A min-max approach to feature selection for nonlinear SVM classification.

Asunción Jiménez-Cordero
asuncionjc@uma.es

Joint work with:
Juan Miguel Morales González
Salvador Pineda Morente

V Congreso de Jóvenes Investigadores de la RSME

January 27th-31st, 2020

What is OASYS?

OASYS
Optimization and Analytics for Sustainable energY Systems

What is OASYS?

Juan Miguel Morales González.
Associate Professor.
juan.morales@uma.es

Salvador Pineda Morente.
Associate Professor.
spinedamorente@gmail.com

OASYS
Optimization and Analytics for Sustainable energY Systems

Adrián Esteban Pérez.
PhD Student.
adrianesteban@uma.es

Asunción Jiménez Cordero.
Postdoc. Researcher.
asuncionjc@uma.es

Miguel Ángel Muñoz Díaz.
PhD Student.
miguelangeljmd@uma.es

Álvaro Porras Cabrera.
PhD Student.
alvaroporras19@gmail.com

Ricardo Fernández-Blanco.
Postdoc. Researcher.
ricardo.fcarramolino@gmail.com

Jesús Huete Cubillo
PhD Student.
j.huetecubillo@gmail.com

What is OASYS?

Juan Miguel Morales González.
Associate Professor.
juan.morales@uma.es

Salvador Pineda Morente.
Associate Professor.
spinedamorente@gmail.com

Adrián Esteban Pérez.
PhD Student.
adrianesteban@uma.es

Asunción Jiménez Cordero.
Postdoc. Researcher.
asuncionjc@uma.es

Miguel Ángel Muñoz Díaz.
PhD Student.
miguelangeljmd@uma.es

Álvaro Porras Cabrera.
PhD Student.
alvaroporras19@gmail.com

Ricardo Fernández-Blanco.
Postdoc. Researcher.
ricardo.fcarramolino@gmail.com

Jesús Huete Cubillo
PhD Student.
j.huetecubillo@gmail.com

OASYS
Optimization and Analytics for Sustainable energY Systems

Combine

- Optimization.
- Data Science.
- Energy.

What is OASYS?

Juan Miguel Morales González.
Associate Professor.
juan.morales@uma.es

Salvador Pineda Morente.
Associate Professor.
spinedamorente@gmail.com

Adrián Esteban Pérez.
PhD Student.
adrianesteban@uma.es

Asunción Jiménez Cordero.
Postdoc. Researcher.
asuncionjc@uma.es

Miguel Ángel Muñoz Díaz.
PhD Student.
miguelangeljmd@uma.es

Álvaro Porras Cabrera.
PhD Student.
alvaroporras19@gmail.com

Ricardo Fernández-Blanco.
Postdoc. Researcher.
ricardo.fcarramolino@gmail.com

Jesús Huete Cubillo
PhD Student.
j.huetecubillo@gmail.com

OASYS
Optimization and Analytics for Sustainable energY Systems

Combine

- Optimization.
- Data Science.
- Energy.

www.oasys.uma.es

What is OASYS?

Juan Miguel Morales González.
Associate Professor.
`juan.morales@uma.es`

Salvador Pineda Morente.
Associate Professor.
`spinedamorente@gmail.com`

Adrián Esteban Pérez.
PhD Student.
`adrianesteban@uma.es`

Asunción Jiménez Cordero.
Postdoc. Researcher.
`asuncionjc@uma.es`

Miguel Ángel Muñoz Díaz.
PhD Student.
`miguelangeljmd@uma.es`

Álvaro Porras Cabrera.
PhD Student.
`alvaroporras19@gmail.com`

Ricardo Fernández-Blanco.
Postdoc. Researcher.
`ricardo.fcarramolino@gmail.com`

Jesús Huete Cubillo
PhD Student.
`j.huetecubillo@gmail.com`

OASYS
Optimization and Analytics for Sustainable energY Systems

Combine

- Optimization.
- Data Science.
- Energy.

`www.oasys.uma.es`

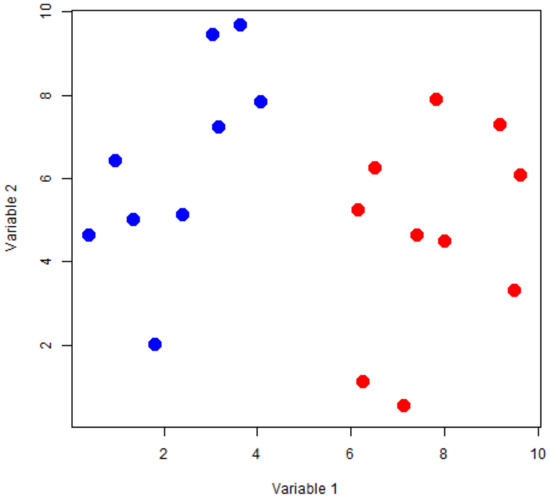
We are always hiring!

`groupoasys@gmail.com`

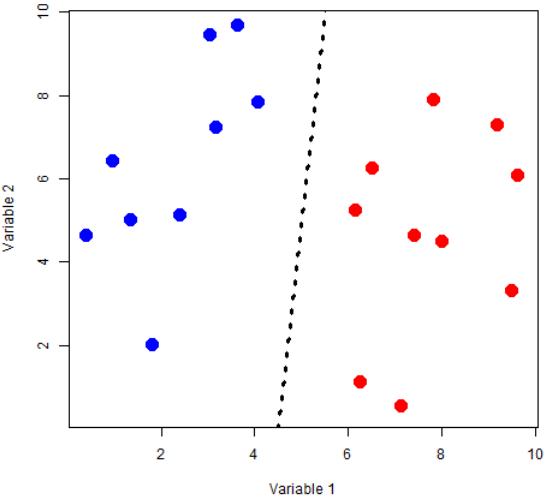
Outline

Outline

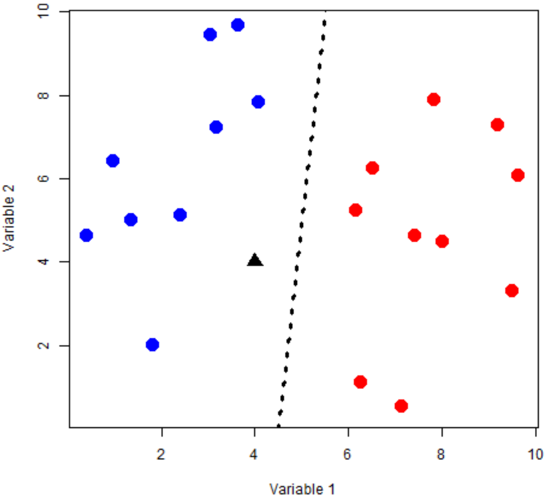
Binary Classification Problem.



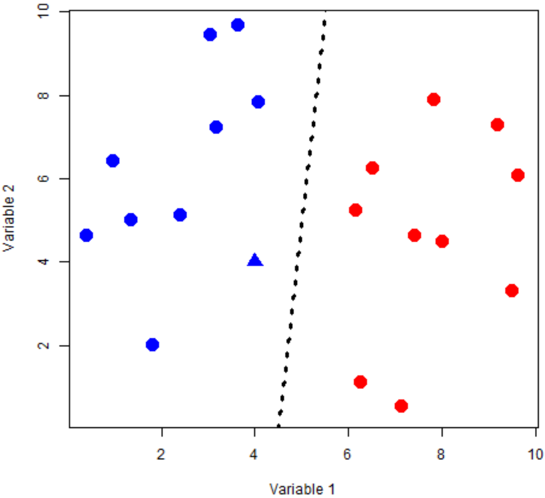
Binary Classification Problem.



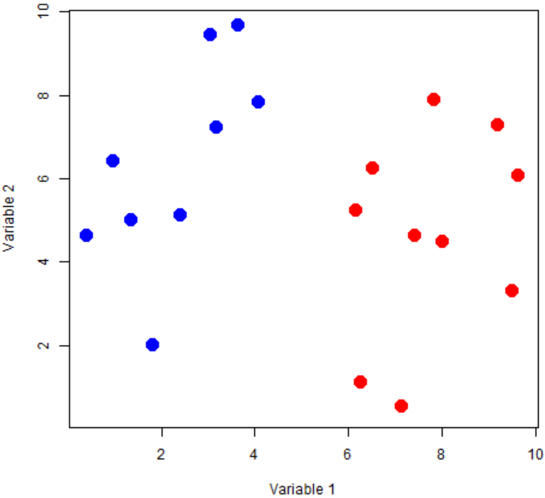
Binary Classification Problem.



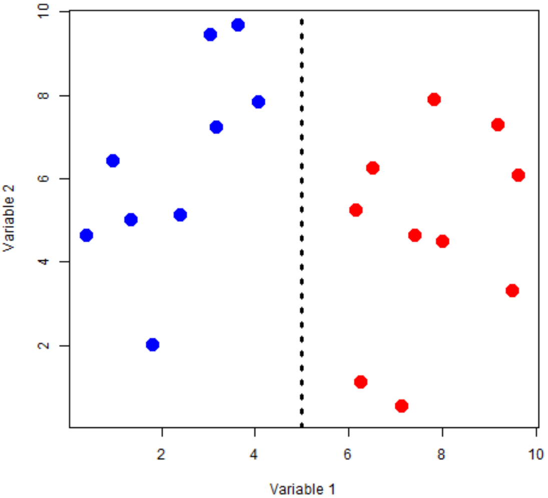
Binary Classification Problem.



Binary Classification Problem. Feature Selection



Binary Classification Problem. Feature Selection



Outline

Aim

- Develop a new mathematical optimization approach to perform feature selection in a binary classification problem.
- Classification tool: nonlinear 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.
        - Complicated ad-hoc approaches.

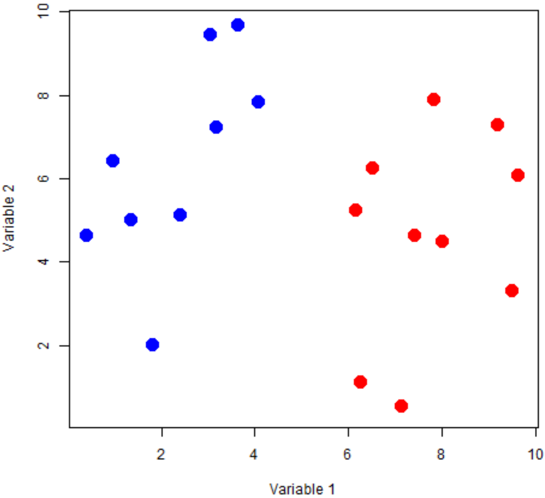
Literature Review

- 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]*):

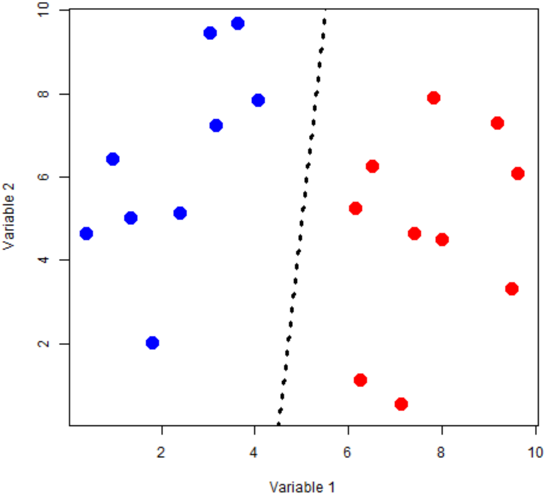
Our contributions

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

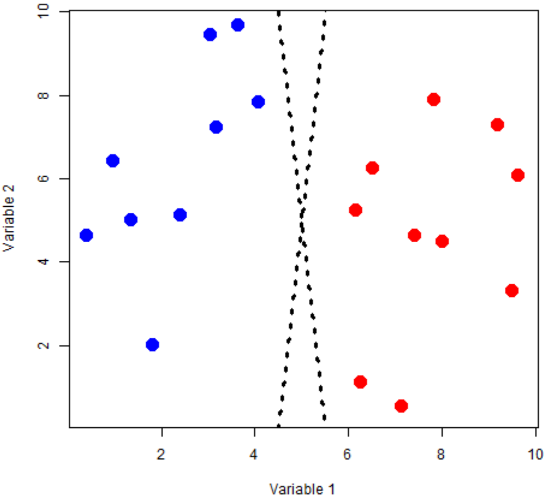
SVM



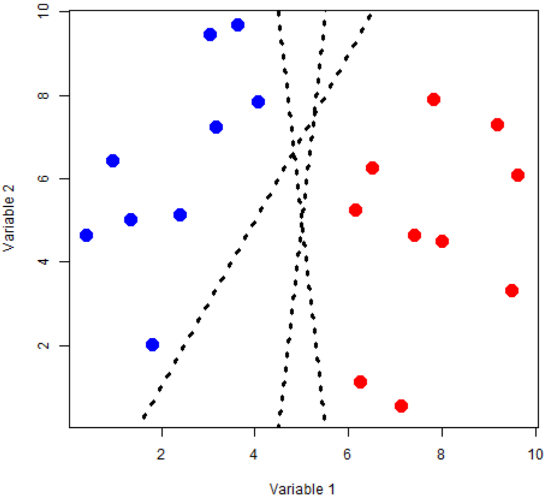
SVM



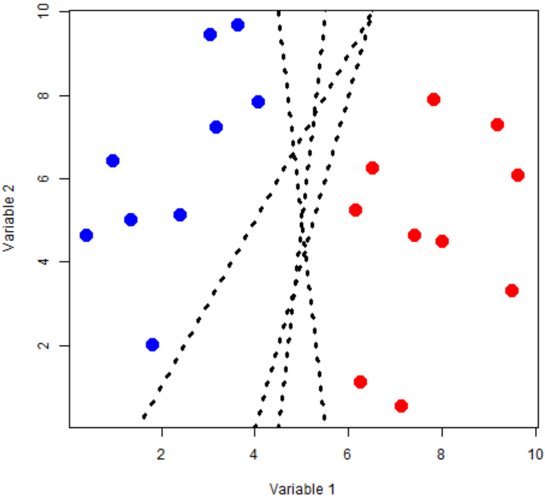
SVM



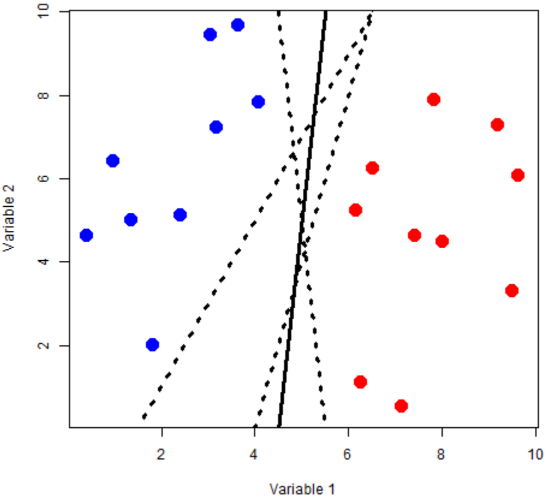
SVM



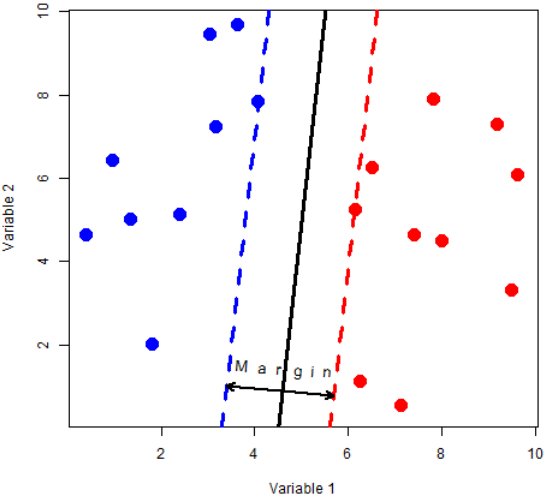
SVM



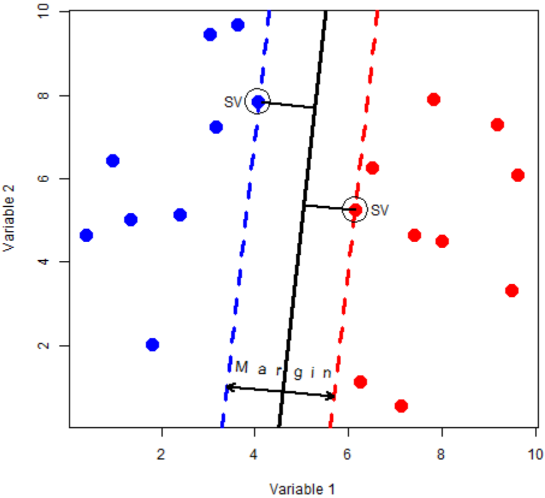
SVM



SVM



SVM



SVM Problem Formulation

Some notation

- Sample $\mathcal{S}$ of individuals.
- Data: $x_i \in \mathbb{R}^M$, $i \in \mathcal{S}$.
- Class label: $y_i \in \{-1, +1\}$, $i \in \mathcal{S}$.
- Hyperplane: $w'x + b = 0$.

SVM Problem Formulation

Some notation

- Sample $\mathcal{S}$ of individuals.
- Data: $x_i \in \mathbb{R}^M$, $i \in \mathcal{S}$.
- Class label: $y_i \in \{-1, +1\}$, $i \in \mathcal{S}$.
- Hyperplane: $w'x + b = 0$.

Optimization Problem (Primal)

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

SVM Problem Formulation

Some notation

- Sample $\mathcal{S}$ of individuals.
- Data: $x_i \in \mathbb{R}^M$, $i \in \mathcal{S}$.
- Class label: $y_i \in \{-1, +1\}$, $i \in \mathcal{S}$.
- Hyperplane: $w'x + b = 0$.

Optimization Problem (Primal)

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

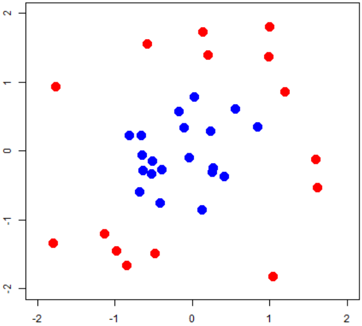
When a new **unseen** individual comes...

Score:

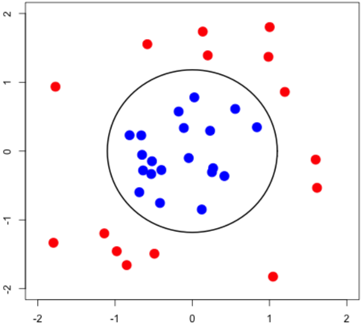
$$\hat{y}(x) = w'x + b$$

x is assigned to class $+1$ iff $\hat{y}(x) > 0$.

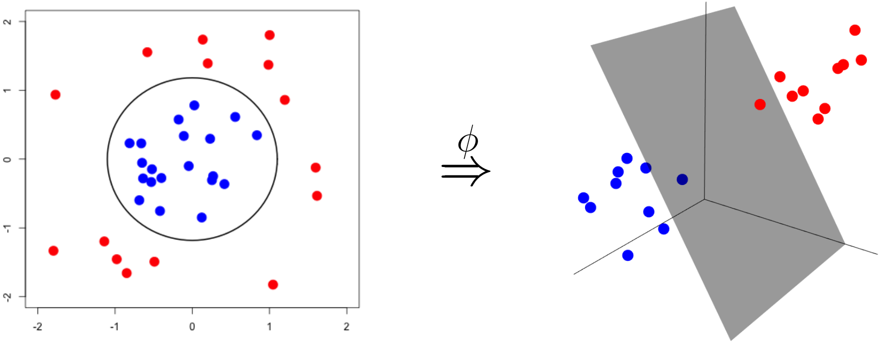
What happens in the nonlinear case?



What happens in the nonlinear case?



What happens in the nonlinear case?



SVM Problem (Primal).

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

SVM Problem (Primal).

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

SVM Problem (Primal). Feature Selection.

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

SVM Problem (Primal). Feature Selection.

$$
\begin{cases}
\min\limits_{\boldsymbol{\gamma}, w, b, \xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i \in \mathcal{S}} \xi_i \\
\text{s.t.} & (w' \phi_{\boldsymbol{\gamma}}(x_i) + b) y_i \geq 1 - \xi_i, \quad i \in \mathcal{S} \\
& \xi_i \geq 0, \quad i \in \mathcal{S} \\
& \boldsymbol{\gamma_j \geq 0, \quad \forall j}
\end{cases}
$$

But...

Thanks to the Mercer's theorem, *Mercer [1909]*

SVM Problem (Primal). Feature Selection.

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

Unfortunately

ϕ function is usually unknown.

SVM Problem (Primal). Feature Selection.

$$
\begin{cases}
\min\limits_{\gamma} \quad \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C \sum\limits_{i \in \mathcal{S}} \xi_i \\
\text{s.t.} & (w'\phi_\gamma(x_i) + b) y_i \geq 1 - \xi_i, \quad i \in \mathcal{S} \\
& \xi_i \geq 0, \quad i \in \mathcal{S} \\
& \gamma_j \geq 0, \quad \forall j
\end{cases}
$$

Unfortunately

ϕ function is usually unknown.

SVM Problem (Dual). Feature Selection.

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

SVM Problem (Dual). Feature Selection.

$$\left\{\begin{array}{lll} \min\limits_{\gamma} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i \in \mathcal{S}} \boldsymbol{\alpha}_i - \frac{1}{2} \sum\limits_{i,\ell \in \mathcal{S}} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_\ell \boldsymbol{y}_i \boldsymbol{y}_\ell \boldsymbol{\phi}(\boldsymbol{x}_i)' \boldsymbol{\phi}(\boldsymbol{x}_\ell) \\ \text{s.t.} & \sum\limits_{i \in \mathcal{S}} \boldsymbol{\alpha}_i \boldsymbol{y}_i = \mathbf{0} \\ & \boldsymbol{\alpha}_i \in [\mathbf{0}, \boldsymbol{C}], \quad \boldsymbol{i} \in \mathcal{S} \\ & \gamma_j \geq 0, \quad \forall j \end{array}\right.$$

Kernel trick

$$K(x_i, x_\ell) = \phi(x_i)' \phi(x_\ell)$$

SVM Problem (Dual). Feature Selection.

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

Kernel trick

$$K(x_i, x_\ell) = \phi(x_i)' \phi(x_\ell)$$

SVM Problem (Dual). Feature Selection.

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

Kernel trick

$$K(x_i, x_\ell) = \phi(x_i)' \phi(x_\ell)$$

Score

$$\hat{y}(x) = \sum_{i \in \mathcal{S}} \alpha_i y_i K(x_i, x)$$

SVM Problem (Dual). Feature Selection.

Anisotropic Gaussian kernel

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

SVM Problem (Dual). Feature Selection.

Anisotropic Gaussian kernel

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

γ_j: importance feature j

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

SVM Problem (Dual). Feature Selection.

Anisotropic Gaussian kernel

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

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 (Dual). Feature Selection.

Anisotropic Gaussian kernel

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

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

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

Problem Formulation

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

Problem Formulation

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

Outline

Min-max optimization problem.

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

Epigraph form.

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

Strong duality lower level problem (SVM).

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

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).

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

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

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

Strong duality lower-level problem (SVM).

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

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

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

Single-level optimization problem

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

Recap

$$\min_{\gamma \geq 0} \left[C_2 \|\gamma\|_p^p + (1 - C_2) \max_{\alpha} \sum_{i \in \mathcal{S}} \alpha_i - \right.$$

$$\left. \frac{1}{2} \sum_{i,\ell \in \mathcal{S}} \alpha_i \alpha_\ell y_i y_\ell K_\gamma(x_i, x_\ell) \right]$$

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

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

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

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

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

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

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).

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).

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

- Simple solving strategy. No ad-hoc approaches.

How to solve the problem?

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

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

- Simple solving strategy. No ad-hoc approaches.
- Grid search (C, C_2) + local solver + k-fold CV (train, validation, test).

Outline

Experimental Setup

- UCI Machine Learning Repository, *Dheeru and Karra Taniskidou [2017]*.
- `diabetes` (768 elements, 8 features).
- `breast` (569 elements, 30 features).
- $p = 2$.
- $C \in \{10^{-4}, \ldots, 10^{4}\}$.
- $C_2 \in \{0, 0.1, \ldots, 0.9, 1\}$.
- `Ipopt`.

Numerical Results

diabetes

	Our approach	Literature benchmark (ad-hoc)	Literature benchmark (off-the-shelf)	No feature selection
% accuracy test	77.59 ± 2.29	76.74 ± 1.9	65.11 ± 0.36	77.61 ± 3.32
# selected features ($< 10^{-5}$)	6	5	0	8

Numerical Results

diabetes

	Our approach	Literature benchmark (ad-hoc)	Literature benchmark (off-the-shelf)	No feature selection
% accuracy test	77.59 ± 2.29	76.74 ± 1.9	65.11± 0.36	77.61 ± 3.32
# selected features ($< 10^{-5}$)	6	5	0	8

breast

	Our approach	Literature benchmark (ad-hoc)	Literature benchmark (off-the-shelf)	No feature selection
% accuracy test	96.83 ±2.29	97.55±0.9	65.23± 6.19	96.85± 2.41
# selected features ($< 10^{-5}$)	22	15	0	30

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 with existing 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 with existing literature results.

Future Research

- Reduce even more the number of features.
- Bigger datasets.
- Extension to regression or clustering.
- Application to Power Systems.

References

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

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.

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.

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. (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.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2001). Feature selection for SVMs. In *Advances in neural information processing systems*, pages 668–674.

A min-max approach to feature selection for nonlinear SVM classification.

Asunción Jiménez-Cordero
asuncionjc@uma.es

Joint work with:
Juan Miguel Morales González
Salvador Pineda Morente

Thank you very much for your attention!

V Congreso de Jóvenes Investigadores de la RSME

January 27th-31st, 2020

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)