Robust Counterfactual Explanations and Mathematical Optimization

Jannis Kurtz

University of Amsterdam
Amsterdam Business School

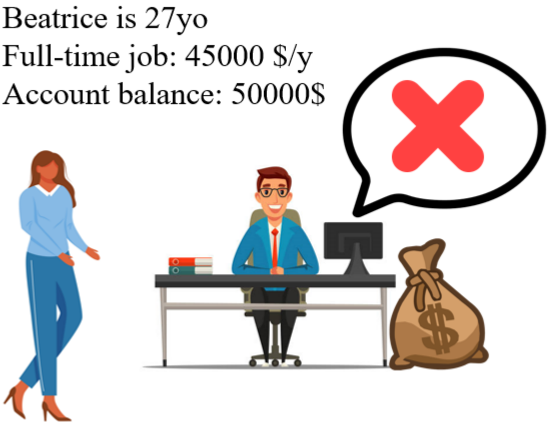
14 May 2025

Overview

Overview

Machine Learning as a Black-Box

- Machine learning models are used to derive decisions, based on individual data.
- Understanding of the models/decisions lacks behind.
- Explanations for the individual have to be provided.



Example: Credit Scoring

Labeled Data

Customer ID	Age	Salary	Current Balance	$\cdots$	Loan Granted
1	28	42.000 EUR	8.200 EUR	$\cdots$	1
2	56	73.800 EUR	22.300 EUR	$\cdots$	1
3	42	35.100 EUR	16.900 EUR	$\cdots$	0
$\vdots$				$\vdots$	$\vdots$

Credit Scoring

Companies decide based on your **individual information** if you should be **granted a loan or not**.

Right to Explanation

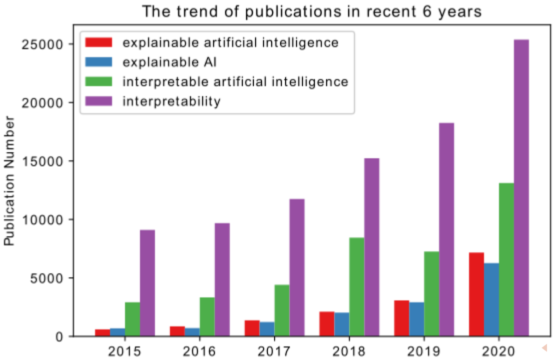
The European Union enacted the **right to explanation** in 2016 which was incorporated in the EU General Data Protection Regulation:

> [...] *In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to* **obtain an explanation of the decision reached after such assessment** *and to challenge the decision.* [...]

Artificial Intelligence Act (European Union)

Article 13

High-risk AI systems shall be designed and developed in such a way to ensure that their operation is **sufficiently transparent** to enable users to **interpret the system's output** and use it appropriately. An appropriate type and degree of transparency shall be ensured, [...].



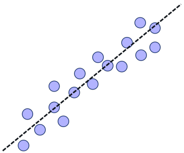
Binary Classification Problems

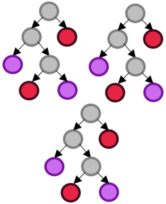
Classification Machine Learning Model

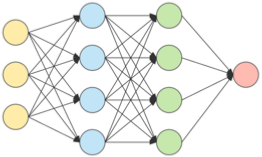
- Train a classifier $h : \mathcal{X} \to [0,1]$ on labeled data
- Assigns "probability" $h(x)$ to each data points $x \in \mathcal{X}$
- Classify x as

$$h_{\text{class}}(x) = \begin{cases} 1 & \text{if } h(x) \geq \tau \\ -1 & \text{if } h(x) < \tau \end{cases}$$

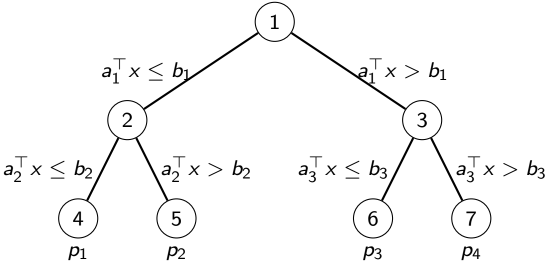
where $\tau \in (0,1)$ is a given threshold parameter (often $\tau = 0.5$).







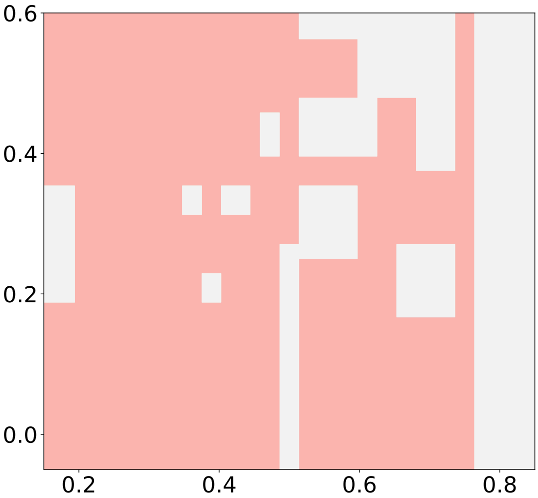
Classification Trees



Properties

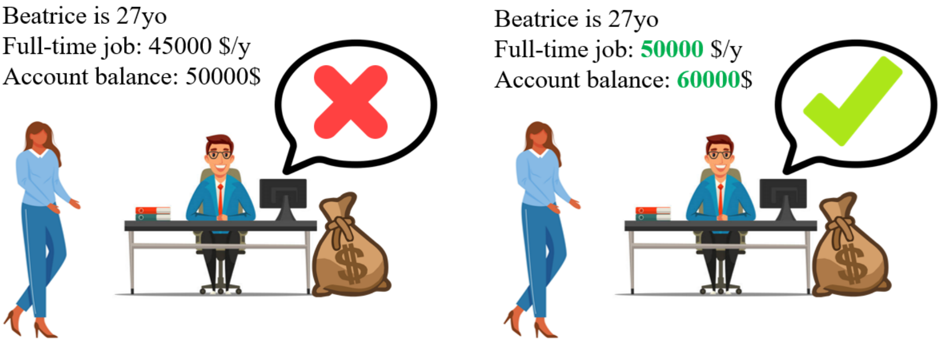
- Each leaf is given by a set of inequalities: $\mathcal{L}_5 = \left\{x \mid a_1^\top x \leq b_1,\ a_2^\top x > b_2\right\}$
- Classifier h assigns fraction $p_i \in [0,1]$ to each point x in the leaf.
- p_i is often the fraction of training data in the leaf with label 1

Classification Trees



Decision region of a classification tree trained by CART on the Iris dataset

Counterfactual Explanation

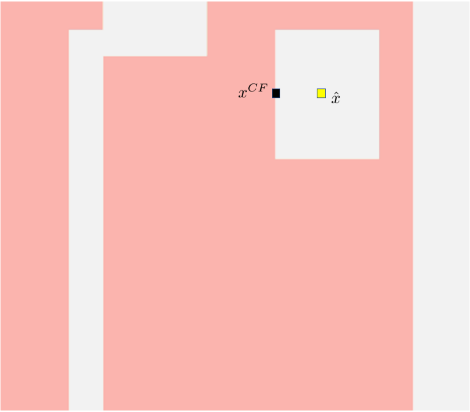


Counterfactual Explanation

If your salary would be 50,000$ and your current balance 60,000$, then you would have been granted a loan.

Counterfactual Explanations

- $\hat{x}$: factual instance (classified as -1)
- x^{CF}: counterfactual explanations (classified as 1)



Counterfactual Explanations

Optimization Problem

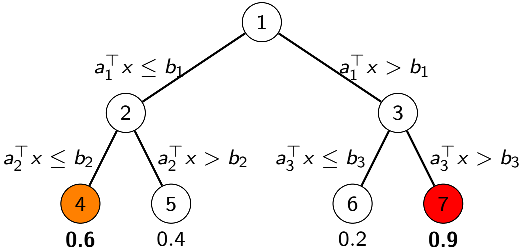
For a given factual instance $\hat{x}$ (classified as -1) a **counterfactual explanation** x^{CF} can be calculated by solving

$$\begin{aligned} \min \quad & d(\hat{x}, x) \\ s.t. \quad & h(x) \geq \tau \\ & x \in \mathcal{X}. \end{aligned}$$

Literature

Wachter et al. (2018), Ustun et al. (2019), Russell (2019), Mahajan et al. (2019), Mothilal et al. (2020), Maragno et al. (2022)

Mixed-Integer Formulation



$$
\begin{aligned}
\min_{x,l} \quad & d(\hat{x}, x) \\
s.t. \quad & a_1^\top x \leq b_1 + M(1 - l_1) \\
& a_2^\top x \leq b_2 + M(1 - l_1) \\
& a_1^\top x \geq b_1 + \varepsilon - M(1 - l_2) \\
& a_3^\top x \geq b_3 + \varepsilon - M(1 - l_2) \\
& l_1 + l_2 = 1 \\
& l_1, l_2 \in \{0, 1\} \\
& x \in \mathcal{X}.
\end{aligned}
$$

For more details:

Maragno, D., Röber, T. E., & Birbil, I. **Counterfactual Explanations Using Optimization With Constraint Learning.** In OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop).

Overview

Example: Credit Scoring

Labeled Data

Customer ID	Age	Salary	Current Balance	⋯	Loan Granted
1	28	42.000 EUR	8.200 EUR	⋯	1
2	56	73.800 EUR	22.300 EUR	⋯	1
3	42	35.100 EUR	16.900 EUR	⋯	0
⋮				⋮	⋮

Credit Scoring

Companies decide based on your **individual information** if you should be **granted a loan or not**.

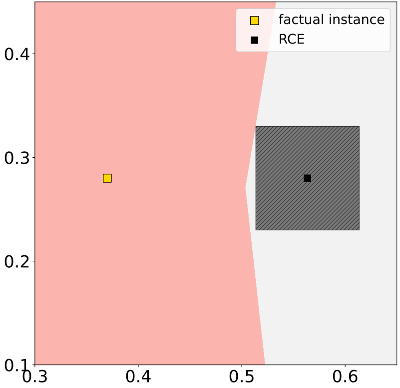
Robust Counterfactual Explanations

Counterfactual Explanation

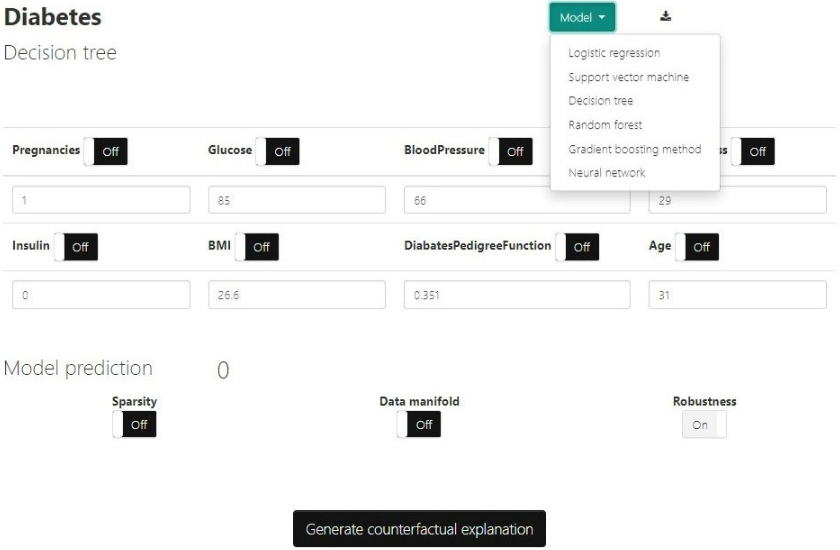
If your salary would be 44.500 EUR and your current balance 10.000, then you would have been granted a loan.

Motivation: Robustness

- But what if my current balance is 9985 EUR, or 10.099 EUR?
- Find counterfactual points which remain counterfactual points after small changes



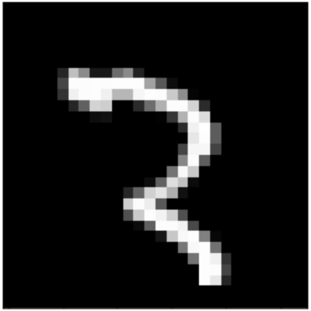
Let the User Decide



Let the User Decide

Decision Boundary

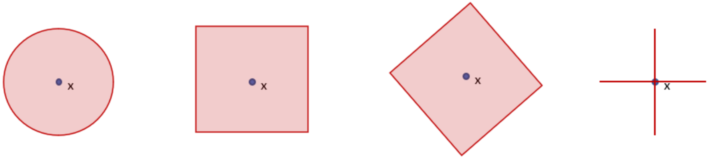
What is the truth on the decision boundary?



Robust Counterfactual Explanations

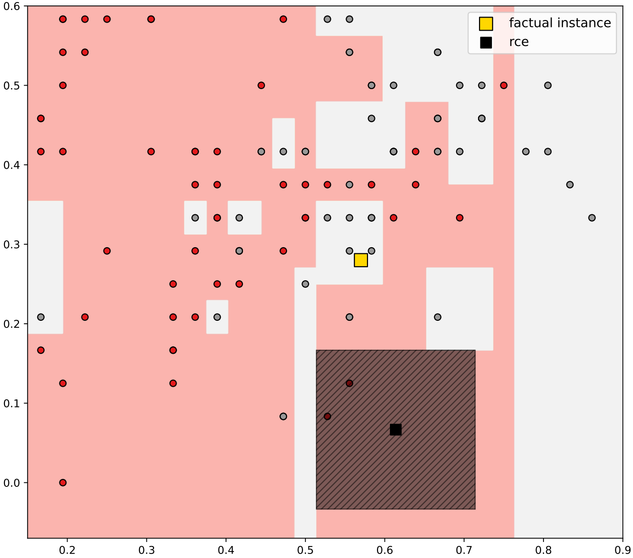
Setup

- Find counterfactual point x^{CE} such that all points in
$$x^{CE} + \mathcal{S}$$
are counterfactual points.
- Uncertainty set $\mathcal{S} = \{s : \|s\| \leq \varepsilon\}$ for a small $\varepsilon > 0$
- E.g. ℓ_1,ℓ_2 or ℓ_∞-norm can be used



ℓ_2 ℓ_∞ ℓ_1 ℓ_0

Example: box uncertainty

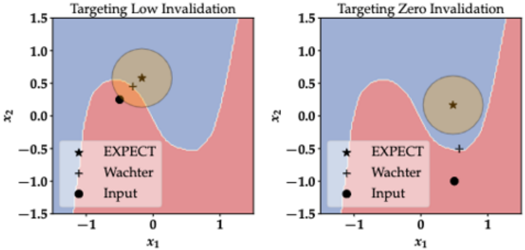


Robust Counterfactual Explanations

Literature

[Pawelczyk et al. (2022)],[Dominguez-Olmedo et al. (2021)]

- Gradient based heuristic
- no full robustness guarantee
- not (directly) applicable to classification trees



Robust Counterfactual Explanations

Goal

Derive a method to generate counterfactuals which

- are as **close as possible** and with **full robustness** guarantee
- which is also applicable to decision trees and random forests

Optimization Problem

Given a factual instance $\hat{x}$ (classified as -1), a **robust counterfactual explanation** is an optimal solution x^{RCF} of the following problem

$$\begin{aligned} \min \quad & d(\hat{x}, x) \\ s.t. \quad & h(x+s) \geq \tau \quad \forall s \in \mathcal{S} \\ & x \in \mathcal{X}. \end{aligned}$$

Algorithm: Adversarial Approach

Master Problem

- The master problem (MP) is a relaxation of the problem with a finite number of scenarios $\mathcal{Z} \subset \mathcal{S}$:

$$\begin{aligned} \min\ & d(\hat{x}, x) \\ s.t.\ \ & h(x+s) \geq \tau \quad \forall s \in \mathcal{Z} \\ & x \in \mathcal{X} \end{aligned}$$

- Provides a **lower bound** for the optimal value.

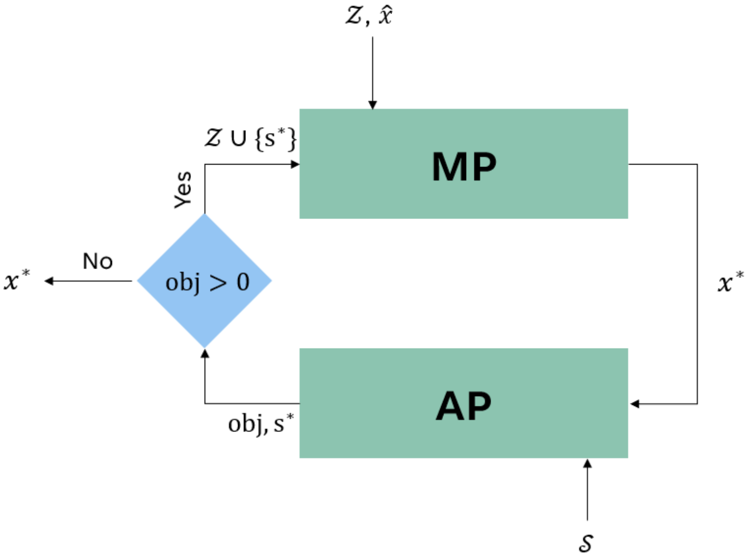
Adversarial Problem

- The adversarial problem (AP) finds a new scenario in $\mathcal{S}$ which maximally violates the constraints for current MP solution x^*:

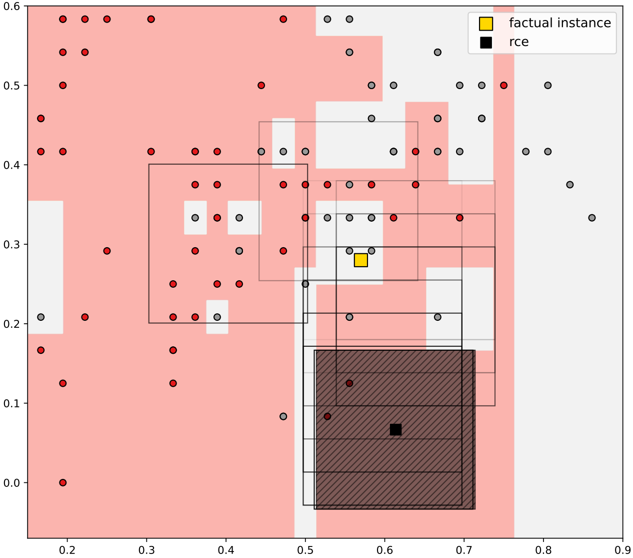
$$\max_{s \in \mathcal{S}} \ \tau - h(x^* + s)$$

- Current solution x^* is cut-off if optimal value > 0

Adversarial Approach



Adversarial Approach



Convergence

Theorem (Mutapcic & Boyd)

If X is bounded and if h is a Lipschitz continuous function, i.e., there exists an $L > 0$ such that

$$|h(x_1) - h(x_2)| \leq L\|x_1 - x_2\|$$

for all $x_1, x_2 \in X$. Then the algorithm terminates after a finite number of steps with a solution x^* such that

$$h(x^* + s) \geq \tau - \varepsilon \quad \forall\, s \in \mathcal{S}.$$

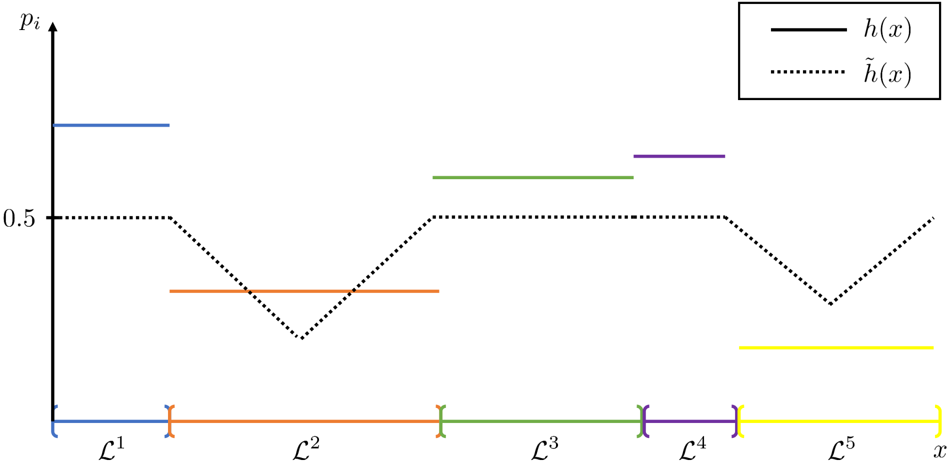
Lipschitz Continuous Classifiers

Lipschitz Continuity

Classifier function h of

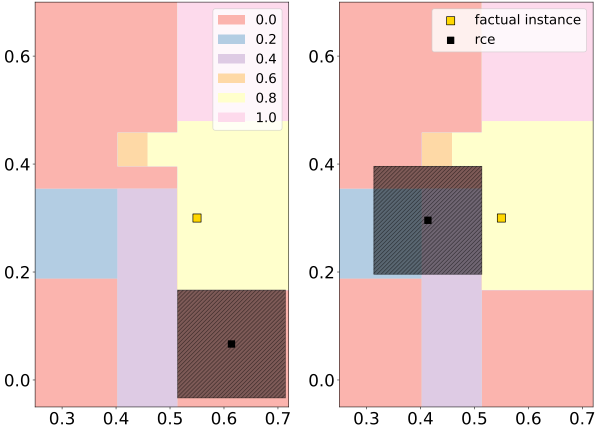
- **Logistic regression** classifier is Lipschitz continuous.
- **Neural network** classifier with ReLU activation functions is Lipschitz continuous.
- **Decision trees** classifier is not even continuous! **We can "Lipschitzify" the function to make it work!**
- **Random Forest** classifier is not even continuous! **We can "Lipschitzify" the function to make it work!**

Lipschitzification of Decision Tree Classifier



Overlapping Leaves

- The uncertainty set $\mathcal{S}$ may overlap with multiple leaves (all of class 1).
- Restricting the set to a single leaf may lead to suboptimal solutions



Master Problem: Decision Trees

General **master problem:**

$$
\begin{aligned}
\min\ & d(\hat{x}, x) \\
s.t.\ & h(x+s) \geq \tau \quad \forall s \in \mathcal{Z} \\
& x \in \mathcal{X}
\end{aligned}
$$

The master problem for a trained **decision tree** can be reformulated as:

$$
\begin{aligned}
\min\ & d(\hat{x}, x) \\
s.t.\ & A^i(x+s) \leq b^i + M(1 - l_i(s)) \quad s \in \mathcal{Z},\ \text{leaf } i \\
& \sum_{\text{leaves } i} l_i(s) = 1 \quad s \in \mathcal{Z} \\
& \sum_{\text{leaves } i} l_i(s) p_i \geq \tau \quad s \in \mathcal{Z} \\
& l_i(s) \in \{0, 1\} \quad s \in \mathcal{Z},\ \text{leaf } i
\end{aligned}
$$

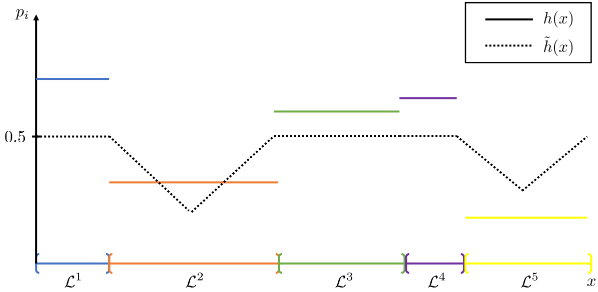
Adversarial Problem

For current MP solution x^* find an s in

$$\arg\max_{s \in \mathcal{S}} \ \tau - h(x^* + s)$$

which is equivalent to

$$\arg\min_{s \in \mathcal{S}} h(x^* + s)$$

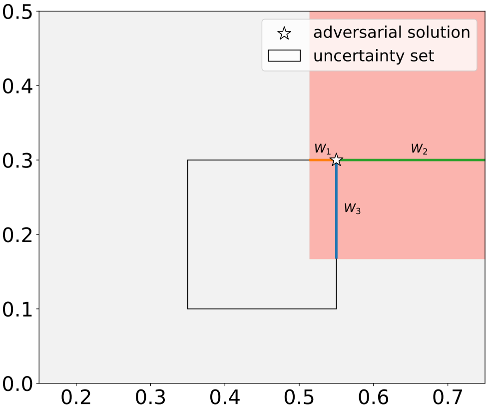


Adversarial Problem: Decision Trees

The **adversarial problem** can be solved by solving

$$\begin{aligned} \max\ & \alpha \\ s.t.\quad & \alpha \leq w \\ & A^i(x^* + s) + w \leq b^i \\ & s \in \mathcal{S}, w \geq 0 \end{aligned}$$

for every leaf i with $p_i < \tau$ and choosing the best optimal value.



Experiments

		BANKNOTE AUTHENTICATION 4 features			DIABETES 8 features			IONOSPHERE 34 features		
Model	**Specs**	**Comp. time (s)**	**# iterations**	**# early stops**	**Comp. time (s)**	**# iterations**	**# early stops**	**Comp. time (s)**	**# iterations**	**# early stops**
					$\rho = 0.01$					
Linear	ElasticNet	0.25 (0.01)	–	–	0.23 (0.01)	–	–	0.25 (0.01)	–	–
DT	max depth: 3	1.53 (0.04)	1.00 (0.00)	–	1.66 (0.07)	1.20 (0.09)	–	1.66 (0.07)	1.10 (0.07)	–
	max depth: 5	1.86 (0.09)	1.10 (0.07)	–	2.29 (0.11)	1.20 (0.09)	–	1.96 (0.08)	1.10 (0.07)	–
	max depth: 10	3.90 (0.88)	2.00 (0.58)	–	5.77 (0.48)	1.40 (0.13)	–	2.86 (0.16)	1.20 (0.09)	–
RF*	# est.: 5	3.50 (0.36)	1.75 (0.22)	–	5.89 (1.98)	2.60 (0.83)	–	3.79 (0.24)	1.70 (0.13)	–
	# est.: 10	6.74 (1.03)	2.60 (0.40)	–	7.20 (0.94)	2.35 (0.29)	–	8.76 (0.89)	2.85 (0.33)	–
	# est.: 20	21.21 (4.61)	4.55 (0.99)	–	33.55 (7.48)	6.15 (0.79)	–	22.33 (2.83)	4.40 (0.51)	–
	# est.: 50	115.79 (34.35)	7.80 (1.65)	–	110.24 (32.43)	6.47 (1.39)	3 ($\bar{\rho} = 0.007$)	137.26 (33.37)	8.20 (1.20)	–
	# est.: 100	214.38 (65.07)	6.44 (1.31)	2 ($\bar{\rho} = 0.009$)	274.09 (71.93)	8.87 (1.49)	5 ($\bar{\rho} = 0.004$)	285.62 (95.57)	8.27 (2.02)	9 ($\bar{\rho} = 0.004$)
GBM**	# est.: 5	2.70 (0.22)	1.20 (0.14)	–	2.76 (0.22)	1.85 (0.17)	–	2.37 (0.15)	1.60 (0.13)	–
	# est.: 10	3.20 (0.30)	1.45 (0.15)	–	2.72 (0.29)	1.50 (0.24)	–	4.35 (0.44)	2.75 (0.30)	–
	# est.: 20	5.94 (0.50)	2.60 (0.23)	–	4.25 (0.45)	2.15 (0.28)	–	9.01 (1.11)	3.85 (0.50)	–
	# est.: 50	18.38 (1.62)	4.05 (0.35)	–	24.60 (8.21)	5.85 (1.41)	–	81.33 (28.39)	8.90 (1.36)	–
	# est.: 100	87.11 (26.24)	7.28 (0.77)	2 ($\bar{\rho} = 0.006$)	164.32 (42.12)	11.63 (2.00)	1 ($\bar{\rho} = 0.004$)	137.98 (22.74)	10.33 (0.97)	2 ($\bar{\rho} = 0.007$)
MLP	(10,)	1.63 (0.04)	1.00 (0.00)	–	1.55 (0.06)	1.00 (0.00)	–	2.22 (0.19)	1.80 (0.21)	–
	(10, 10, 10)	2.98 (0.15)	1.15 (0.08)	–	2.40 (0.12)	1.15 (0.08)	–	13.01 (3.57)	2.30 (0.40)	–
	(50,)	2.60 (0.14)	1.00 (0.00)	–	2.09 (0.12)	1.05 (0.05)	–	5.75 (0.75)	1.20 (0.12)	–
	(100,)	3.53 (0.15)	1.00 (0.00)	–	4.36 (0.62)	1.10 (0.07)	–	61.31 (33.11)	1.50 (0.22)	10 ($\bar{\rho} = 0.000$)

* max depth of each decision tree equal to 3.; ** max depth of each decision tree equal to 2.

Table 1. Generation of robust CEs for 20 factual instances using ℓ_∞-norm as uncertainty set.

Experiments

	Dominguez-Olmedo et al. (2021)		Our algorithm	
ρ	**robustness**	**validity**	**robustness**	**validity**
		Banknote NN(50)		
0.1	80%	100%	100%	100%
0.2	0%	100%	100%	100%
		Diabetes NN(50)		
0.1	60%	100%	100%	100%
0.2	0%	100%	80%	100%
		Ionosphere NN(10)		
0.1	70%	90%	100%	100%
0.2	40%	90%	100%	100%
		Ionosphere NN(50)		
0.1	50%	100%	100%	100%
0.2	30%	100%	100%	100%

Conclusion

Summary

- Robust counterfactuals can be calculated by robust adversarial approach.
- Works for most important classifiers and uncertainty sets.

Future Work

- Speed up calculations of master problem
- Include categorical Features
- Model robustness?

For more details

Maragno, D., Kurtz, J., Röber, T. E., Goedhart, R., Birbil, Ş. I., & den Hertog, D. (2024). **Finding regions of counterfactual explanations via robust optimization.** INFORMS Journal on Computing, 36(5), 1316-1334.

Overview

Motivation: Diet Problem

Problem Description

Given a set of **nourishments** with corresponding **nutrition values**, decide how much of each product to purchase, such that all daily nutrition requirements are fulfilled and the costs are minimal.

(World Food Programme (WFP) [Peters et al. (2021)])

$$
\begin{aligned}
\min\ & 800x_{\text{beans}} + 1003x_{\text{rice}} + 300x_{\text{wheat}} + \cdots \\
s.t.\ & 335x_{\text{beans}} + 360x_{\text{rice}} + 330x_{\text{wheat}} + \cdots \geq 2100, && \text{(Energy(kcal))} \\
& 20x_{\text{beans}} + 7x_{\text{rice}} + 12x_{\text{wheat}} + \cdots \geq 52.5, && \text{(Protein(g))} \\
& \vdots \\
& x \geq 0.
\end{aligned}
$$

Motivation: Diet Problem

Optimal solution of the diet problem

$50.5g$

$20g$

$5g$

$20.6g$

$135.8g$

$277.1g$

$72g$

Motivation: Diet Problem

Hypothetical situation:

- The local supplier of the WFP only sells Lentils, Rice and Chickpeas
- No product of this supplier is purchased

Supplier may ask:

"Why don't you buy anything from me?"

Counterfactual Question

"How much do I have to change my prizes and nutrition values such that you buy at least 200g of my products?"

Explanations are needed!

Facility location problem

- Optimize geospatial **accessibility to healthcare centers.** [Krishnakumari et al. (2024)]
- Question: "Why did you not build a new healthcare centers in our district?"

Scheduling trains in the Netherlands

- Netherlands Railways introduced a completely **new timetable** by using sophisticated operations research techniques. [Kroon et al. (2009)]
- Question: "Why do I have to work three times in a row on a Saturday?"

Boston public school transport

- Create better **bus routes for Boston Public Schools.** [Bertsimas et al. (2020)]
- Lead to reduced costs but 85% of the schools' start times would have been changed.
- Question: "Why does the school of my child start earlier now?"

Explainable Optimization

- **Sensitivity Analysis or Parametric Optimization**
- **Interpretable Optimization**
  - build tree which assigns solutions to problem instance [Goerigk and Hartisch (2023)]
- **Explainable Optimization**
  - Counterfactual Explanations: [Korikov et al. (2021) and Korikov and Beck (2023, 2021)]
  - Counterfactual Explanations in the context space: [Forel et al. (2023)]
  - Data-driven explanation based on historical solutions [Aigner et al. (2024)]

Present Problem

- The decision maker solves the **present problem**

$$\begin{aligned} \min\ & \hat{c}^\top x \\ s.t.\ \ & \hat{A}x \geq \hat{b}, \\ & x \geq 0, \end{aligned} \tag{PP}$$

  and implements an **optimal solution** $\hat{x}$.
- A stakeholder (influenced by the decision) wants to receive an explanation for $\hat{x}$.

Example: Diet Problem

- Decision maker: World Food Programme
- Stakeholder: Supplier

Counterfactual Explanations: Setup

The stakeholder has a

- **mutable parameter space** $\mathcal{H}$: contains all problem parameters (c, A, b) which the stakeholder can reach
- **favored solution space** $\mathcal{D}(\hat{x})$: the set of all solutions x which the stakeholder favors

Example: Diet Problem

- $\mathcal{H}$: contains all prizes (objective parameters) and nutrition values (constraint parameters) which differ at most 5% from the current value
- $\mathcal{D}(\hat{x}) = \{x \geq 0 : x_{\text{Lentils}} + x_{\text{Rice}} + x_{\text{Chickpeas}} \geq 200g\}$

Counterfactual Explanations

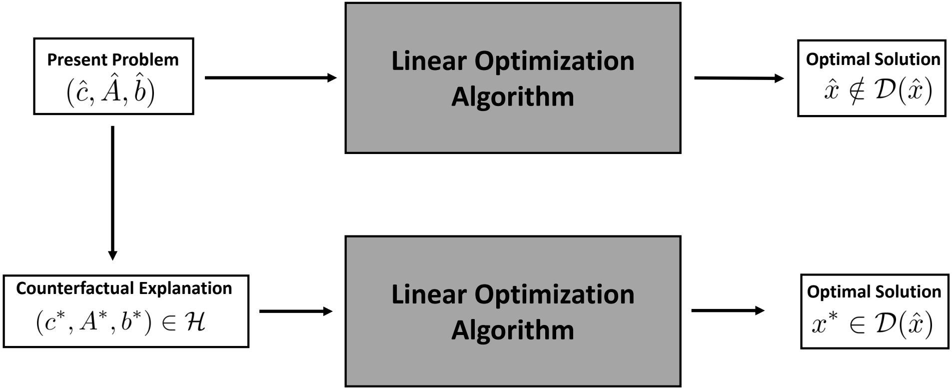
Counterfactual Question

"What is the minimal change of the mutable parameters I have to make such that a solution from the favored solution space is optimal?"

Counterfactual Explanation

"If you decrease your prize for Lentils by 50% *and the prize of Wheat would be increased by* 25.2%, *then a solution you favor would be optimal!"*

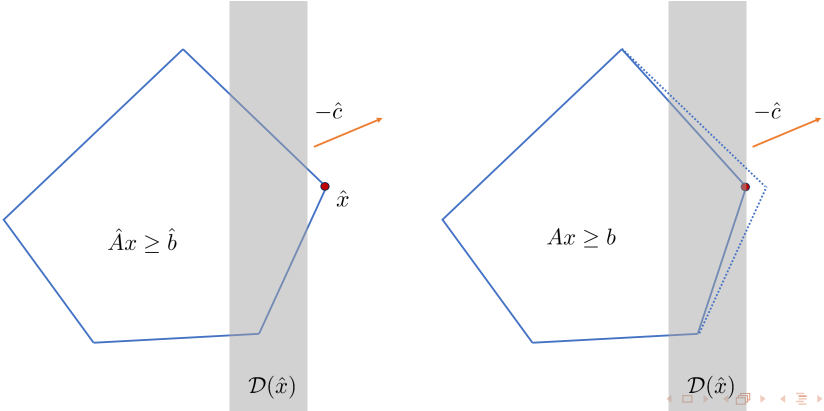
Counterfactual Explanations



Weak Counterfactual Explanations

Weak Counterfactual Explanation

A point $(c, A, b) \in \mathcal{H}$ such that there **exists an optimal solution** x^* of the corresponding LP which lies in the *favored solution space* $\mathcal{D}(\hat{x})$.



Weak Counterfactual Explanations

Consider any distance measure $\delta : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$.

Weak Counterfactual Explanation Problem

$$
\begin{aligned}
(\text{WCEP}): \quad & \inf_{x,c,A,b} \; \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \\
s.t. \quad & x \in \operatorname*{arg\,min}_{z: Az \geq b, z \geq 0} \; c^\top z, \\
& x \in D(\hat{x}), \\
& (c,A,b) \in \mathcal{H},
\end{aligned}
$$

Related to

- optimistic bilevel optimization,
- inverse optimization

Weak Counterfactual Explanations

Theorem

Problem WCEP is equivalent to the following problem:

$$
\begin{aligned}
(WCEP') : \quad \inf_{x,y,c,A,b} \quad & \delta\left((c, A, b), (\hat{c}, \hat{A}, \hat{b})\right) \\
\text{s.t.} \quad & c^\top x \le b^\top y, \\
& A^\top y \le c, \\
& Ax \ge b, \\
& x \in D(\hat{x}), \\
& (c, A, b) \in \mathcal{H}, \\
& x, y \ge 0.
\end{aligned}
$$

Weak Counterfactual Explanations

The latter formulation is bilinear and has some undesired properties:

Properties

- Projection of the feasible region on the (c, A, b)-space may be **open**.
- Projection of the feasible region on the (c, A, b)-space may be **non-convex and even disconnected**.
- Projection of the feasible region on the x-space can be **non-convex**.

Strong Counterfactual Explanations

Counterfactual Explanation

"If you decrease your prize for Lentils by 50% *and the prize of Wheat would be increased by* 25.2%, *then a solution you favor would be optimal!"*

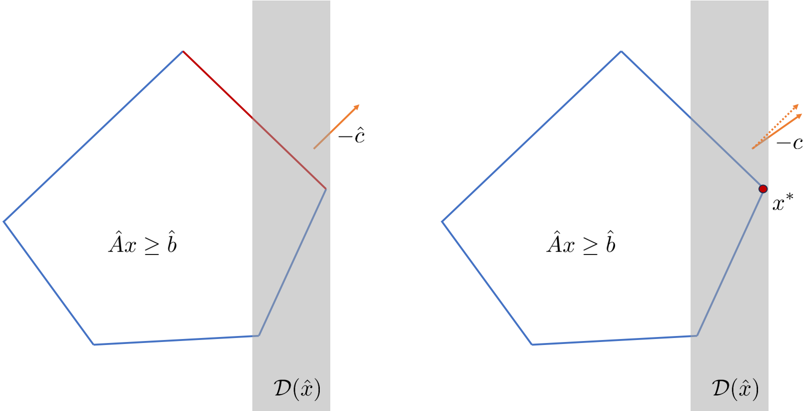
Problem

- But what if multiple optimal solutions exist? Are all of them in my favored solution space?
- If not, it depends on the solution algorithm the WFP uses which solution is implemented.

Strong Counterfactual Explanations

Strong Counterfactual Explanation

A point $(c, A, b) \in \mathcal{H}$ such that the **whole set of optimal solutions** $\mathcal{X}^*$ of the corresponding LP lies in the *favored solution space* $\mathcal{D}(\hat{x})$.



Strong Counterfactual Explanations

Consider any distance measure $\delta : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$.

Strong Counterfactual Explanation Problem

$$
\begin{aligned}
\text{(SCEP)}: \quad & \inf_{x,c,A,b} \ \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \\
& s.t. \quad x \in D(\hat{x}), \quad \forall\, x \in \operatorname*{arg\,min}_{z: Az \geq b, z \geq 0} c^\top z, \\
& \qquad (c,A,b) \in \mathcal{H},
\end{aligned}
$$

Relative Counterfactual Explanations

Counterfactual Explanation

"If you decrease your prize for Lentils by 50% *and the prize of Wheat would be increased by* 25.2%, *then a solution you favor would be optimal!"*

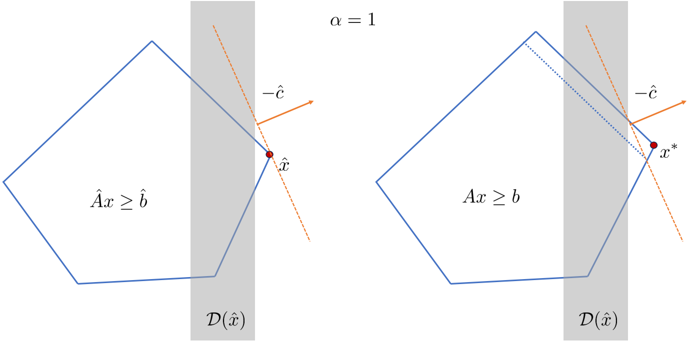
Problem

- Do we actually need optimality? What if my parameter changes lead to a cost reduction?

Relative Counterfactual Explanations

Relative Counterfactual Explanation

A point $(c, A, b) \in \mathcal{H}$ such that there exists a **feasible solution** which has costs at most α times the current costs (for a given factor $\alpha \in [0, \infty)$).



Relative Counterfactual Explanations

Relative Counterfactual Explanation Problem

$$
\begin{aligned}
\text{(RCEP)}: \quad \min_{x,c,A,b} \quad & \delta\Big((c, A, b), (\hat{c}, \hat{A}, \hat{b})\Big) \\
s.t. \quad & c^\top x \le \alpha \hat{c}^\top \hat{x}, \\
& Ax \ge b, \\
& x \in D(\hat{x}), \\
& (c, A, b) \in \mathcal{H}, \\
& x \ge 0.
\end{aligned}
$$

Computations: Diet Problem

Supplier asks:

"How much do the prices (and nutrition values) have to change such that at least 200g of the food basket are products from the local market?"

Mutable Parameters	**relative CE**		**weak CE**		**strong CE**	
	CE	t in s (opt. gap)	CE	t in s (opt. gap)	CE	t in s (opt. gap)
only prizes	$p_{\text{Lentils}}: \downarrow 41.9\%$	1.4	$p_{\text{Lentils}}: \downarrow 50.0\%$ $p_{\text{Wheat}}: \uparrow 25.2\%$	600.0 (11.4%)	$p_{\text{Lentils}}: \downarrow 50.0\%$ $p_{\text{Wheat}}: \uparrow 25.2\%$	600.0 (100%)
prizes and nutrition values	$p_{\text{Lentils}}: \downarrow 41.9\%$	4.52	$p_{\text{Lentils}}: \downarrow 24.9\%$ RiboflavinB2(mg) in Milk : $\downarrow 28.9\%$ RiboflavinB2(mg) in WSB : $\downarrow 1.7\%$	600.0 (68.1%)	$p_{\text{Milk}}: \uparrow 2.2\%$ $p_{\text{Lentils}}: \downarrow 16.4\%$ $p_{\text{WSB}}: \uparrow 30.8\%$ Fat(g) in Oil : $\uparrow 18.6\%$ RiboflavinB2(mg) in Milk : $\downarrow 49.6\%$ RiboflavinB2(mg) in WSB : $\downarrow 5.0\%$	600.0 (100%)

Conclusion

Summary

- Three different types of counterfactual explanations: weak, strong and relative
- Weak and strong CEs not tractable: open and non-convex feasible regions possible
- Relative CEs can be calculated by solving a convex/linear optimization problem
- Solution time of the same order of magnitude as the linear problem

Future Research Directions

- Faster and more stable solution methods for strong and weak CEs
- Algorithm specific CEs
- CEs for integer optimization problems with mutable constraint parameters

For more details

Kurtz, J., Birbil, Ş. İ., & Hertog, D. D. (2024). **Counterfactual Explanations for Linear Optimization**. arXiv preprint arXiv:2405.15431.

Robust Optimization Webinar

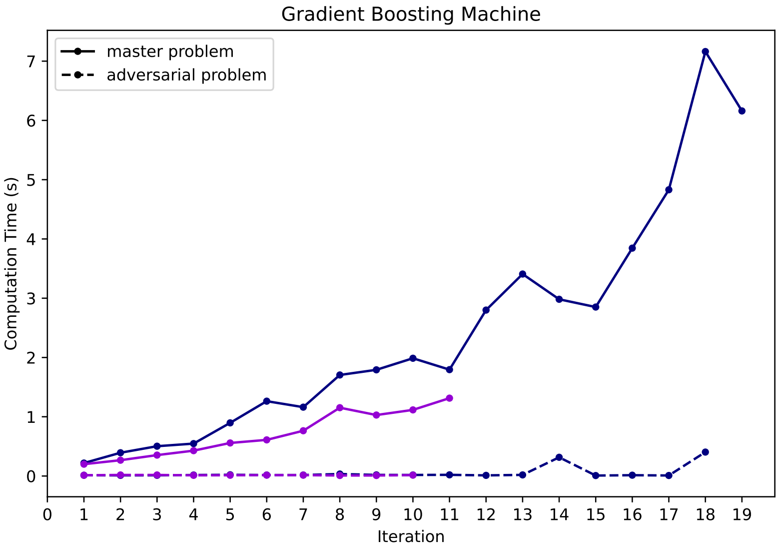
January 24	**Peyman Mohajerin Esfahani**	Delft University of Technology
February 7	**Lasse Wulf**	Technical University of Denmark
February 21	**Eojin Han**	University of Notre Dam
March 7	**Dorothee Henke**	University of Passau
	Yasmine Beck	ESSEC Business School
March 21	**Francesco Leofante**	Imperial College London
April 4	**Marco C. Campi**	University of Brescia
May 2	**Francesca Maggioni**	University of Bergamo
May 16	**Bahar Taskesen**	University of Chicago Booth School of Business
May 30	**Zhao Kang**	Eindhoven University of Technology
	Abbas Khademi	University of Tehran
June 13	**Ian Yihang Zhu**	NUS Business School
June 27	**Zihang Qiu**	University of Amsterdam
	Taylan Kargin	Caltech

Webpage

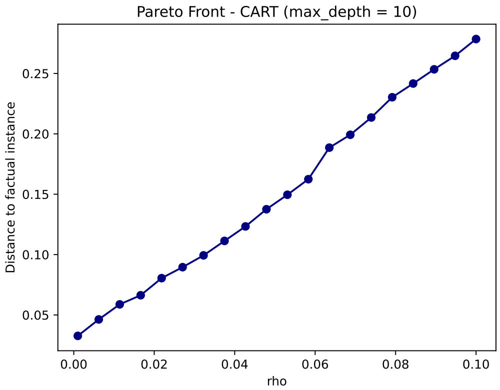
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

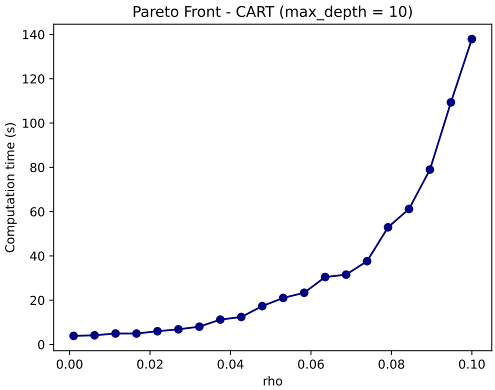
UNIVERSITY OF AMSTERDAM

Time per Iteration



Experiments





Strong Counterfactual Explanations

Theorem

Assume the favorable solution space is given by

$$D(\hat{x}) = \{x \geq 0 : Wx \leq h\}.$$

Then, SCEP has the same objective value as the following problem:

$$\begin{aligned}
(SCEP'): \inf_{c,A,b,\Lambda,\Gamma,\tau} \quad & \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \\
s.t. \quad & -\Lambda b + \Gamma c \leq h, \\
& c\tau^\top - A^\top \Lambda^\top \geq W^\top, \\
& -b\tau^\top + A\Gamma \geq 0, \\
& (c,A,b) \in \mathcal{H}, \\
& \Lambda \in \mathbb{R}_+^{q\times m}, \Gamma \in \mathbb{R}_+^{q\times n}, \tau \in \mathbb{R}_+^{q}.
\end{aligned}$$

Strong Counterfactual Explanations

The latter formulation is bilinear and has some undesired properties:

Properties

- Projection of the feasible region on the (c, A, b)-space may be **open**.
- Projection of the feasible region on the (c, A, b)-space may be **non-convex and disconnected**.

Relative Counterfactual Explanations

The latter formulation is bilinear and has some more positive properties:

Properties

- Projection of the feasible region on the (c, A, b)-space is **closed**.
- **Hidden convexity**, leading to a convex problem for certain classes of objective functions.

Relative Counterfactual Explanations

Assumption 1

The mutable parameter space $\mathcal{H}$ is compact, convex and columnwise, i.e., $\mathcal{H} = \mathcal{H}_1 \times \dots \mathcal{H}_n \times \mathcal{H}_{n+1}$.

Assumption 2

The favored solution space $\mathcal{D}(\hat{x})$ is a compact and convex set.

Assumption 3

The distance measure δ is continuous and columnwisely separable, i.e.,

$$\delta\left((c, A, b), (\hat{c}, \hat{A}, \hat{b})\right) = \sum_{j=1}^{n} \delta_j\left((c_j, A_j), (\hat{c}_j, \hat{A}_j)\right) + \delta_{n+1}(b),$$

Relative Counterfactual Explanations

Linearization

Apply the following variable transformation:

$$\begin{aligned} w_j &= c_j x_j, \\ u_{ij} &= a_{ij} x_j, \quad \forall i \in [m], \end{aligned}$$

Theorem

After the variable transformation the RCEP can be modeled as a convex optimization problem for distance measures of one of the following forms:

- $\delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b})$
- $\max_{j \in [n]} \delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b})$
- $\sum_{j=1}^{n} x_j \delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b})$
- $\dots$

Relative CEs for NETLIB Instances

n	m	# mut. col.	Bilinear formulation with obj. $\sum_{j=1}^{n} \delta_j((c, A), (\hat{c}, \hat{A}))$ $\|c-\hat{c}\|_0$	$\|A-\hat{A}\|_0$	Linearized formulation with obj. $\sum_{j=1}^{n} \boxed{x_j} \delta_j((c, A), (\hat{c}, \hat{A}))$ $\|c-\hat{c}\|_0$	$\|A-\hat{A}\|_0$
small	small	1	0.25	2.24	0.18	1.14
		5	0.46	4.39	0.20	1.20
		10	0.42	6.08	0.17	1.34
	medium	1	0.63	2.30	0.55	1.04
		5	0.84	4.56	0.45	2.03
		10	0.76	5.31	0.34	2.88
medium	small	1	0.49	2.13	0.38	1.66
		5	0.50	4.08	0.09	2.17
		10	1.06	10.92	0.00	1.99
	medium	1	0.33	2.59	0.31	0.92
		5	0.42	3.23	0.31	0.98
		10	0.58	4.31	0.32	0.96
	large	1	0.49	1.37	0.47	0.75
		5	0.63	2.02	0.48	1.02
		10	0.80	3.60	0.53	1.16
large	small	1	1.00	0.00	1.00	0.00
		5	1.26	0.00	1.26	0.00
		10	1.45	0.05	1.20	0.05
	medium	1	0.44	1.05	0.42	0.93
		5	0.55	2.09	0.48	0.97
		10	0.75	6.15	0.71	1.25
	large	1	0.42	1.73	0.41	1.64
		5	0.56	1.57	0.47	1.39
		10	0.89	1.78	0.69	1.64

Number of parameters to change

Relative CEs for NETLIB Instances

			Bilinear formulation with obj. $\sum_{j=1}^{n} \delta_j((c,A),(\hat{c},\hat{A}))$			Linearized formulation with obj. $\sum_{j=1}^{n} \boxed{x_j}\delta_j((c,A),(\hat{c},\hat{A}))$			Present Problem	
n	m	# mut. col.	Hit TL (in %)	t (in sec.)	t infeas. (in sec.)	Hit TL (in %)	t (in sec.)	t infeas. (in sec.)	t (in sec.)	t setup (in sec.)
small	small	1	0.00	0.19	0.13	0.00	0.14	0.13		
		5	1.00	29.19	0.13	0.00	0.15	0.13	0.13	0.12
		10	2.00	43.97	0.14	0.00	0.17	0.15		
	medium	1	0.00	0.42	0.31	0.00	0.31	0.30		
		5	1.00	16.09	0.31	0.00	0.33	0.31	0.32	0.30
		10	3.00	53.40	0.32	0.00	0.36	0.33		
medium	small	1	0.00	0.79	0.78	0.00	0.53	0.77		
		5	0.00	1.49	nan	0.00	0.54	nan	0.53	0.49
		10	4.00	109.16	nan	0.00	0.58	nan		
	medium	1	0.00	0.99	0.45	0.00	0.50	0.44		
		5	0.00	11.28	0.44	0.00	0.49	0.44	0.50	0.45
		10	1.00	32.76	0.45	0.00	0.51	0.45		
	large	1	0.00	1.64	1.03	0.00	1.04	1.03		
		5	1.00	21.10	1.00	0.00	1.05	1.01	1.14	1.06
		10	2.00	51.35	1.01	0.00	1.08	1.04		
large	small	1	0.00	11.91	8.40	0.00	11.10	8.27		
		5	0.00	11.43	8.48	0.00	10.50	8.28	8.22	7.98
		10	0.00	12.17	7.27	0.00	11.08	7.19		
	medium	1	0.00	2.17	1.40	0.00	0.90	1.36		
		5	0.00	3.48	1.39	0.00	0.92	1.35	1.34	1.24
		10	0.00	15.65	1.42	0.00	0.95	1.38		
	large	1	1.00	94.58	28.78	0.00	4.90	5.36		
		5	3.00	117.39	24.37	0.00	4.42	5.40	17.35	3.18
		10	6.00	202.81	27.19	0.00	4.95	4.90		