Finding Regions of Counterfactual Explanations via Robust Optimization

Donato Maragno
Jannis Kurtz
Tabea Röber
Rob Goedhart
Ilker Birbil
Dick den Hertog

University of Amsterdam
Amsterdam Business School

January 2023

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

Given **individual information** of a new customer, decide if the customer should be **granted a loan or not**.

Binary Classification Problems

Classification Problem

Given a data point x in a data space $\mathcal{X} \subseteq \mathbb{R}^n$, **classify** it as 1 or -1.

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

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

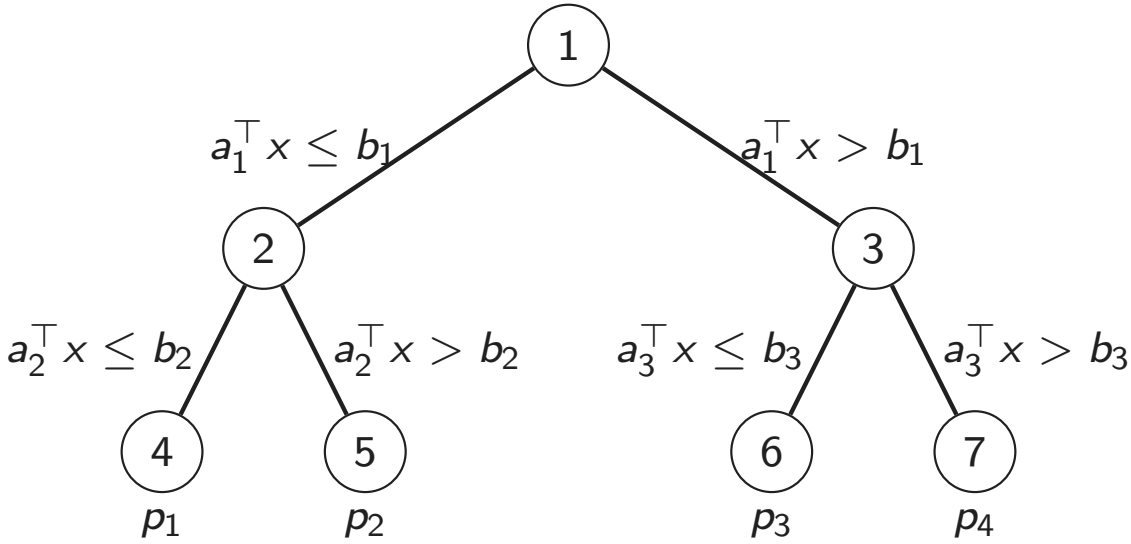
where $\tau \in (0, 1)$ is a given threshold parameter (often $\tau = 0.5$).

Models

Classifier h can be a

- Logistic Regression Function
- Classification Tree
- Neural Network

Classification Trees



Training Classification Trees

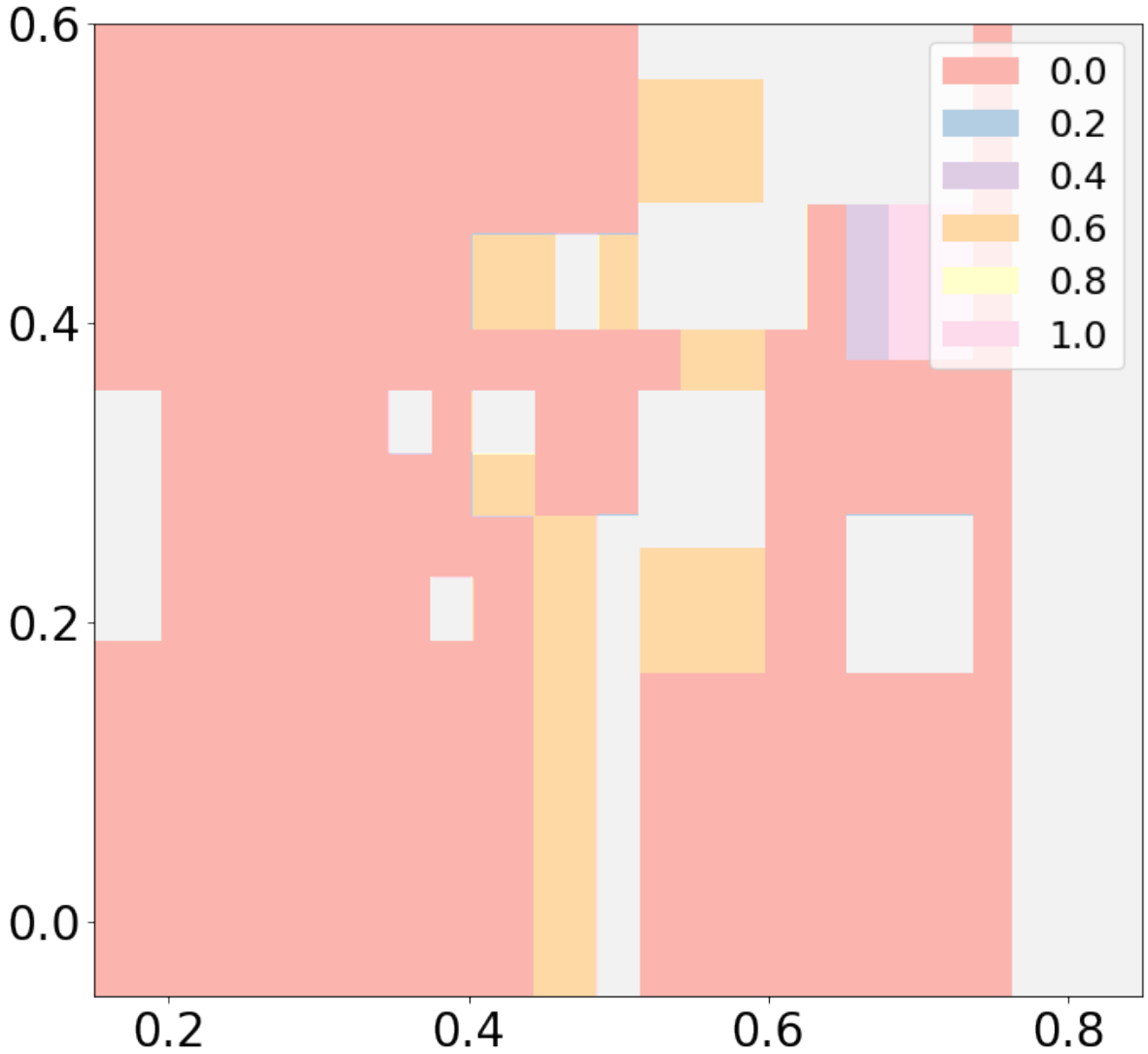
- CART (single feature splits) [Breiman et al. (1984)]
- Optimal Decision Trees (hyperplane splits) [Bertsimas, Dunn (2015)]

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

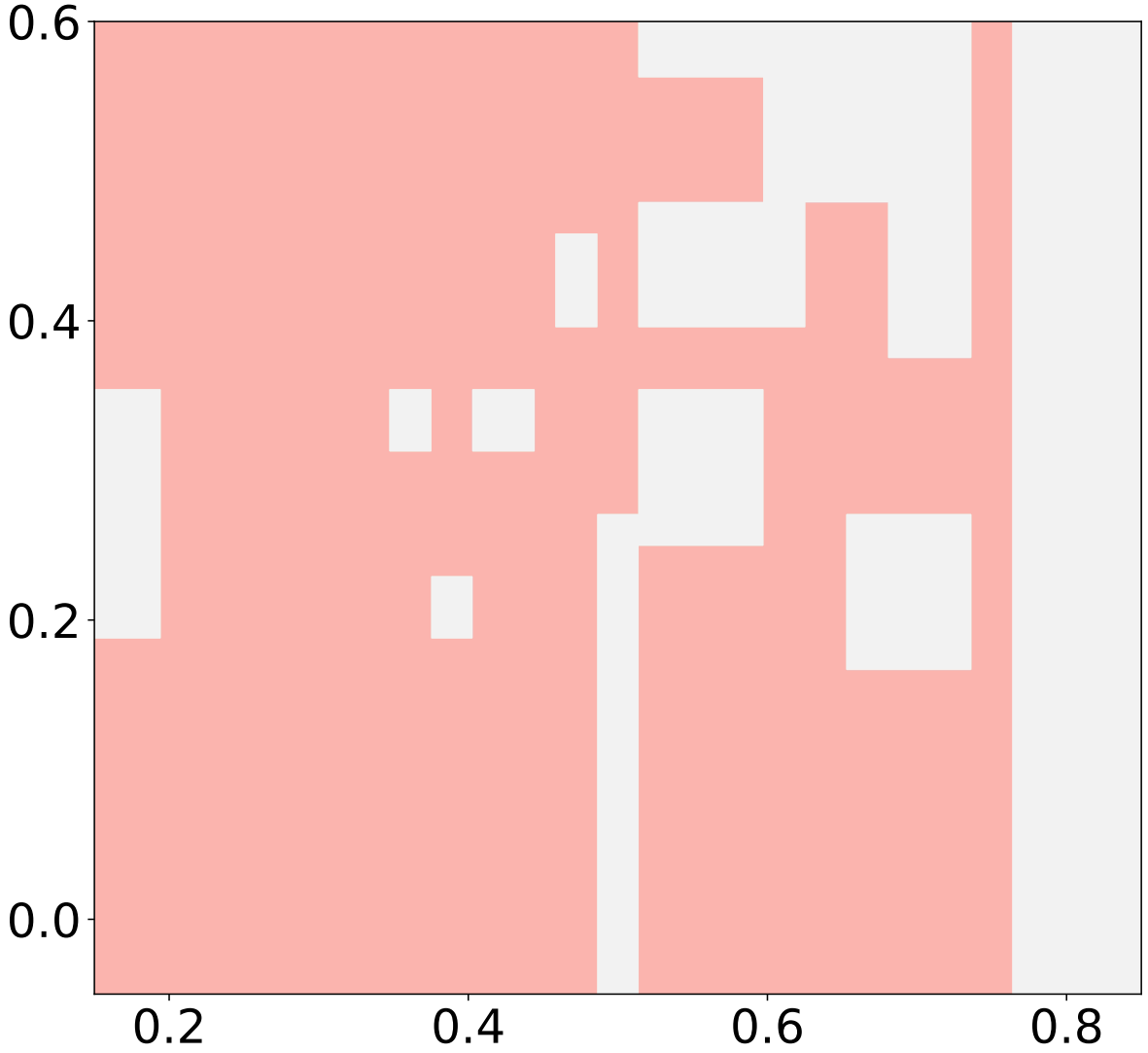
Classification Trees

Prediction values of a classification tree trained by CART on the Iris dataset



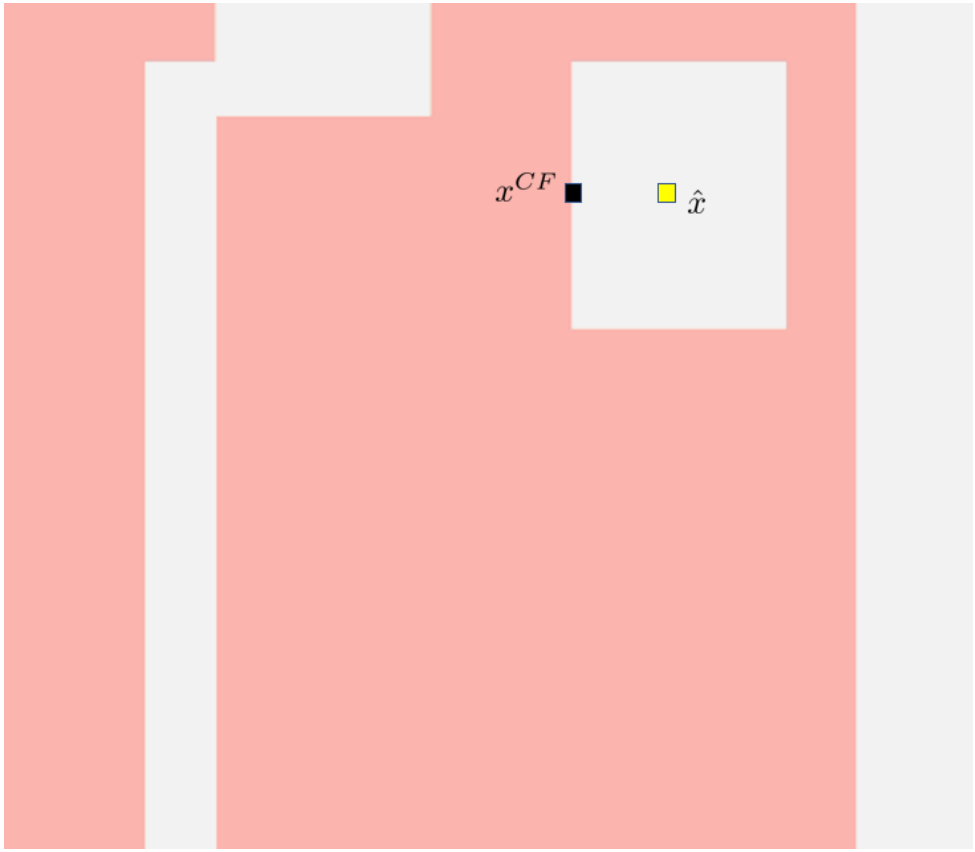
Classification Trees

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



Closest Counterfactual Point

Given a trained classifier h and a **factual point** $\hat{x}$ predicted as -1, we are looking for the **closest counterfactual point**, i.e. the closest x predicted as 1.



Optimization Problem

Counterfactual can be calculated by solving

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

Heuristic Method

$$\min_{x} \; \max_{\lambda} \; d(\hat{x}, x) + \lambda(h(x) - \tau)$$

[Wachter et al. (2017)]

Example

Credit Scoring

- Assume a person with individual information $\hat{x}$ does **not** get granted a loan.

Age	Salary	Current Balance	$\cdots$
28	42.000 EUR	8.200 EUR	$\cdots$

- Calculated counterfactual point:

Age	Salary	Current Balance	$\cdots$
28	44.500 EUR	10.000 EUR	$\cdots$

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

Right to Explanation

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

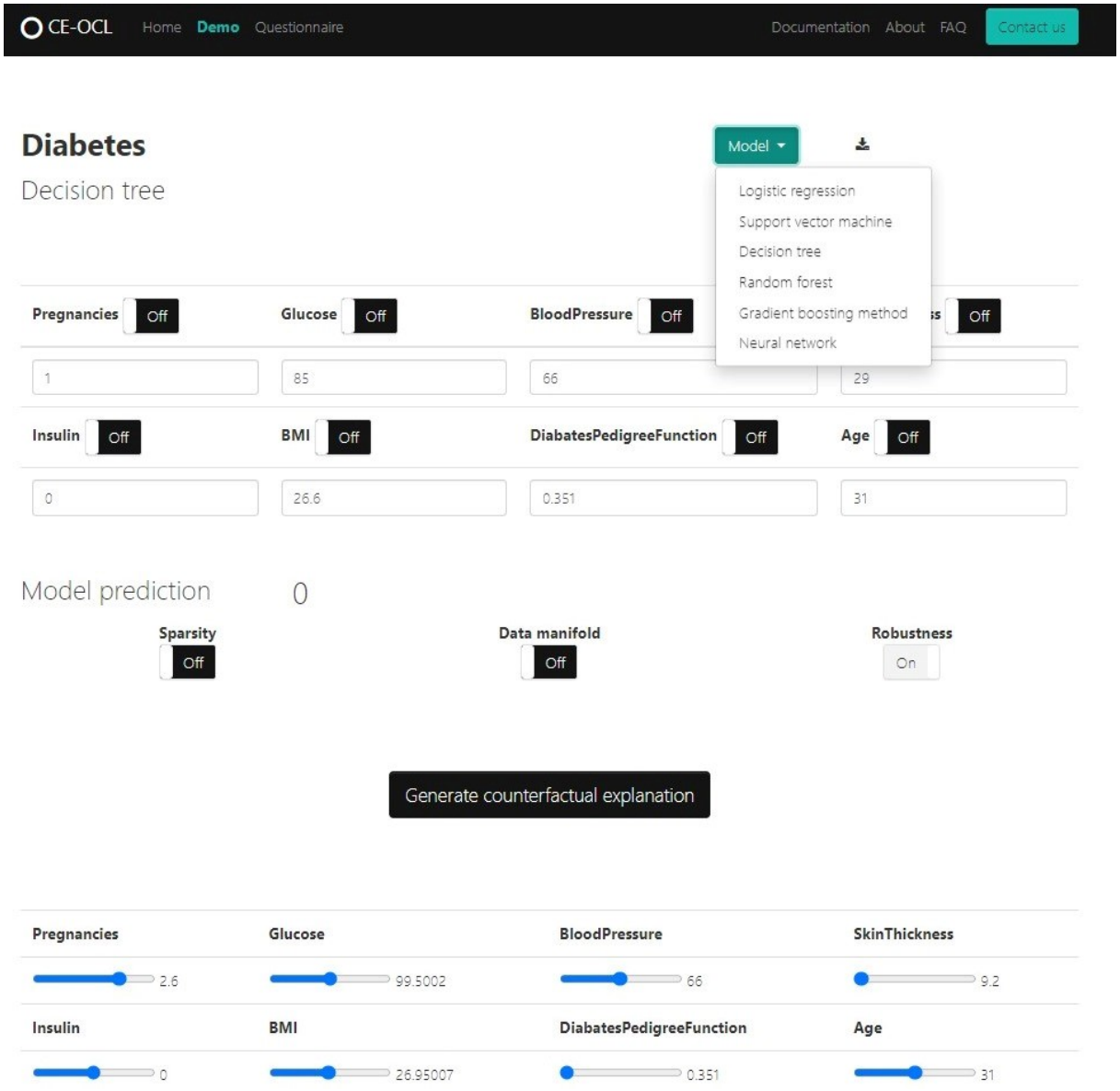
> [...] In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **obtain an explanation of the decision reached after such assessment** and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

Robust Counterfactual Explanations

Motivation

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

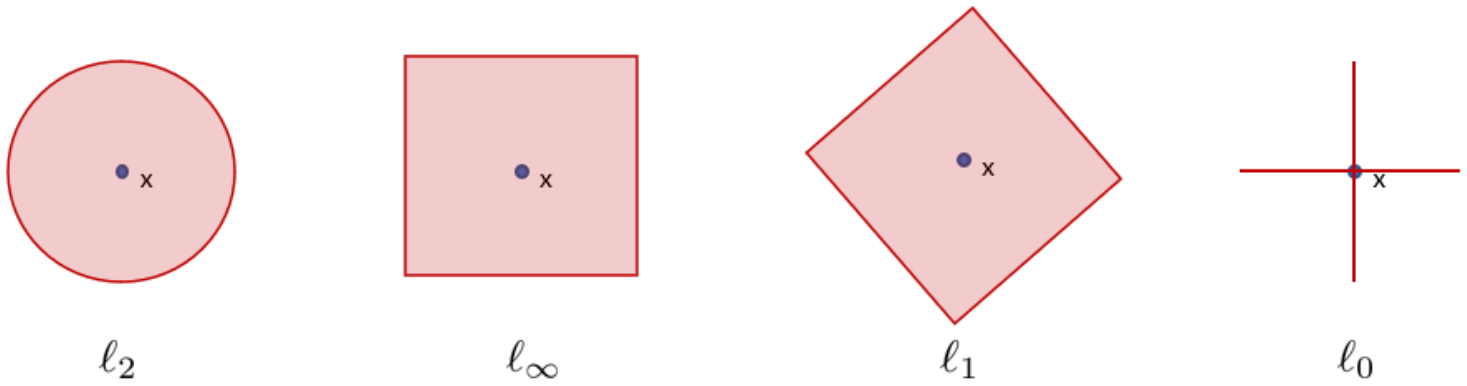


Setup

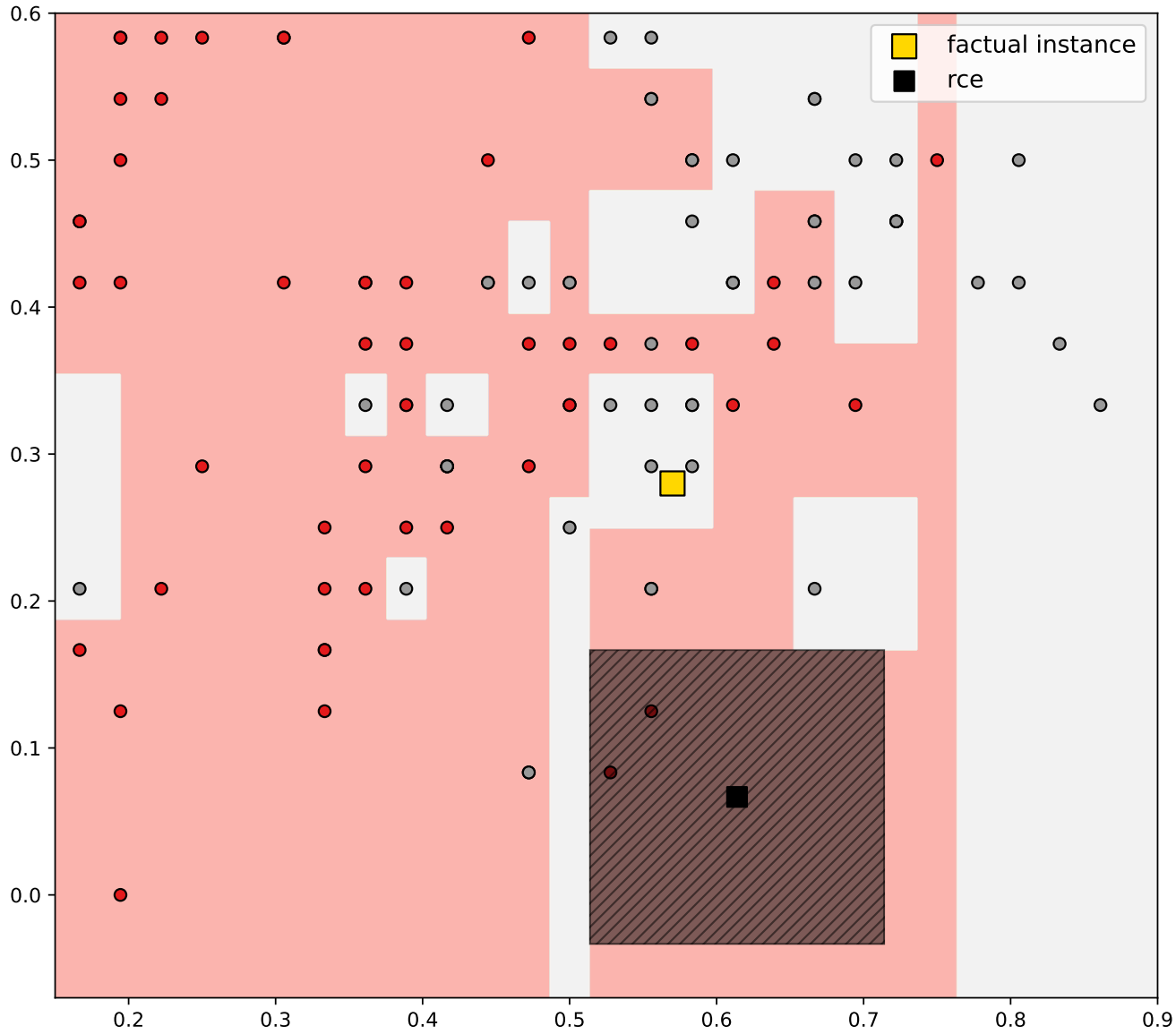
- Find counterfactual point x^{CE} such that all points in

$$x^{CE} + \mathcal{S}$$

  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



Example: box uncertainty



Robust Counterfactual Explanations

Literature

- Gradient based heuristic with probabilistic robustness guarantees [Pawelczyk et al. (2022)]
  - heuristic solution
  - no full robustness guarantee
  - not applicable to classification trees
- Gradient based heuristic for robust CE problem [Dominguez-Olmedo et al. (2021)]
  - heuristic solution
  - no global optimal solutions, hence no robustness guarantee
  - not applicable to classification trees

Goal

Derive a method

- which guarantees full robustness
- which is also applicable to decision trees and random forests

Optimization Problem

Given a factual instance $\hat{x}$ a **robust counterfactual explanation** can be derived by solving

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

where $\hat{x}$ is the factual instance.

Difficulties

- Infinitely many constraints
- Duality trick from robust optimization does not work for all h.
- Iterative approach needed: **adversarial approach**

Adversarial Approach

Master Problem

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

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

- 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^*:

$$\begin{aligned} \max\ & \tau - h(x^* + s) \\ s.t.\quad & s \in \mathcal{S} \end{aligned}$$

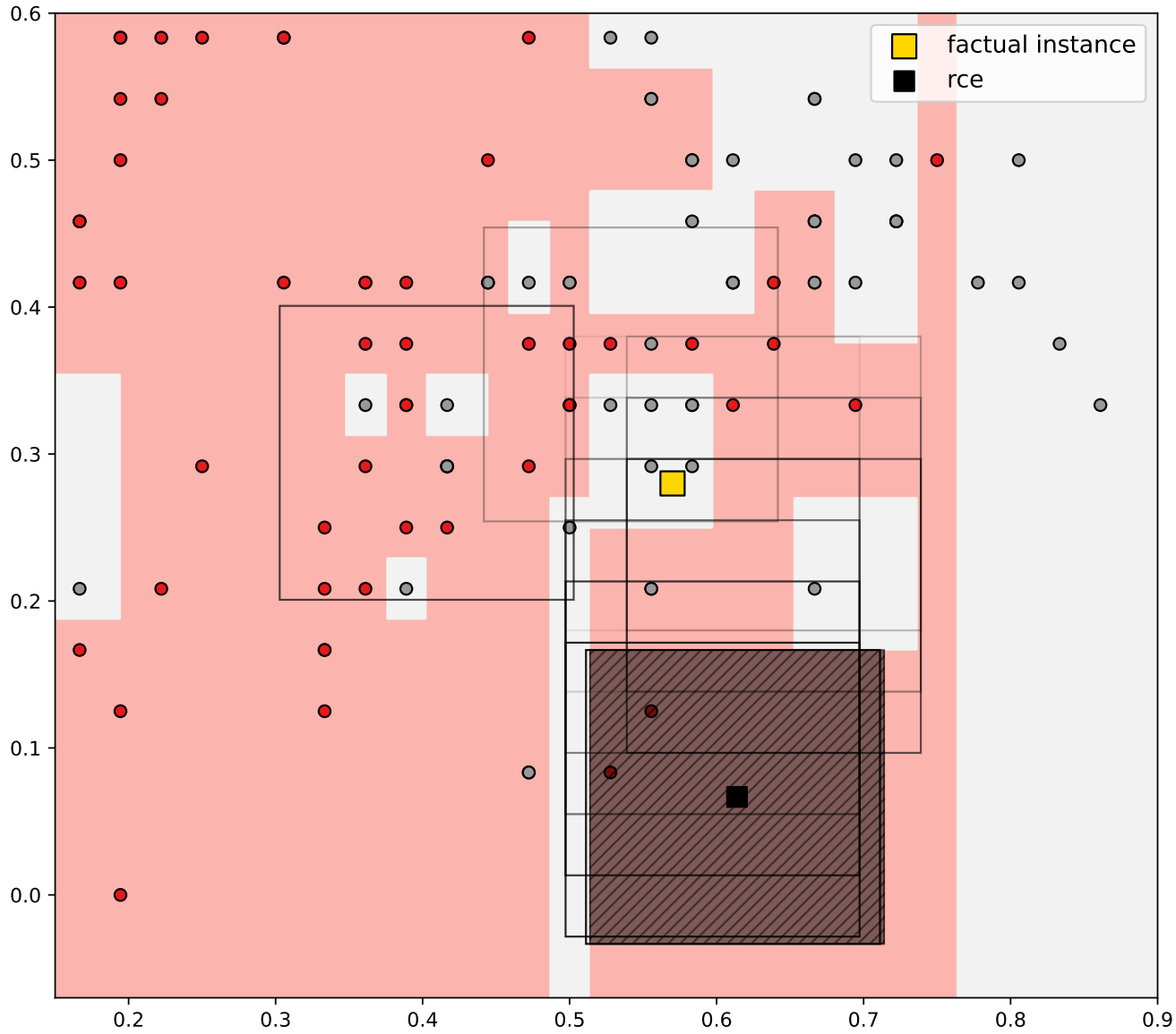
- Current solution x^* is cut-off if optimal value > 0

Adversarial Approach

Algorithm

1. **Input**: $\hat{x}$, $\mathcal{S}$, $\varepsilon > 0$
2. $\mathcal{Z} = \{0\}$
3. **Repeat:**
4. $x^* \leftarrow \mathit{MasterProblem}(\mathcal{Z})$
5. $s^*, \text{opt} \leftarrow \mathit{AdversarialProblem}(x^*, \mathcal{S})$
6. $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{s^*\}$
7. **Until:** $\text{opt} \leq \varepsilon$
8. **Return**: x^*

Adversarial Approach



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

$$h(x^* + s) \geq \tau - \varepsilon$$

for all $s \in \mathcal{S}$.

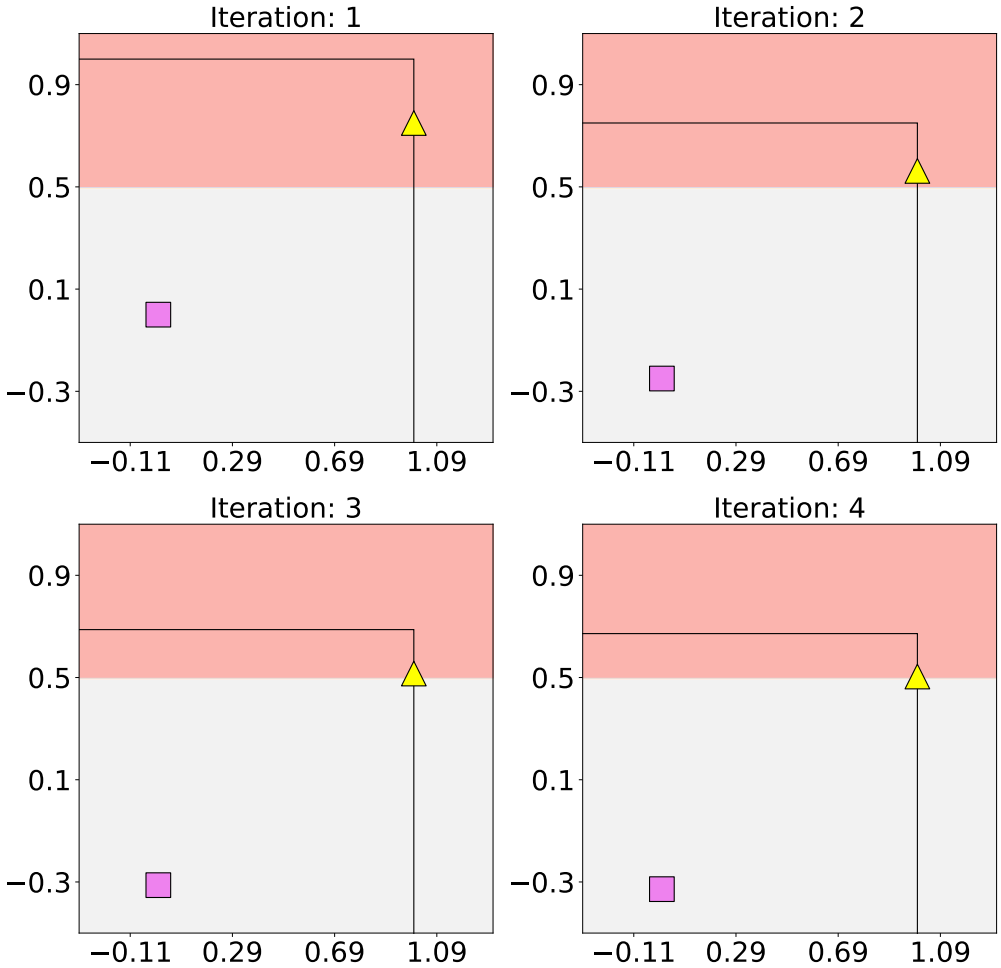
Lipschitz Continuity

Classifier function h of

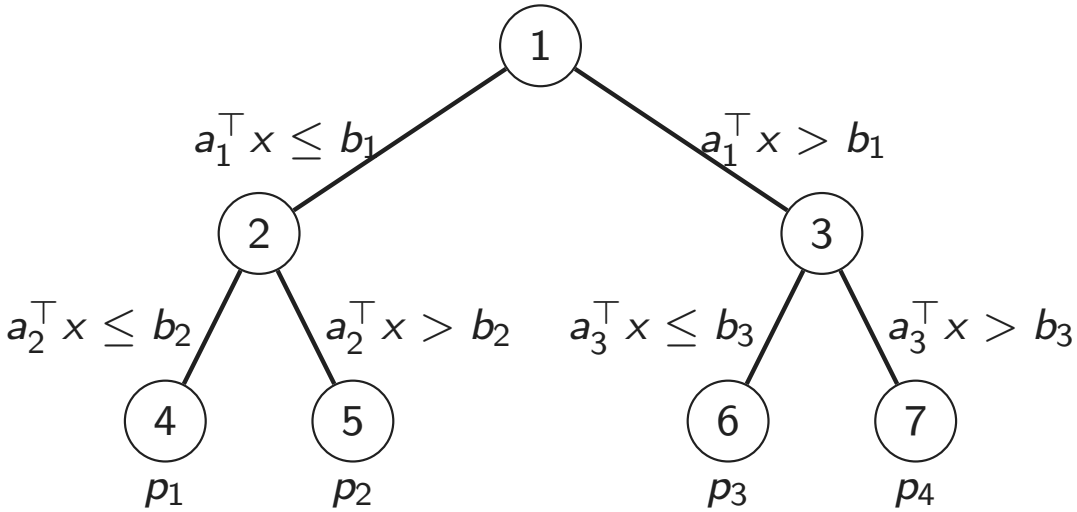
- **logistic regression** is Lipschitz continuous.
- **neural networks** with ReLU activation functions is Lipschitz continuous.
- **decision trees** is **not even continuous**!

Lipschitz Continuity is Needed

- Classifier $h : \mathbb{R}^2 \to [0, 1]$ with $h(x) = 0$ if $x_2 \geq \frac{1}{2}$ and $h(x) = 1$ otherwise.
- Factual instance is $\hat{z} = (0, 2)$
- Optimal MP solution in iteration i: $x^i = (0, -\sum_{j=1}^{i} \left(\frac{1}{4}\right)^j)$
- Optimal AP solution in iteration i: $s^i = (1, \frac{1}{2} + \sum_{j=1}^{i} \left(\frac{1}{4}\right)^j)$



Decision Trees



Define the continuous function

$$\tilde{h}(x) := \begin{cases} \tau, & x \in \mathcal{L}_i, p_i \ge \tau; \\ \max\limits_{j \in \mathcal{N}^i_{\le} \cup \mathcal{N}^i_{<}} \{\tau + a_j^\top x - b_j\}, & x \in \mathcal{L}_i, p_i < \tau. \end{cases}$$

where $\mathcal{L}_i = \left\{ x : a_j^\top x \le b_j, a_k^\top x < b_k, j \in \mathcal{N}^i_{\le}, k \in \mathcal{N}^i_{<} \right\}$.

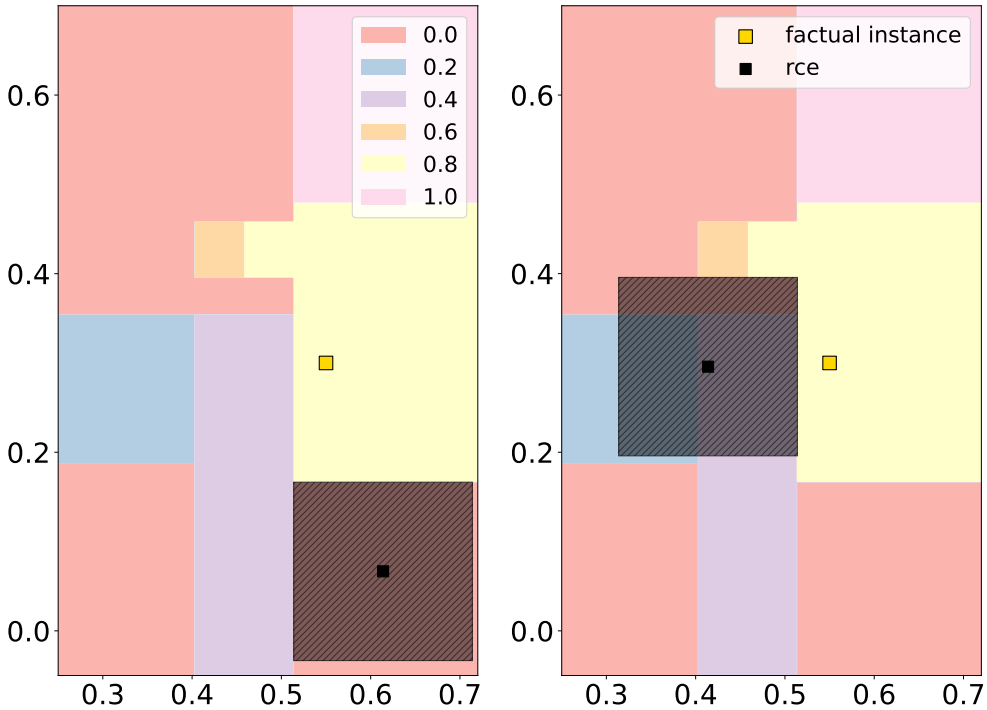
Lemma

The function $\tilde{h}$ is Lipschitz continuous and classifies each point which is not on the boundary of a leaf the same as the original decision tree.

Master Problem: Decision Trees

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

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

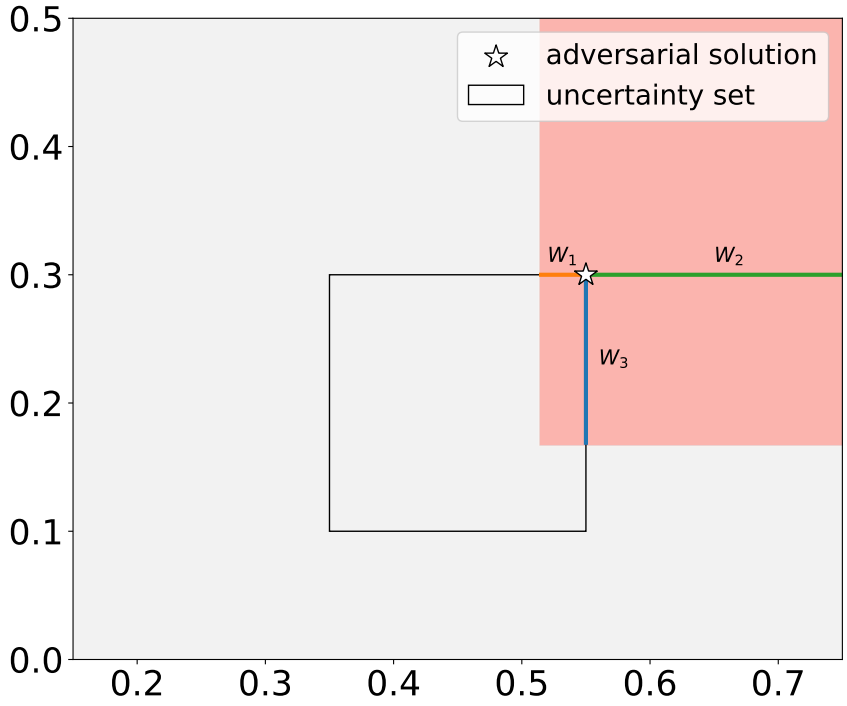


Adversarial Problem: Decision Trees

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.



Adversarial Approach: Decision Trees