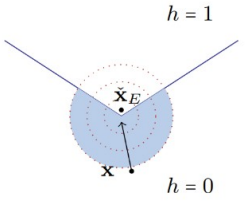
Probabilistically Robust Recourse: Navigating the Tradeoffs Between Costs and Robustness in Algorithmic Recourse

Authors: Martin Pawelczyk, Teresa Datta, Johannes van-den-Heuvel, Gjergji Kasneci, Himabindu Lakkaraju (2022)

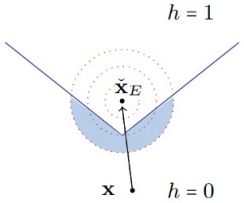
Presentation: Dan Ley, Tessa Han, Yasha Ektefaie

Motivation

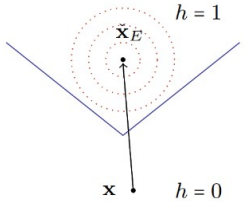
- Algorithmic recourse aims to provide users with actionable changes to move from a negative to a positive prediction in a machine learning (ML) model
- Typically, the minimum cost change is computed
- In the real-world, these changes are often implemented noisily
  - E.g. an individual who was asked to increase their salary by $500 may get a promotion which comes with a raise of $505 or even $499.95
  - Or, as seen on Monday, changing certain factors may cause others to change



(a) Low recourse cost and low recourse robustness



(b) Medium recourse cost and medium recourse robustness

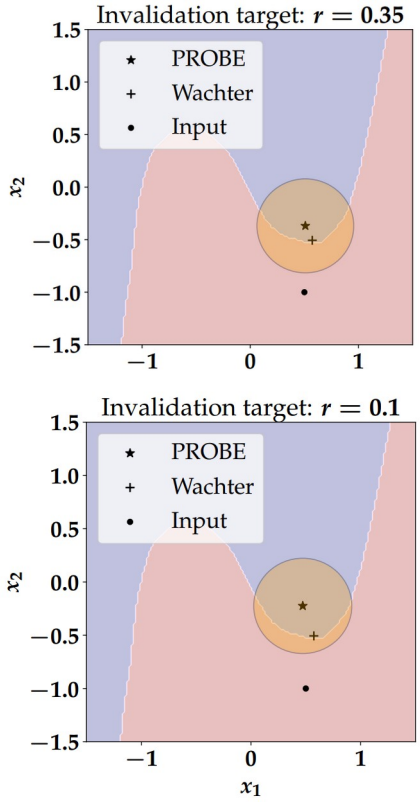


(c) High recourse cost and high recourse robustness

Related Work

- Counterfactual Explanations and Recourse
  - A plethora of papers including Wachter et al. (2018)
- Summarised in Verma et al. (2020)
  - Type of the underlying predictive model (e.g. tree based vs. differentiable classifier)
  - Whether they encourage sparsity in counterfactuals (i.e. a small number of features)
  - Whether counterfactuals should lie on the data manifold
  - Whether the underlying causal relationships should be accounted for when generating counterfactuals
- Robustness to model/data shift:
  - ROAR - Robust Algorithmic Recourse, Upadhyay et al. (2021)
- Robustness to input perturbations:
  - Causal Algorithmic Recourse, Dominguez-Olmedo et al. (2022)
- All approaches generate recourses assuming that the prescribed recourses will be correctly implemented by users

Solution



- PROBE - Probabilistically Robust Recourse
- Allows users to manage the recourse cost vs. robustness trade-offs
- Users can choose the **recourse invalidation rate**
  - Probability with which a recourse could get invalidated
- Baselines include:
  - Upadhyay et al. (2021) - ROAR, targeting recourses that are robust to model changes
  - Dominguez-Olmedo et al. (2022) - targeting recourses that are robust to input changes
- PROBE recourses are, compared to the baselines:
  - Less costly than previous methods
  - More robust to noisy implementations
  - Able to provide recourses at various invalidation rates

Notation

$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ input space

$\mathcal{Y} = \{0, 1\}$ output space

- 0: unfavourable outcome
- 1: favourable outcome

$h : \mathcal{X} \to \mathcal{Y}$ classifier

$h(\mathbf{x}) = g(f(\mathbf{x}))$

$f : \mathcal{X} \to \mathbb{R}$ inputs → logits

$g : \mathbb{R} \to \mathcal{Y}$ logits → binary labels

General formulation of algorithmic recourse

$$\check{\mathbf{x}} = \underset{\mathbf{x}' \in \mathcal{A}}{\arg\min}\ \underbrace{\ell\big(h(\mathbf{x}'), 1\big)}_{\text{make CF have favourable outcome}} + \lambda \cdot \underbrace{d_c(\mathbf{x}, \mathbf{x}')}_{\text{low cost}}$$

- Accounts for CF having favourable outcome and low cost
- Does not account for potential noise in the implemented counterfactual
  - Addressed by this paper

Recourse invalidation rate

$$\Delta(\check{\mathbf{x}}_E) = \mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\underbrace{h(\check{\mathbf{x}}_E)}_{\textit{CF class}} - \underbrace{h(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon})}_{\textit{class after response}}\Big]$$

where $\boldsymbol{\varepsilon} \sim p_{\boldsymbol{\varepsilon}}$ → probability distribution that captures noise in response

e.g. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

Recourse invalidation rate aware objective function

$$\mathcal{L} = R(\mathbf{x}'; r, \sigma^2\mathbf{I}) + \ell\Big(f(\mathbf{x}'), s\Big) + \lambda d_c(\mathbf{x}', \mathbf{x})$$

make IR of CF close to target IR (new term)

make CF have favourable outcome (seen before)

low cost (seen before)

where $R(\mathbf{x}'; r, \sigma^2\mathbf{I}) = \max(0, \Delta(\mathbf{x}'; \sigma^2\mathbf{I}) - r)$

CF's IR

target IR

Approximation of recourse invalidation rate (Theorem 1)

Problem: $\Delta(\mathbf{x}')$ is not differentiable (used in objective function)

Solution: use a first order approximation of $\Delta(\mathbf{x}')$

$$\tilde{\Delta}(\check{\mathbf{x}}_E; \sigma^2\mathbf{I}) = 1 - \Phi\left(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)^\top \sigma^2\mathbf{I} \nabla f(\check{\mathbf{x}}_E)}}\right)$$

where

$\check{\mathbf{x}}_E$: counterfactual

$f(\check{\mathbf{x}}_E)$: logit at counterfactual

Proof sketch:

1. solve $\mathbb{P}\left(f(\check{\mathbf{x}}_E + \boldsymbol{\epsilon}) > 0\right)$
2. Use first order Taylor series to approximate logit
3. Calculate probability that normal r.v. is less than a value → CDF

Approximation of recourse invalidation rate (Theorem 1): further analyses

- Proposition 1 — Wachter et al. (2018) method for logistic regression
  - Derive closed form solution for IR of CF
  - Show how to make CF more robust

- Proposition 2 — PROBE recourse incurs an additional cost (linear regression)

- Proposition 3 — Upperbound on IR

Experimental Evaluation: Datasets considered

- Adult Dataset: Predict whether an individual has an income greater than 50,000 USD/year
- Give Me Some Credit: Predict whether an individual will experience financial distress within the next two years or not
- COMPAS: Predict if criminal is high or low risk of re-offending

Experimental Evaluations: Baselines considered

- Baseline Methods
  - Growing Spheres (GS)
    - Random search algorithm: generate observations until decision boundary is crossed then move greedily toward decision boundary
  - AR (-LIME) Local interpretable model-agnostic explanations (LIME)
    - Actional Recourse in Linear Models (AR) → use integer programming
  - Diverse Counterfactual Explanations (DICE)
    - Promote diversity of counterfactual explanations
  - Gradient
    - General formulation of algorithmic recourse
- Adversarial Minmax Objectives Methods
  - Robust Algorithm Recourse (ROAR)
    - Recourse robust to model changes by generating counterfactuals that minimize worst-case loss over plausible model shifts
  - **A**dversarial **R**obustness of Casual **A**lgorithmic **R**ecourse (ARAR)
    - Recourse robust to features of individual (in the time individual seeks implement recourse other features may change)

Experimental Evaluation: Measures Used

- Average Cost (AC): Average cost for all prescribed recourses for the test set (implemented as l1-norm)
- Recourse Accuracy (RA): For all prescribed recourses what fraction results in the desired prediction
- Average IR (AIR): Average recourse invalidation rate for all prescribed recourses in test set

Experimental Evaluation: Results

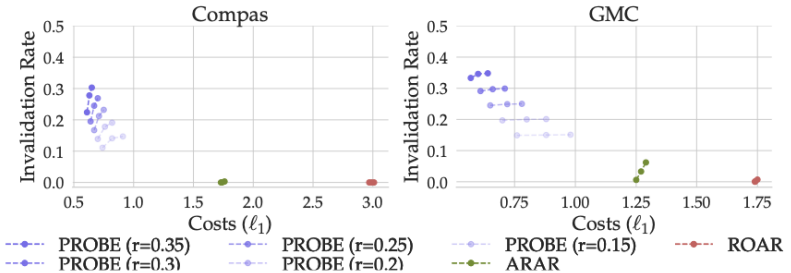
		Adult				Compas				GMC			
	Measures	AR	Wachter	GS	PROBE	AR	Wachter	GS	PROBE	AR	Wachter	GS	PROBE
LR	RA (↑)	0.98	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	AIR (↓)	0.5 ± 0.01	0.46 ± 0.02	0.35 ± 0.11	**0.34 ± 0.02**	0.48 ± 0.04	0.47 ± 0.02	0.3 ± 0.18	**0.28 ± 0.02**	0.47 ± 0.06	0.45 ± 0.03	0.48 ± 0.04	**0.24 ± 0.01**
	AC (↓)	**0.55 ± 0.4**	0.62 ± 0.43	2.12 ± 1.05	1.56 ± 0.92	**0.16 ± 0.17**	0.22 ± 0.17	0.73 ± 0.45	0.63 ± 0.39	**0.29 ± 0.27**	0.49 ± 0.51	**0.28 ± 0.31**	0.60 ± 0.56
NN	RA (↑)	0.38	**1.0**	**1.0**	**0.99**	0.84	**1.0**	**1.0**	**1.0**	0.38	**1.0**	**1.0**	**1.0**
	AIR (↓)	0.49 ± 0.03	0.5 ± 0.02	0.48 ± 0.02	**0.35 ± 0.01**	0.34 ± 0.09	0.46 ± 0.02	0.43 ± 0.07	**0.33 ± 0.02**	0.34 ± 0.07	0.43 ± 0.03	0.45 ± 0.03	**0.25 ± 0.03**
	AC (↓)	1.05 ± 0.22	**0.3 ± 0.19**	2.99 ± 1.51	1.43 ± 0.49	1.15 ± 0.52	**0.2 ± 0.16**	0.81 ± 0.45	0.8 ± 0.34	0.2 ± 0.19	0.26 ± 0.18	**0.12 ± 0.09**	0.47 ± 0.21

(a) Comparing PROBE to baseline recourse methods.

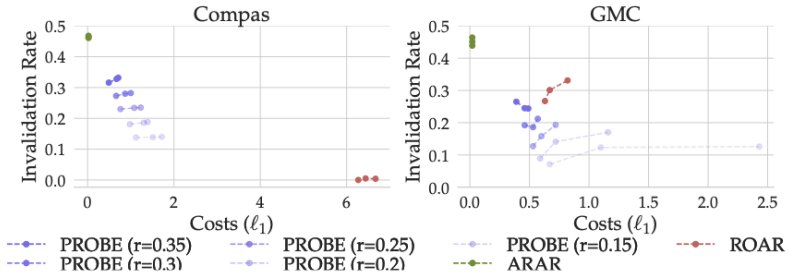
		Adult			Compas			GMC		
	Measures	ROAR	ARAR	PROBE	ROAR	ARAR	PROBE	ROAR	ARAR	PROBE
LR	RA(↑)	**1.0**	**1.0**	**1.0**	**1.0**	0.99	**1.0**	**1.0**	**1.0**	**1.0**
	AIR (↓)	**0.0 ± 0.0**	0.02 ± 0.01	0.34 ± 0.02	**0.0 ± 0.0**	**0.0 ± 0.0**	0.28 ± 0.02	**0.0 ± 0.0**	0.35 ± 0.01	0.24 ± 0.01
	AC (↓)	3.56 ± 0.8	2.68 ± 0.79	**1.56 ± 0.92**	2.99 ± 0.31	1.74 ± 0.3	**0.63 ± 0.39**	1.74 ± 0.45	1.27 ± 0.45	**0.60 ± 0.56**
NN	RA(↑)	0.94	0.03	**0.99**	0.97	0.02	**1.0**	0.06	0.06	**1.0**
	AIR (↓)	**0.0 ± 0.0**	0.51 ± 0.0	0.35 ± 0.01	**0.01 ± 0.06**	0.46 ± 0.0	0.33 ± 0.02	0.3 ± 0.21	0.45 ± 0.01	**0.25 ± 0.03**
	AC (↓)	19.8 ± 3.39	0.04 ± 0.0*	**1.43 ± 0.49**	6.41 ± 1.07	0.02 ± 0.0*	**0.8 ± 0.34**	0.67 ± 0.94	0.02 ± 0.0*	**0.47 ± 0.21**

(b) Comparing PROBE to adversarially robust recourse methods.

Experimental Evaluation: Results



(a) Logistic Regression



(b) Neural Network

Conclusions, Future work

- First work to navigate tradeoff between recourse cost and robustness, give control to the user
- Experimental evidence demonstrates usefulness of framework and the existence of tradeoff
- Future work: generate recourse that is simultaneously robust to noise in inputs and shifts in model parameters

Discussions/Limitations

- “Generate recourse that is simultaneously robust to noise in inputs and shifts in model parameters”
  - Is there not a tradeoff still? Would the user have to provide different robustness thresholds for the different aspects the method is robust with respect to? Or is there one threshold to rule them all?

- Would users be okay with this risk of defining r% invalidation? Bias? How does recourse relate to bias?

- In general this idea of cost being described as l1 distance across methods seems to be a major limitation. Not all features incur similar contributions to cost (i.e. getting a degree versus putting more money into your savings account, one is harder than the other). Cost is a difficult concept to quantify.