

# Using Mathematical Programming for Supervised Classification: new SVM based models

Luisa I. Martínez-Merino

Departamento de Estadística e Investigación Operativa, Universidad de Sevilla  
Subprograma Juan de la Cierva Formación (2019)

Advances on logistics and transportation problems  
on complex networks: Evaluation and conclusions  
June 23-25, 2021, Fuengirola (Spain).

# Outline

- 1 Introduction to Support Vector Machines (SVM)
- 2 Including feature selection in SVM: FS-SVM
- 3 Ramp loss SVM based model: RL- $\ell_p$ -M
- 4 Combining feature selection and ramp loss: RL-FS-M
- 5 Ordered Weighted Average in SVM: OWA-SVM
- 6 Some conclusions

# Introduction to Support Vector Machine (SVM)

- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.

$$(\mathbf{y}, \mathbf{x}) = \begin{pmatrix} \textcolor{red}{y_1} & x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \textcolor{red}{y_i} & x_{i1} & \dots & x_{ij} & \dots & x_{id} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \textcolor{red}{y_n} & x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$

# Introduction to Support Vector Machine (SVM)

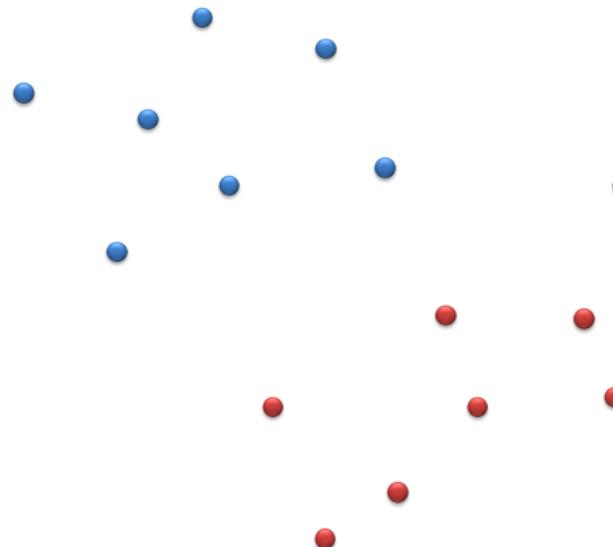
- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.

$$(\mathbf{y}, \mathbf{x}) = \begin{pmatrix} \color{red}{y_1} & x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \color{red}{y_i} & x_{i1} & \dots & x_{ij} & \dots & x_{id} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \color{red}{y_n} & x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$

$$\left( \color{red}{?} \quad \tilde{x}_1 \quad \dots \quad \tilde{x}_j \quad \dots \quad \tilde{x}_d \right)$$

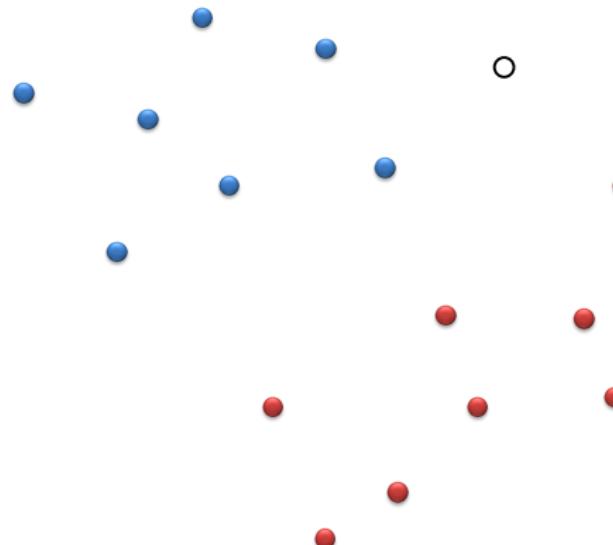
# Introduction to Support Vector Machine (SVM)

- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.



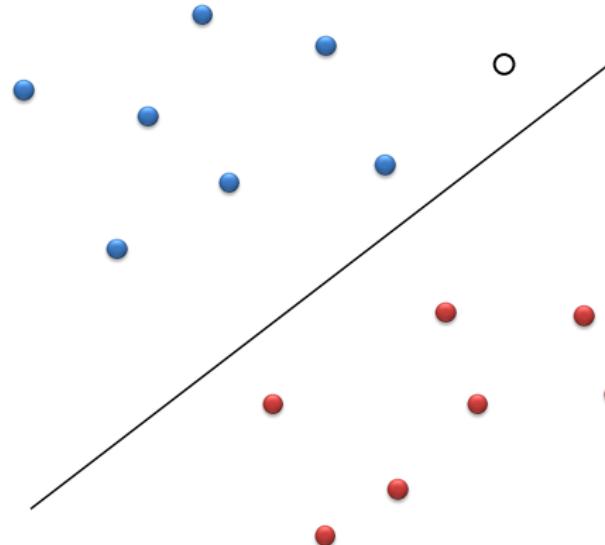
# Introduction to Support Vector Machine (SVM)

- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.



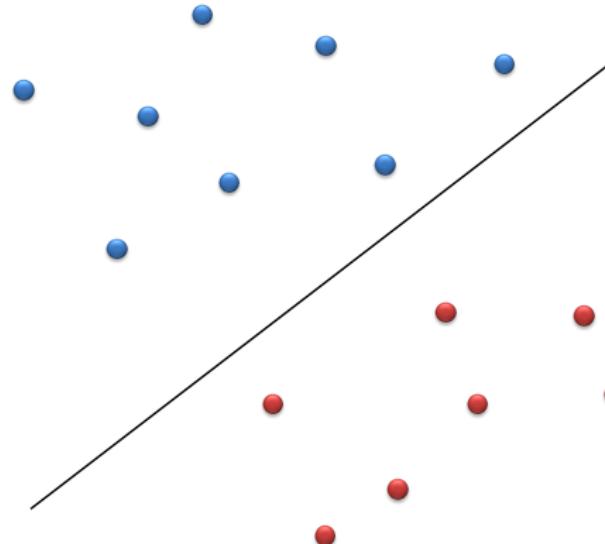
# Introduction to Support Vector Machine (SVM)

- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.

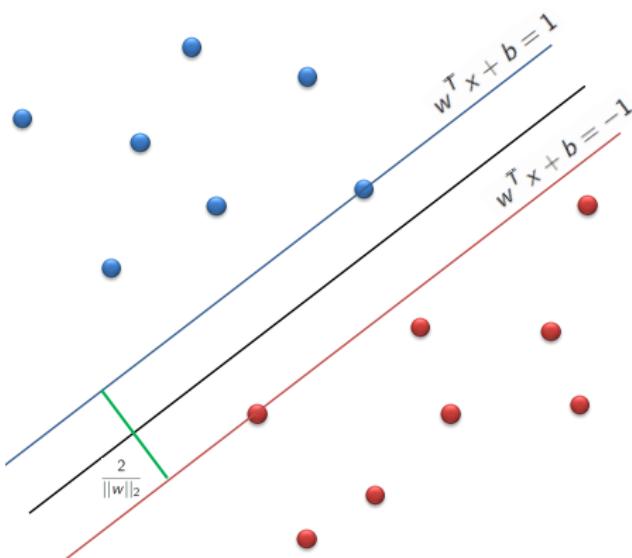


# Introduction to Support Vector Machine (SVM)

- $\mathbf{x}_i \in \mathbb{R}^d$  training sample  $i = 1, \dots, n$ .
- $y_i \in \{-1, +1\}$  their labels.
- $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  to optimally separate the training sample.

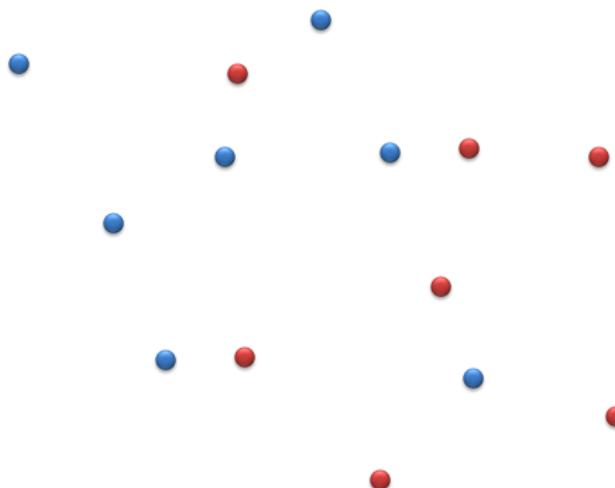


# Introduction to Support Vector Machine (SVM)

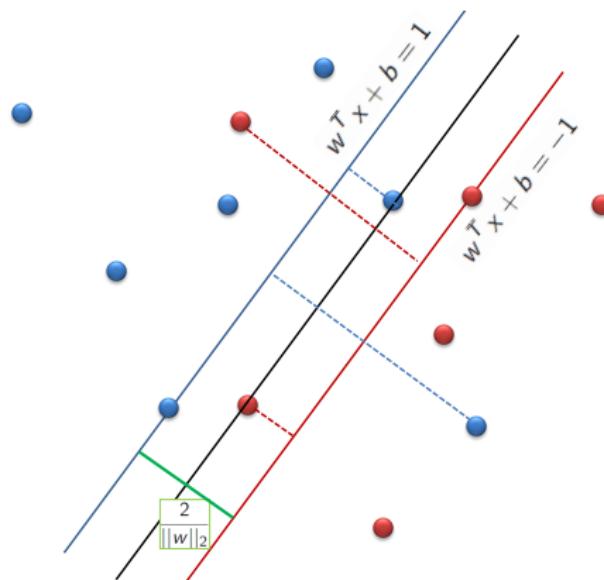


$$\left. \begin{array}{l} y_i = 1 \implies \mathbf{w}^T \mathbf{x}_i + b \geq 1 \\ y_i = -1 \implies \mathbf{w}^T \mathbf{x}_i + b \leq -1 \end{array} \right\} \implies y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$$

# Introduction to Support Vector Machine (SVM)



# Introduction to Support Vector Machine (SVM)



$$\left. \begin{array}{l} y_i = 1 \implies \mathbf{w}^T \mathbf{x}_i + b \geq 1 - \xi_i \\ y_i = -1 \implies \mathbf{w}^T \mathbf{x}_i + b \leq -1 + \xi_i \end{array} \right\} \implies y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$$

# Classic SVM model (Bradley and Mangasarian, 1998)

$$\begin{aligned} \min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

# Classic SVM model (Bradley and Mangasarian, 1998)

$$\begin{aligned}
 \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i \\
 \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

In its dual form:

$$\begin{aligned}
 \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\
 \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\
 & 0 \leq \alpha_i \leq C, \quad i \in N.
 \end{aligned}$$

# Classic SVM model (Bradley and Mangasarian, 1998)

$$\begin{aligned}
 & \min_{\boldsymbol{w}, b, \xi} \quad \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^n \xi_i \\
 & s.t. \quad y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & \quad \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

In its dual form:

$$\begin{aligned}
 & \max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
 & s.t. \quad \sum_{i=1}^m \alpha_i y_i = 0, \\
 & \quad 0 \leq \alpha_i \leq C, \quad i \in N.
 \end{aligned}$$

# New approaches to SVM

$$\begin{aligned} \min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{p} \|\boldsymbol{w}\|_p^p + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

# New approaches to SVM

$$\begin{aligned} \min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{p} \|\boldsymbol{w}\|_p^p + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

- Considered norms:  $\ell_1$ -norm and  $\ell_2$ -norm.

# New approaches to SVM

$$\begin{aligned} \min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{p} \|\boldsymbol{w}\|_p^p + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

- Considered norms:  $\ell_1$ -norm and  $\ell_2$ -norm.
- Feature selection.

# New approaches to SVM

$$\begin{aligned} \min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{p} \|\boldsymbol{w}\|_p^p + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

- Considered norms:  $\ell_1$ -norm and  $\ell_2$ -norm.
- Feature selection.
- Penalization of deviations associated with misclassified individuals:
  - Ramp Loss.
  - Ordered Weighted Average.

# New approaches to SVM

-  M. Labb  , L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, *Discrete Applied Mathematics* 261 (2019) 276–304.
-  M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Tightening big ms in integer programming formulations for support vector machines with ramp loss, *European Journal of Operational Research* 286 (1) (2020) 84–100.
-  M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, A robust SVM-based approach with feature selection and outliers detection for classification problems, *Expert Systems with Applications* 178 (2021) 15017.
-  A. Mar  n, L. I. Mart  nez-Merino, J. Puerto, A. M. Rodr  guez-Ch  a, The soft-margin Support Vector Machine with ordered weighted average, (submitted).

# Feature selection

-  M. Labb , L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, *Discrete Applied Mathematics* 261 (2019) 276–304.
-  M. Baldomero-Naranjo, L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, Tightening big Ms in integer programming formulations for support vector machines with ramp loss, *European Journal of Operational Research* 286 (1) (2020) 84–100.
-  M. Baldomero-Naranjo, L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, A robust SVM-based approach with feature selection and outliers detection for classification problems, *Expert Systems with Applications* 178 (2021) 15017.
-  A. Mar n, L. I. Mart nez-Merino, J. Puerto, A. M. Rodr guez-Ch a, The soft-margin Support Vector Machine with ordered weighted average, (submitted).

# Feature selection



M. Labb  , L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, Discrete Applied Mathematics 261 (2019) 276  304.

$$\begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{id} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$

$$d >>> n$$

# Including feature selection

$$\begin{aligned} \text{(FS-SVM)} \quad & \min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{v}} \quad \|\boldsymbol{w}\|_1 + C \sum_{i=1}^n \xi_i, \\ \text{s.t.} \quad & y_i \left( \sum_{j=1}^d w_j x_{ij} + b \right) \geq 1 - \xi_i, \quad i \in N, \\ & w \leq \hat{u}_j v_j, \quad j \in D, \\ & \sum_{j=1}^n v_j \leq B, \\ & v_j \in \{0, 1\}, \quad j \in D, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

# Including feature selection

$$\begin{aligned}
 (\text{FS-SVM}) \quad & \min_{\mathbf{w}^+, \mathbf{w}^-, b, \boldsymbol{\xi}, \mathbf{v}} \quad \sum_{j=1}^d (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i, \\
 \text{s.t.} \quad & y_i \left( \sum_{j=1}^d (w_j^+ - w_j^-) x_{ij} + b \right) \geq 1 - \xi_i, \quad i \in N, \\
 & w_j^+ \leq u_j v_j, \quad j \in D, \\
 & w_j^- \leq l_j v_j, \quad j \in D, \\
 & \sum_{j=1}^n v_j \leq B, \\
 & v_j \in \{0, 1\}, \quad j \in D, \\
 & w_j^+, w_j^- \geq 0, \quad j \in D, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

# Including feature selection

$$\begin{aligned}
 (\text{FS-SVM}) \quad & \min_{\mathbf{w}^+, \mathbf{w}^-, b, \boldsymbol{\xi}, \mathbf{v}} \quad \sum_{j=1}^d (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i, \\
 \text{s.t.} \quad & y_i \left( \sum_{j=1}^d (w_j^+ - w_j^-) x_{ij} + b \right) \geq 1 - \xi_i, \quad i \in N, \\
 & w_j^+ \leq u_j \mathbf{v}_j, \quad j \in D, \\
 & w_j^- \leq l_j \mathbf{v}_j, \quad j \in D, \\
 & \sum_{j=1}^n \mathbf{v}_j \leq B, \\
 & \mathbf{v}_j \in \{0, 1\}, \quad j \in D, \\
 & w_j^+, w_j^- \geq 0, \quad j \in D, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

# Strategies to tighten $l$ and $u$

- **Algorithm I:** Based on solving a linear model derived from the original one.

$$\begin{aligned}
 (\text{LP}) \quad & \max_{\mathbf{w}^+, \mathbf{w}^-, b, \xi, v} \quad w_k^+ + w_k^- \\
 \text{s.t.} \quad & \text{FS-SVM constraints,} \\
 & \sum_{j=1}^d (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i \leq \text{UB}, \\
 & 0 \leq v_j \leq 1, \quad \forall j \in D.
 \end{aligned}$$

# Strategies to tighten $l$ and $u$

- **Algorithm I:** Based on solving a linear model derived from the original one.

$$\begin{aligned} (\text{LP}) \quad & \max_{\mathbf{w}^+, \mathbf{w}^-, b, \xi, v} \quad w_k^+ + w_k^- \\ \text{s.t.} \quad & \text{FS-SVM constraints,} \end{aligned}$$

$$\begin{aligned} & \sum_{j=1}^d (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i \leq \text{UB}, \\ & 0 \leq v_j \leq 1, \quad \forall j \in D. \end{aligned}$$

- **Algorithm II** Based on a Lagrangian relaxation of the model.

$$\bar{u}_{j_0}^+ := \min \left\{ \frac{\text{UB} - z_{LB}^{LP}}{1 - \sum_{i=1}^m \alpha_i y_i x_{ij_0}}, u_{j_0} (B - \sum_{j=1}^n \bar{v}_j) \right\}$$

$$\bar{u}_{j_0}^- := \min \left\{ \frac{\text{UB} - z_{LB}^{LP}}{1 + \sum_{i=1}^m \alpha_i y_i x_{ij_0}}, -l_{j_0} (B - \sum_{j=1}^n \bar{v}_j) \right\}$$

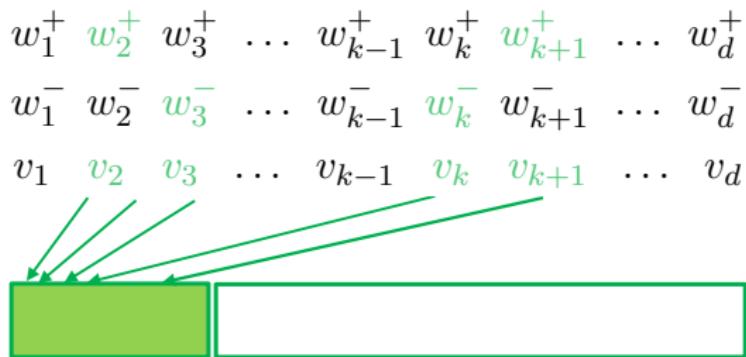
# A heuristic approach for FS-SVM: the Kernel Search

- Kernel Search (KS): Solve a sequence of restricted MILP derived from the original problem to obtain a bound.
- Applications:
  - Location problems. (Guastaroba and Speranza, 2012)
  - Portfolio selection. (Angelelli et al., 2012)
- Kernel ( $K$ ): Set of “promising” variables  $v$ ,  $w^+$  and  $w^-$ .

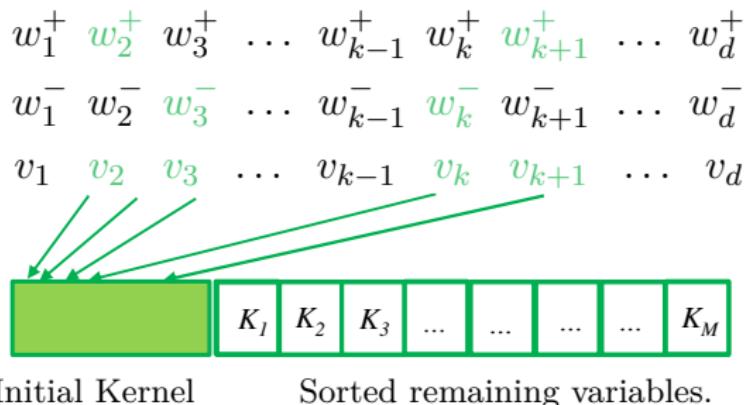
# A heuristic approach for FS-SVM: the Kernel Search

$$\begin{array}{cccccccccc} w_1^+ & w_2^+ & w_3^+ & \dots & w_{k-1}^+ & w_k^+ & w_{k+1}^+ & \dots & w_n^+ \\ w_1^- & w_2^- & w_3^- & \dots & w_{k-1}^- & w_k^- & w_{k+1}^- & \dots & w_d^- \\ v_1 & v_2 & v_3 & \dots & v_{k-1} & v_k & v_{k+1} & \dots & v_d \end{array}$$

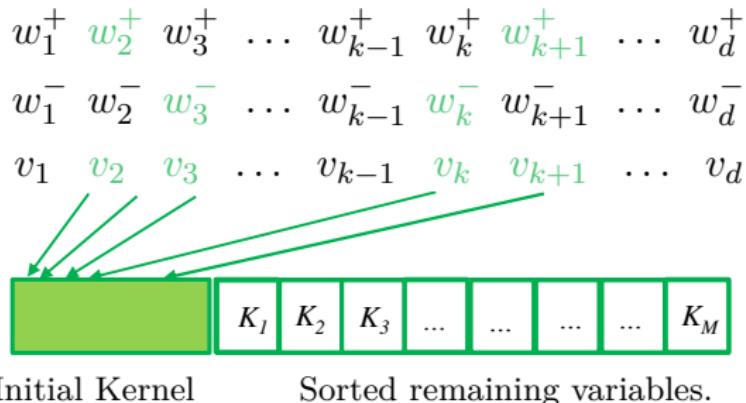
# A heuristic approach for FS-SVM: the Kernel Search



# A heuristic approach for FS-SVM: the Kernel Search



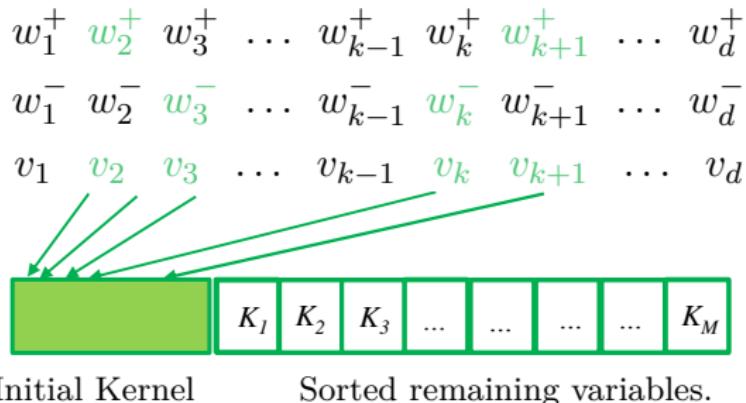
# A heuristic approach for FS-SVM: the Kernel Search



$$\mathcal{K} = K \cup K_1$$

- $\sum_{j \in \mathcal{K}} (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i \leq \text{UB},$
- $\sum_{j \in K_1} v_j \geq 1.$

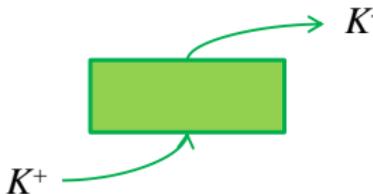
# A heuristic approach for FS-SVM: the Kernel Search



$$\mathcal{K} = K \cup K_1$$

- $\sum_{k \in \mathcal{K}} (w_j^+ + w_j^-) + C \sum_{i=1}^n \xi_i \leq \text{UB},$
- $\sum_{j \in K_1} v_j \geq 1.$

- Update UB.
- Update  $K$ .



# Computational results on FS-SVM

- Intel(R) Core(TM) i7-4790K CPU 32 GB RAM. Cplex 12.6.3.

Small number of features			
Name	m	n	Class(%)
BUPA	345	6	42/58
PIMA	768	8	65/35
Cleveland	297	13	42/58
Housing	506	13	51/49
Australian	690	14	44/56
GC	1000	24	30/70
WBC	569	30	37/63
Ionosphere	351	33	64/36

Big number of features							
Small sample size			Big sample size				
Name	m	n	Class(%)	Name	m	n	Class(%)
Colon	62	2000	35/65	Lepiota	8124	109	52/48
Leukemia	72	5327	47/53	Arrythmia	420	258	57/43
DLBCL	77	7129	75/25	Madelon	2000	500	50/50
Carcinoma	36	7457	53/47	Mfeat	2000	649	10/90

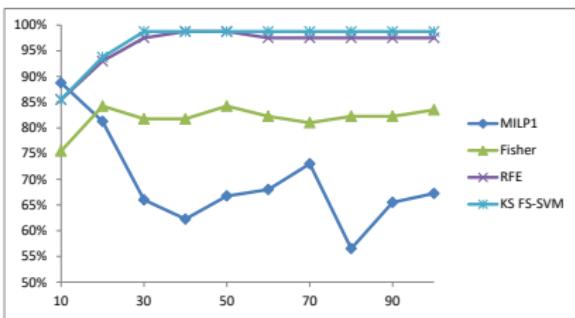
## Part 1: Validation of the model

- Ten-fold-cross-validation,
- Accuracy:  $\frac{TP+TN}{TP+TN+FP+FN}$  and AUC:  $\frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$ .

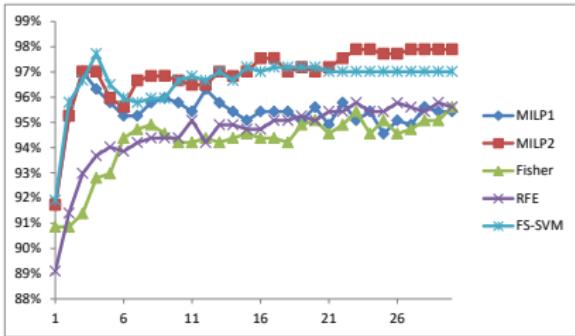
## Part 2: Strategies I and II and heuristic approach

# Validation of the FS-SVM model

	WBC n=569 d=30				
Form.	Av. ACC	Av. AUC	Av. F.	B	C
$\ell_1$ -SVM	97.37%	96.92%	10	-	1
$\ell_2$ -SVM	<b>98.07%</b>	<b>97.58%</b>	30	-	$2^2$
LP-SVM	97.89%	97.44%	30	-	$2^{-5}$
MILP1	97.02%	96.45%	3	3	-
MILP2	97.89%	97.54%	23	23	$2^{-5}$
FSV	42.35%	54.03%	20	-	-
Fisher-SVM	95.60%	96.19%	30	30	$2^6$
RFE-SVM	95.78%	96.33%	23	23	$2^6$
<b>FS-SVM</b>	97.72%	97.20%	4	4	$2^4$
<b>KS FS-SVM</b>	97.72%	97.20%	4	4	$2^4$



	DLBCL n=77 d=7129				
Form.	Av. ACC	Av. AUC	Av. F.	B	C
$\ell_1$ -SVM	<b>98.75%</b>	97.50%	32	-	1
$\ell_2$ -SVM	96.25%	94.17%	7129	-	1
LP-SVM	97.50%	95.00%	7129	-	$2^{-4}$
MILP1	88.75%	85.83%	10	10	-
FSV	49.75%	66.67%	23	-	-
Fisher-SVM	83.50%	88.17%	100	100	$2^{-5}$
RFE-SVM	<b>98.75%</b>	<b>99.17%</b>	40	40	$2^7$
<b>KS FS-SVM</b>	<b>98.75%</b>	97.50%	30	30	$2^6$



# Strategies and heuristic approach

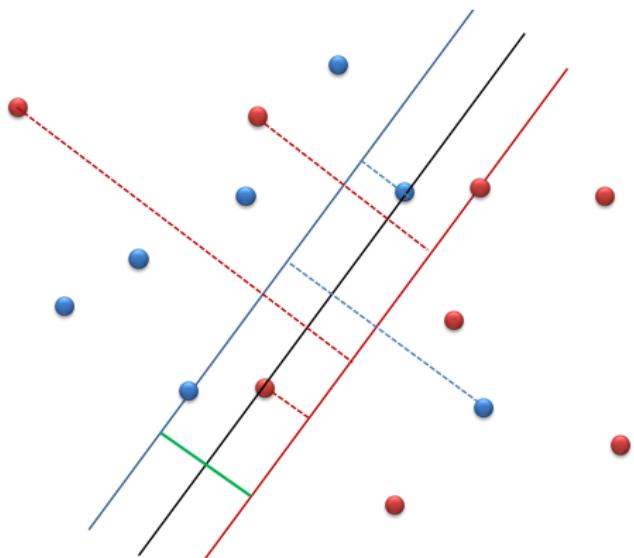
Colon n=62 d=2000						
B/C	FS-SVM		St. + FS-SVM			$t_{total}$
	Gap	Time	Gap	$t_{st}$	$t_{solv}$	
20/2 <sup>0</sup>	0.0	579.94	<b>0.0</b>	31.91	93.35	<b>125.26</b>
20/2 <sup>1</sup>	0.3	<u>7200.19</u>	<b>0.0</b>	35.65	1911.11	<b>1946.76</b>
20/2 <sup>2</sup>	0.3	<u>7200.10</u>	<b>0.0</b>	35.30	1634.42	<b>1669.72</b>
20/2 <sup>3</sup>	0.3	<u>7200.10</u>	<b>0.0</b>	34.34	1652.39	<b>1686.73</b>
20/2 <sup>4</sup>	0.3	<u>7200.10</u>	<b>0.0</b>	34.80	1549.91	<b>1584.71</b>
20/2 <sup>5</sup>	0.3	<u>7200.1</u>	<b>0.0</b>	34.32	1380.88	<b>1415.20</b>
20/2 <sup>6</sup>	0.4	<u>7200.10</u>	<b>0.0</b>	34.23	1098.26	<b>1132.48</b>
20/2 <sup>7</sup>	0.5	<u>7200.11</u>	<b>0.0</b>	34.76	1009.93	<b>1044.69</b>
10/2 <sup>0</sup>	2.5	<u>7203.00</u>	<b>0.0</b>	40.03	4598.49	<b>4638.52</b>
10/2 <sup>1</sup>	12.4	<u>7204.81</u>	<b>11.2</b>	46.16	<u>7203.86</u>	<u>7250.02</u>
10/2 <sup>2</sup>	<b>10.5</b>	<u>7200.94</u>	13.6	51.20	<u>7206.82</u>	<u>7258.02</u>
10/2 <sup>3</sup>	15.2	<u>7201.77</u>	<b>12.3</b>	46.35	<u>7206.01</u>	<u>7252.36</u>
10/2 <sup>4</sup>	16.9	<u>7203.00</u>	<b>13.5</b>	47.03	<u>7206.16</u>	<u>7253.19</u>
10/2 <sup>5</sup>	17.1	<u>7206.07</u>	<b>13.0</b>	49.17	<u>7206.29</u>	<u>7255.46</u>
10/2 <sup>6</sup>	16.6	<u>7204.87</u>	<b>12.3</b>	46.80	<u>7205.97</u>	<u>7252.77</u>
10/2 <sup>7</sup>	17.0	<u>7201.27</u>	<b>10.8</b>	46.39	<u>7206.23</u>	<u>7252.62</u>

Heuristic results				
B/C	Colon	Gap	$t_{KS}$	$t_{Best}$
20/1		0.0	5.77	125.26
20/2		0.0	20.87	1388.57
20/2 <sup>2</sup>		0.0	12.90	1564.57
20/2 <sup>3</sup>		0.0	10.85	1528.63
20/2 <sup>4</sup>		0.0	12.96	1580.48
20/2 <sup>5</sup>		0.0	11.82	1415.20
20/2 <sup>6</sup>		0.0	11.15	1087.20
20/2 <sup>7</sup>		0.0	20.51	712.30
10/1		0.0	18.12	4447.71
10/2		6.3/0.0	317.79	7396.14 <sup>(8.5)</sup>
10/2 <sup>2</sup>		7.5/0.0	453.18	7442.52 <sup>(9.3)</sup>
10/2 <sup>3</sup>		7.5/0.0	581.85	7470.23 <sup>(9.1)</sup>
10/2 <sup>4</sup>		7.8/0.0	489.36	7481.98 <sup>(9.2)</sup>
10/2 <sup>5</sup>		8.0/0.0	508.48	7434.46 <sup>(9.1)</sup>
10/2 <sup>6</sup>		7.3/0.0	532.52	7485.48 <sup>(9.1)</sup>
10/2 <sup>7</sup>		7.4/0.0	499.11	7437.58 <sup>(9.1)</sup>

# Introduction to ramp loss SVM

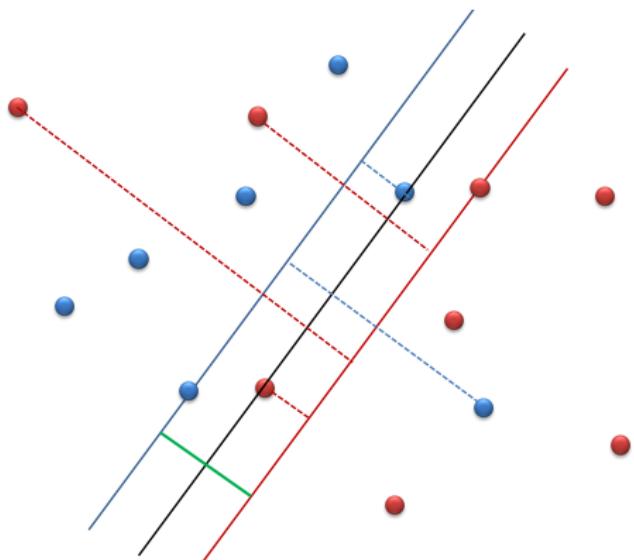
-  M. Labb , L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, *Discrete Applied Mathematics* 261 (2019) 276–304.
-  M. Baldomero-Naranjo, L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, Tightening big Ms in integer programming formulations for support vector machines with ramp loss, *European Journal of Operational Research* 286 (1) (2020) 84–100.
-  M. Baldomero-Naranjo, L. I. Mart nez-Merino, A. M. Rodr guez-Ch a, A robust SVM-based approach with feature selection and outliers detection for classification problems, *Expert Systems with Applications* 178 (2021) 15017.
-  A. Mar n, L. I. Mart nez-Merino, J. Puerto, A. M. Rodr guez-Ch a, The soft-margin Support Vector Machine with ordered weighted average, (submitted).

# Introduction to ramp loss SVM



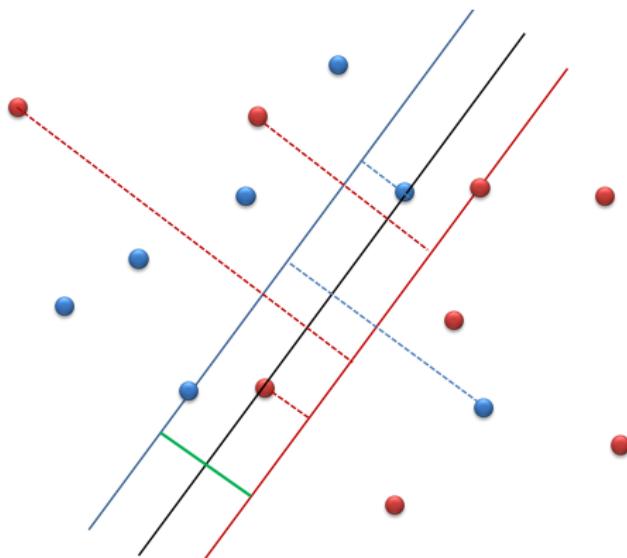
$\text{error}_i = 0 \Leftrightarrow (x_i, y_i)$  is correctly classified.

# Introduction to ramp loss SVM



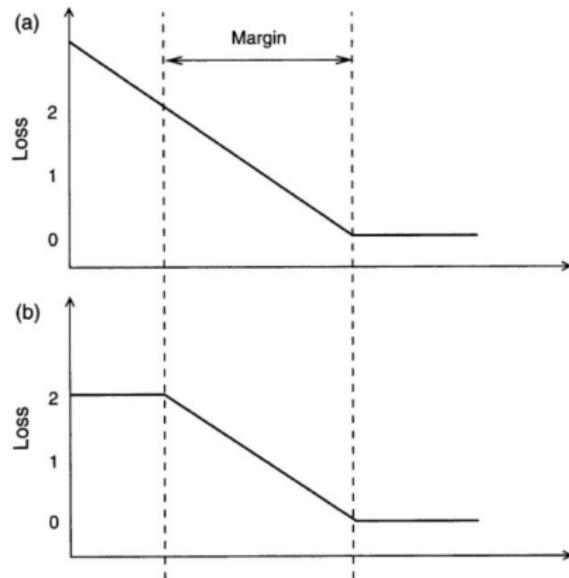
$0 < \text{error}_i < 2 \Leftrightarrow (x_i, y_i)$  is in the margin.

# Introduction to ramp loss SVM



$\text{error}_i = 2 \Leftrightarrow (x_i, y_i)$  is outside the margin and misclassified.

# Introduction to ramp loss SVM



A general ramp loss model

$$\begin{aligned}
(\text{RL-}\ell_p) \quad & \min \quad \frac{1}{p} (\|\boldsymbol{w}\|_p)^p + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
\text{s.t.} \quad & \text{if } z_i = 0, \quad y_i \left( \sum_{k=1}^d w_k x_{ik} + b \right) \geq 1 - \xi_i, \quad i \in N, \\
& 0 \leq \xi_i \leq 2, \quad i \in N, \\
& z_i \in \{0, 1\}, \quad i \in N.
\end{aligned}$$

First introduced by (Brooks, 2011).

# A general ramp loss model

$$\begin{aligned} \text{(RL-}\ell_p\text{-M)} \quad \min \quad & \frac{1}{p}(\|\mathbf{w}\|_p)^p + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\ \text{s.t.} \quad & y_i \left( \sum_{k=1}^d w_k x_{ik} + b \right) \geq 1 - \xi_i - Mz_i, \quad i \in N, \\ & 0 \leq \xi_i \leq 2, \quad i \in N, \\ & z_i \in \{0, 1\}, \quad i \in N. \end{aligned}$$

First introduced by (Brooks, 2011).

# A general ramp loss model

$$\begin{aligned}
 (\text{RL-}\ell_p\text{-M}) \quad & \min \quad \frac{1}{p}(\|\mathbf{w}\|_p)^p + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t.} \quad & y_i \left( \sum_{k=1}^d w_k x_{ik} + b \right) \geq 1 - \xi_i - M z_i, \quad i \in N, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

Our contributions:

- Valid inequalities for RL- $\ell_p$ -M.
- Algorithms for tightening  $M$  parameters in formulations:
  - RL- $\ell_1$ -M
  - RL- $\ell_2$ -M
- Computational results.

# Ramp loss SVM for $\ell_1$ - and $\ell_2$ -norm cases

$$\begin{aligned}
 (\text{RL- } \ell_1\text{-M}) \text{ min} \quad & \sum_{k=1}^d (w_k^+ + w_k^-) + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t:} \quad & y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, \quad i \in N, \\
 & w_k^+ \geq 0, w_k^- \geq 0, \quad k \in D, \\
 & \xi_i \leq 2(1 - z_i), \quad i \in N, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

# Ramp loss SVM for $\ell_1$ - and $\ell_2$ -norm cases

$$\begin{aligned}
 (\text{RL- } \ell_2\text{-M}) \text{ min} \quad & \frac{1}{2} \sum_{k=1}^d w_k^2 + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t.:} \quad & y_i \left( \sum_{k=1}^d w_k \cdot x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, \quad i \in N, \\
 & \xi_i \leq 2(1 - z_i), \quad i \in N, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

# Strategies to tighten $M$ parameter

$$y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, i = 1, \dots, n,$$

# Strategies to tighten $M$ parameter

$$M_i z_i \geq 1 - \xi_i - y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right), i = 1, \dots, n,$$

# Strategies to tighten $M$ parameter

$$M_i \geq 1 - \xi_i - y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right), i = 1, \dots, n,$$

# Strategies to tighten $M$ parameter

$$M_i \geq 1 - \xi_i - y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right), i = 1, \dots, n,$$

$$(\text{UB}_{M_i}) \quad \max \quad 1 - \xi_i - y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) x_{ik} + b \right),$$

s.t: constraints of RL- $\ell_1$ -SVM,

$$\sum_{k=1}^d (w_k^+ + w_k^-) + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right) \leq UB_{RL-\ell_1},$$

$$0 \leq z_i \leq 1, \quad i \in N.$$

# Strategies to tighten $M$ parameter

$$M_i \geq 1 - \xi_i - y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) \cdot x_{ik} + b \right), i = 1, \dots, n,$$

$$(\text{UB}_{M_i}) \quad \max \quad 1 - \xi_i - y_i (w_k x_{ik} + b),$$

s.t: constraints of RL- $\ell_2$ -SVM,

$$\frac{1}{2} \sum_{k=1}^d w_k^2 + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right) \leq UB_{RL-\ell_2},$$

$$0 \leq z_i \leq 1, \quad i \in N.$$

## Some computational results

- Comparison of several solution approaches for solving the models RL- $\ell_1$  and (RL- $\ell_1$ -M).
- Intel(R) Xeon(R) W-2135 CPU 3.70 GHz 32 GB RAM computer, using CPLEX 12.7.0. in Concert Technology C++ .
- Data:

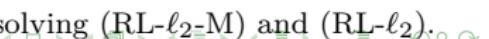
Datasets	$n$	$d$	Class(%)
SONAR	208	60	54/46
SPECT	267	22	79/21
IONO	351	33	64/36
Wdbc	569	30	63/37
WBC	683	9	65/35
Ijcnn1	35000	22	91/9

Table: Real-life datasets

Data	$C$	St.+RL- $\ell_1$ -M				Ind. Const. (LIC)	
		M's Impr.	$t_{st}$	$t_{total}$	GAP	$t$	GAP
SONAR	100	97.92%	11.01	11.90	0.00%	<b>2.20</b>	<b>0.00%</b>
SONAR	10	<b>90.83%</b>	<b>8.76</b>	<b>2686.02</b>	<b>0.00%</b>	7201.62	29.98%
SONAR	1	<b>78.07%</b>	<b>4.59</b>	<b>7218.52</b>	<b>43.16%</b>	7202.05	69.25%
SONAR	0.1	<b>64.72%</b>	<b>2.42</b>	<b>7214.13</b>	<b>8.94%</b>	7207.08	81.12%
SONAR	0.01	<b>73.80%</b>	<b>1.53</b>	<b>2.42</b>	<b>0.00%</b>	7206.15	85.57%
SPECT	100	63.21%	4.83	7211.33	51.18%	<b>7202.47</b>	<b>34.16%</b>
SPECT	10	<b>64.22%</b>	<b>4.75</b>	<b>7206.52</b>	<b>31.43%</b>	7203.11	36.40%
SPECT	1	<b>60.44%</b>	<b>4.95</b>	<b>7207.00</b>	<b>14.51%</b>	7205.19	42.86%
SPECT	0.1	<b>53.77%</b>	4.00	<b>16.31</b>	<b>0.00%</b>	7203.19	43.64%
SPECT	0.01	<b>99.55%</b>	<b>0.28</b>	<b>0.31</b>	<b>0.00%</b>	7204.44	32.73%
IONO	100	94.92%	29.38	159.09	0.00%	7201.85	34.23%
IONO	10	<b>90.97%</b>	<b>25.61</b>	<b>7227.58</b>	<b>18.93%</b>	7202.60	57.00%
IONO	1	<b>85.06%</b>	<b>22.71</b>	<b>7236.20</b>	<b>42.66%</b>	7201.71	65.34%
IONO	0.1	<b>78.95%</b>	<b>18.55</b>	<b>7232.33</b>	<b>22.22%</b>	7204.15	81.53%
IONO	0.01	<b>77.23%</b>	<b>16.76</b>	<b>19.15</b>	<b>0.00%</b>	7205.09	88.89%
Wdbc	100	99.08%	38.97	39.74	0.00%	<b>11.62</b>	<b>0.00%</b>
Wdbc	10	<b>98.12%</b>	<b>38.37</b>	<b>43.90</b>	<b>0.00%</b>	2432.62	0.00%
Wdbc	1	<b>101.33%</b>	<b>40.90</b>	<b>41.18</b>	<b>0.00%</b>	7204.85	63.00%
Wdbc	0.1	<b>102.22%</b>	<b>29.47</b>	<b>29.58</b>	<b>0.00%</b>	7202.43	86.06%
Wdbc	0.01	<b>38.52%</b>	<b>19.80</b>	<b>23.72</b>	<b>0.00%</b>	7204.90	93.87%
WBC	100	<b>95.88%</b>	<b>50.90</b>	<b>163.51</b>	<b>0.00%</b>	7212.06	39.09%
WBC	10	<b>95.78%</b>	<b>44.25</b>	<b>135.87</b>	<b>0.00%</b>	7207.31	30.46%
WBC	1	<b>95.89%</b>	<b>34.13</b>	<b>183.21</b>	<b>0.00%</b>	7207.86	39.13%
WBC	0.1	<b>97.54%</b>	<b>32.11</b>	<b>32.77</b>	<b>0.00%</b>	7203.27	51.23%
WBC	0.01	<b>97.69%</b>	<b>21.76</b>	<b>21.82</b>	<b>0.00%</b>	7205.73	82.77%

Table: Performance of exact approaches on real data for solving RL- $\ell_1$ -M and RL- $\ell_1$ .

Data	C	St.+RL- $\ell_2$ -M				Ind. Const. (LIC)	
		M's Impr.	$t_{st}$	$t_{total}$	GAP	$t$	GAP
Wpbc	100	75.03%	83.20	7283.31	61.50%	7200.30	74.89%
Wpbc	10	73.12%	91.57	7298.90	54.19%	7200.22	79.39%
Wpbc	1	62.64%	78.19	7280.04	56.31%	7200.12	79.00%
Wpbc	0.1	61.73%	64.19	7265.96	54.23%	7200.46	78.06%
Wpbc	0.01	63.82%	53.60	7255.50	37.44%	7201.47	76.30%
SONAR	100	90.09%	162.79	173.73	0.00%	511.82	0.00%
SONAR	10	78.74%	149.84	7350.08	17.15%	7200.22	62.90%
SONAR	1	73.85%	119.74	7321.56	39.72%	7200.24	75.27%
SONAR	0.1	64.39%	115.54	7317.44	50.75%	7200.23	82.49%
SONAR	0.01	50.34%	147.20	7348.83	47.62%	7200.24	88.27%
SPECT	100	70.21%	189.01	836.86	0.00%	7200.23	76.40%
SPECT	10	70.13%	158.47	7361.31	35.42%	7200.29	78.39%
SPECT	1	69.89%	150.57	7353.08	51.00%	7200.35	79.05%
SPECT	0.1	65.31%	170.15	7376.85	36.75%	7200.19	79.81%
SPECT	0.01	55.10%	95.92	7298.00	53.97%	7209.16	80.60%
IONO	100	88.33%	266.56	7467.05	22.43%	7200.40	58.52%
IONO	10	85.44%	245.85	7446.34	35.13%	7200.96	65.65%
IONO	1	81.51%	228.17	7429.80	42.81%	7200.51	78.51%
IONO	0.1	78.84%	207.92	7411.75	40.79%	7201.60	83.29%
IONO	0.01	69.17%	214.34	7430.09	48.28%	7200.25	90.34%
Wdbc	100	97.51%	227.89	230.00	0.00%	249.77	0.00%
Wdbc	10	98.62%	319.62	325.70	0.00%	7200.44	34.59%
Wdbc	1	103.31%	398.20	398.62	0.00%	7200.28	68.70%
Wdbc	0.1	103.59%	410.37	410.53	0.00%	7200.49	87.19%
Wdbc	0.01	101.29%	403.58	403.63	0.00%	7200.27	95.10%
WBC	100	93.37%	244.73	313.64	0.00%	7200.33	40.99%
WBC	10	93.65%	211.91	271.58	0.00%	7200.50	43.27%
WBC	1	94.49%	214.41	277.75	0.00%	7200.71	43.56%
WBC	0.1	99.24%	355.86	356.35	0.00%	7200.31	60.36%
WBC	0.01	105.06%	364.61	364.69	0.00%	7200.27	80.36%

Table: Performance of exact approaches on real data for solving (RL- $\ell_2$ -M) and (RL- $\ell_2$ ). 

# A new model combining feature selection and ramp loss.



M. Labb  , L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, *Discrete Applied Mathematics* 261 (2019) 276–304.



M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Tightening big Ms in integer programming formulations for support vector machines with ramp loss, *European Journal of Operational Research* 286 (1) (2020) 84–100.



M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, A robust SVM-based approach with feature selection and outliers detection for classification problems, *Expert Systems with Applications* 178 (2021) 15017.



A. Mar  n, L. I. Mart  nez-Merino, J. Puerto, A. M. Rodr  guez-Ch  a, The soft-margin Support Vector Machine with ordered weighted average, (submitted).

# A new model with feature selection and ramp loss.

$$\begin{aligned}
 (\text{RL-FS-M}) \quad \min \quad & \sum_{k=1}^d (w_k^+ + w_k^-) + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t.} \quad & y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, \quad i \in N, \\
 & w_k^+ \leq u_k v_k, \quad k \in D, \\
 & w_k^- \leq l_k v_k, \quad k \in D, \\
 & \sum_{k=1}^d v_k \leq B, \\
 & v_k \in \{0, 1\}, \quad k \in D, \\
 & w_k^+ \geq 0, w_k^- \geq 0, \quad k \in D, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

# A new model with feature selection and ramp loss.

$$\begin{aligned}
 (\text{RL-FS-M}) \quad \min \quad & \sum_{k=1}^d (w_k^+ + w_k^-) + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t.} \quad & \textcolor{red}{y}_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, \quad i \in N, \\
 & w_k^+ \leq u_k v_k, \quad k \in D, \\
 & w_k^- \leq l_k v_k, \quad k \in D, \\
 & \sum_{k=1}^d v_k \leq B, \\
 & v_k \in \{0, 1\}, \quad k \in D, \\
 & w_k^+ \geq 0, w_k^- \geq 0, \quad k \in D, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

# A new model with feature selection and ramp loss.

$$\begin{aligned}
 (\text{RL-FS-M}) \quad \min \quad & \sum_{k=1}^d (w_k^+ + w_k^-) + C \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n z_i \right), \\
 \text{s.t.} \quad & y_i \left( \sum_{k=1}^d (w_k^+ - w_k^-) x_{ik} + b \right) \geq 1 - \xi_i - M_i z_i, \quad i \in N, \\
 & w_k^+ \leq u_k v_k, \quad k \in D, \\
 & w_k^- \leq l_k v_k, \quad k \in D, \\
 & \sum_{k=1}^d v_k \leq B, \\
 & v_k \in \{0, 1\}, \quad k \in D, \\
 & w_k^+ \geq 0, w_k^- \geq 0, \quad k \in D, \\
 & 0 \leq \xi_i \leq 2, \quad i \in N, \\
 & z_i \in \{0, 1\}, \quad i \in N.
 \end{aligned}$$

# Tightening big M parameters

---

**Algorithm 1:** Variant 1 and 2. Computation of big M parameters.
 

---

- 1 Solve the problem (SVM- $\ell_1$ ). From its optimal solution, build and solve  $(\text{SVM-}\ell_1)_{\bar{v}, \bar{z}}$ . From its solution build a feasible solution of (RL-FS-M) and obtain an upper bound (UB).
  - 2 **for**  $i \in N$  **do**
  - 3    $M_i = \max_{j \in N} \{\|x_i - x_j\|_\infty : y_i = y_j\} \cdot \text{UB}$ .
  - 4 **for**  $k \in D$  **do**
  - 5    $u_k = \text{UB}$ ,  $l_k = \text{UB}$ .
  - 6 Solve (UB-w). Let UB<sub>w</sub> be the optimal objective value of this problem.
  - 7 Update  $M_i = \max_{j \in N} \{\|x_i - x_j\|_\infty : y_i = y_j\} \cdot \text{UB}_w$ ,  $u_k = \text{UB}_w$ , and  $l_k = \text{UB}_w$ . Add the obtained bounds to the formulation (RL-FS-M) including the set of constraints associated with w bounds.
  - 8 Obtain LB<sub>b</sub> and UB<sub>b</sub>, lower and upper bounds respectively, of the b-variable by solving (LB-b) and (UB-b).
  - 9 Include the constraint LB<sub>b</sub>  $\leq b \leq$  UB<sub>b</sub> in the formulation (RL-FS-M).
  - 10 **while** an improvement of the bounds is obtained **do**
  - 11   Repeat Steps 6 and 7 including constraints related to bounds and LB<sub>b</sub>  $\leq b \leq$  UB<sub>b</sub> in model (UB-w).
    - Case Variant I:
      - for**  $i \in N$  **do**
      - Update  $M_i$  as the optimal value of the problem (UB<sub>M<sub>i</sub></sub>).
    - Case Variant II:
      - For  $i \in N$ , when  $y_i = 1$ , update  $M_i$  as the optimal value of the problem (UB<sub>M<sub>+</sub></sub>).
      - For  $i \in N$ , when  $y_i = -1$ , update  $M_i$  as the optimal value of the problem (UB<sub>M<sub>-</sub></sub>).
-

# Adaptative Kernel Search-based heuristic (DAKS)

- **Initial step:** Algorithm 5 is applied obtaining:
  - Initial UB on the opt. obj. val. and initial values of big Ms.
  - $\hat{z}_i \begin{cases} 0, & z_i = 0 \text{ in the next iteration.} \\ 1, & z_i = 1 \text{ in the next iteration.} \\ 2, & z_i \text{ is a binary variable in the next iteration.} \end{cases}$
- **Second step:**
  - Sort  $v$  variables and define the Initial kernel  $\mathcal{K}_0$ .
  - Solve RL-FS-M( $\mathcal{K}$ ) $_{\hat{z}}$  and update  $\hat{z}$  is updated.
- **Third step:** Iterative process:
  - In each iteration,  $K_i \subseteq D$  is added to the kernel.
  - Solve RL-FS-M( $\mathcal{K} \cup B_{it}$ ) $_{\hat{z}}$ , update Kernel set and update  $\hat{z}$ .
  - After some iterations, the second phase is repeated to reorder  $v$ -variables using the information of the current  $z$ -variables values.

**Other characteristics:** different stop criteria, updating kernel set sizes, updating time limits, etc.

# Computational results

- Intel(R) Xeon(R) W-2135 CPU 3.70 GHz 32 GB RAM computer, using CPLEX 12.7.0. in Concert Technology C ++ .
- Data:

Label	n	d	Class(%)
Colon	62	2000	35/65
Leukemia	72	5327	47/53
DLBCL	77	7129	75/25
SONAR	208	60	54/46
IONO	351	33	64/36
Arrhythmia	420	258	57/43
Wdbc	569	30	63/37
Mfeat	2000	649	10/90
Lepiota	8124	109	52/48

- Validation of (RL-FS-M) and results of DASKS.

# DASKS results

Data	C	$t_e$	GAP	$t_h$	%BS
Colon	100	7225.01	23.59%	49.06	0.00%
	10	7219.15	21.95%	34.66	2.40%
	1	7218.79	5.35%	15.26	1.54%
	0.1	2856.23	0.00%	6.69	0.00%
	0.01	1.25	0.00%	1.24	0.00%
Leukemia	100	7230.38	11.45%	120.14	3.23%
	10	7241.98	12.77%	101.67	2.05%
	1	7281.67	14.82%	92.79	-1.90%
	0.1	41.10	0.00%	43.90	0.00%
	0.01	4.75	0.00%	4.71	0.00%
DLBCL	100	7250.71	10.42%	73.15	0.37%
	10	7266.44	10.76%	82.42	0.37%
	1	7326.72	10.35%	113.93	2.33%
	0.1	2437.37	0.00%	58.74	0.00%
	0.01	7.10	0.00%	7.36	0.00%
SONAR	100	7217.41	94.46%	302.70	4.83%
	10	7211.48	90.50%	202.88	1.91%
	1	7205.42	59.18%	578.35	0.56%
	0.1	7207.64	55.09%	1487.38	0.00%
	0.01	7207.57	1.02%	2.76	0.00%
IONO	100	7222.98	84.66%	77.35	-8.55%
	10	7226.58	72.06%	173.36	0.00%
	1	7237.98	63.71%	117.58	-3.07%
	0.1	7221.80	46.56%	305.99	-0.07%
	0.01	7222.49	12.13%	378.45	0.00%

Data	C	$t_e$	GAP	$t_h$	%BS
Arrhythmia	100	7244.87	98.89%	1844.90	-10.00%
	10	7268.66	98.38%	1860.49	-8.91%
	1	7246.76	89.60%	1844.65	-1.91%
	0.1	7255.37	83.78%	1832.81	0.19%
	0.01	7258.67	59.64%	537.19	0.00%
Wdbc	100	1954.27	0.00%	148.65	1.92%
	10	204.05	0.00%	44.96	0.00%
	1	41.17	0.00%	39.31	0.00%
	0.1	29.74	0.00%	29.17	0.00%
	0.01	7221.52	16.27%	378.81	0.00%
Mfeat	100	7223.05	9.31%	131.33	2.21%
	10	7218.46	8.04%	96.09	2.21%
	1	7228.61	7.75%	89.22	0.80%
	0.1	1060.93	0.00%	70.27	0.00%
	0.01	7233.85	57.84%	101.85	0.00%
Lepiota	100	63.53	0.00%	26.52	0.00%
	10	41.76	0.00%	29.15	0.00%
	1	55.67	0.00%	30.75	0.00%
	0.1	7249.02	11.64%	58.62	0.00%
	0.01	7403.56	1.75%	276.12	0.00%

# Validation of (RL-FS-M)

	5% Label noise			5% SVM outliers				
	Av. (Max.)	ACC impr.	Av. (Max.)	AUC impr.	Av. (Max.)	ACC impr.	Av. (Max.)	AUC impr.
(FS-SVM)		0.44 (2.81)		0.32 (3.08)		2.71(10.13)		3.70 (14.58)
(Fisher-SVM)		5.50 (23.58)		5.00 (22.78)		3.39 (14.46)		3.15 (11.22)
(RFE-SVM)		5.49 (8.08)		7.01 (19.35)		1.11 (2.54)		1.26 (2.82)
(RL- $\ell_1$ -M)		0.51 (3.88)		0.75 (4.29)		0.50 (3.32)		0.64 (3.64)
(SVM- $\ell_1$ )		1.19 (3.17)		1.66 (3.21)		2.58 (5.97)		3.06 (6.53)

**Table:** Improvement of (RL-FS-M) with respect to the rest of the models.

# Ordered Weighted Average in SVM

-  M. Labb  , L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, *Discrete Applied Mathematics* 261 (2019) 276–304.
-  M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, Tightening big Ms in integer programming formulations for support vector machines with ramp loss, *European Journal of Operational Research* 286 (1) (2020) 84–100.
-  M. Baldomero-Naranjo, L. I. Mart  nez-Merino, A. M. Rodr  guez-Ch  a, A robust SVM-based approach with feature selection and outliers detection for classification problems, *Expert Systems with Applications* 178 (2021) 15017.
-  A. Mar  n, L. I. Mart  nez-Merino, J. Puerto, A. M. Rodr  guez-Ch  a, The soft-margin Support Vector Machine with ordered weighted average, (submitted).

# Ordered weighted average

Let  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$  weight vector with  $\lambda_i \geq 0$  for  $i \in N$ ,

$$\begin{aligned} \min_{(\boldsymbol{w}, b, \boldsymbol{\xi})} \quad & \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^n \lambda_i \xi_{(i)}, \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

# Ordered weighted average

Let  $\lambda = (\lambda_1, \dots, \lambda_n)$  weight vector with  $\lambda_i \geq 0$  for  $i \in N$ ,

$$\begin{aligned} \min_{(\boldsymbol{w}, b, \xi)} \quad & \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^n \lambda_i \xi_{(i)}, \\ \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\ & \xi_i \geq 0, \quad i \in N. \end{aligned}$$

- It allows the decision maker to decide how much to penalize big or small deviations.
- Previous approaches considering OWA in SVMs:
  -  S. Maldonado, and J. Merigó, and J. Miranda.  
Redefining support vector machines with the ordered weighted average  
*Knowledge-Based Systems*, 148, 41-46, 2018.
- Our contribution:
  - Formulation for non-decreasing weights (QP)
  - Formulation for general weights (MIQP)

# Formulation for non-decreasing weights

Let  $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ ,

$$\begin{aligned}
 & \min_{(\mathbf{w}, b, \xi, z)} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 + \max C \sum_{i=1}^n \lambda_i \xi_i z_{ij}, \\
 \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & \sum_{i=1}^n z_{ij} = 1, \quad j \in N, \\
 & \sum_{i=j}^n z_{ij} = 1, \quad i \in N, \\
 & 0 \leq z_{ij} \leq 1, \quad i, j \in N, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

# Formulation for non-decreasing weights

Let  $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ ,

$$\begin{aligned}
 (\text{C-OWA-SVM}) \quad & \min_{(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{u}, \boldsymbol{v})} \quad \frac{1}{2} \|\boldsymbol{w}\|_2^2 + \sum_{j=1}^n u_j + \sum_{i=1}^n v_i, \\
 \text{s.t.} \quad & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & u_j + v_i \geq C \lambda_j \xi_i, \quad i, j \in N, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

# Formulation for non-decreasing weights

Let  $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ ,

$$\begin{aligned}
 \text{(C-OWA-SVM)} \quad & \min_{(\boldsymbol{w}, b, \xi, \boldsymbol{u}, \boldsymbol{v})} \quad \frac{1}{2} \|\boldsymbol{w}\|_2^2 + \sum_{j=1}^n u_j + \sum_{i=1}^n v_i, \\
 \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & u_j + v_i \geq C \lambda_j \xi_i, \quad i, j \in N, \\
 & \xi_i \geq 0, \quad i \in N.
 \end{aligned}$$

Non linear Kernels can be applied by using the dual form of (C-OWA-SVM).

# Formulation for general weights

Let  $0 \leq \lambda_1, \lambda_2, \dots, \lambda_n$  general weights.

$$\begin{aligned}
 (\text{NC-OWA-SVM}) \quad & \min_{(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta})} \quad \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{k=1}^n \lambda_k \theta_k, \\
 \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & \sum_{i=1}^n z_{ik} = 1, \quad k \in N, \\
 & \sum_{k=1}^n z_{ik} = 1, \quad i \in N, \\
 & \theta_k \geq \xi_i - M(1 - \sum_{\substack{j=1 \\ j \leq k}}^n z_{ij}), \quad i, k \in N, \\
 & \xi_i \geq 0, \quad i \in N, \\
 & z_{ik} \in \{0, 1\}, \quad i, k \in N, \\
 & \theta_k \geq 0, \quad k \in N.
 \end{aligned}$$

# Formulation for general weights

Let  $0 \leq \lambda_1, \lambda_2, \dots, \lambda_n$  general weights.

$$\begin{aligned}
 (\text{NC-OWA-SVM}) \quad & \min_{(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta})} \quad \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{k=1}^n \lambda_k \theta_k, \\
 \text{s.t.} \quad & y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i \in N, \\
 & \sum_{i=1}^n \textcolor{red}{z}_{ik} = 1, \quad k \in N, \\
 & \sum_{k=1}^n \textcolor{red}{z}_{ik} = 1, \quad i \in N, \\
 & \theta_k \geq \xi_i - M(1 - \sum_{\substack{j=1 \\ j \leq k}}^n \textcolor{red}{z}_{ij}), \quad i, k \in N, \\
 & \xi_i \geq 0, \quad i \in N, \\
 & \textcolor{red}{z}_{ik} \in \{0, 1\}, \quad i, k \in N, \\
 & \theta_k \geq 0, \quad k \in N.
 \end{aligned}$$

Non linear Kernels can be applied by using the dual form of (NC-OWA-SVM).

$$\begin{aligned}
 (\text{NC-OWA-SVM}_K) \quad & \min_{(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta})} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + C \sum_{k=1}^n \lambda_k \theta_k, \\
 \text{s.t.} \quad & y_i \left( \sum_{j=1}^n y_j \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \geq 1 - \xi_i, \quad i \in N, \\
 & \sum_{i=1}^n z_{ik} = 1, \quad k \in N, \\
 & \sum_{k=1}^n z_{ik} = 1, \quad i \in N, \\
 & \theta_k \geq \xi_i - M \left( 1 - \sum_{\substack{j=1 \\ j \leq k}}^n z_{ij} \right), \quad i, k \in N, \\
 & \xi_i \geq 0, \quad i \in N, \\
 & z_{ik} \in \{0, 1\}, \quad i, k \in N, \\
 & \theta_k \geq 0, \quad k \in N, \\
 & \alpha_i \geq 0,
 \end{aligned}$$

Non linear Kernels can be applied by using the dual form of (NC-OWA-SVM).

$$\begin{aligned}
 (\text{NC-OWA-SVM}_K) \quad & \min_{(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta})} \quad \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) + C \sum_{k=1}^n \lambda_k \theta_k, \\
 \text{s.t.} \quad & y_i \left( \sum_{j=1}^n y_j \alpha_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq 1 - \xi_i, \quad i \in N, \\
 & \sum_{i=1}^n z_{ik} = 1, \quad k \in N, \\
 & \sum_{k=1}^n z_{ik} = 1, \quad i \in N, \\
 & \theta_k \geq \xi_i - M \left( 1 - \sum_{\substack{j=1 \\ j \leq k}}^n z_{ij} \right), \quad i, k \in N, \\
 & \xi_i \geq 0, \quad i \in N, \\
 & z_{ik} \in \{0, 1\}, \quad i, k \in N, \\
 & \theta_k \geq 0, \quad k \in N, \\
 & \alpha_i \geq 0, \quad i \in N.
 \end{aligned}$$

# Computational results

- Intel (R) Xeon (R) W-2245 CPU 3.90 GHz 3.91 GHz 256 GB RAM
- Python+Cplex 20.1
- Real life datasets (UCI repository):

Dataset	Samples	% Class	Features
IONO	351	64.1/35.9	34
WBC	569	62.7/37.3	30
AUS	690	55.5/44.5	14
DIA	768	65.1/34.9	8
GC	1000	70.0/30.0	24
SPL	1000	51.7/48.3	60

- Weights based on linguistic quantifiers.

# ACC and AUC comparisons

Data	ACC (%)					
	$\ell_2$ -SVM	app-OWA-SVM	ex-OWA-SVM	$\ell_2$ -SVM <sub>K</sub>	app-OWA-SVM <sub>K</sub>	ex-OWA-SVM <sub>K</sub>
IONO	90.60%	90.89%	90.89%	95.44%	<b>95.72%</b>	<b>95.72%</b>
WBC	98.07%	98.07%	97.71%	98.24%	98.42%	<b>98.77%</b>
AUS	85.51%	86.09%	85.51%	86.38%	87.25%	<b>87.39%</b>
DIA	77.60%	77.73%	77.73%	77.34%	<b>78.38%</b>	78.12%
GC	76.90%	77.50%	77.30%	77.30%	<b>77.50%</b>	<b>77.50%</b>
SPL	81.30%	82.00%	81.70%	88.40%	<b>89.80%</b>	89.40%

Data	AUC (%)					
	$\ell_2$ -SVM	app-OWA-SVM	ex-OWA-SVM	$\ell_2$ -SVM <sub>K</sub>	app-OWA-SVM <sub>K</sub>	ex-OWA-SVM <sub>K</sub>
IONO	88.67%	88.89%	88.89%	94.55%	<b>95.15%</b>	<b>95.15%</b>
WBC	97.70%	97.70%	97.14%	97.84%	98.08%	<b>98.44%</b>
AUS	86.20%	86.67%	86.20%	86.54%	87.46%	<b>87.71%</b>
DIA	72.32%	72.96%	73.00%	72.19%	<b>73.30%</b>	<b>73.30%</b>
GC	68.98%	71.00%	69.19%	68.74%	<b>71.67%</b>	69.90%
SPL	81.41%	82.02%	81.78%	88.41%	<b>89.86%</b>	<b>89.86%</b>

# Time comparisons

Data	$\ell_2$ -SVM	Time (s)							
		app-OWA-SVM		ex-OWA-SVM	$\ell_2$ -SVM <sub>K</sub>	app-OWA-SVM <sub>K</sub>		ex-OWA-SVM <sub>K</sub>	
		Step 1	Step 2			Step 1	Step 2		
IONO	0.023	0.014	0.212	4.725	0.006	0.003	1.900	4.967	
WBC	0.003	0.003	0.494	12.628	0.003	0.002	4.915	9.345	
AUS	0.011	0.425	0.297	19.900	0.014	0.014	8.105	26.609	
DIA	0.013	0.009	0.286	25.133	0.013	0.008	9.944	21.536	
GC	0.033	0.016	0.560	41.788	0.024	0.025	17.090	31.707	
SPL	0.022	0.522	0.815	40.561	0.057	0.038	16.387	30.866	

# Some conclusions

## New considerations in the classical SVM model

- Feature selection (FS-SVM):
  - Better interpretation.
  - Dealing with costly features.
- SVM with ramp loss. ( $\text{RL-}\ell_1\text{-M}$ ,  $\text{RL-}\ell_2\text{-M}$ )
  - Robust in presence of outliers.
- Feature selection and ramp loss. ( $\text{RL-FS-SVM}$ )
- Including Ordered Weighted Average in SVM. ( $\text{OWA-SVM}$ )
  - Different penalization to errors depending on their sizes.

# Bibliography

-  E. Angelelli, R. Mansini, and M. Speranza. Kernel search: A new heuristic framework for portfolio selection. *Computational Optimization and Applications*, 51(1):345–361, 2012.
-  P. S. Bradley, O. L. Mangasarian, Feature selection via concave minimization and support vector machines, *ICML 98* (1998) 82–90.
-  J. P. Brooks, Support vector machines with the ramp loss and the hard margin loss, *Operations Research* 59 (2) (2011) 467–479.
-  G. Guastaroba and M. Speranza. Kernel search for the capacitated facility location problem. *Journal of Heuristics*, 18(6):877–917, 2012.

# Thank you for your attention!