Série dos Seminários de Acompanhamento à Pesquisa

DEI
DEPARTAMENTO DE ENGENHARIA INDUSTRIAL

Número 25 | 07 2021

Predictive PolieDRO:

A novel classification and regression framework with non-parametric regularization

Autor:

Tomás Gutierrez

PUC RIO

Série dos Seminários de Acompanhamento à Pesquisa

Número 25 | 07 2021

Predictive PolieDRO:

A novel classification and regression framework with non-parametric regularization

Autor:

Tomás Gutierrez

Orientador: Davi Michel Valladão
Coorientador: Bernardo Pagnoncelli (UAI – Chile)

CRÉDITOS:

SISTEMA MAXWELL / LAMBDA
https://www.maxwell.vrac.puc-rio.br/

Organizadores: Fernanda Baião / Soraida Aguilar

Layout da Capa: Aline Magalhães dos Santos

Histórico

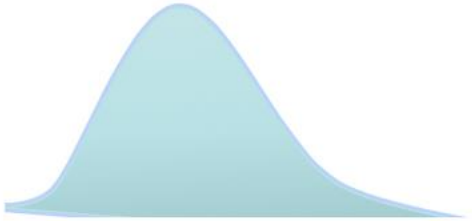
- Graduação em Engenharia de Produção (2013 - 16) – PUC Rio
  - Bolsista de Iniciação Científica no DEI
- Mestrado em Engenharia de Produção (2017 - 18) – PUC Rio
  - Área de concentração: Finanças e Análise de Investimentos
- Doutorando em Engenharia de Produção (2019 -) – PUC Rio
  - Área de concentração: Pesquisa Operacional

Sumário

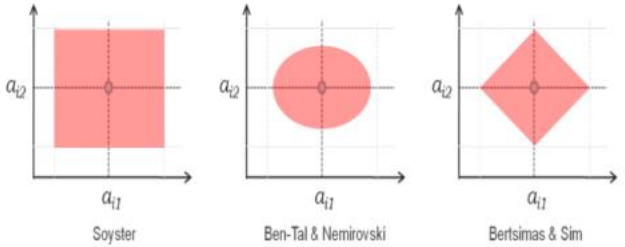
- Contexto
- Motivação
- Metodologia proposta
- Resultados
- Extensões
- Referências

Contexto

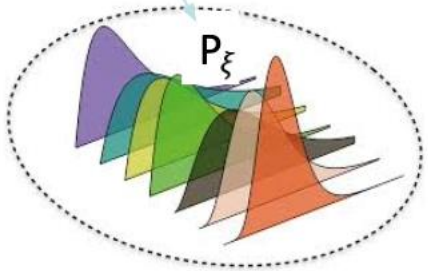
- Otimização
  - Estocástica $\max_x E[V(x,\xi)]$



  - Robusta $\max_x \{\min_{\xi\in\Xi} V(x,\xi)\}$



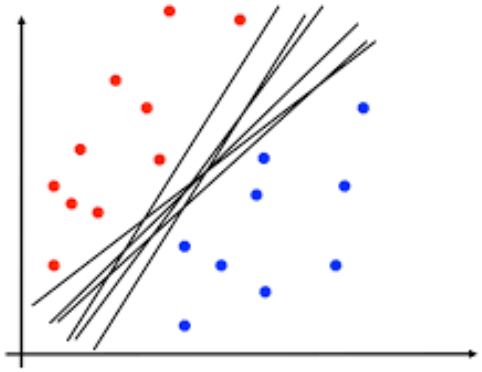
  - *Distributionally Robust* $\max_x \{\min_{\mathbb{P}\in\mathbf{P}_\xi} E_{\mathbb{P}}[V(x,\xi)]\}$



Contexto

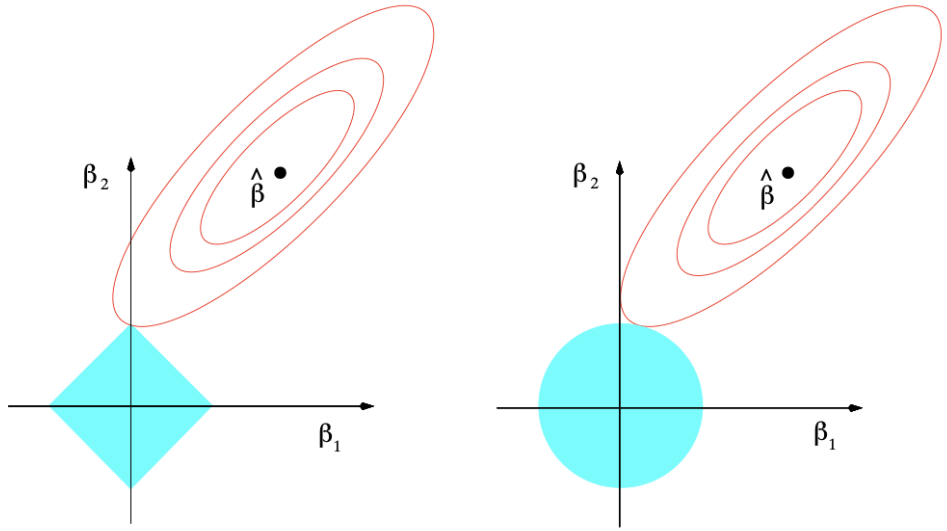
- *Machine & Statistical Learning*
  - *Classificação:*
    - *Support Vector Machine* (SVM)
  - Regressão:
    - LASSO, Ridge

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to} \quad \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i.$$



$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}}\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}.$$

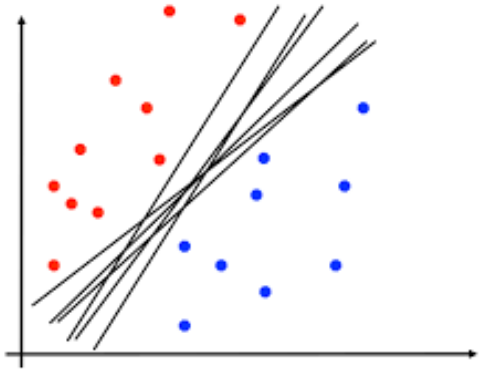


Contexto

- *Machine & Statistical Learning*
  - Otimização
  - **Hiper parâmetros**

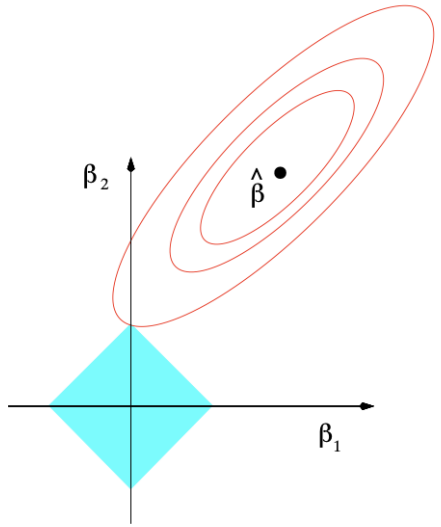
$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

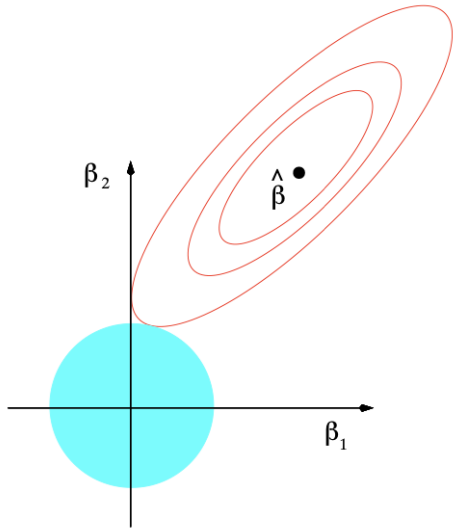
$$\text{subject to} \quad \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i$$



$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}}\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}.$$



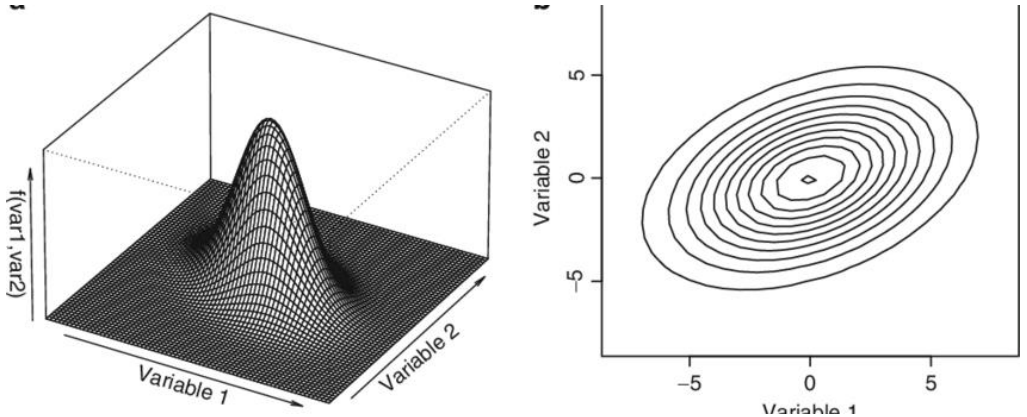


Motivação

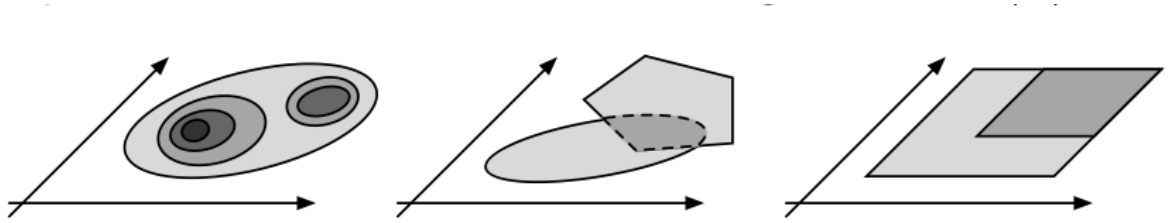
- *Distributionally Robust Convex Optimization – Wiesemann, Kuhn, Sim (2018)*
  - Como representar o conjunto de ambiguidades ?
  - Formulação tratável?

$$\max_{x}\{\min_{\mathbb{P}\in\mathbf{P}_\xi} E_{\mathbb{P}}[V(x,\xi)]\}$$

$$\mathcal{P}=\left\{\mathbb{P}\in\mathcal{P}_0(\mathbb{R}^P\times\mathbb{R}^Q):\ \begin{array}{r}\mathbb{E}_{\mathbb{P}}[\mathbf{A}\tilde{\mathbf{z}}+\mathbf{B}\tilde{\mathbf{u}}]=\mathbf{b},\\ \mathbb{P}[(\tilde{\mathbf{z}},\tilde{\mathbf{u}})\in\mathcal{C}_i]\in[\underline{\mathbf{p}}_i,\bar{\mathbf{p}}_i]\\ \forall i\in\mathcal{I}\end{array}\right\}$$



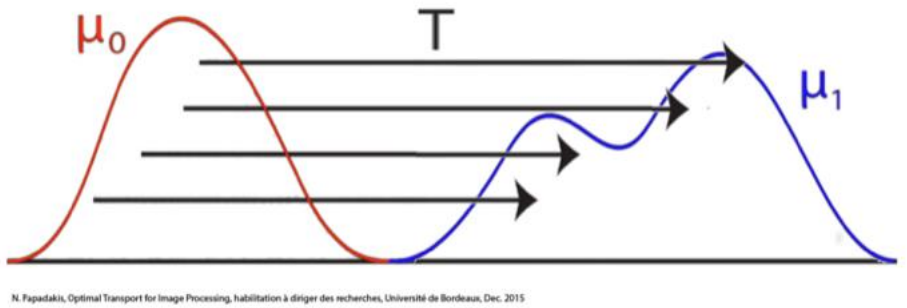
$$\mathcal{C}_i=\{(\mathbf{z},\mathbf{u})\in\mathbb{R}^P\times\mathbb{R}^Q:\ \mathbf{C}_i\mathbf{z}+\mathbf{D}_i\mathbf{u}\preccurlyeq_{\mathcal{K}_i}\mathbf{c}_i\}$$



Motivação

- *Robust Wasserstein Profile Inference and Applications to Machine Learning – Blanchet, Kang, Murthy (2019)*

  - Machine Learning como DRO!
  - Distância de Wassertein
  - Reinterpretação dos hiper parâmetros

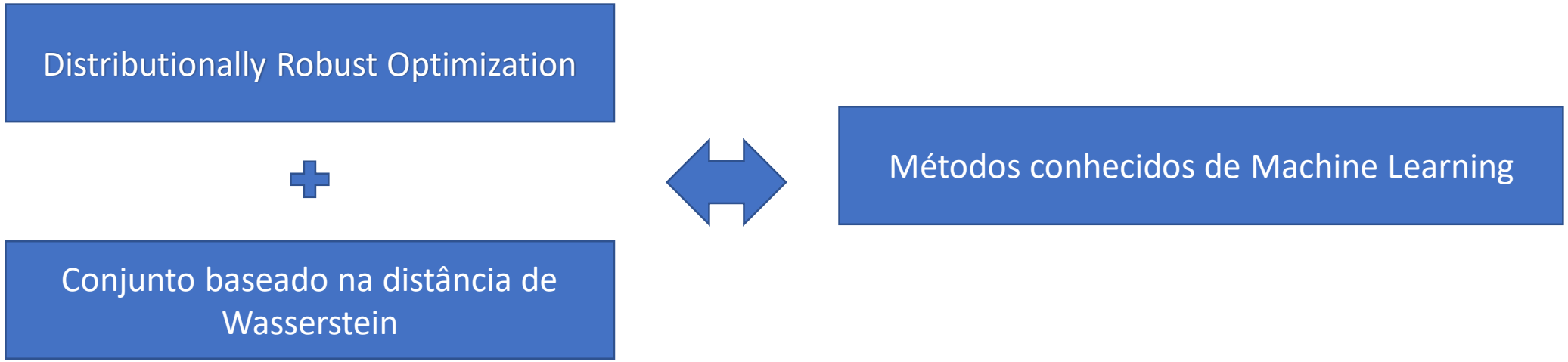


$$l(x, y; \beta) = (y - \beta^T x)^2$$

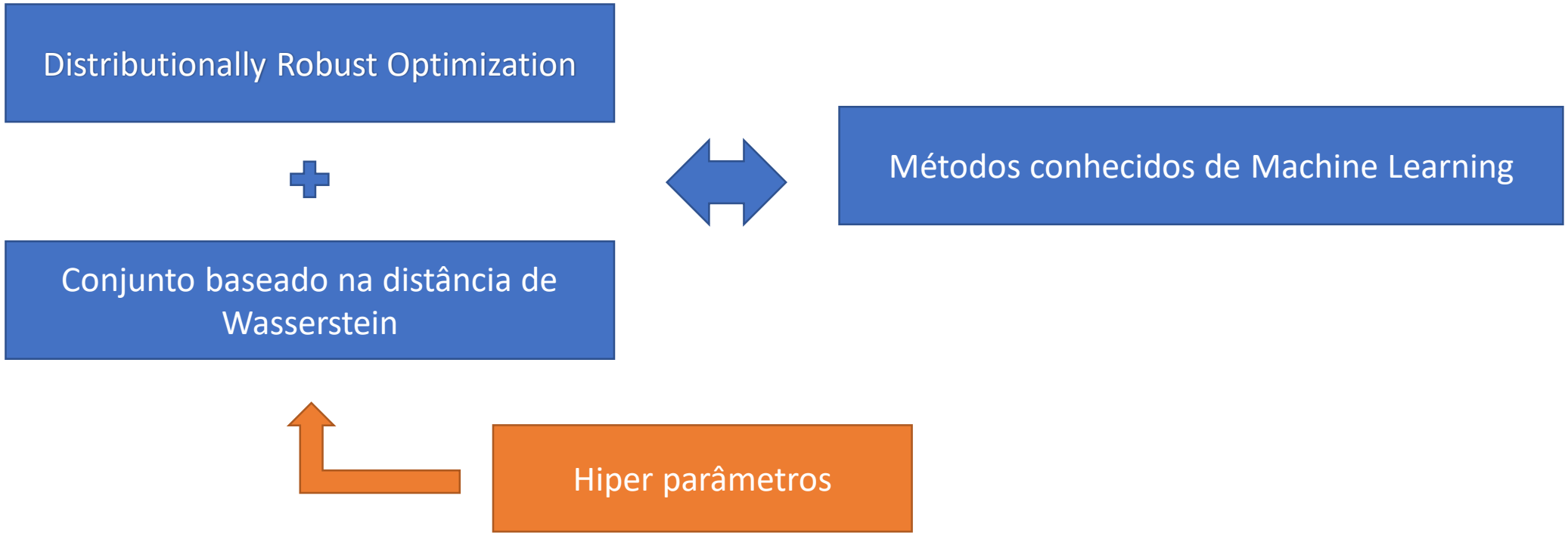
$$\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{E_{P_n}\left[l\left(X, Y; \beta\right)\right]} + \lambda \left\|\beta\right\|_p \right\}^2 = \min_{\beta \in \mathbb{R}^d} \max_{P:\ \mathcal{D}_c(P, P_n) \leq \delta} E_P\left[l(X, Y; \beta)\right] \qquad \lambda = \sqrt{\delta}$$

$$\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \beta^T X_i)^+ + \ \delta \left\|\beta\right\|_p \ = \ \inf_{\beta \in \mathbb{R}^d} \sup_{P:\ D_c(P, P_n) \leq \delta} E_P\left[(1 - Y \beta^T X)^+\right]$$

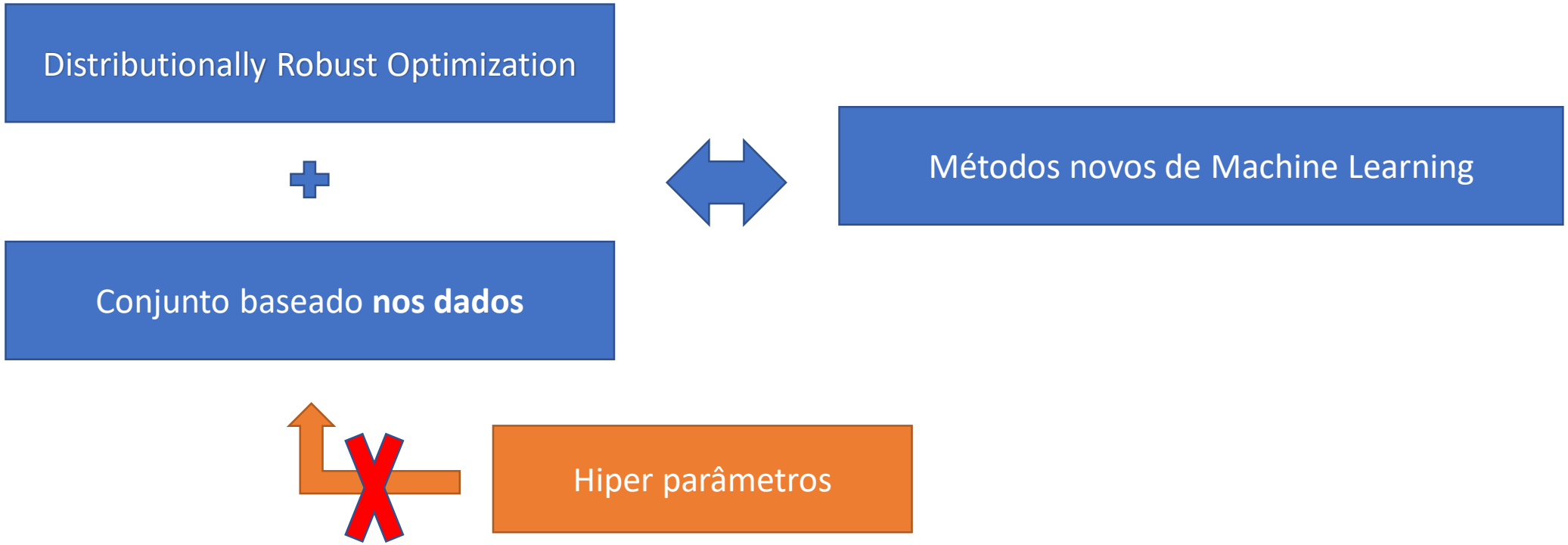
Motivação


Distributionally Robust Optimization
+
Conjunto baseado na distância de Wasserstein
Métodos conhecidos de Machine Learning

Motivação

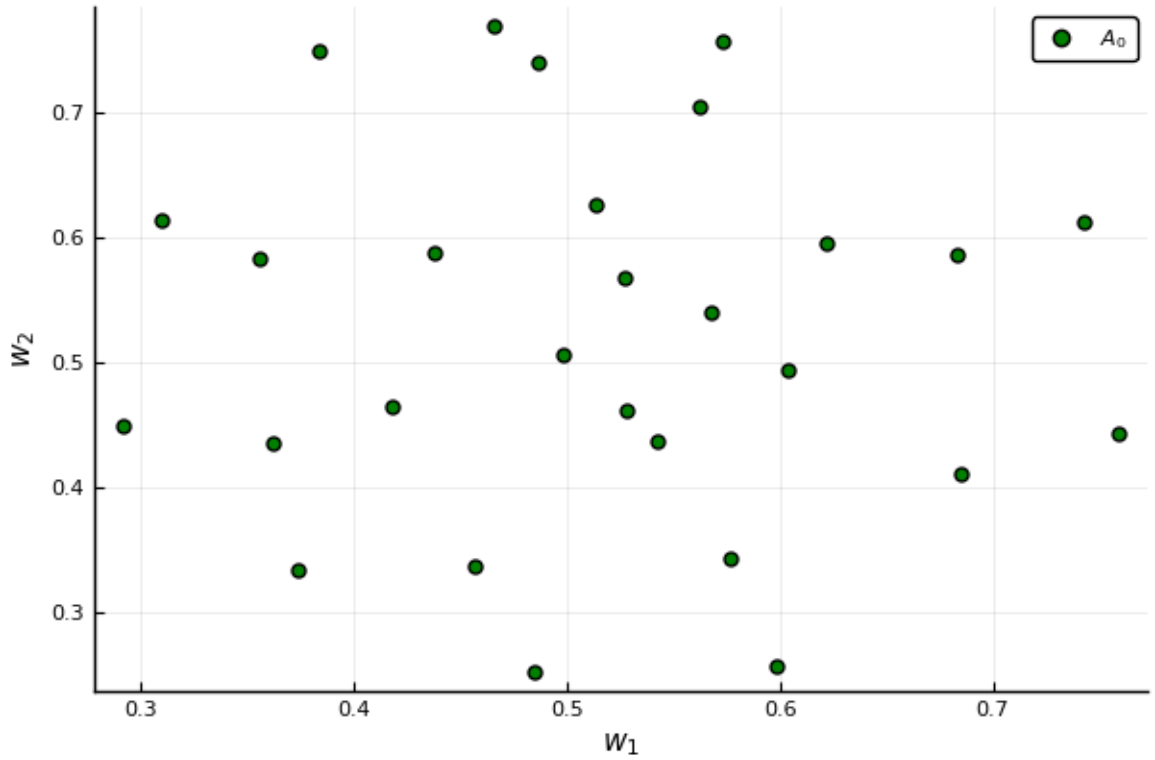

Distributionally Robust Optimization
+
Conjunto baseado na distância de Wasserstein
Métodos conhecidos de Machine Learning
Hiper parâmetros

Motivação


Distributionally Robust Optimization
+
Conjunto baseado **nos dados**
Métodos novos de Machine Learning
Hiper parâmetros

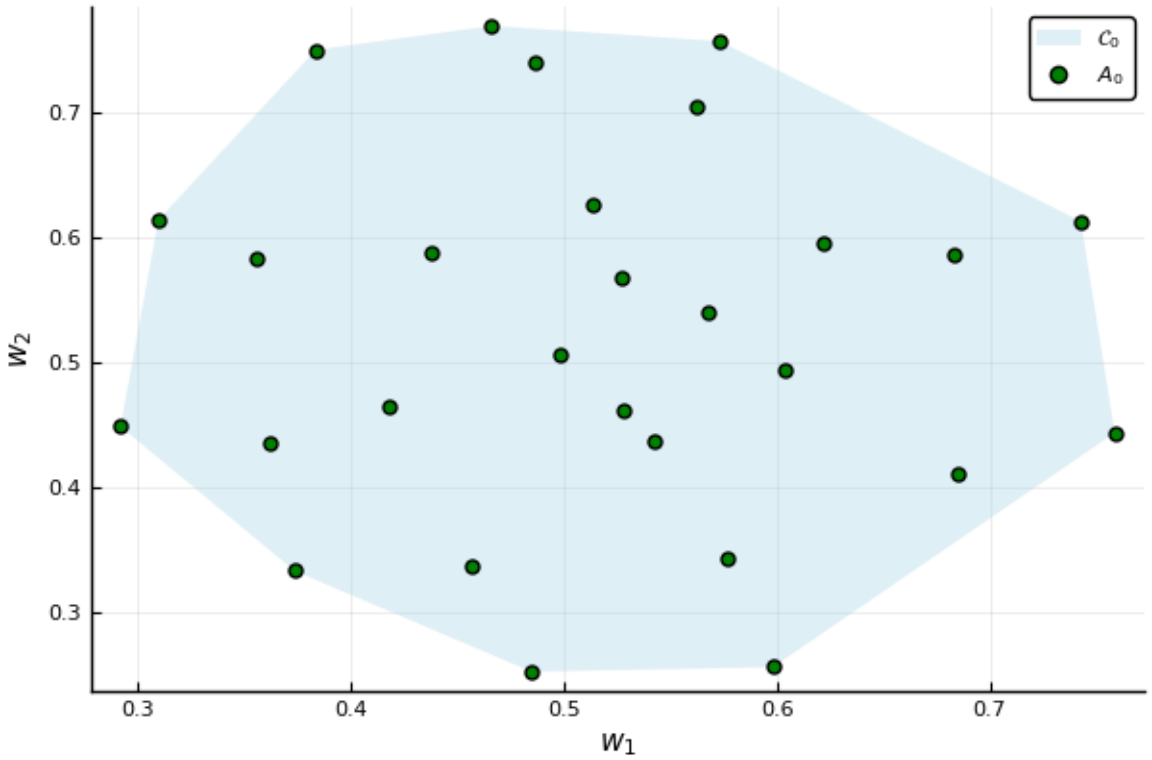
Metodologia proposta

- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros



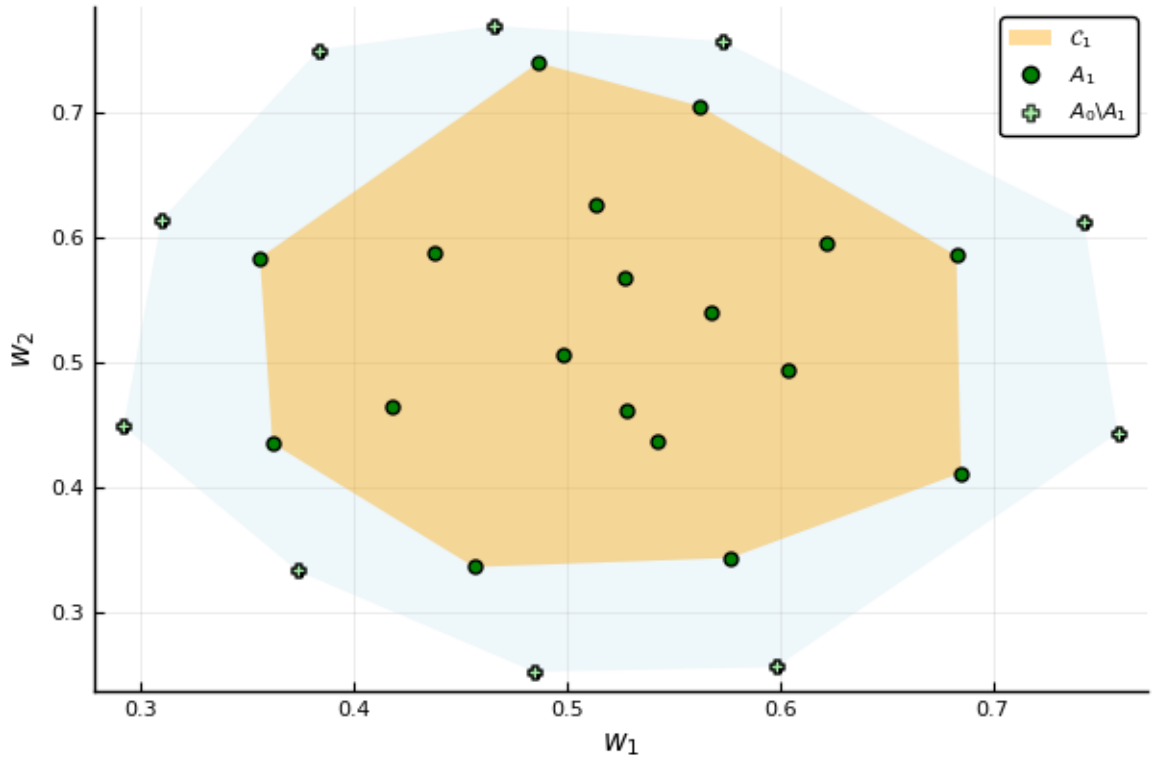
Metodologia proposta

- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros



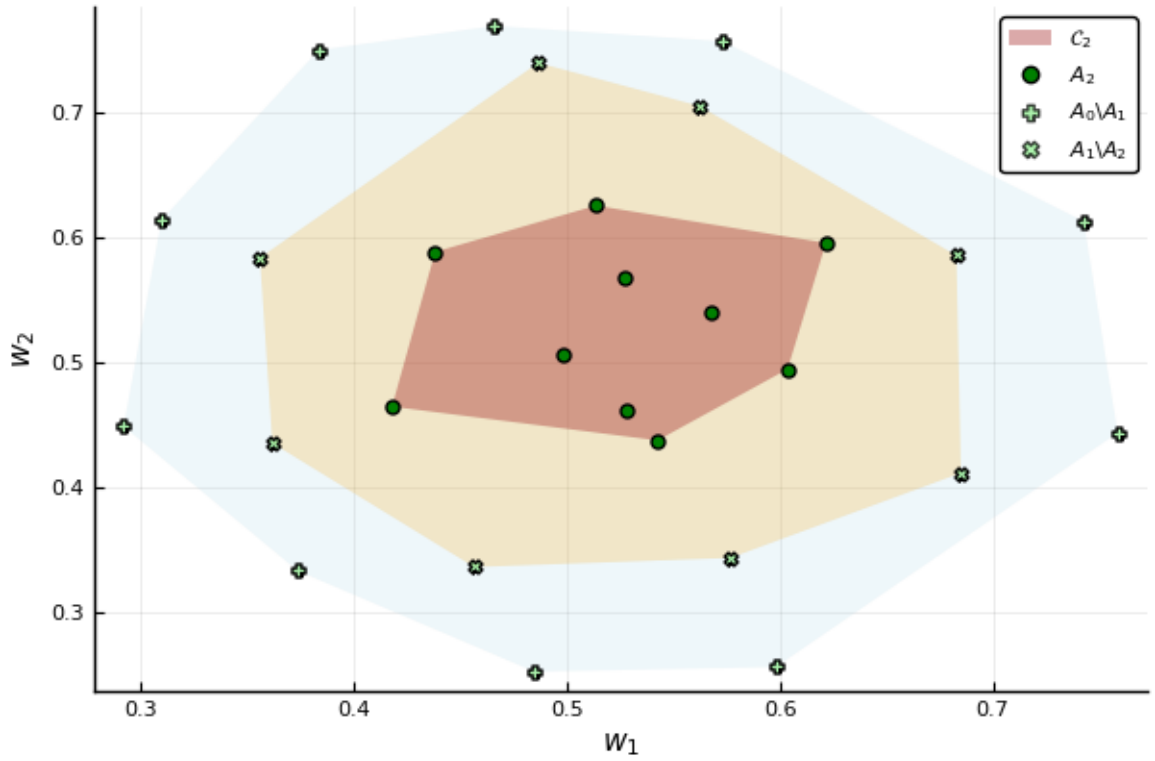
Metodologia proposta

- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros



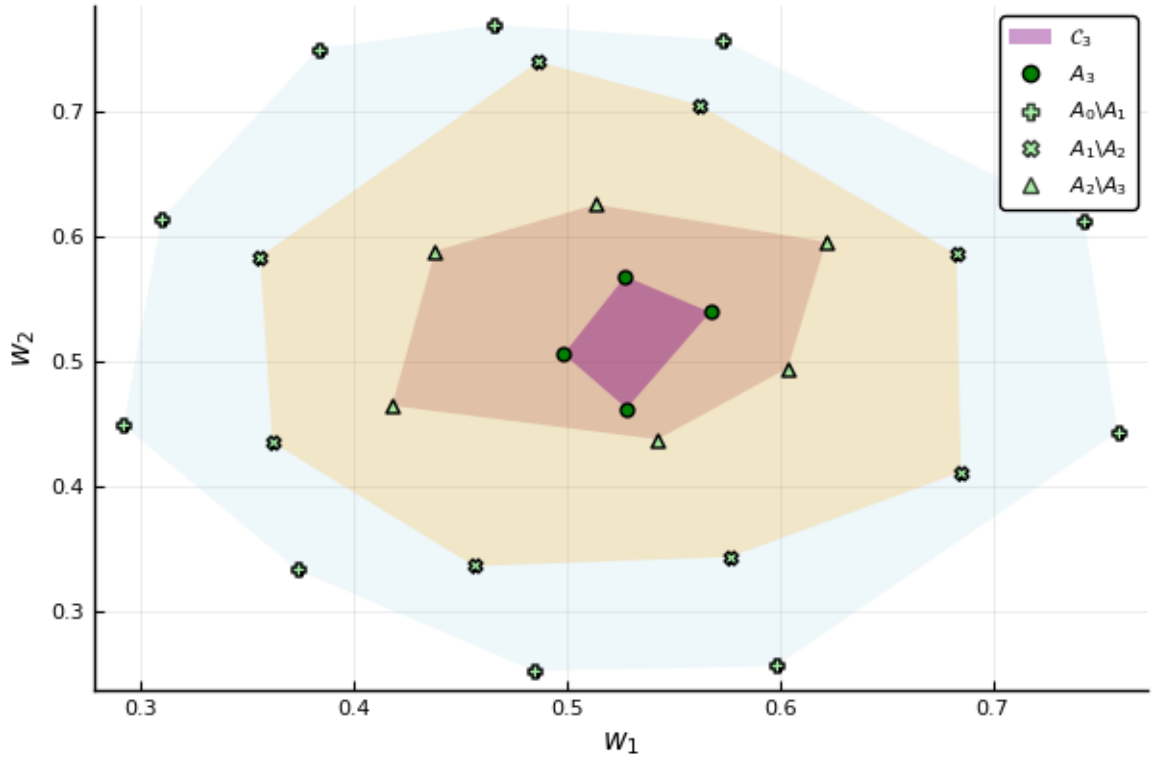
Metodologia proposta

- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros



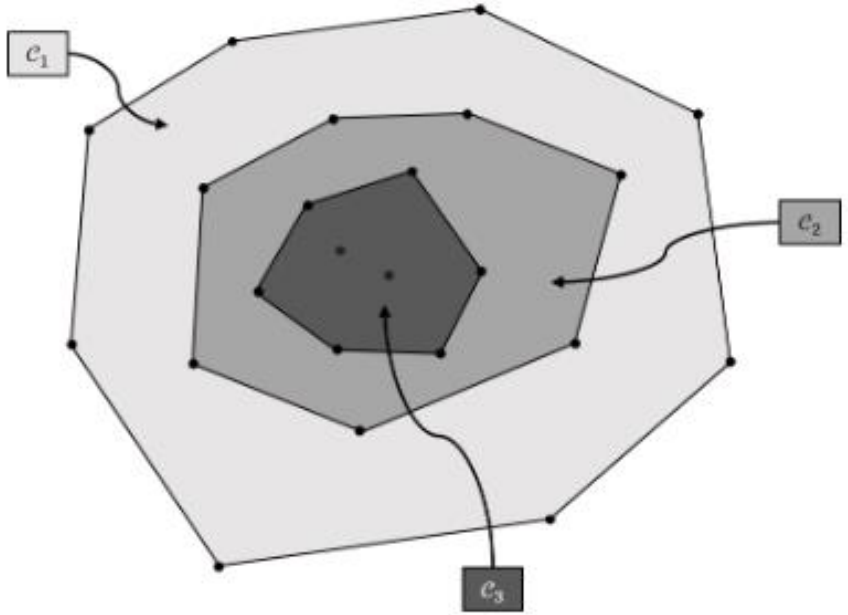
Metodologia proposta

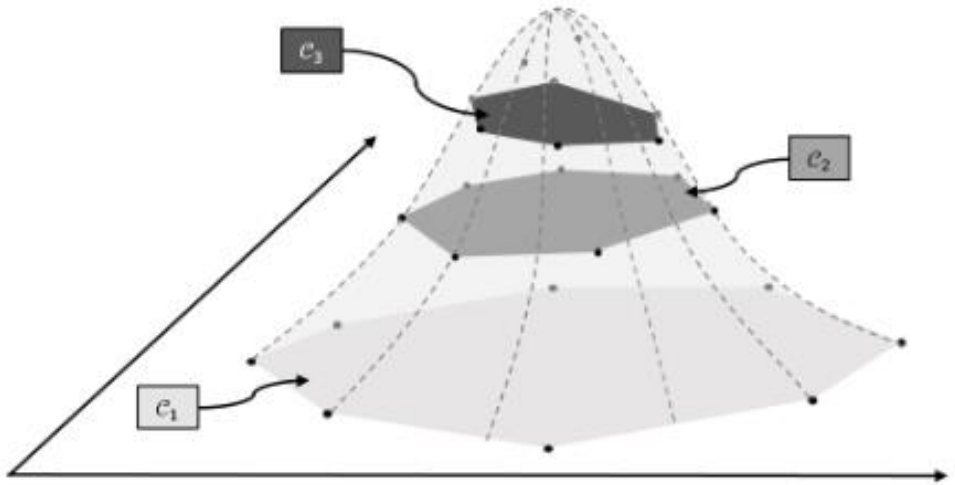
- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros



Metodologia proposta

- Conjunto de ambiguidade formulado a partir dos dados
- Sem calibração de hiper parâmetros





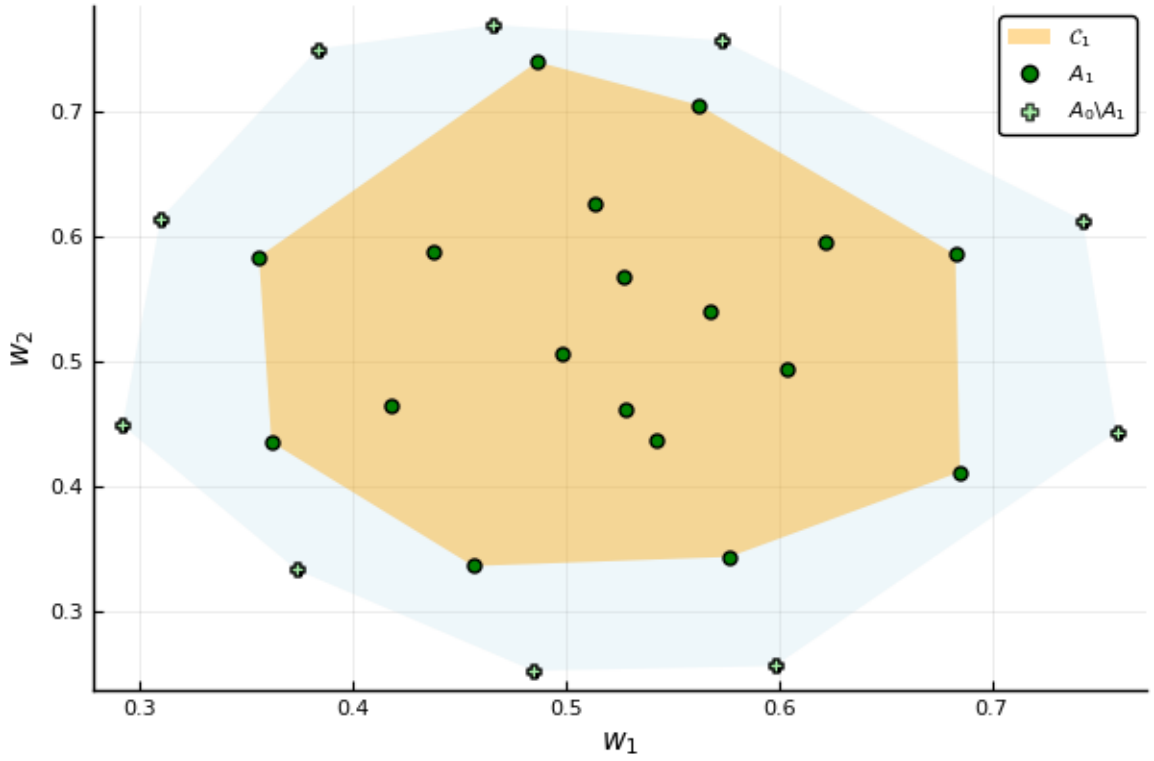
Metodologia proposta

- Intervalo de probabilidade de cada subconjunto calculado a partir dos dados utilizando o Teorema Central do Limite

$$\hat{p}_i = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}_{\mathcal{V}_i}(w_j) \xrightarrow{D} \mathcal{N}\left(\hat{p}_i, \frac{\hat{p}_i(1-\hat{p}_i)}{N}\right)$$

$$\underline{p_i} = \hat{p}_i - \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}$$

$$\overline{p_i} = \hat{p}_i + \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}$$



Metodologia proposta

- A partir desse conjunto de ambiguidades, podemos reformular o problema original como um problema de programação linear

$$\begin{aligned} \min_{\beta \in \mathcal{B}} \quad & \max_{P} \quad \mathbb{E}_P[h(W, \beta)] \\ & s.t. \quad P \in \left\{ P \in \mathcal{M}_+(\mathbb{R}^d) \middle| P(W \in \mathcal{C}_i) \in [\underline{p_i}, \overline{p_i}], \forall i \in \mathcal{F} \right\} \end{aligned}$$

Metodologia proposta

- A partir desse conjunto de ambiguidades, podemos reformular o problema original como um problema de programação linear

$$\min_{\beta \in \mathcal{B}} \quad \max_{P} \quad \mathbb{E}_P[h(W, \beta)]$$
$$s.t. \quad P \in \left\{ P \in \mathcal{M}_+(\mathbb{R}^d) \middle| P(W \in \mathcal{C}_i) \in [\underline{p_i}, \overline{p_i}], \forall i \in \mathcal{F} \right\}$$

Kuhn et al (2018)

$$\min_{\lambda, \kappa} \quad \sum_{i \in \mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w; \beta) - \sum_{i' \in \mathcal{A}(i)} (\kappa_i - \lambda_i) \leq 0, \quad \forall i \in \mathcal{F}, w \in \mathcal{C}_i$$
$$\lambda_i \geq 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \geq 0, \quad \forall i \in \mathcal{F}$$

Metodologia proposta

- A partir desse conjunto de ambiguidades, podemos reformular o problema original como um problema de programação linear

$$\min_{\beta \in \mathcal{B}} \quad \max_{P} \quad \mathbb{E}_P[h(W, \beta)]$$
$$s.t. \quad P \in \left\{ P \in \mathcal{M}_+(\mathbb{R}^d) \middle| P(W \in \mathcal{C}_i) \in [\underline{p_i}, \overline{p_i}], \forall i \in \mathcal{F} \right\}$$

Kuhn et al (2018)

$$\min_{\lambda, \kappa} \quad \sum_{i \in \mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w; \beta) - \sum_{i' \in \mathcal{A}(i)} (\kappa_i - \lambda_i) \leq 0, \quad \forall i \in \mathcal{F}, w \in \mathcal{C}_i$$
$$\lambda_i \geq 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \geq 0, \quad \forall i \in \mathcal{F}$$

Metodologia proposta

- A partir desse conjunto de ambiguidades, podemos reformular o problema original como um problema de programação linear

$$\min_{\beta\in\mathcal{B}} \quad \max_{P} \quad \mathbb{E}_P[h(W,\beta)]$$
$$s.t. \quad P \in \left\{P \in \mathcal{M}_+(\mathbb{R}^d) \middle| P(W \in \mathcal{C}_i) \in [\underline{p_i}, \overline{p_i}], \forall i \in \mathcal{F}\right\}$$

Kuhn et al (2018)

$$\min_{\lambda,\kappa} \quad \sum_{i\in\mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w;\beta) - \sum_{i'\in\mathcal{A}(i)} (\kappa_i - \lambda_i) \leq 0, \quad \forall i \in \mathcal{F}, w \in \mathcal{C}_i$$
$$\lambda_i \geq 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \geq 0, \quad \forall i \in \mathcal{F}$$

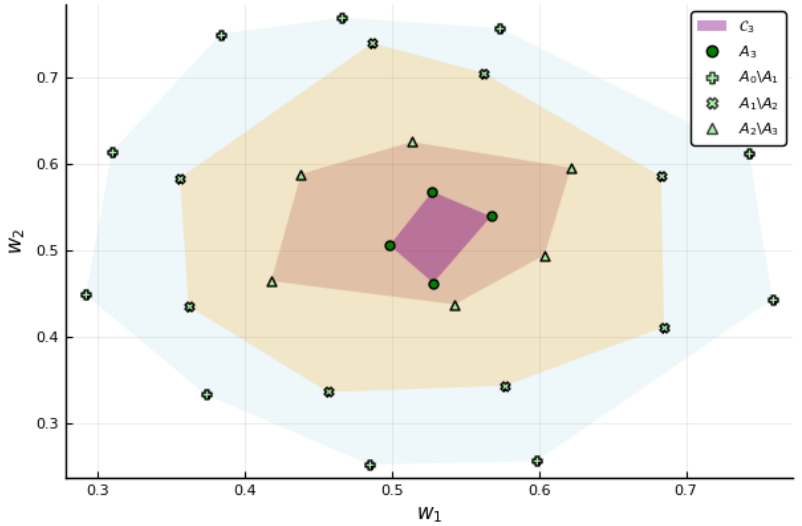
Data Driven Ambiguity Set

$$\min_{\beta,\lambda,\kappa} \quad \sum_{i\in\mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w;\beta) - \sum_{i'\in\mathcal{A}(i)} (\kappa_i - \lambda_i) \leq 0, \quad \forall w \in \mathcal{V}_i, \forall i \in \mathcal{F}$$
$$\lambda_i \geq 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \geq 0, \quad \forall i \in \mathcal{F}$$
$$\beta \in \mathcal{B}$$

Metodologia proposta – Predictive PolieDRO

- A partir desse conjunto de ambiguidades, podemos reformular o problema original como um problema de programação linear

$$\min_{\beta\in\mathcal{B}} \quad \max_{P} \quad \mathbb{E}_P[h(W,\beta)]$$
$$s.t. \quad P \in \left\{P \in \mathcal{M}_+(\mathbb{R}^d) \middle| P(W \in \mathcal{C}_i) \in [\underline{p_i}, \overline{p_i}], \forall i \in \mathcal{F}\right\}$$



Kuhn et al (2018)

$$\min_{\lambda,\kappa} \quad \sum_{i\in\mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w;\beta) - \sum_{i'\in\mathcal{A}(i)} (\kappa_i - \lambda_i) \le 0, \quad \forall i \in \mathcal{F}, w \in \mathcal{C}_i$$
$$\lambda_i \ge 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \ge 0, \quad \forall i \in \mathcal{F}$$

Data Driven Ambiguity Set

$$\min_{\beta,\lambda,\kappa} \quad \sum_{i\in\mathcal{F}} (\kappa_i \overline{p_i} - \lambda_i \underline{p_i})$$
$$\text{s.t.} \quad h(w;\beta) - \sum_{i'\in\mathcal{A}(i)} (\kappa_i - \lambda_i) \le 0, \quad \forall w \in \mathcal{V}_i, \forall i \in \mathcal{F}$$
$$\lambda_i \ge 0, \quad \forall i \in \mathcal{F}$$
$$\kappa_i \ge 0, \quad \forall i \in \mathcal{F}$$
$$\beta \in \mathcal{B}$$

Resultados

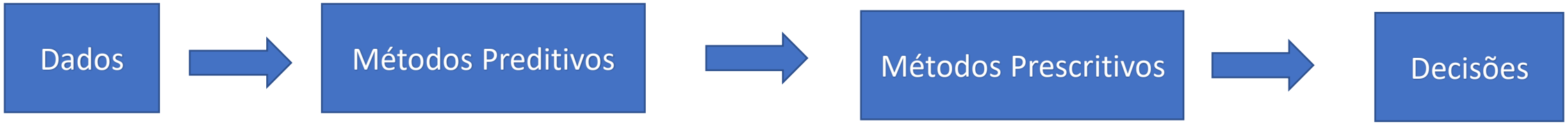
Data set name	n	d	DRO Hinge	SVM	C's
accute-inflamations-1	120	6	**1.0000**	**1.0000**	2
accute-inflamations-2	120	6	**1.0000**	0.9833	2
balance-scale	625	4	**0.9456**	**0.9456**	3
banknote-authentication	1,372	4	**0.9883**	0.9810	3
blood-transfusion	748	4	**0.7760**	0.7638	4
breast-cancer-prognostic	194	32	0.5564	**0.5897**	4
car-evaluation	1,728	21	**0.9520**	**0.9520**	3
climate-model-crashes	540	20	**0.9537**	0.9490	2
connectionist-bench-sonar	208	60	0.7000	**0.7317**	3
contraceptive-method-choice	1473	9	0.6786	**0.6836**	2
dermatology	358	34	**0.9905**	**0.9905**	1
fertility	100	9	**0.9000**	**0.9000**	1
glass-identification	214	9	**0.6465**	0.6093	3
haberman-survival	306	3	**0.7508**	0.7409	2
hayes-roth	132	4	0.6307	**0.6846**	3
indian-liver-patient	583	10	**0.6905**	**0.6905**	1
ionosphere	351	34	**0.8685**	**0.8685**	3
iris	150	4	**1.0000**	**1.0000**	1
letter-recognition	20,000	16	**0.9867**	0.9772	2
libras-movement	360	90	0.9138	**0.9777**	2
magic-gamma-telescope	19,020	11	**0.9995**	**0.9995**	1
mammography	830	5	**0.8193**	**0.8193**	1
monks-problems-1	556	6	**0.6702**	**0.6702**	1
monks-problems-2	601	6	**0.6533**	**0.6533**	1
monks-problems-3	554	6	**0.7829**	**0.7829**	1

Resultados

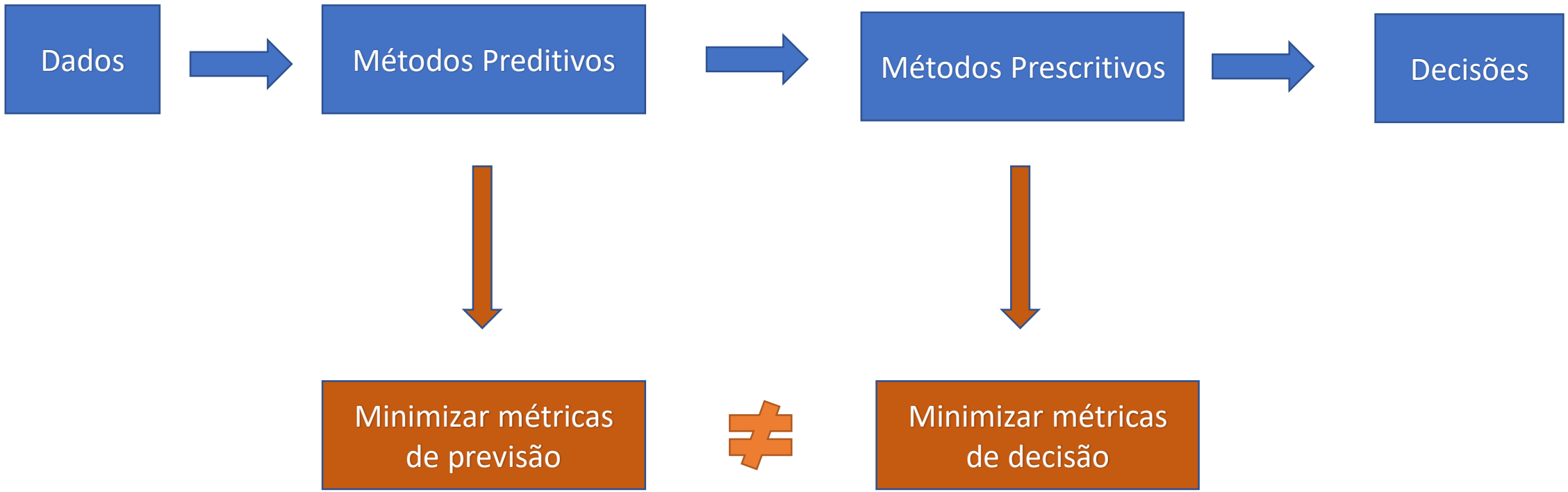
Data set name	n	d	DRO Hinge	SVM	C's
optical recognition	5,620	64	**0.9857**	0.9774	2
ozone-level-detection-eight	2,534	72	**0.9420**	0.9364	3
parkinsons	195	22	0.8461	**0.8512**	3
pima-indians-diabetes	768	8	**0.7818**	0.7784	3
planning-relax	182	12	**0.7777**	**0.7777**	1
qsar-biodegradation	1,055	41	**0.8739**	0.8663	3
seeds	210	7	**0.9380**	0.8714	3
seismic-bumps	210	7	**0.9571**	0.8857	2
spect-heart	267	22	**0.7924**	0.7888	3
spectf-heart	267	22	0.7782	**0.7814**	2
teaching-assistant-evaluation	151	5	**0.7200**	0.6866	3
thoracic-surgery	470	16	**0.8638**	**0.8638**	1
wine	178	13	0.9533	**0.9657**	3
yeast-1	2,417	103	**0.7871**	0.7842	2
yeast-2	2,417	103	**0.6327**	0.6272	3
yeast-3	2,417	103	0.7279	**0.7289**	2
yeast-4	2,417	103	**0.7436**	0.7297	3
yeast-5	2,417	103	**0.7706**	**0.7706**	2
yeast-6	2,417	103	**0.7631**	0.7502	1
yeast-7	2,417	103	**0.8256**	**0.8256**	1
yeast-8	2,417	103	**0.8082**	**0.8082**	1
yeast-9	2,417	103	**0.9304**	0.9102	
yeast-10	2,417	103	**0.8993**	0.8991	2
yeast-11	2,417	103	**0.8853**	**0.8853**	1
yeast-12	2,417	103	**0.7540**	0.7537	1
yeast-13	2,417	103	**0.7524**	0.7519	1
yeast-14	2,417	103	0.9511	**0.9867**	2

- 25 vitórias
- 18 empates
- 10 derrotas

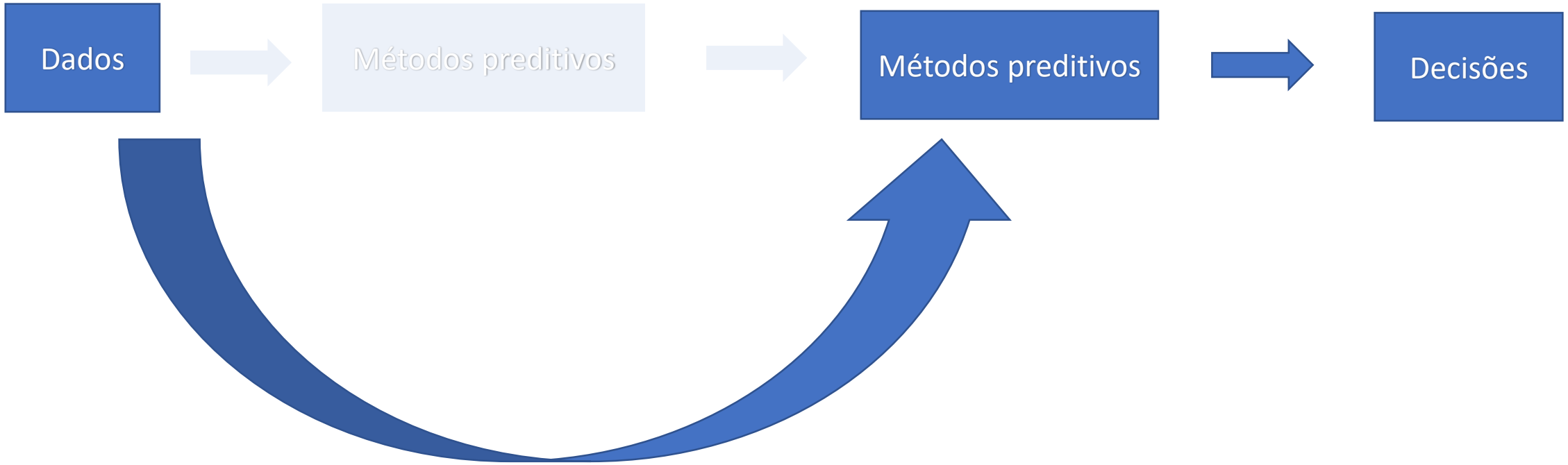
Extensões



Extensões


Dados
Métodos Preditivos
Métodos Prescritivos
Decisões
Minimizar métricas de previsão
≠
Minimizar métricas de decisão

Extensões – Prescriptive PolieDRO



Referências

[Blanchet et al., 2019] Blanchet, J., Kang, Y., and Murthy, K. (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857.

[Bykat, 1978] Bykat, A. (1978). Convex hull of a finite set of points in two dimensions. *Information Processing Letters*, 7(6):296–298.

[Casella and Berger, 2001] Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Cengage Learning, 2nd edition.

[Chen and Paschalidis, 2018] Chen, R. and Paschalidis, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13).

[Eddy, 1977] Eddy, W. F. (1977). A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, 3(4):398–403.

[Esfahani and Kuhn, 2018] Esfahani, P. M. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166.

[Fernandes et al., 2016] Fernandes, B., Street, A., Valladão, D., and Fernandes, C. (2016). An adaptive robust portfolio optimization model with loss constraints based on data-driven polyhedral uncertainty sets. *European Journal of Operational Research*, 255(3):961 – 970.

[Shafieezadeh-Abadeh et al., 2015] Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *arXiv preprint arXiv:1509.09259*.

[Shafieezadeh-Abadeh et al., 2019] Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68.

[Sivaprasad et al., 2020] Sivaprasad, P. T., Mai, F., Vogels, T., Jaggi, M., and Fleuret, F. (2020). Optimizer benchmarking needs to account for hyperparameter tuning. In *International Conference on Machine Learning*, pages 9036–9045. PMLR.

[Uryasev and Pardalos, 2013] Uryasev, S. and Pardalos, P. M. (2013). *Stochastic optimization: algorithms and applications*, volume 54. Springer Science & Business Media.

[Wiesemann et al., 2014] Wiesemann, W., Kuhn, D., and Sim, M. (2014). Distributionally robust convex optimization. *Oper. Res.*, 62(6):1358–1376.

[Wilson, 1927] Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212.