a_1 \in \{2^0, \ldots, 2^{10}\}$		*RSVM* $a_1 \in \{2^0, \ldots, 2^{10}\}$		*RR* $a_1 \in \{2^0, \ldots, 2^{10}\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
adult	90.43	79.27	75.81	100.00	82.33	97.36
mushroom	100.00	91.60	52.01	100.00	99.47	97.56
german	75.32	70.95	69.24	98.89	69.28	93.02
ijcnn1	90.17	100.00	90.18	98.18	90.17	96.59
careval	96.21	59.52	82.03	62.86	87.61	81.19
gamma	79.09	50.00	77.95	80.00	78.31	73.00
abalone	78.28	70.00	77.19	72.00	76.91	75.50
shuttle	97.32	63.33	84.30	75.56	93.75	68.33
cod-rna	78.17	87.50	71.05	88.75	76.13	93.75
calhous	81.98	31.25	78.60	50.00	78.17	65.00

Performance of the DILSVM$^{(1)}$ classifier

Name	*SVM*		*CKZ* $C = \infty, a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
adult	84.89	0.00	82.93	55.93
mushroom	99.98	0.00	99.98	49.08
german	72.96	4.12	72.52	51.43
ijcnn1	91.37	0.00	90.18	81.82
careval	95.05	0.00	86.92	30.00
gamma	79.32	0.00	70.37	31.00
abalone	79.52	0.00	74.28	15.00
shuttle	98.17	0.00	80.83	15.55
cod-rna	93.86	0.00	66.56	87.50
calhous	83.59	0.00	63.57	0.00

Name	*MILP* vs *SVM* $a_1 \in \{2^0, \dots, 2^{10}\}$		*RSVM* $a_1 \in \{2^0, \dots, 2^{10}\}$		*RR* $a_1 \in \{2^0, \dots, 2^{10}\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
adult	**B**	**B**	75.81	100.00	82.33	97.36
mushroom		**B**	52.01	100.00	99.47	97.56
german	**B**	**B**	69.24	98.89	69.28	93.02
ijcnn1	**W**	**B**	90.18	98.18	90.17	96.59
careval	**B**	**B**	82.03	62.86	87.61	81.19
gamma		**B**	77.95	80.00	78.31	73.00
abalone	**W**	**B**	77.19	72.00	76.91	75.50
shuttle		**B**	84.30	75.56	93.75	68.33
cod-rna	**W**	**B**	71.05	88.75	76.13	93.75
calhous	**W**	**B**	78.60	50.00	78.17	65.00

Performance of the DILSVM$^{(2)}$ classifier

Name	*SVM*		*CKZ* $C = \infty,\ a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
adult	84.89	0.00	82.93	55.93
mushroom	99.98	0.00	99.98	49.08
german	72.96	4.12	72.52	51.43
ijcnn1	91.37	0.00	90.18	81.82
careval	95.05	0.00	86.92	30.00
gamma	79.32	0.00	70.37	31.00
abalone	79.52	0.00	74.28	15.00
shuttle	98.17	0.00	80.83	15.55
cod-rna	93.86	0.00	66.56	87.50
calhous	83.59	0.00	63.57	0.00

Name	*MILP* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RSVM* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RR* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*fixing* $\bar{C}, \bar{a}_1$ $a_2 \in \{\bar{a}_1/2, \bar{a}_1/2^2\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
adult	90.56	77.40	77.88	90.33	89.04	93.33	89.08	67.07
mushroom	100.00	91.60	99.41	95.80	100.00	100.00	100.00	91.60
german	75.96	55.87	70.92	57.78	73.00	54.76	71.60	35.56
ijcnn1	90.18	95.46	90.46	65.91	90.25	70.45	90.17	100.00
careval	98.71	35.72	89.81	43.34	98.76	53.81	96.84	51.90
gamma	79.31	50.00	78.28	52.00	79.19	57.00	79.36	32.00
abalone	79.05	55.00	78.28	57.00	77.64	65.00	78.43	61.00
shuttle	97.59	45.55	92.16	43.33	96.30	43.33	97.32	47.78
cod-rna	90.64	42.50	73.94	82.50	80.10	62.50	81.32	46.25
calhous	83.03	17.50	80.14	25.00	80.47	17.50	82.25	20.00

Performance of the DILSVM$^{(2)}$ classifier

Name	SVM		CKZ $C = \infty, a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
adult	84.89	0.00	82.93	55.93
mushroom	99.98	0.00	99.98	49.08
german	72.96	4.12	72.52	51.43
ijcnn1	91.37	0.00	90.18	81.82
careval	95.05	0.00	86.92	30.00
gamma	79.32	0.00	70.37	31.00
abalone	79.52	0.00	74.28	15.00
shuttle	98.17	0.00	80.83	15.55
cod-rna	93.86	0.00	66.56	87.50
calhous	83.59	0.00	63.57	0.00

Name	MILP vs SVM $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		RSVM $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		RR $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		fixing $\bar{C}, \bar{a}_1$ $a_2 \in \{\bar{a}_1/2, \bar{a}_1/2^2\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
adult	B	B	77.88	90.33	89.04	<u>93.33</u>	<u>89.08</u>	67.07
mushroom		B	99.41	95.80	<u>100.00</u>	<u>100.00</u>	<u>100.00</u>	91.60
german	B	B	70.92	<u>57.78</u>	73.00	54.76	71.60	35.56
ijcnn1	W	B	<u>90.46</u>	65.91	90.25	70.45	90.17	<u>100.00</u>
careval	B	B	89.81	43.34	<u>98.76</u>	<u>53.81</u>	96.84	51.90
gamma		B	78.28	52.00	79.19	<u>57.00</u>	<u>79.36</u>	32.00
abalone		B	78.28	57.00	77.64	<u>65.00</u>	78.43	61.00
shuttle		B	92.16	43.33	96.30	43.33	97.32	<u>47.78</u>
cod-rna	W	B	73.94	<u>82.50</u>	80.10	62.50	81.32	46.25
calhous		B	80.14	<u>25.00</u>	80.47	17.50	82.25	20.00

Performance of the DILSVM$^{(2)}$ classifier

Name	*SVM*		*CKZ* $C = \infty, a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
`adult`	84.89	0.00	82.93	55.93
`mushroom`	99.98	0.00	99.98	49.08
`german`	72.96	4.12	72.52	51.43
`ijcnn1`	91.37	0.00	90.18	81.82
`careval`	95.05	0.00	86.92	30.00
`gamma`	79.32	0.00	70.37	31.00
`abalone`	79.52	0.00	74.28	15.00
`shuttle`	98.17	0.00	80.83	15.55
`cod-rna`	93.86	0.00	66.56	87.50
`calhous`	83.59	0.00	63.57	0.00

Name	*MILP* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RSVM* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RR* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*fixing* vs *SVM* $\bar{C}, \bar{a}_1$ $a_2 \in \{\bar{a}_1/2, \bar{a}_1/2^2\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
`adult`	90.56	77.40	77.88	90.33	89.04	93.33	**B**	**B**
`mushroom`	100.00	91.60	99.41	95.80	100.00	100.00		**B**
`german`	75.96	55.87	70.92	57.78	73.00	54.76	**W**	**B**
`ijcnn1`	90.18	95.46	90.46	65.91	90.25	70.45	**W**	**B**
`careval`	98.71	35.72	89.81	43.34	98.76	53.81	**B**	**B**
`gamma`	79.31	50.00	78.28	52.00	79.19	57.00		**B**
`abalone`	79.05	55.00	78.28	57.00	77.64	65.00	**W**	**B**
`shuttle`	97.59	45.55	92.16	43.33	96.30	43.33		**B**
`cod-rna`	90.64	42.50	73.94	82.50	80.10	62.50	**W**	**B**
`calhous`	83.03	17.50	80.14	25.00	80.47	17.50	**W**	**B**

Performance of the DILSVM$^{(2)}$ classifier

Name	*SVM*		*CKZ* $C = \infty, a_1 = 1$	
	Accuracy	Sparsity	Accuracy	Sparsity
adult	84.89	0.00	82.93	55.93
mushroom	99.98	0.00	99.98	49.08
german	72.96	4.12	72.52	51.43
ijcnn1	91.37	0.00	90.18	81.82
careval	95.05	0.00	86.92	30.00
gamma	79.32	0.00	70.37	31.00
abalone	79.52	0.00	74.28	15.00
shuttle	98.17	0.00	80.83	15.55
cod-rna	93.86	0.00	66.56	87.50
calhous	83.59	0.00	63.57	0.00

Name	*MILP* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RSVM* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*RR* $a_1 \in \{2^0, \ldots, 2^{10}\}$ $a_2 \in \{a_1/2, a_1/2^2\}$		*fixing* vs *SVM* $\bar{C}, \bar{a}_1$ $a_2 \in \{\bar{a}_1/2, \bar{a}_1/2^2\}$	
	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
adult	90.56	77.40	77.88	90.33	89.04	93.33	**B**	**B**
mushroom	100.00	91.60	99.41	95.80	100.00	100.00		**B**
german	75.96	55.87	70.92	57.78	73.00	54.76		**B**
ijcnn1	90.18	95.46	90.46	65.91	90.25	70.45		**B**
careval	98.71	35.72	89.81	43.34	98.76	53.81	**B**	**B**
gamma	79.31	50.00	78.28	52.00	79.19	57.00		**B**
abalone	79.05	55.00	78.28	57.00	77.64	65.00		**B**
shuttle	97.59	45.55	92.16	43.33	96.30	43.33		**B**
cod-rna	90.64	42.50	73.94	82.50	80.10	62.50	**W**	**B**
calhous	83.03	17.50	80.14	25.00	80.47	17.50		**B**

SVM and categorical features

When we have categorical features ...

Suppose $\mathrm{x} = (\mathrm{x}', \mathrm{x}'')$, with m_{cont} continuous and m_{cat} categorical features.

- each feature of x' : continuous
- each feature j of x'' :
  - takes K_j different values
  - which we codify as K_j dummy $(0-1)$ variables $x''_{j1}, \dots, x''_{jK_j}$
- ...and we may have many categorical variables: Carrizosa, Martín-Barragán, Romero Morales, *IJOC*, 2010; Liu, Hussain, Tan, Dash, *DMKD*, 2002, ...

SVM and categorical features

When we have categorical features ...

- Suppose $\mathrm{x} = (\mathrm{x}', \mathrm{x}'')$, with m_{cont} continuous and m_{cat} categorical features.
  - each feature of x' : continuous
  - each feature j of x'' :
    - takes K_j different values
    - which we codify as K_j dummy $(0-1)$ variables $x''_{j1}, \ldots, x''_{jK_j}$
  - score function:

$$\sum_{j=1}^{m_{cont}} \nu_j x'_j + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{jk} + \beta$$

- Optimization problem:

$$\begin{array}{lll} \min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k\in K_j} \omega_{jk}^2 + C \sum_{i\in\Omega_{tr}} \xi_i & \\ s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k\in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\ & \xi_i \geq 0 & i \in \Omega_{tr} \end{array}$$

SVM and categorical features

When we have categorical features ...

- Suppose $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features.
  - each feature of x' : continuous
  - each feature j of x'' :
    - takes K_j different values
    - which we codify as K_j dummy $(0-1)$ variables $x''_{j1}, \ldots, x''_{jK_j}$
  - score function:

$$\sum_{j=1}^{m_{cont}} \nu_j x'_j + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{jk} + \beta$$

- Optimization problem:

$$\begin{array}{lll} \min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\ s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\ & \xi_i \geq 0 & i \in \Omega_{tr} \end{array}$$

SVM and categorical features

When we have categorical features ...

- Suppose $\mathrm{x} = (\mathrm{x}', \mathrm{x}'')$, with m_{cont} continuous and m_{cat} categorical features.
  - each feature of x' : continuous
  - each feature j of x'' :
    - takes K_j different values
    - which we codify as K_j dummy $(0-1)$ variables $x''_{j1}, \ldots, x''_{jK_j}$
  - score function:

$$\sum_{j=1}^{m_{cont}} \nu_j x'_j + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{jk} + \beta$$

- Optimization problem:

$$\begin{array}{lll} \min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\ s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\ & \xi_i \geq 0 & i \in \Omega_{tr} \end{array}$$

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Again, convex quadratic problem with linear constraints
- Hard to interprete if many categorical values in features
- Convenient to reduce complexity (= reduce number of categorical values in features) without dammaging accuracy of classifier

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Again, convex quadratic problem with linear constraints
- Hard to interprete if many categorical values in features
- Convenient to reduce complexity (= reduce number of categorical values in features) without dammaging accuracy of classifier

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Again, convex quadratic problem with linear constraints
- Hard to interprete if many categorical values in features
- Convenient to reduce complexity (= reduce number of categorical values in features) without dammaging accuracy of classifier

Reducing the complexity

$$\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta$$

Strategies

- Feature selection
- Clustering values in certain categorical features

SVM and categorical features

We have ...

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes K_j different values

We wish ...

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes L_j different values, with $L_j \leq K_j$ fixed, $L_j = 2$, *say*
- For each categurical feature, needed:

  cluster the K_j values into L_j clusters

SVM and categorical features

We have . . .

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes K_j different values

We wish . . .

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes L_j different values, with $L_j \leq K_j$ fixed, $L_j = 2$, *say*
- For each categorical feature, needed:
  - cluster the K_j values into L_j clusters
  - Assign one weight to each , i.e., assign a common weight to all values in the same cluster

SVM and categorical features

We have ...

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes K_j different values

We wish ...

- $x = (x', x'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes L_j different values, with $L_j \leq K_j$ fixed, $L_j = 2$, *say*
- For each categorical feature, needed:
  - cluster the K_j values into L_j clusters
  - Assign one weight to each cluster, i.e., assign a common weight to all values in the same cluster

SVM and categorical features

We have . . .

- $\mathrm{x} = (\mathrm{x}', \mathrm{x}'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes K_j different values

We wish . . .

- $\mathrm{x} = (\mathrm{x}', \mathrm{x}'')$, with m_{cont} continuous and m_{cat} categorical features, each j of x'' : takes L_j different values, with $L_j \leq K_j$ fixed, $L_j = 2$, *say*
- For each categorical feature, needed:
    - cluster the K_j values into L_j clusters
    - Assign one weight to each cluster, i.e., assign a common weight to all values in the same cluster

We had ...

$$\begin{array}{lll} \min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\ s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\ & \xi_i \geq 0 & i \in \Omega_{tr} \end{array}$$

Now we have ...

- Define z_{jkl} :

$$z_{jkl} = \begin{cases} 1, & \text{if, in feature } j\text{, value } k \text{ is assigned to cluster } l \\ 0, & \text{else} \end{cases}$$

- Seek $\bar{\omega}_{jl}$: weight of all values of feature j allocated to cluster l

We had . . .

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{k \in K_j} \omega_{jk} x''_{ijk} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

Now we have... . . .

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl} \sum_{k=1}^{K_j} z_{jkl} x''_{ijk} + \beta \right) & \\
 & \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\
 & z_{jkl} \in \{0, 1\} & j, k, l \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl} \sum_{k=1}^{K_j} z_{jkl} x''_{ijk} + \beta \right) & \\
 & \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\
 & z_{jkl} \in \{0, 1\} & j, k, l \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Mixed Intenger Nonlinear Nonconvex Optimization Problem
- Very hard to solve, even with state-of-the-art software, e.g. Couenne
- Structural properties are obtained

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl} \sum_{k=1}^{K_j} z_{jkl} x''_{ijk} + \beta \right) & \\
 & \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\
 & z_{jkl} \in \{0, 1\} & j, k, l \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Mixed Intenger Nonlinear Nonconvex Optimization Problem
- Very hard to solve, even with state-of-the-art software, e.g. Couenne
- Structural properties are obtained

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl} \sum_{k=1}^{K_j} z_{jkl} x''_{ijk} + \beta \right) & \\
 & \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\
 & z_{jkl} \in \{0, 1\} & j, k, l \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Mixed Intenger Nonlinear Nonconvex Optimization Problem
- Very hard to solve, even with state-of-the-art software, e.g. Couenne
- Structural properties are obtained

Property

If, for some categorical feature j^*, an optimal solution exists with $z_{j^*kl^*=1}$ for all $k = 1, \ldots, K_j$ for some $l^* \in \{1, \ldots, L_{j^*}\}$, i.e., if the trivial cluster is obtained, then $\bar{\omega}_{j^*l} = 0$ for all $l \in \{1, \ldots, L_{j^*}\}$, i.e., feature j is irrelevant.

Property

Suppose $L_j = 2$. Then, any optimal solution satisfies

$$\bar{\omega}_{j1}\bar{\omega}_{j2} \leq 0$$

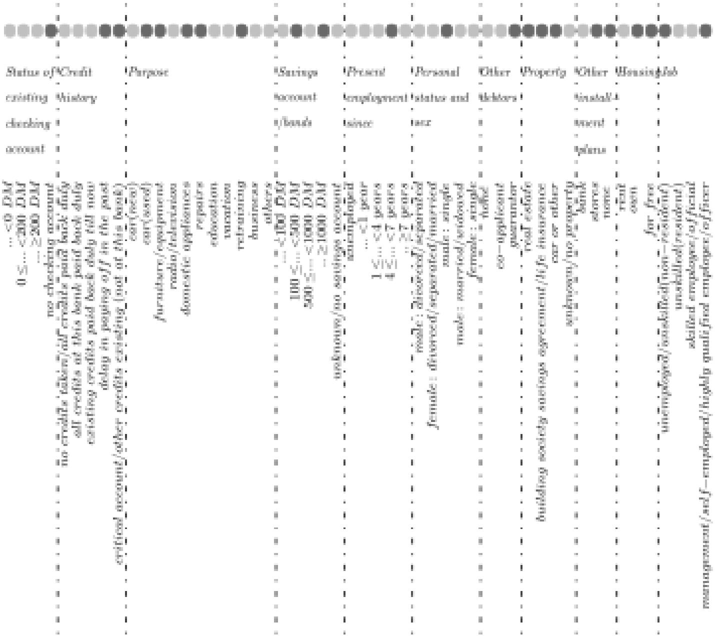
Property

If, for some categorical feature j^*, an optimal solution exists with $z_{j^*kl^*=1}$ for all $k = 1, \ldots, K_j$ for some $l^* \in \{1, \ldots, L_{j^*}\}$, i.e., if the trivial cluster is obtained, then $\bar{\omega}_{j^*l} = 0$ for all $l \in \{1, \ldots, L_{j^*}\}$, i.e., feature j is irrelevant.

Property

Suppose $L_j = 2$. Then, any optimal solution satisfies

$$\bar{\omega}_{j1}\bar{\omega}_{j2} \leq 0$$


Status of existing checking account
Credit history
Purpose
Savings account/bonds
Present employment since
Personal status and sex
Other debtors
Property
Other install-ment plans
Housing
Job
... < 0 DM
0 ≤ ... < 200 DM
... ≥ 200 DM
no checking account
no credits taken/all credits paid back duly
all credits at this bank paid back duly
existing credits paid back duly till now
delay in paying off in the past
critical account/other credits existing (not at this bank)
car(new)
car(used)
furniture/equipment
radio/television
domestic appliances
repairs
education
vacation
retraining
business
others
... < 100 DM
100 ≤ ... < 500 DM
500 ≤ ... < 1000 DM
... ≥ 1000 DM
unknown/no savings account
unemployed
... < 1 year
1 ≤ ... < 4 years
4 ≤ ... < 7 years
... ≥ 7 years
male : divorced/separated
female : divorced/separated/married
male : single
male : married/widowed
female : single
none
co-applicant
guarantor
real estate
building society savings agreement/life insurance
car or other
unknown/no property
bank
stores
none
rent
own
for free
unemployed/unskilled(non-resident)
unskilled(resident)
skilled employee/official
management/self-employed/highly qualified employee/officer

Alternative (big-M) formulation

- Define variables $\widetilde{\omega}_{jkl}$:

$$\widetilde{\omega}_{jkl} = \begin{cases} \bar{\omega}_{jl}, & \text{if value } k \text{ of feature } j \text{ goes to cluster } l \\ 0, & \text{else} \end{cases}$$

$$\begin{array}{lll} \min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\ s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl} \sum_{k=1}^{K_j} z_{jkl} x''_{ijk} + \beta \right) & \\ & \geq 1 - \xi_i & i \in \Omega_{tr} \\ & \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\ & z_{jkl} \in \{0, 1\} & j, k, l \\ & \xi_i \geq 0 & i \in \Omega_{tr} \end{array}$$

Alternative (big-M) formulation

- Define variables $\widetilde{\omega}_{jkl}$:

$$\widetilde{\omega}_{jkl} = \begin{cases} \bar{\omega}_{jl}, & \text{if value } k \text{ of feature } j \text{ goes to cluster } l \\ 0, & \text{else} \end{cases}$$

$$\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \widetilde{\omega}_{jk(i)l} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
& \widetilde{\omega}_{jkl} \leq \bar{\omega}_{jl} + M(1 - z_{jkl}) & j, k, l \\
& \widetilde{\omega}_{jkl} \geq \bar{\omega}_{jl} - M(1 - z_{jkl}) & j, k, l \\
& \widetilde{\omega}_{jkl} \leq M z_{jkl} & j, k, l \\
& \widetilde{\omega}_{jkl} \geq -M z_{jkl} & j, k, l \\
& \sum_{l=1}^{L_j} z_{jkl} = 1 & j, k \\
& z_{jkl} \in \{0, 1\} & j, k, l \\
& \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}$$

Alternative (big-M) formulation

$$
\begin{array}{lll}
\min & \sum_{j=1}^{m_{cont}} \nu_j^2 + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \bar{\omega}_{jl}^2 + C \sum_{i \in \Omega_{tr}} \xi_i & \\
s.t. & y_i \left(\sum_{j=1}^{m_{cont}} \nu_j x'_{ij} + \sum_{j=1}^{m_{cat}} \sum_{l=1}^{L_j} \widetilde{\omega}_{jk(i)l} + \beta \right) \geq 1 - \xi_i & i \in \Omega_{tr} \\
 & \widetilde{\omega}_{jkl} \leq \bar{\omega}_{jl} + M(1 - z_{jkl}) & j,k,l \\
 & \widetilde{\omega}_{jkl} \geq \bar{\omega}_{jl} - M(1 - z_{jkl}) & j,k,l \\
 & \widetilde{\omega}_{jkl} \leq M z_{jkl} & j,k,l \\
 & \widetilde{\omega}_{jkl} \geq -M z_{jkl} & j,k,l \\
 & \sum_{l=1}^{L_j} z_{jkl} = 1 & j,k \\
 & z_{jkl} \in \{0,1\} & j,k,l \\
 & \xi_i \geq 0 & i \in \Omega_{tr}
\end{array}
$$

- Many more variables ($\widetilde{\omega}_{jkl}$) and constraints
- Convex quadratic objective + linear constraints
- Solvable (in principle) by off-the-shelf software such as Cplex

(Optimization-based) strategies

Strategies

- Exact (solving the nonlinear problem with nonconvex constraints)
- Heuristics
  - $\mathrm{SVM}^{\mathrm{C}}$
  - $\mathrm{SVM}^{\mathrm{RR}}$
  - $\mathrm{SVM}^{\mathrm{MRR}}$

SVM^{C}

Centroid-based

1. Solve original SVM^{O}
2. From the weights, perform 1-dim clustering (1d L_j-means)
3. Return SVM with clustered values

SVM^{RR}

Randomized rounding

1. Solve the continuous relaxation of the nonlinear nonconvex formulation
2. Round (at random) the allocation variables z_{jkl}
3. Return SVM with clustered values

Big M Randomized rounding

1. Solve the continuous relaxation of the nonlinear convex (big M) formulation
2. Round (at random) the allocation variables z_{jkl}
3. Return SVM with clustered values

Experiments

10 data sets from the UCI ML Repository

Name	Name in Repository	$\lvert\Omega\rvert$	n	Class split (in %)	J	J'	$\sum_{j=1}^{J} K_j$	K_j
census income	Census-Income (KDD) Data Set	95130	5000	94/6	31	9	491	9,52,47,17,3,7,24,15,5,10,3,6,8,6,6, 50,38,8,9,8,9,3,3,5,42,42,42,5,3,3,3
adult	Adult	30956	5000	24/76	11	3	117	5,8,5,16,5,7,14,6,5,5,41
mushrooms	Mushrooms	8124	5000	48/52	17	4	111	6,4,10,9,4,3,12,4,4,9,9,4,3,8,9,6,7
coil 2000	Insurance Company Benchmark (COIL 2000)	5822	3900	94/6	5	80	77	41,6,10,10,10
abalone	Abalone	4177	2800	50/50	1	7	3	3
molecular	Molecular Biology (Splice junction Gene Sequences)	3190	2200	52/48	60	0	480	8,8,8,...
careval	Car Evaluation	1728	1200	30/70	6	0	21	4,4,4,3,3,3
solar-c	Solar Flare Data Set	1066	800	83/17	5	5	23	7,6,4,3,3
german	Statlog (German Credit Data)	1000	700	30/70	11	9	52	4,5,11,5,5,5,3,4,3,3,4
australian	Statlog (Australian Credit Approval)	690	500	56/44	4	10	29	3,14,9,3

Results

Accuracy

- In 7 datasets (`census income`, `mushrooms`, `coil 2000`, `abalone`, `molecular`, `solar-c`, `german`), at least 1 strategy: comparable to SVM^{O}
- In 2 (`adult`, `australian`) SVM^{O} : outperformed
- Just in 1 (`careval`) SVM^{O} : winner.

Results

Complexity

Dramatic reduction in nonzero weights:

- smallest: `coil 2000` (8.12 p.p.)
- remaining: at least 30 p.p. reduction (even 85.25 p.p.)

Running times

- Heursitics: just slightly higher cost than SVM^{O}
- Exact: rather expensive, but yielding better results.
- Trade-off accuracy/running time: needed

Running times

- Heursitics: just slightly higher cost than SVM^O
- Exact: rather expensive, but yielding better results.
- Trade-off accuracy/running time: needed

Running times

- Heursitics: just slightly higher cost than SVM^O
- Exact: rather expensive, but yielding better results.
- Trade-off accuracy/running time: needed

Name	SVM^O					
	Accuracy			Reduction		
	mean	std	med	mean	std	med
census income	94.90	0.00	94.90	36.86	0.00	36.86
adult	84.57	0.22	84.63	16.07	3.26	15.38
mushrooms	100.00	0.00	100.00	28.83	0.00	28.83
coil 2000	100.00	0.00	100.00	98.70	0.00	98.70
abalone	79.87	1.18	79.72	0.00	0.00	0.00
molecular	94.22	0.80	94.04	42.96	3.12	41.46
careval	96.74	1.34	96.97	0.95	2.86	0.00
solar-c	83.53	1.23	83.46	47.83	34.51	30.43
german	74.60	2.71	75.66	5.38	1.88	4.81
australian	84.11	3.17	84.73	13.10	4.83	10.34

SVM^C

	Accuracy			Reduction		
	mean	std	med	mean	std	med
census income	94.85	0.00	94.85	89.82	0.00	89.82
adult	88.22	2.44	89.59	86.24	2.21	86.32
mushrooms	100.00	0.00	100.00	76.58	0.00	76.58
coil 2000	100.00	0.00	100.00	98.70	0.00	98.70
abalone	79.90	1.05	79.44	33.33	0.00	33.33
molecular	93.94	0.73	93.84	100.00	0.00	100.00
careval	82.80	5.15	82.95	55.24	8.57	61.90
solar-c	83.61	1.38	83.46	87.83	10.43	91.30
german	74.80	2.36	75.00	57.69	0.00	57.69
australian	84.42	3.32	84.73	80.69	9.66	75.86

SVM^{RR}

	Accuracy			Reduction		
	mean	std	med	mean	std	med
census income	94.84	0.04	94.82	91.69	0.33	91.45
adult	83.44	0.37	83.44	83.42	1.34	83.76
mushrooms	**100.00**	0.00	**100.00**	80.72	4.56	81.98
coil 2000	**100.00**	0.00	**100.00**	**98.70**	0.00	**98.70**
abalone	79.86	1.02	79.44	**33.33**	0.00	**33.33**
molecular	93.40	1.28	93.53	75.13	0.19	75.00
careval	92.23	1.28	92.42	58.10	12.2	61.90
solar-c	**83.83**	1.08	**83.46**	94.78	11.14	**100.00**
german	74.60	3.12	74.00	61.73	4.59	58.65
australian	84.53	3.12	84.73	84.83	11.03	86.21

SVM^{MRR}

	Accuracy			Reduction		
	mean	std	med	mean	std	med
census income	94.40	0.04	94.37	91.45	1.62	91.65
adult	88.75	2.96	89.63	89.83	4.36	91.45
mushrooms	98.58	0.77	98.59	78.29	5.87	77.93
coil 2000	100.00	0.00	100.00	98.70	0.00	98.70
abalone	79.87	0.96	79.66	33.33	0.00	33.33
molecular	88.04	1.68	88.38	77.29	0.90	77.50
careval	83.94	4.91	85.41	70.48	10.82	71.43
solar-c	83.76	1.02	83.46	99.13	2.61	100.00
german	72.53	3.77	71.66	63.08	4.28	63.46
australian	84.53	3.05	84.73	94.48	2.76	93.10

Many thanks!!!

ecarrizosa@us.es