Columbia University Statistics Seminar, November 13, 2023

The Risks of Recourse in Explainable Machine Learning

Tim van Erven

University of Amsterdam
Korteweg de Vries Institute for Mathematics

Joint work with:

Hidde Fokkema
(University of Amsterdam)

Damien Garreau
(Université Côte d'Azur)

Outline

Explainable Machine Learning

The Need for Explanations:

Why did the machine learning system

- Classify my company as high risk for money laundering?
- Reject my bank loan?
- Predict this patient can safely leave the intensive care?
- Mistake a picture of a husky for a wolf?
- Reject the profile picture I uploaded to get a public transport card?[1]
- ...

[1] Personal experience

Explainable Machine Learning

The Need for Explanations:

Why did the machine learning system

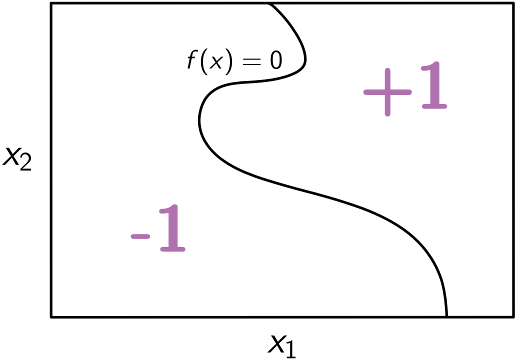
- Classify my company as high risk for money laundering?
- Reject my bank loan?
- Predict this patient can safely leave the intensive care?
- Mistake a picture of a husky for a wolf?
- Reject the profile picture I uploaded to get a public transport card?[1]
- ...

Information-Theoretic Constraints:

- Cannot communicate millions of parameters!
- Can communicate only some **relevant aspects** and/or need **high-level concepts** in common with user

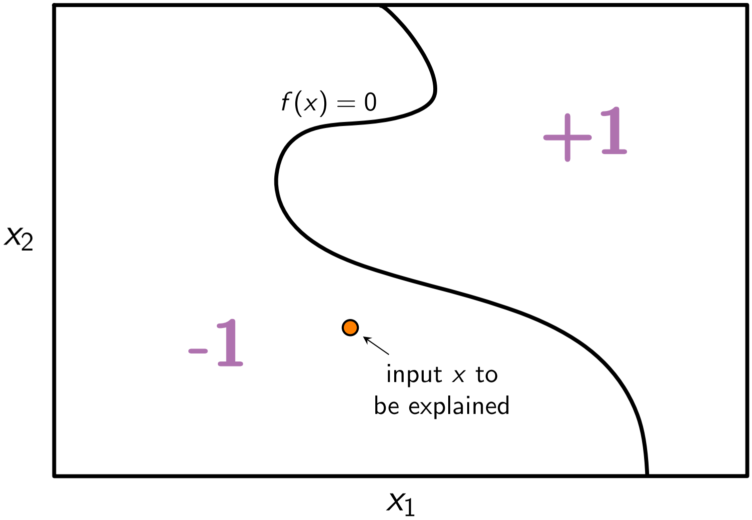
[1] Personal experience

Machine Learning: Binary Classification



- Goal: classify an input $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ as class -1 or class $+1$
- Usually by **thresholding a real-valued classifier** $f : \mathbb{R}^d \to \mathbb{R}$, e.g. predicted class is $\mathrm{sign}(f(x))$
- Classifier f obtained by minimizing error on **training data**

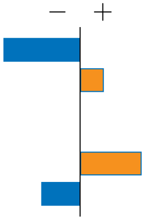
Local Post-hoc Explanations



- **Local:** only explain the part of f that is **(most) relevant for** x
- **Post-hoc:** ignore explainability concerns when estimating f

Local Explanations via Attributions

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d-1} \\ x_d \end{bmatrix}$$

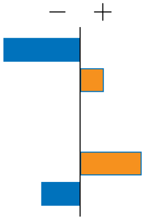


$$\begin{bmatrix} \varphi_f(x)_1 \\ \varphi_f(x)_2 \\ \vdots \\ \varphi_f(x)_{d-1} \\ \varphi_f(x)_d \end{bmatrix} = \varphi_f(x)$$

$\phi_f(x) \in \mathbb{R}^d$ attributes a **weight to each feature**, which explains **how important** the feature is **for the classification of x by f**.

Local Explanations via Attributions

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d-1} \\ x_d \end{bmatrix}$$



$$\begin{bmatrix} \varphi_f(x)_1 \\ \varphi_f(x)_2 \\ \vdots \\ \varphi_f(x)_{d-1} \\ \varphi_f(x)_d \end{bmatrix} = \varphi_f(x)$$

$\phi_f(x) \in \mathbb{R}^d$ attributes a **weight to each feature**, which explains **how important** the feature is **for the classification of x by f**.

Example: low d, linear f

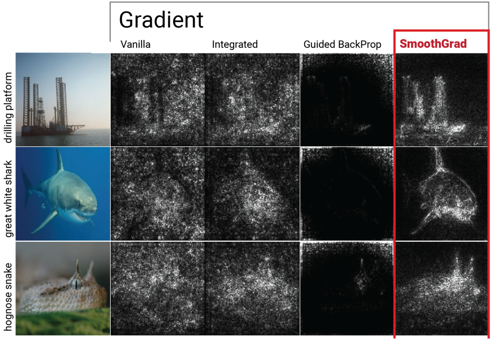
$$f(x) = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

$\phi_f(x)_i = \theta_i$ could be **coefficient** of x_i

- NB This example is **too simple!** In general $\phi_f(x)$ will depend on x. But many methods can be viewed as local linearizations of f.

Example: Gradient-based Explanations

Various gradient methods[2]



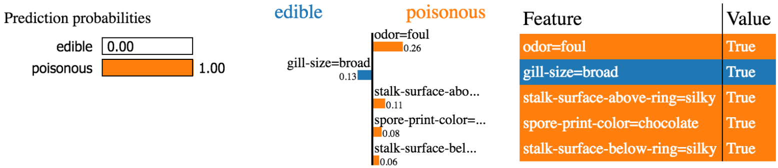
- Vanilla gradient: $\phi_f(x) = \nabla f(x)$
- SmoothGrad: $\phi_f(x) = \mathbb{E}_{Z \sim \mathcal{N}(x, \Sigma)}[\nabla f(Z)]$ (Smilkov et al., 2017)
- ...

[2]Image source: (Smilkov et al., 2017)

Example: LIME

LIME (Ribeiro, Singh, and Guestrin, 2016): Do local linear approximation of f near x (optionally in dimensionality reduced space), and report coefficients

LIME for tabular data:[3]



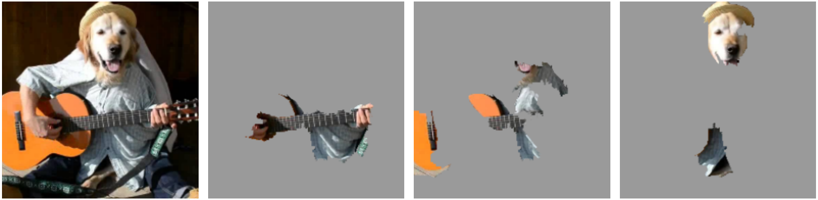
(classifying edibility of mushrooms)

[3]Image source: `https://github.com/marcotcr/lime`

Example: LIME

LIME (Ribeiro, Singh, and Guestrin, 2016): Do local linear approximation of f near x (optionally in dimensionality reduced space), and report coefficients

LIME for images:[3]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[3]Image by Ribeiro, Singh, and Guestrin (2016)

Exciting Times to Work on Explainability

Lots of open issues:

- Easily **manipulated**
- Explanation methods often **disagree**
- Plausible looking explanations may **not represent** model being explained (Adebayo et al., 2018)
- Unclear for **which goal** approximation methods are useful

Image by Dombrowski et al., 2019

LIME Method

SHAP Method

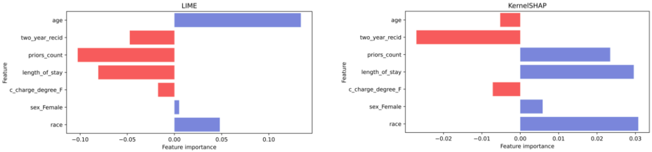
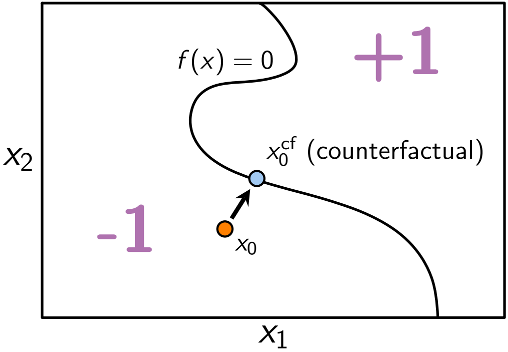


Image by Krishna et al., 2022

Outline

Counterfactual Explanations

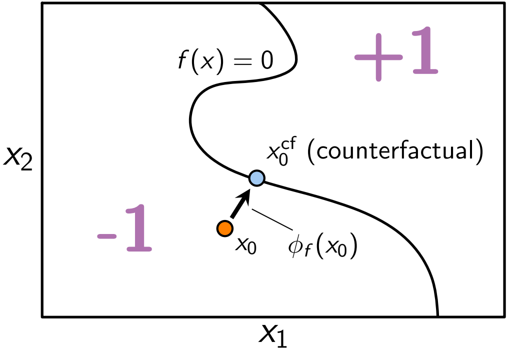
"If you would have had an income of **€40 000** instead of **€35 000**, your loan request would have been approved."



Counterfactual explanation: $x_0^{cf} = \underset{x:\mathrm{sign}(f(x))=+1}{\arg\min} \mathrm{dist}(x, x_0)$

Counterfactual Explanations

"If you would have had an income of **€40 000** instead of **€35 000**, your loan request would have been approved."



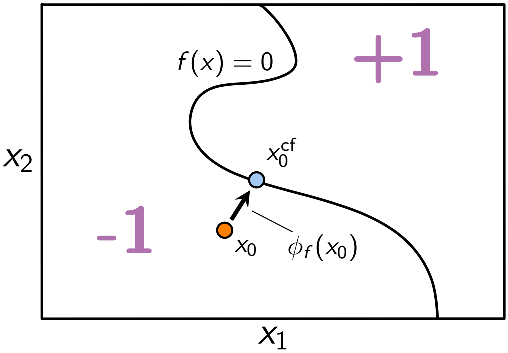
Counterfactual explanation: $$x_0^{\text{cf}} = \underset{x:\text{sign}(f(x))=+1}{\arg\min}\ \text{dist}(x, x_0)$$

Viewed as attribution method[4]: $$\phi_f(x_0) = x_0^{\text{cf}} - x_0$$

[4] Gives scaled coefficients $\phi_f(x_0)_i = \frac{\text{dist}(x_0^{\text{cf}}, x_0)}{\|\theta\|}\theta_i$ if f is linear

Explanations with Recourse as their Goal

"If you change your current income of **€35 000** to **€40 000**,
then your loan request will be approved."



- Counterfactual methods **provide recourse** by telling the user how to **change their features** such that f **takes their desired value**.

More Realistic Variations

Literature background:

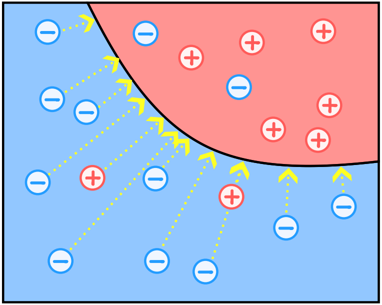
- Original counterfactuals (Wachter, Mittelstadt, and Russell, 2017)
- Robust counterfactuals: if users implement recourse approximately, they should still switch class (Ustun, Spangher, and Liu, 2019)
- Causal models:
  - User can only changes features indirectly via causal model of their actions (Karimi et al., 2021)
  - Steer towards actions that truly improve probability of desired class, not just classifier decision (König, Freiesleben, and Grosse-Wentrup, 2023)

Most discussion in the literature at the **level of individuals**.

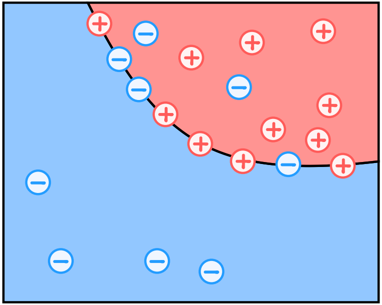
What is the effect at the population level?

Outline

Effect of Recourse on the Population



Before recourse



After recourse

What happens to the accuracy of the classifier?

Accuracy matters!

Example: incorrect $+1$ classifications $=$ users defaulting on loans

Effect of Recourse

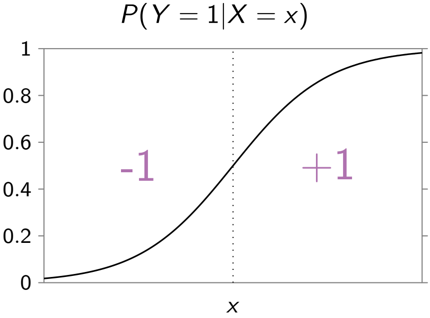
Situation before Recourse:

- User distribution: $(X_0, Y) \sim P$
- Classifier $f : \mathcal{X} \to \{-1, +1\}$
- Risk: $R_P(f) = P(f(X_0) \neq Y)$

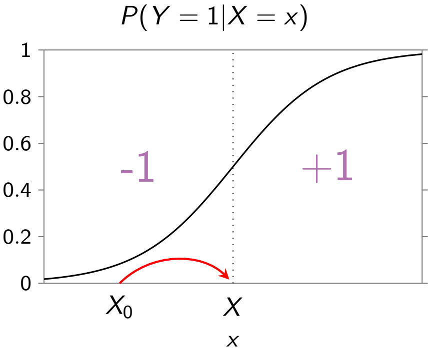
Effect of Recourse:

- User features change from X_0 to X
- Need to model use behavior: how does distribution of Y change?

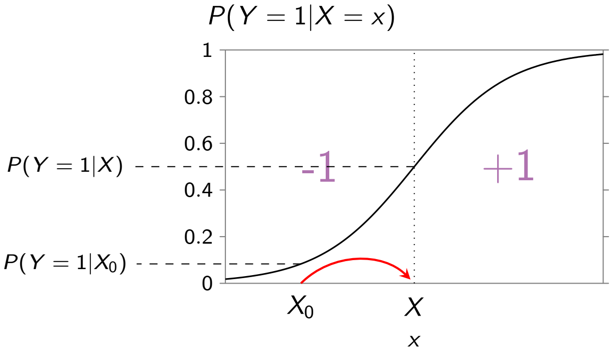
Modeling User Behavior



Modeling User Behavior



Modeling User Behavior



- **Compliant users:** probability of Y after recourse is $P(Y|X)$
- **Defiant users:** probability of Y after recourse is $P(Y|X_0)$

Modeling User Behavior

Examples:

- Credit loan application:
  - Compliant: Applicant improves risky behaviour
  - Defiant: Applicant tries to "game the system"
- Medical Diagnosis:
  - Compliant: Patient improves their health
  - Defiant: Patient takes medicine to reduce symptoms
- Job applications:
  - Compliant: Applicant improves their skills
  - Defiant: Applicant improves their CV

- **Compliant users:** probability of Y after recourse is $P(Y|X)$
- **Defiant users:** probability of Y after recourse is $P(Y|X_0)$

Learning-theoretic Framework

Situation before Recourse:

- User distribution: $(X_0, Y) \sim P$
- Classifier $f : \mathcal{X} \to \{-1, +1\}$
- Risk: $R_P(f) = P(f(X_0) \neq Y)$

Learning-theoretic Framework

Situation before Recourse:

- User distribution: $(X_0, Y) \sim P$
- Classifier $f : \mathcal{X} \to \{-1, +1\}$
- Risk: $R_P(f) = P(f(X_0) \neq Y)$
- Users' choice to accept recourse is $B \in \{0, 1\}$ with $\Pr(B = 1 | X_0) = r(X_0)$.

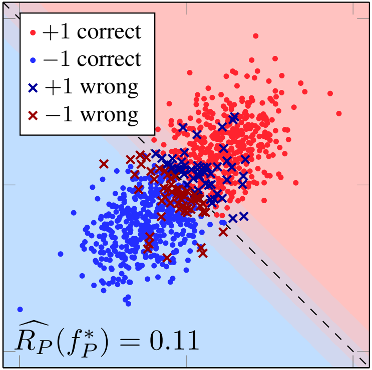
Situation with Recourse:

- Users arrive as before: $X_0 \sim P$
- Recourse proposal: $X_0^{\mathrm{cf}} = \arg\min_{x : f(x) = +1} \|x - X_0\|$
- Users' choice to accept is $B \in \{0, 1\}$ with $\Pr(B = 1 | X_0) = r(X_0)$:

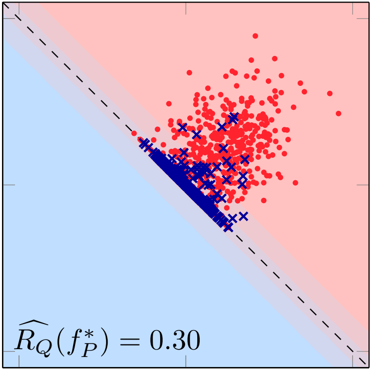
$$X = (1 - B)X_0 + BX_0^{\mathrm{cf}}$$

- Q is the resulting distribution of X_0, B, X, Y
- Risk: $R_Q(f) = Q(f(X) \neq Y)$

Effect of Recourse on Population-level Accuracy



Before recourse



After recourse
(compliant users)

- Simulation with Gaussian data
- **Average nr. of mistakes goes up / accuracy goes down**
- Many more customers defaulting on their loans!

Recourse Increases the Risk

Bayes-optimal classifier under P:

$$f_P^* = \arg\min_f R_P(f)$$

$$f_P^*(x) = \begin{cases} +1 & \text{if } P(Y = 1 | X_0 = x) \geq 1/2, \\ -1 & \text{otherwise.} \end{cases}$$

Recourse Increases the Risk

Bayes-optimal classifier under P:

$$f_P^* = \arg\min_f R_P(f)$$

$$f_P^*(x) = \begin{cases} +1 & \text{if } P(Y = 1 | X_0 = x) \geq 1/2, \\ -1 & \text{otherwise.} \end{cases}$$

Regularity conditions:

- Well-defined setup: $\{x \in \mathcal{X} : f_P^*(x) = +1\}$ is closed
- Continuous conditional probabilities: $P(Y = 1 | X_0 = x) = 1/2$ for all x on the decision boundary of f_P^*

Theorem

Then, both if the users are **defiant** *and if the users are* **compliant**, **recourse always increases the risk***:*

$$R_Q(f_P^*) \geq R_P(f_P^*).$$

The inequality is strict if the probability of recourse in the negative class is non-zero: $P(B = 1, f_P^*(X_0) = -1) > 0$.

Recourse Increases the Risk

Regularity conditions:

- Well-defined setup: $\{x \in \mathcal{X} : f_P^*(x) = +1\}$ is closed
- Continuous conditional probabilities: $P(Y = 1 | X_0 = x) = 1/2$ for all x on the decision boundary of f_P^*

Theorem

Then, both if the users are **defiant** *and if the users are* **compliant***,* **recourse always increases the risk***:*

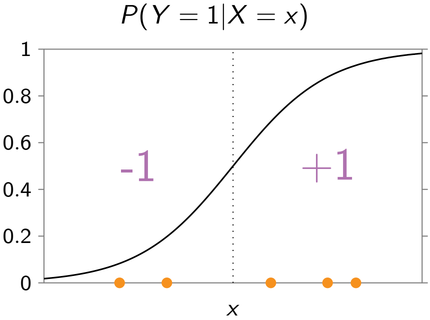
Defiant case:

$$
\begin{aligned}
&R_Q(f_P^*) - R_P(f_P^*) \\
&= P(B = 1, f_P^*(X_0) = -1, Y = -1) - P(B = 1, f_P^*(X_0) = -1, Y = +1) \\
&\geq 0.
\end{aligned}
$$

Compliant case:

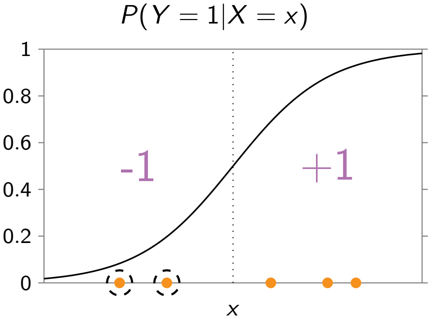
$$
\begin{aligned}
&R_Q(f_P^*) - R_P(f_P^*) \\
&= \tfrac{1}{2} P(B = 1, f_P^*(X_0) = -1) - P(B = 1, f_P^*(X_0) = -1, Y = 1) \geq 0.
\end{aligned}
$$

Proof Idea: Defiant Case



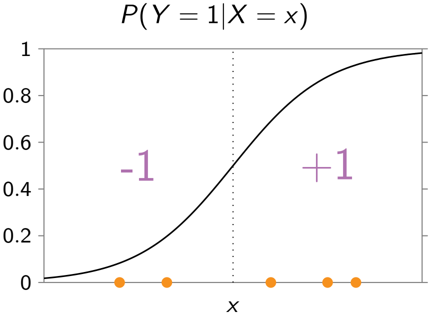
- Defiant case: $Q(Y|X, X_0) = P(Y|X_0)$

Proof Idea: Defiant Case

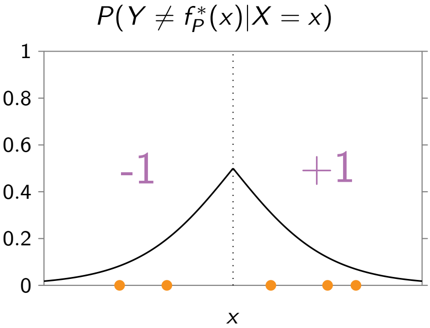


- Defiant case: $Q(Y|X, X_0) = P(Y|X_0)$
- Recourse misclassifies users from class -1 as class $+1$

Proof Idea: Compliant Case

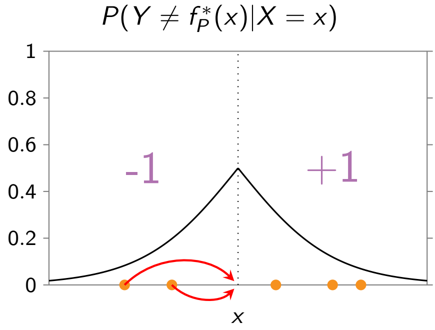


Proof Idea: Compliant Case



- Compliant case: $Q(Y|X, X_0) = P(Y|X)$

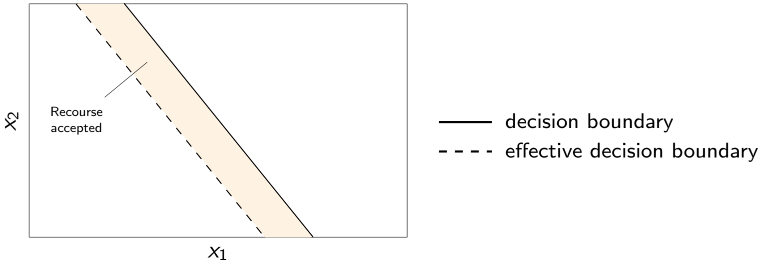
Proof Idea: Compliant Case



- Compliant case: $Q(Y|X, X_0) = P(Y|X)$
- Recourse moves users from high certainty to lowest certainty region

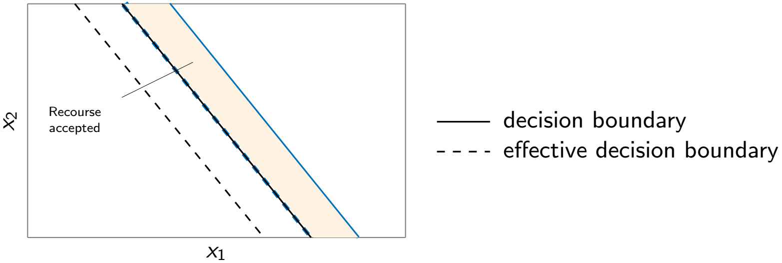
Outline

Strategic Classification



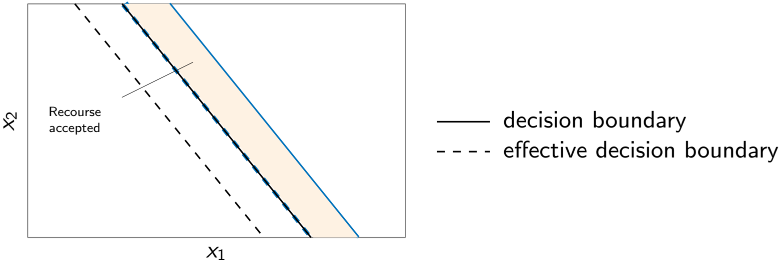
- Suppose recourse accepted deterministically within distance D of decision boundary

Strategic Classification



- Suppose recourse accepted deterministically within distance D of decision boundary
- **Cancel effect of recourse** by moving decision boundary back by distance D

Strategic Classification



- Suppose recourse accepted deterministically within distance D of decision boundary
- **Cancel effect of recourse** by moving decision boundary back by distance D

Definition

A set of classifiers $\mathcal{F}$ is **invariant under recourse** if for any $f \in \mathcal{F}$ there exists a **unique** $f' \in \mathcal{F}$ such that the decision boundary for f without recourse is equal to the effective decision boundary of f' with recourse.

Strategic Classification

Assumptions:

- $\mathcal{F}$ invariant under recourse

Theorem (Defiant Case)

Recourse has no effect:

$$\min_{f \in \mathcal{F}} R_{Q_f}(f) = \min_{f \in \mathcal{F}} R_P(f).$$

- Write Q_f instead of Q to emphasize dependence of the effect of recourse on f.

Strategic Classification

Assumptions:

- $\mathcal{F}$ invariant under recourse

Theorem (Compliant Case)

Recourse may have positive effect:

Let $\bar{f} \in \arg\min_{f \in \mathcal{F}} R_P(f)$ *with corresponding* $f' \in \mathcal{F}$ *that has the same effective decision boundary after recourse. Then*

$$\min_{f \in \mathcal{F}} R_{Q_f}(f) \leq R_{Q_{f'}}(f') = \min_{f \in \mathcal{F}} R_P(f) - \Delta,$$

where $\Delta = \Pr_{(X_0,Y)\sim P}(\bar{f}(X_0) \neq Y) - \Pr_{(X_0,Y)\sim Q_{f'}}(\bar{f}(X_0) \neq Y)$.

- Think of $Q_{f'}$ as moving users away from the decision boundary compared to P, so likely that $\Delta > 0$.
- Only case where we find that recourse is **beneficial** in terms of accuracy.
- But cancels the effect of recourse and does not help any users from the original -1 class. Not really what we imagined...

Summary

Algorithmic Recourse:

- Provides explanations that help users overturn an unfavorable decision by a machine learning system
- Standard example: rejected loan application

Effects of Providing Algorithmic Recourse:

- Classifier **accuracy gets (much) worse**
  - Not just for defiant users, but also for compliant users
- Strategizing may avoid reduced accuracy
  - But effect is: same customers get a loan, but some have to **jump through more hoops** to get it
  - Does not help any customers who originally did not get a loan

Discussion

Conclusion: Algorithmic recourse is **not reliably beneficial**

Remark:

- This seems **inherent to the goal**, so changing the method will not fix it

Discussion

Conclusion: Algorithmic recourse is **not reliably beneficial**

Remark:

- This seems **inherent to the goal**, so changing the method will not fix it

Possible ways forward:

1. Identify applications in which **classifier accuracy is less important** (for the people receiving recourse)
   - Not: the standard loan application example
   - Alternative: journal paper acceptance, profile picture acceptance for public transport card, ...

Discussion

Conclusion: Algorithmic recourse is **not reliably beneficial**

Remark:

- This seems **inherent to the goal**, so changing the method will not fix it

Possible ways forward:

1. Identify applications in which **classifier accuracy is less important** (for the people receiving recourse)
   - Not: the standard loan application example
   - Alternative: journal paper acceptance, profile picture acceptance for public transport card, . . .
2. Replace recourse by something else
   - For instance: **contestability**, which allows users to appeal incorrect decisions

References I

Fokkema, Garreau, Van Erven
The Risks of Recourse in Binary Classification
ArXiv::2306.00497, 2023

- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity Checks for Saliency Maps". In: *Advances in Neural Information Processing Systems*. Vol. 31.
- Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel (2019). "Explanations can be manipulated and geometry is to blame". In: *Advances in Neural Information Processing Systems*. Vol. 32.
- Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera (2021). "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects". In: *arXiv preprint arXiv:2010.04050*.
- Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup (2023). "Improvement-Focused Causal Recourse (ICR)". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10, pp. 11847–11855. DOI: 10.1609/aaai.v37i10.26398.

References II

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju (2022). "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In: *ArXiv 2202.01602 preprint*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin (2016). ""Why should I trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg (2017). "SmoothGrad: removing noise by adding noise". In: *ArXiv:1706.03825*.

Berk Ustun, Alexander Spangher, and Yang Liu (2019). "Actionable recourse in linear classification". In: *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31, p. 841.