A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification

Asunción Jiménez-Cordero

asuncionjc@uma.es

Joint work with:

Juan Miguel Morales González

Salvador Pineda Morente

oasys.uma.es

Premio *Ramiro Melendreras*. Jornadas SEIO 2021

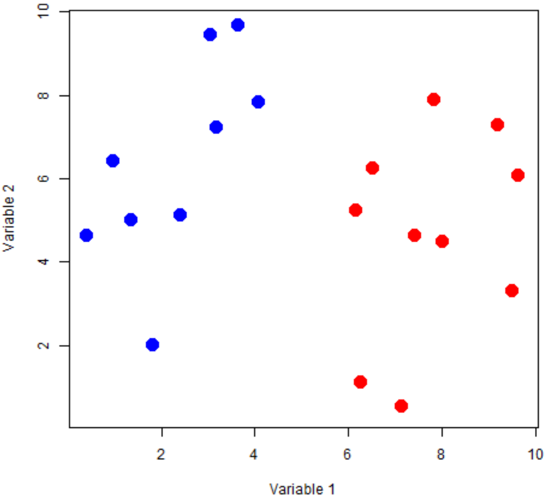
June 10th, 2021

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 755705)

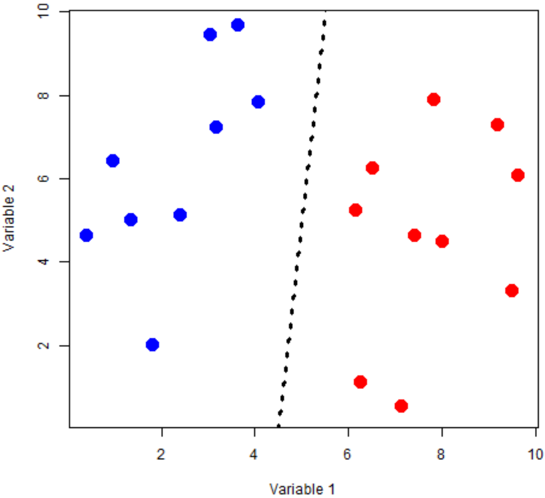
Outline

Outline

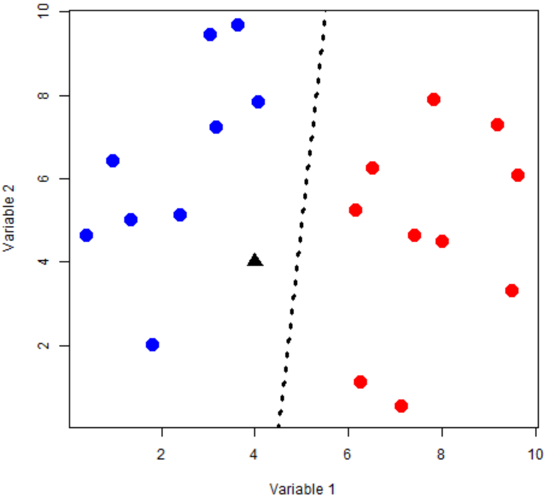
Binary Classification Problem.



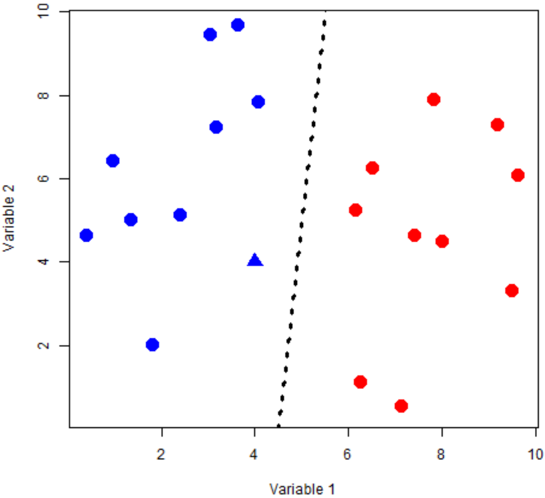
Binary Classification Problem.



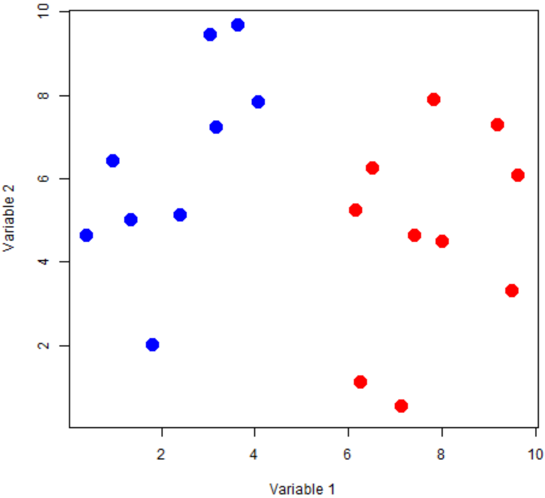
Binary Classification Problem.



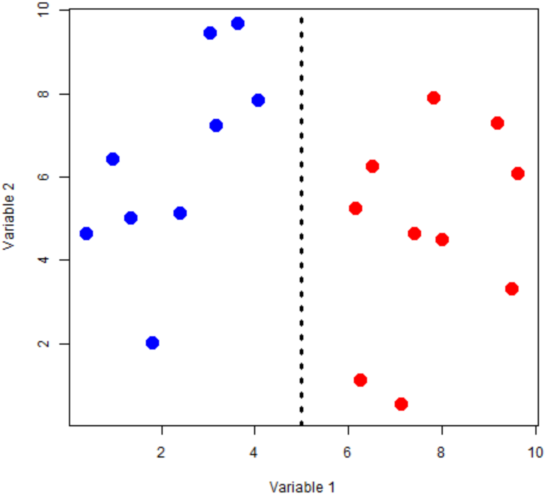
Binary Classification Problem.



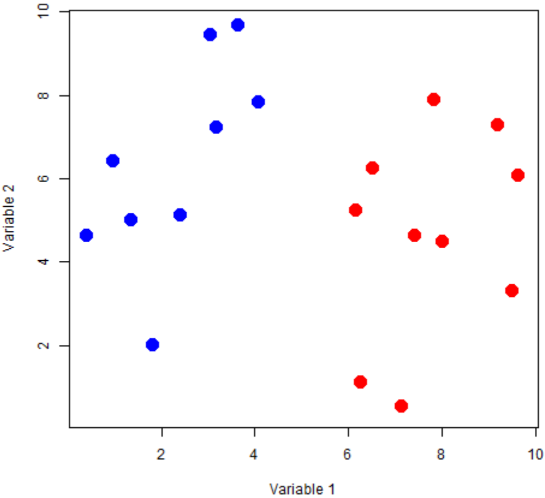
Binary Classification Problem. Feature Selection



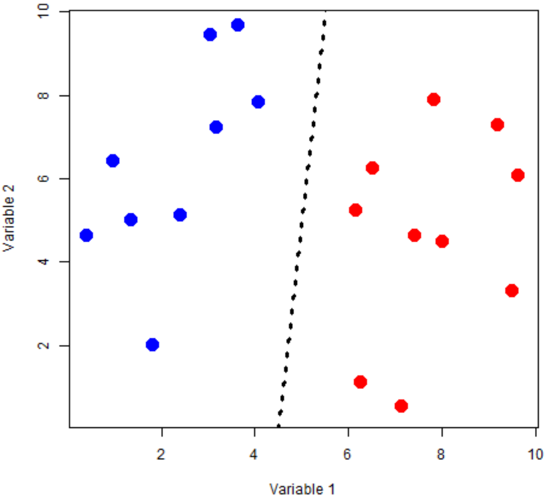
Binary Classification Problem. Feature Selection



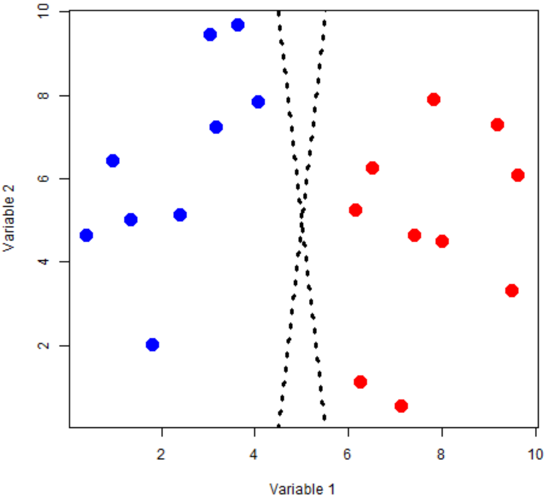
Support Vector Machines (SVM)



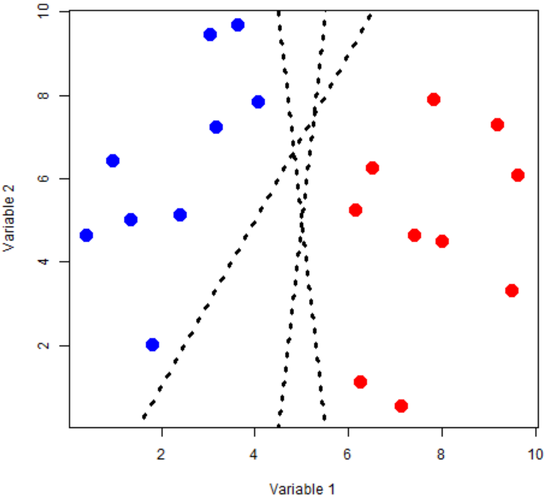
Support Vector Machines (SVM)



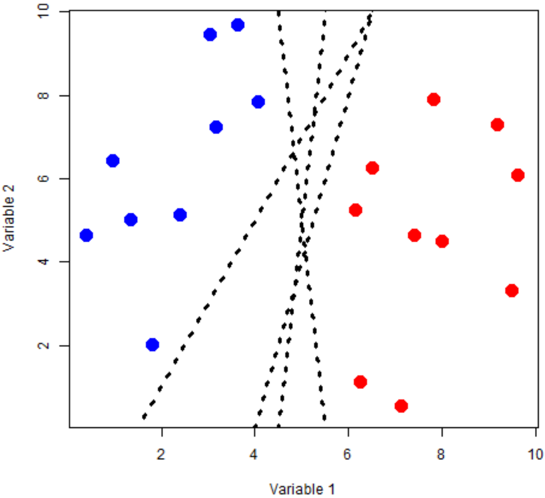
Support Vector Machines (SVM)



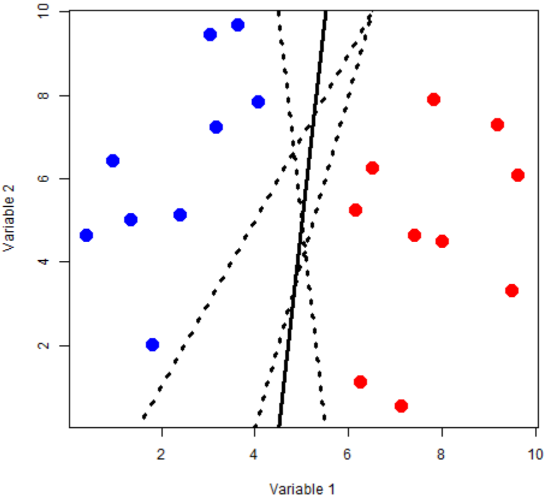
Support Vector Machines (SVM)



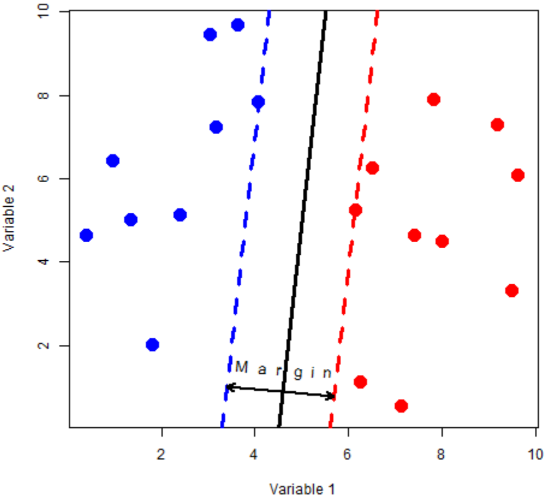
Support Vector Machines (SVM)



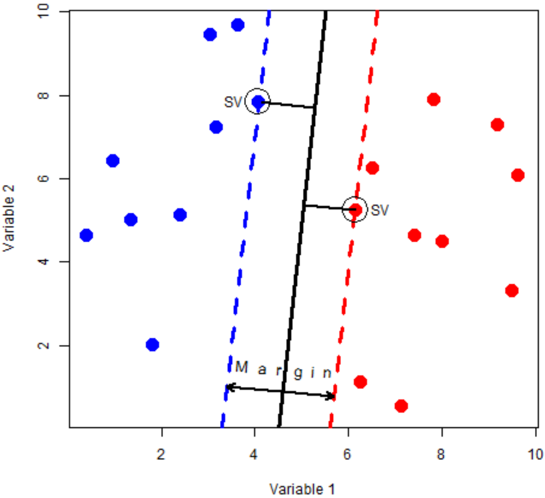
Support Vector Machines (SVM)



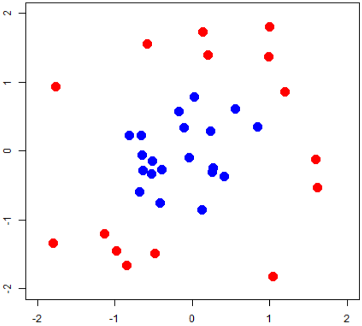
Support Vector Machines (SVM)



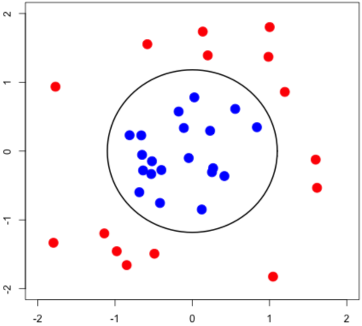
Support Vector Machines (SVM)



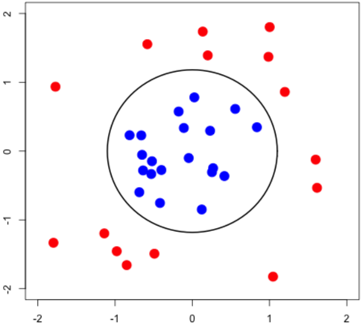
What happens in the nonlinear case?



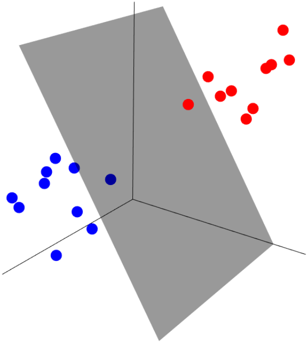
What happens in the nonlinear case?



What happens in the nonlinear case?



$$\overset{\phi}{\Longrightarrow}$$



Aim

- Develop a new Mathematical Optimization approach to perform feature selection in a binary classification problem.
- Classification tool: Support Vector Machine (SVM).
- Min-max approach.

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
    - Most separated classes in high-dimensional space.
    - Fast, but not take into account classifier information.

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):
    - Min-max approach with binary variables.
    - Fixed # of selected features.
    - Equivalent RFE-SVM which sequentially removes features.

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):
  - Embedded

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):
  - Embedded (KP-SVM, *Maldonado et al. [2011]*):

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):
  - Embedded (KP-SVM, *Maldonado et al. [2011]*):
    - ℓ_0-(pseudo)norm approximation to dual SVM.
    - Large number of hyperparameters.
    - Ad-hoc approaches.

Literature Review

- Linear: *Gaudioso et al. [2017]; Labbé et al. [2019]; Maldonado et al. [2014]*; ...
- **Nonlinear:**
  - Filter (kernel polarization, *Wang et al. [2010]*):
  - Wrapper (min-max RFE, *Onel et al. [2019]*):
  - Embedded (KP-SVM, *Maldonado et al. [2011]*):

Our contributions

- Embedded feature selection method.
- # selected features is not fixed, but provided by our methodology.
- No ad-hoc strategies. Off-the-shelf solvers.

Outline

SVM Problem (Primal).

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{ \begin{array}{rl} \min\limits_{w,\,b,\,\xi} & \frac{1}{2}\|w\|^2 + C \sum\limits_{i \in \mathcal{S}} \xi_i \\ \text{s.t.} & (w' \quad x_i \ + b) y_i \geq 1 - \xi_i, \quad i \in \mathcal{S} \\ & \xi_i \geq 0, \quad i \in \mathcal{S} \end{array} \right.$$

SVM Problem (Primal).

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{ \begin{array}{rl} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C \sum\limits_{i \in \mathcal{S}} \xi_i \\ \text{s.t.} & (w' \phi(x_i) + b) y_i \geq 1 - \xi_i, \quad i \in \mathcal{S} \\ & \xi_i \geq 0, \quad i \in \mathcal{S} \end{array} \right.$$

SVM Problem (Primal).

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{\begin{array}{lll} \boldsymbol{\min\limits_{\gamma}} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i \in \mathcal{S}} \xi_i & \\ \text{s.t.} & (w'\boldsymbol{\phi(x_i)} + b)y_i \geq 1 - \xi_i, & i \in \mathcal{S} \\ & \xi_i \geq 0, \quad i \in \mathcal{S} & \\ & \boldsymbol{\phi \in \mathcal{F}_\gamma} & \end{array}\right.$$

SVM Problem (Primal).

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{ \begin{array}{lll} \min\limits_{\gamma} & \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C \sum\limits_{i \in \mathcal{S}} \xi_i \\ & \text{s.t.} & (w'\phi(x_i) + b)y_i \geq 1 - \xi_i, \quad i \in \mathcal{S} \\ & & \xi_i \geq 0, \quad i \in \mathcal{S} \\ & & \phi \in \mathcal{F}_\gamma \end{array} \right.$$

Unfortunately

ϕ and $\mathcal{F}_\gamma$ are usually unknown.

SVM Problem (Dual). Feature Selection.

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{ \begin{array}{rl} \min\limits_{\gamma} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell \phi(x_i)' \phi(x_\ell) \\ \text{s.t.} & \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\ & \alpha_i \in [0, C], \quad i \in \mathcal{S} \end{array} \right.$$

SVM Problem (Dual). Feature Selection.

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$
\left\{
\begin{array}{rl}
\min\limits_{\gamma} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell \phi(x_i)' \phi(x_\ell) \\
\text{s.t.} & \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\
 & \alpha_i \in [0, C], \quad i \in \mathcal{S}
\end{array}
\right.
$$

Kernel trick

$$K_\gamma(x_i, x_\ell) = \phi(x_i)' \phi(x_\ell)$$

SVM Problem (Dual). Feature Selection.

Data: $x_i \in \mathbb{R}^M$ Class label: $y_i \in \{-1, 1\}$

$$\left\{\begin{array}{rl} \min\limits_{\gamma} \max\limits_{\alpha} & \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell \boldsymbol{K_\gamma(x_i, x_\ell)} \\ \text{s.t.} & \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\ & \alpha_i \in [0, C], \quad i \in \mathcal{S} \end{array}\right.$$

Kernel trick

$$K_\gamma(x_i, x_\ell) = \phi(x_i)'\phi(x_\ell)$$

SVM Problem. Kernel.

Anisotropic Gaussian kernel

$$K_\gamma(x_i, x_\ell) = \exp\left(-\sum_{j=1}^{M} \boldsymbol{\gamma_j}(x_{ij} - x_{\ell j})^2\right)$$

SVM Problem. Kernel.

Anisotropic Gaussian kernel

$$K_\gamma(x_i, x_\ell) = \exp\left(-\sum_{j=1}^{M} \boldsymbol{\gamma_j}(x_{ij} - x_{\ell j})^2\right)$$

γ_j: importance feature j

- $\gamma_j \to 0$: no role in the classification.
- large γ_j: critical when classifying.

SVM Problem. Kernel.

Anisotropic Gaussian kernel

$$K_\gamma(x_i, x_\ell) = \exp\left(- \sum_{j=1}^{M} \boldsymbol{\gamma_j} (x_{ij} - x_{\ell j})^2 \right)$$

γ_j: importance feature j

- $\gamma_j \to 0$: no role in the classification.
- large γ_j: critical when classifying.
- $\gamma_j = 0, \forall j \Rightarrow$ all points classified same class.
- $\gamma_j \to \infty$: overfitting.

SVM Problem. Kernel.

Anisotropic Gaussian kernel

$$K_\gamma(x_i, x_\ell) = \exp\left(-\sum_{j=1}^{M} \boldsymbol{\gamma_j}(x_{ij} - x_{\ell j})^2\right)$$

γ_j: importance feature j

- $\gamma_j \to 0$: no role in the classification.
- large γ_j: critical when classifying.
- $\gamma_j = 0, \forall j \Rightarrow$ all points classified same class.
- $\gamma_j \to \infty$: overfitting.

Trade-off

- Model complexity.
- Classification accuracy.

Problem Formulation

$$
\begin{cases}
\min\limits_{\gamma \geq 0} \left[\max\limits_{\boldsymbol{\alpha}} \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right] \\
\text{s.t.} \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\
0 \leq \alpha_i \leq C, \forall i
\end{cases}
$$

Classification accuracy

Problem Formulation

$$\begin{cases} \min\limits_{\gamma \geq 0} \left[\quad \|\gamma\|_p^p + \qquad \begin{array}{l} \max\limits_{\alpha} \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \end{array} \right] \\ \qquad\qquad\qquad\qquad \text{s.t.} \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\ \qquad\qquad\qquad\qquad\qquad 0 \leq \alpha_i \leq C, \forall i \end{cases}$$

Model complexity

Classification accuracy

Problem Formulation

$$\begin{cases} \min\limits_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max\limits_{\alpha} \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right] \\ \qquad \text{s.t.} \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\ \qquad 0 \leq \alpha_i \leq C, \forall i \end{cases}$$

Model complexity

Classification accuracy

Trade-off

Outline

Min-max optimization problem.

$$\begin{cases} \min\limits_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max\limits_{\alpha} \sum\limits_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right] \\ \qquad\qquad\qquad\qquad \text{s.t.} \quad \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\ \qquad\qquad\qquad\qquad\qquad 0 \leq \alpha_i \leq C, \forall i \end{cases}$$

Epigraph form.

$$
\begin{cases}
\min\limits_{\gamma \geq 0,\, \boldsymbol{z}} \; C_2 \|\gamma\|_p^p + (1 - C_2)\boldsymbol{z} \\[2ex]
\text{s.t.} \quad \boldsymbol{z \geq \max\limits_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i, \ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell)} \\[2ex]
\qquad\qquad\qquad\quad \textbf{s.t.} \; \boldsymbol{\sum_{i \in \mathcal{S}} \alpha_i y_i = 0} \\[2ex]
\qquad\qquad\qquad\qquad \boldsymbol{0 \leq \alpha_i \leq C,\, \forall i}
\end{cases}
$$

Strong duality lower level problem (SVM).

$$
\begin{cases}
\min\limits_{\gamma \geq 0, z} \; C_2 \|\gamma\|_p^p + (1 - C_2) z \\
\text{s.t.} \quad z \geq \max\limits_{\alpha} \sum\limits_{i \in \mathcal{S}} \alpha_i - \dfrac{1}{2} \sum\limits_{i, \ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \\
\qquad\qquad \text{s.t.} \; \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 & \boldsymbol{(\nu)} \\
\qquad\qquad\qquad 0 \leq \alpha_i \leq C, \, \forall i & \boldsymbol{(\lambda_i^0, \lambda_i^C)}
\end{cases}
$$

Strong duality lower level problem (SVM).

Dual lower-level problem. Lagrangian. $[G_\gamma = diag(y) K_\gamma diag(y)]$

$$\begin{cases} \min\limits_{\alpha,\nu,\lambda^0,\lambda^C} & -\dfrac{1}{2}\alpha' G_\gamma \alpha + (e - \nu y + \lambda^0 - \lambda^C)'\alpha + C(\lambda^C)'e \\ \text{s.t.} & G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0 \\ & \lambda^0, \lambda^C \geq 0 \\ & 0 \leq \alpha \leq C \end{cases}$$

Strong duality lower level problem (SVM).

$$
\begin{cases}
\min\limits_{\gamma \geq 0, z} \; C_2 \|\gamma\|_p^p + (1 - C_2) z \\
\text{s.t.} \quad z \geq \max\limits_{\alpha} \sum\limits_{i \in \mathcal{S}} \alpha_i - \dfrac{1}{2} \sum\limits_{i, \ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \\
\qquad\qquad \text{s.t.} \quad \sum\limits_{i \in \mathcal{S}} \alpha_i y_i = 0 \\
\qquad\qquad\qquad 0 \leq \alpha_i \leq C, \; \forall i
\end{cases}
$$

Dual lower-level problem. Lagrangian. $[G_\gamma = diag(y) K_\gamma diag(y)]$

$$
\begin{cases}
\min\limits_{\alpha, \nu, \lambda^0, \lambda^C} \; -\dfrac{1}{2} \alpha' G_\gamma \alpha + (e - \nu y + \lambda^0 - \lambda^C)' \alpha + C (\lambda^C)' e \\
\quad \text{s.t.} \quad G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0 \\
\qquad\qquad \lambda^0, \lambda^C \geq 0 \\
\qquad\qquad 0 \leq \alpha \leq C
\end{cases}
$$

Strong duality lower level problem (SVM).

$$
\begin{cases}
\min\limits_{\gamma \geq 0,\, z} \; C_2 \|\gamma\|_p^p + (1 - C_2) z \\
\text{s.t.} \quad z \geq \min\limits_{\boldsymbol{\alpha}, \boldsymbol{\nu}, \boldsymbol{\lambda}^{\mathbf{0}}, \boldsymbol{\lambda}^{\boldsymbol{C}}} \; -\frac{1}{2} \boldsymbol{\alpha}' \boldsymbol{G}_{\boldsymbol{\gamma}} \boldsymbol{\alpha} + (\boldsymbol{e} - \boldsymbol{\nu} \boldsymbol{y} + \boldsymbol{\lambda}^{\mathbf{0}} - \boldsymbol{\lambda}^{\boldsymbol{C}})' \boldsymbol{\alpha} + \boldsymbol{C} (\boldsymbol{\lambda}^{\boldsymbol{C}})' \boldsymbol{e} \\
\qquad\qquad \textbf{s.t.} \quad \boldsymbol{G}_{\boldsymbol{\gamma}} \boldsymbol{\alpha} - (\boldsymbol{e} - \boldsymbol{\nu} \boldsymbol{y} + \boldsymbol{\lambda}^{\mathbf{0}} - \boldsymbol{\lambda}^{\boldsymbol{C}}) = \mathbf{0} \\
\qquad\qquad\qquad \boldsymbol{\lambda}^{\mathbf{0}}, \boldsymbol{\lambda}^{\boldsymbol{C}} \geq \mathbf{0} \\
\qquad\qquad\qquad \mathbf{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{C}
\end{cases}
$$

Dual lower-level problem. Lagrangian. $[G_\gamma = diag(y) K_\gamma diag(y)]$

$$
\begin{cases}
\min\limits_{\alpha, \nu, \lambda^0, \lambda^C} \; -\frac{1}{2} \alpha' G_\gamma \alpha + (e - \nu y + \lambda^0 - \lambda^C)' \alpha + C (\lambda^C)' e \\
\qquad \text{s.t.} \quad G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0 \\
\qquad\qquad \lambda^0, \lambda^C \geq 0 \\
\qquad\qquad 0 \leq \alpha \leq C
\end{cases}
$$

Single-level optimization problem

$$\begin{cases} \min\limits_{\gamma,\,\alpha,\nu,\lambda^0,\lambda^C} C_2\|\gamma\|_p^p - (1-C_2)\left(\frac{1}{2}\alpha' G_\gamma \alpha - (e-\nu y+\lambda^0-\lambda^C)'\alpha - C(\lambda^C)'e\right) \\ \quad\text{s.t.}\ \ G_\gamma\alpha - (e-\nu y+\lambda^0-\lambda^C) = 0 \\ \qquad\quad \gamma,\lambda^0,\lambda^C \geq 0 \\ \qquad\quad 0 \leq \alpha \leq C \end{cases}$$

Recap

$$\min_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right]$$

$$\text{s.t.} \sum_{i \in \mathcal{S}} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \forall i$$

Recap

$$\min_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right]$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{S}} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \forall i$$

$$\min_{\gamma \geq 0, z} C_2 \|\gamma\|_p^p + (1 - C_2) z$$

$$\text{s.t.} \quad z \geq \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{S}} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \forall i$$

Recap

$$\min_{\gamma\geq 0}\left[C_2\|\gamma\|_p^p+(1-C_2)\max_{\alpha}\sum_{i\in\mathcal{S}}\alpha_i-\frac{1}{2}\sum_{i,\ell\in\mathcal{S}}\alpha_i\alpha_\ell y_i y_\ell K_\gamma(x_i,x_\ell)\right]$$
$$\text{s.t.}\quad \sum_{i\in\mathcal{S}}\alpha_i y_i=0$$
$$0\leq\alpha_i\leq C,\forall i$$

$$\min_{\gamma\geq 0,z} C_2\|\gamma\|_p^p+(1-C_2)z$$
$$\text{s.t.}\quad z\geq\max_{\alpha}\sum_{i\in\mathcal{S}}\alpha_i-\frac{1}{2}\sum_{i,\ell\in\mathcal{S}}\alpha_i\alpha_\ell y_i y_\ell K_\gamma(x_i,x_\ell)$$
$$\text{s.t.}\quad \sum_{i\in\mathcal{S}}\alpha_i y_i=0$$
$$0\leq\alpha_i\leq C,\ \forall i$$

$$\min_{\gamma\geq 0,\ z} C_2\|\gamma\|_p^p+(1-C_2)z$$
$$\text{s.t.}\quad z\geq\min_{\alpha,\nu,\lambda^0,\lambda^C}-\frac{1}{2}\alpha' G_\gamma\alpha+(e-\nu y+\lambda^0-\lambda^C)'\alpha+C(\lambda^C)'e$$
$$\text{s.t.}\quad G_\gamma\alpha-(e-\nu y+\lambda^0-\lambda^C)=0$$
$$\lambda^0,\lambda^C\geq 0$$
$$0\leq\alpha\leq C$$

Recap

$$\min_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right]$$
$$\text{s.t.} \quad \sum_{i \in \mathcal{S}} \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \forall i$$

$$\min_{\gamma \geq 0, z} C_2 \|\gamma\|_p^p + (1 - C_2) z$$
$$\text{s.t.} \quad z \geq \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell)$$
$$\text{s.t.} \quad \sum_{i \in \mathcal{S}} \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \forall i$$

$$\min_{\gamma \geq 0,\, z} C_2 \|\gamma\|_p^p + (1 - C_2) z$$
$$\text{s.t.} \quad z \geq \min_{\alpha, \nu, \lambda^0, \lambda^C} -\frac{1}{2} \alpha' G_\gamma \alpha + (e - \nu y + \lambda^0 - \lambda^C)' \alpha + C (\lambda^C)' e$$
$$\text{s.t.} \quad G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0$$
$$\lambda^0, \lambda^C \geq 0$$
$$0 \leq \alpha \leq C$$

$$\min_{\gamma,\, \alpha, \nu, \lambda^0, \lambda^C} C_2 \|\gamma\|_p^p - (1 - C_2) \left(\tfrac{1}{2} \alpha' G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C)' \alpha - C (\lambda^C)' e \right)$$
$$\text{s.t.} \quad G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0$$
$$\gamma, \lambda^0, \lambda^C \geq 0$$
$$0 \leq \alpha \leq C$$

Outline

How to solve the problem?

- Problem very hard to solve (nonlinear, nonconvex, C, C_2).

$$
\begin{aligned}
\min_{\gamma,\, \alpha, \nu, \lambda^0, \lambda^C} \; & C_2 \|\gamma\|_p^p - (1 - C_2)\left(\tfrac{1}{2}\alpha' G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C)'\alpha - C(\lambda^C)' e\right) \\
\text{s.t. } & G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0 \\
& \gamma, \lambda^0, \lambda^C \geq 0 \\
& 0 \leq \alpha \leq C
\end{aligned}
$$

How to solve the problem?

- Problem very hard to solve (nonlinear, nonconvex, C, C_2).

$$\begin{aligned}
\min_{\gamma,\, \alpha, \nu, \lambda^0, \lambda^C} \quad & C_2\|\gamma\|_p^p - (1-C_2)\left(\tfrac{1}{2}\alpha' G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C)'\alpha - C(\lambda^C)'e\right) \\
\text{s.t.} \quad & G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0 \\
& \gamma, \lambda^0, \lambda^C \geq 0 \\
& 0 \leq \alpha \leq C
\end{aligned}$$

- Simple solving strategy. No ad-hoc approaches.

How to solve the problem?

- Problem very hard to solve (nonlinear, nonconvex, C, C_2).

$$\min_{\gamma,\,\alpha,\nu,\lambda^0,\lambda^C} C_2\|\gamma\|_p^p - (1-C_2)\left(\tfrac{1}{2}\alpha' G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C)'\alpha - C(\lambda^C)'e\right)$$
$$\text{s.t. } G_\gamma \alpha - (e - \nu y + \lambda^0 - \lambda^C) = 0$$
$$\gamma, \lambda^0, \lambda^C \geq 0$$
$$0 \leq \alpha \leq C$$

- Simple solving strategy. No ad-hoc approaches.
- Grid search + local solver (`Ipopt`)+ k-fold CV (train, validation, test).

Experimental Setup

Data set	# individuals	# features	% predominant class
breast	569	30	63%
diabetes	768	8	65%
lymphoma	96	4026	64%
colorectal	62	2000	65%

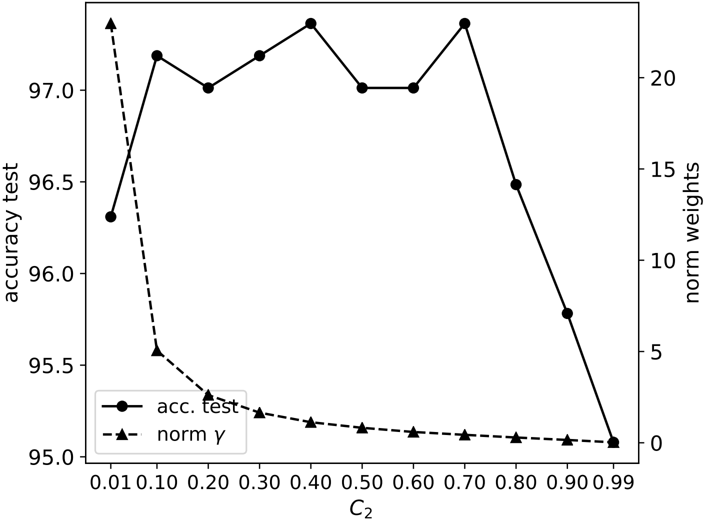
UCI Machine Learning Repository

Experimental Setup

Data set	# individuals	# features	% predominant class
breast	**569**	**30**	**63%**
diabetes	768	8	65%
lymphoma	96	4026	64%
colorectal	62	2000	65%

UCI Machine Learning Repository

Numerical Results



Numerical Results

	breast
MM-FS	97.35%
NO-FS	97.89%
ℓ_1-SVM	96.83%
Bench. ad-hoc	97.55%
Bench. off-the-shelf	62.74%

Numerical Results

	breast
MM-FS	97.35%
NO-FS	97.89%

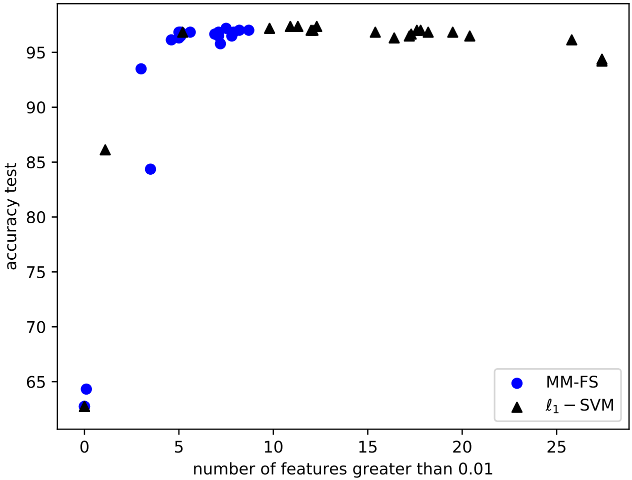
Numerical Results

	breast
MM-FS	97.35%
Bench. ad-hoc	97.55%
Bench. off-the-shelf	62.74%

Numerical Results

	breast
MM-FS	97.35%
ℓ_1-SVM	96.83%

Numerical Results



Numerical Results

C_2	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
	21	21	21	21	21	21	21	21	21	21	28
	11	22	22	22	22	22	22	22	22	28	21
breast	22	11	8	25	29	28	25	25	28	22	8
	25	7	25	29	25	29	29	29	25	8	23
	30	25	11	8	28	25	28	28	8	7	22

Further info

European Journal of Operational Research 293 (2021) 24–35

Contents lists available at ScienceDirect

European Journal of Operational Research

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

ELSEVIER

Continuous Optimization

A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification

Asunción Jiménez-Cordero*, Juan Miguel Morales, Salvador Pineda

OASYS Group, University of Málaga, Málaga, Spain

ARTICLE INFO

Article history:
Received 22 April 2020
Accepted 5 December 2020
Available online 17 December 2020

Keywords:
Machine learning
Min-max optimization
Duality theory
Feature selection
Nonlinear Support Vector Machine classification

ABSTRACT

In recent years, feature selection has become a challenging problem in several machine learning fields, such as classification problems. Support Vector Machine (SVM) is a well-known technique applied in classification tasks. Various methodologies have been proposed in the literature to select the most relevant features in SVM. Unfortunately, all of them either deal with the feature selection problem in the linear classification setting or propose ad-hoc approaches that are difficult to implement in practice. In contrast, we propose an embedded feature selection method based on a min-max optimization problem, where a trade-off between model complexity and classification accuracy is sought. By leveraging duality theory, we equivalently reformulate the min-max problem and solve it without further ado using off-the-shelf software for nonlinear optimization. The efficiency and usefulness of our approach are tested on several benchmark data sets in terms of accuracy, number of selected features and interpretability.

Published in EJOR

Available at
`https://www.sciencedirect.com/science/article/pii/S0377221720310195`

Outline

Conclusions

- Min-max optimization problem for SVM classification and feature selection.
- Single-level reformulation based on strong duality.
- Simple but efficient solving strategy. No ad-hoc.
- Competitive literature results.

Conclusions

- Min-max optimization problem for SVM classification and feature selection.
- Single-level reformulation based on strong duality.
- Simple but efficient solving strategy. No ad-hoc.
- Competitive literature results.

Future Research

- Extension to regression or clustering.
- Other real-life applications (Physically-aware approach).

References

Agor, J. and Özaltın, O. Y. (2019). Feature selection for classification models via bilevel optimization. *Computers & Operations Research*, 106:156 – 168.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`, University of California, Irvine, School of Information and Computer Sciences.

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

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

Labbé, M., Martínez-Merino, L. I., and Rodríguez-Chía, A. M. (2019). Mixed integer linear programming for feature selection in Support Vector Machine. *Discrete Applied Mathematics*, 261:276 – 304.

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

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

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209(441-458):415–446.

Nguyen, M. H. and de la Torre, F. (2010). Optimal feature selection for support vector machines. *Pattern Recognition*, 43(3):584 – 591.

Onel, M., Kieslich, C. A., and Pistikopoulos, E. N. (2019). A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the tennessee eastman process. *AIChE Journal*, 65(3):992–1005.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, New York.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons.

Wang, T., Huang, H., Tian, S., and Xu, J. (2010). Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels. *Expert Systems with Applications*, 37(9):6663 – 6668.

A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification

Asunción Jiménez-Cordero
asuncionjc@uma.es

Joint work with:
Juan Miguel Morales González
Salvador Pineda Morente

Thank you very much for your attention!

oasys.uma.es

Premio *Ramiro Melendreras*. Jornadas SEIO 2021

June 10th, 2021

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 755705)