Interpretable Learning in Power Systems Operations

Asunción Jiménez-Cordero
asuncionjc@uma.es

Joint work with:
Juan Miguel Morales González
Salvador Pineda Morente

UNIVERSIDAD DE MÁLAGA

OASYS Optimization and Analytics for Sustainable energY Systems

erc European Research Council

oasys.uma.es - groupoasys@gmail.com

INFORMS Annual Meeting 2020

November 7th-13th, 2020

Outline

Outline

Motivation of the problem

Unit Commitment

$$\left\{\begin{array}{ll} \min\limits_{\Xi} & \text{total cost} \\ \text{s.t.} & \text{balance constraints} \\ & \text{flow and power generations bounds} \\ & \text{binary variables on-off status generators} \end{array}\right.$$

- One of the most important problems in Power Systems.
- MILP problem. NP-hard. Intractable for large systems.
- Line congestion identification is on trend.
- Interpretability and relevance of nodes is desired.

Main Goal

Aim

Develop a **Mathematical Optimization** approach which **selects** the most important **nodes** in the **congestion line identification** problem by taking advantage of both the **physical and data** information.

Line congestion identification→ Binary classification problem.

- From input data (demand, d_i).
- Find classification rule for congested $(+1)$ and non congested lines (-1).

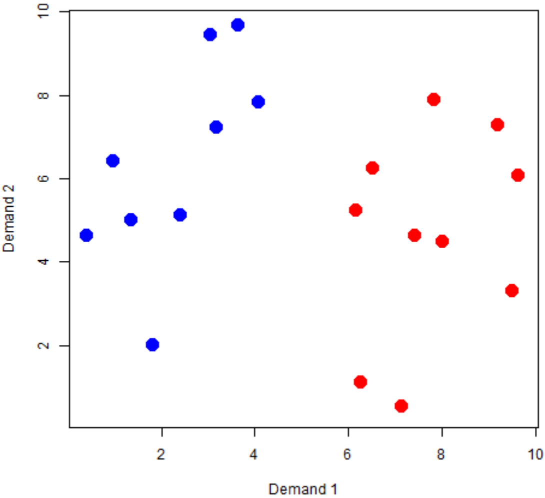
Line congestion identification→ Binary classification problem.

- From input data (demand, d_i).
- Find classification rule for congested (+1) and non congested lines (-1).

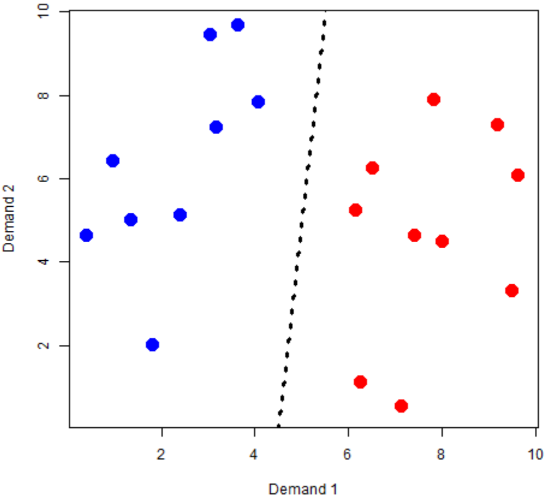
Support Vector Machines (SVM)

Outline

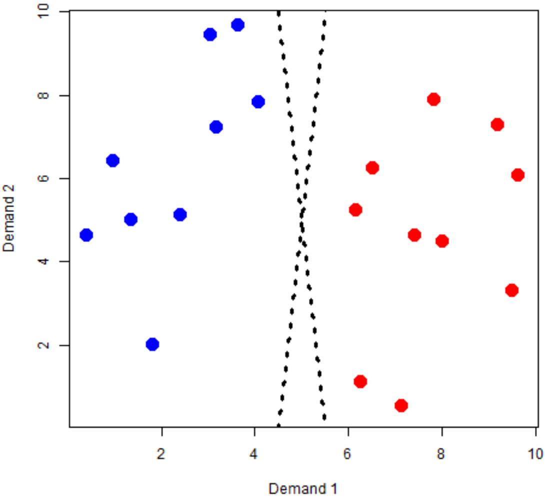
SVM



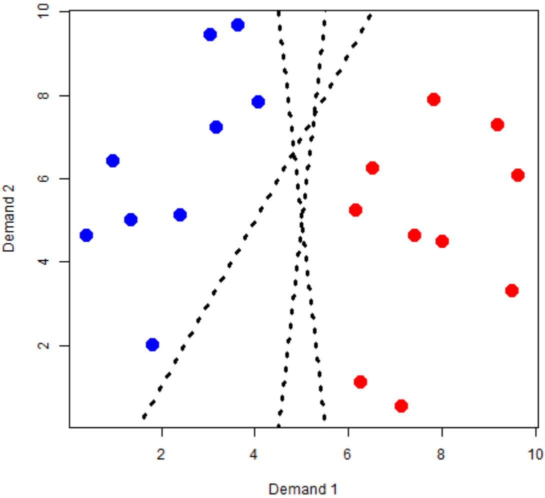
SVM



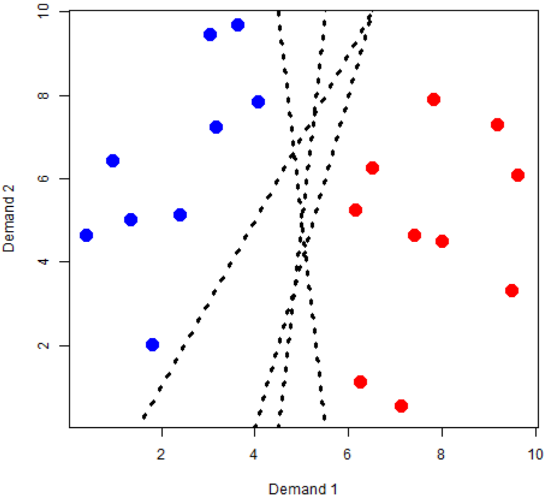
SVM



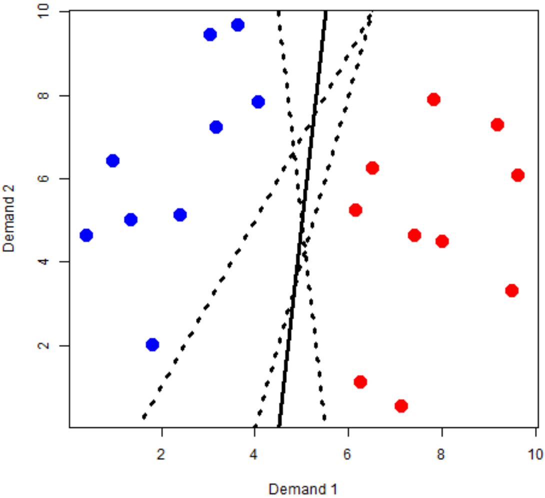
SVM



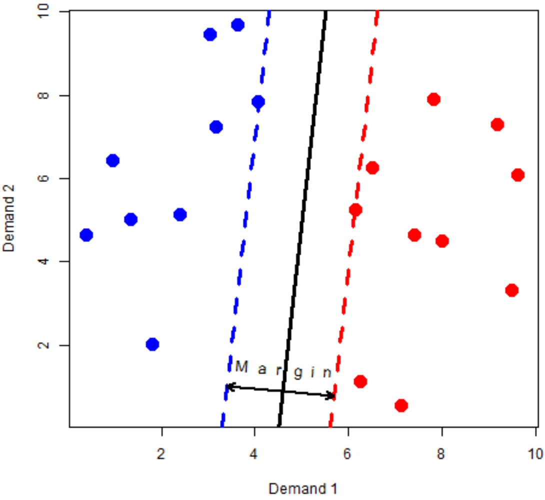
SVM



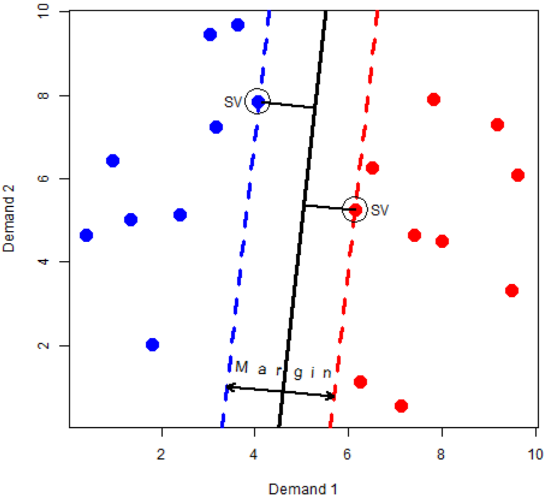
SVM



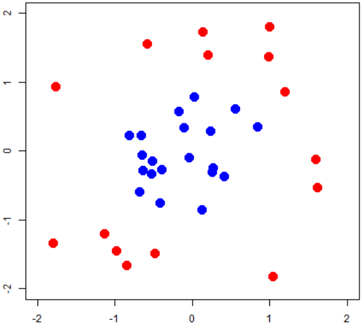
SVM



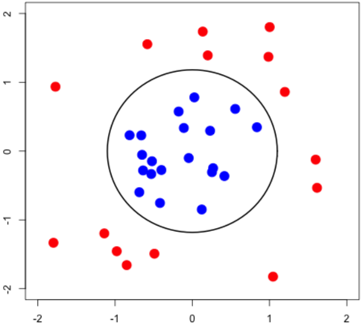
SVM



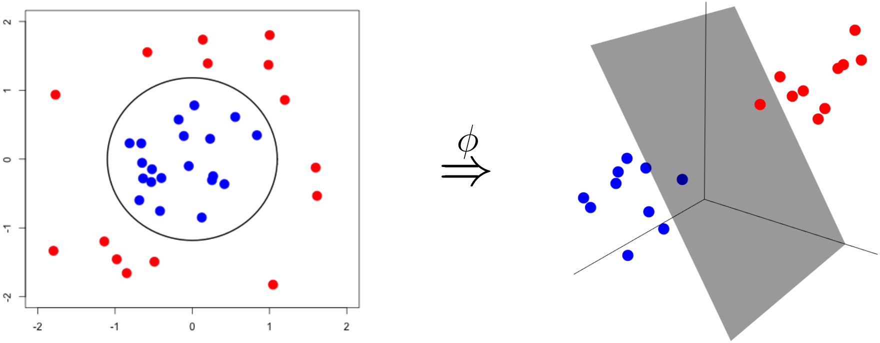
What happens in the nonlinear case?



What happens in the nonlinear case?



What happens in the nonlinear case?



Primal formulation

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i\in\mathcal{S}}\xi_i & \text{[margin + penalization]} \\ \text{s.t.} & (w' \quad d_i \quad + b)y_i \geq 1-\xi_i, \quad i\in\mathcal{S} & \text{[data - correct label]} \\ & \xi_i \geq 0, \quad i\in\mathcal{S} & \\ & & \end{cases}$$

Primal formulation

$$
\begin{cases}
\min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i\in\mathcal{S}} \xi_i & & \text{[margin + penalization]} \\
\text{s.t.} & (w'\phi(d_i) + b)y_i \geq 1 - \xi_i, & i \in \mathcal{S} & \text{[data - correct label]} \\
 & \xi_i \geq 0, \quad i \in \mathcal{S} & &
\end{cases}
$$

Primal formulation

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i\in\mathcal{S}} \xi_i & \text{[margin + penalization]} \\ \text{s.t.} & (w'\phi(d_i)+b)y_i \geq 1-\xi_i, \quad i\in\mathcal{S} & \text{[data - correct label]} \\ & \xi_i \geq 0, \quad i\in\mathcal{S} & \end{cases}$$

Dual Formulation

$$\begin{cases} \max\limits_{\alpha} & \sum\limits_{i\in\mathcal{S}} \alpha_i - \frac{1}{2}\sum\limits_{i,\ell\in\mathcal{S}} \alpha_i\alpha_\ell y_i y_\ell \boldsymbol{\phi(d_i)'\phi(d_\ell)} \\ \text{s.t.} & \sum\limits_{i\in\mathcal{S}} \alpha_i y_i = 0 \\ & \alpha_i \in [0,C], \quad i\in\mathcal{S} \end{cases}$$

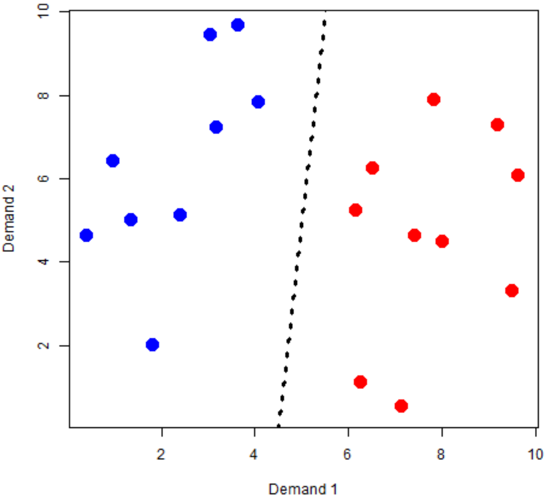
Primal formulation

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i\in\mathcal{S}} \xi_i & \text{[margin + penalization]} \\ \text{s.t.} & (w'\phi(d_i)+b)y_i \geq 1-\xi_i, \quad i\in\mathcal{S} & \text{[data - correct label]} \\ & \xi_i \geq 0, \quad i\in\mathcal{S} & \end{cases}$$

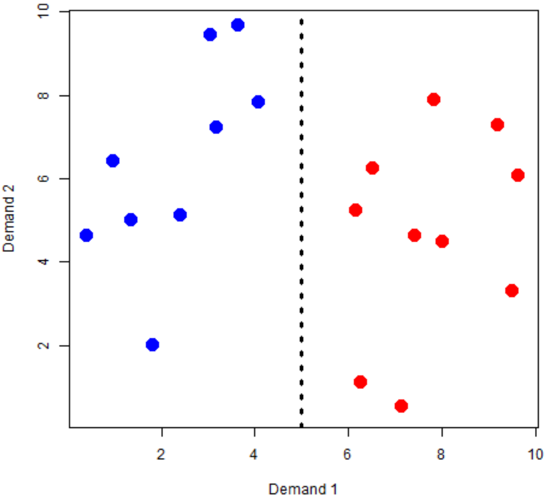
Dual Formulation

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

Feature Selection



Feature Selection



Outline

SVM Problem (Dual). Feature Selection.

Gaussian kernel

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

SVM Problem (Dual). Feature Selection.

Gaussian kernel

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

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.

Gaussian kernel

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

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.

Gaussian kernel

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

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

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

Problem Formulation

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

Model complexity

Problem Formulation

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

Classification accuracy

Problem Formulation

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

Trade-off

Benefits of the Proposed Approach

Our contributions

- Method which simultaneously selects features and classify.
- # selected features is not fixed, but provided by our approach.
- No ad-hoc strategies. Strong duality reformulation. Off-the-shelf solvers. C and C_2 tuned with cross-validation.

Further info

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

Abstract

In recent years, feature selection has become a challenging problem in several machine learning fields, particularly in classification problems. Support Vector Machine (SVM) is a well-known technique applied in (nonlinear) classification. 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.

Keywords: Machine learning, min-max optimization, duality theory, feature selection, nonlinear Support Vector Machine classification

*Corresponding author

Email addresses: asuncionjc@uma.es (Asunción Jiménez-Cordero), juanmi82mg@gmail.com (Juan Miguel Morales), spinedamorente@gmail.com (Salvador Pineda)

Under review (major changes)

Available at ResearchGate (https://www.researchgate.net/publication/340826631_A_novel_embedded_min-max_approach_for_feature_selection_in_nonlinear_Support_Vector_Machine_classification)

So far...

- Develop a pure data-driven method.

So far...

- Develop a pure data-driven method.
- How to incorporate physical information to our proposal?

Physically-and-data-aware approach

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

Physically-and-data-aware approach

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

- Non convex optimization problem.
- Just local optimal solution are guaranteed.
- Initial solutions play a key role.

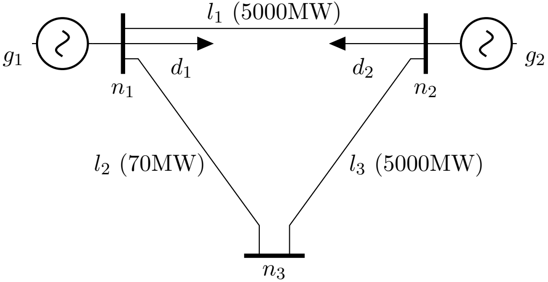
Physically-and-data-aware approach

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

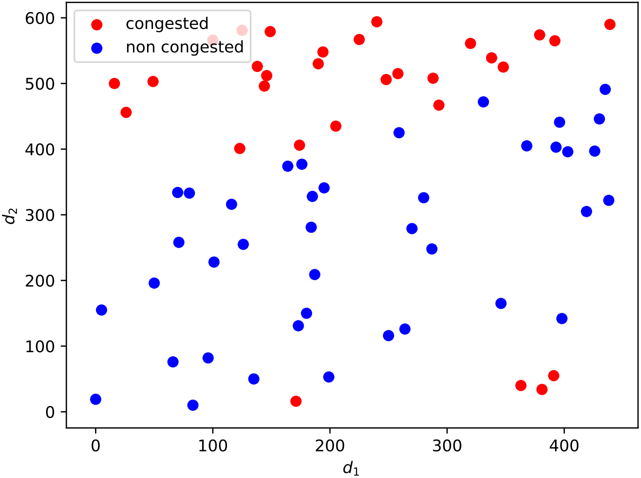
- Non convex optimization problem.
- Just local optimal solution are guaranteed.
- Initial solutions play a key role.
- **Our proposal:** physically-and-data-aware initial solutions.

Outline

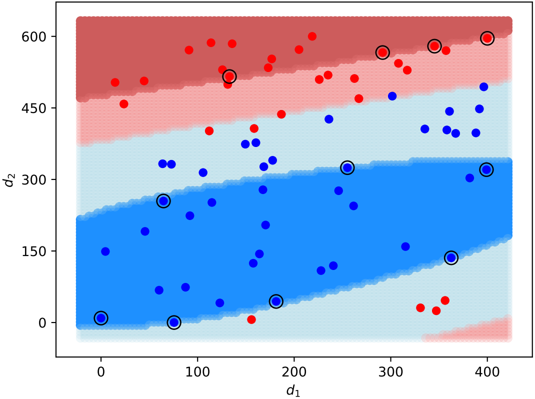
Illustrative example



Line congestion l_2

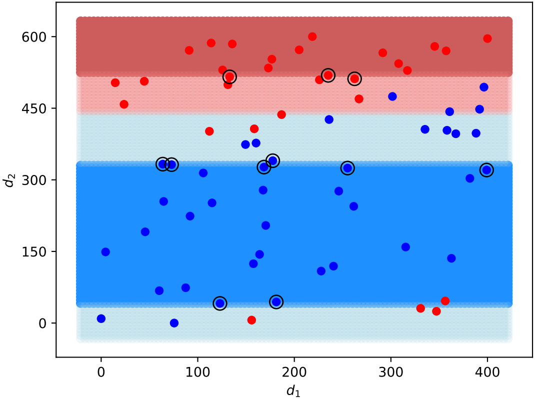


$\gamma = [0.1, 0.1]$



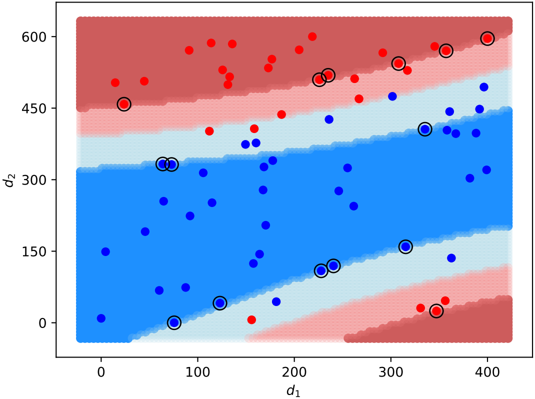
- Pure data-aware initial solution.
- No features selected.

$\gamma = [0, 0.17]$ (ptdf)



- Pure physically-aware initial solution.
- Dependent on slack node (n_1).
- Data are ignored. Possible errors selected features.

$\gamma = [0.11, 0.26]$



- Physically-and-data-aware initial solution.
- Regression: demand - flow. Linear relationship ptdf.
- Interpretability.

Outline

Conclusions and Future Research

Conclusions

- Design a Mathematical Optimization approach which leverages data and physical information to learn line congestion in an UC problem.
- A min-max optimization problem which simultaneously classify data and select features.
- Physical and data information is introduced via initial solutions.
- Illustrative example shows promising results.

Conclusions and Future Research

Conclusions

- Design a Mathematical Optimization approach which leverages data and physical information to learn line congestion in an UC problem.
- A min-max optimization problem which simultaneously classify data and select features.
- Physical and data information is introduced via initial solutions.
- Illustrative example shows promising results.

Future Research

- Bigger case study.
- How affects relationship among features, e.g. multicollinearity.

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.

Interpretable Learning in Power Systems Operations

Asunción Jiménez-Cordero
asuncionjc@uma.es

Joint work with:
Juan Miguel Morales González
Salvador Pineda Morente

Thank you very much for for your attention!

oasys.uma.es - groupoasys@gmail.com

INFORMS Annual Meeting 2020

November 7th-13th, 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)