University of Amsterdam
Korteweg de Vries Institute for Mathematics

Performativity and Risks of Algorithmic Recourse

2025-11-25

Performativity and Risks of Algorithmic Recourse

- Based on 2 papers:
  - **Risks of Recourse in Binary Classification,** presented at AISTATS 2024, together with Tim van erven & Damien Garreau
  - **Performative Validity of Recourse Explanations**, will be presented at NeurIPS 2025, together with Gunnar König, Timo Freiesleben, Celestine Mendler-Dünner and Ulrike von Luxburg

Programme of today

- Introduction to XAI and counterfactual explanations
- Modelling the setting: A statistical point of view
- Optimal classifiers & Strategising
- Modelling the setting: A causal point of view
- Performative validity and how to ensure it
- Conclusion

Introduction to XAI and the problem

With a peculiar example

Call for eXplainable AI (XAI)

A couple reasons

- Fairness:
  - Biases in your model could be detected earlier
- Could increase trust
- Could increase reliability
- Regulation

EU Artificial Intelligence Act

Explainability methods

Explosion

Methods
CAM with global average pooling [42], [91]
+ Grad-CAM [43] generalizes CAM, utilizing gradient
+ Guided Grad-CAM and Feature Occlusion [68]
+ Respond CAM [44]
+ Multi-layer CAM [92]
LRP (Layer-wise Relevance Propagation) [13], [53]
+ Image classifications. PASCAL VOC 2009 etc [45]
+ Audio classification. AudioMNIST [47]
+ LRP on DeepLight. fMRI data from Human Connectome Project [48]
+ LRP on CNN and on BoW(bag of words)/SVM [49]
+ LRP on compressed domain action recognition algorithm [50]
+ LRP on video deep learning, *selective relevance method* [52]
+ BiLRP [51]
DeepLIFT [57]
Prediction Difference Analysis [58]
Slot Activation Vectors [41]
PRM (Peak Response Mapping) [59]
LIME (Local Interpretable Model-agnostic Explanations) [14]
+ MUSE with LIME [85]
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [93]
Also listed elsewhere: [56], [69], [71], [94]
Others. Also listed elsewhere: [95]
+ Direct output labels. Training NN via multiple instance learning [65]
+ Image corruption and testing Region of Interest statistically [66]
+ Attention map with autofocus convolutional layer [67]
DeconvNet [72]
Inverting representation with natural image prior [73]
Inversion using CNN [74]
Guided backpropagation [75], [91]
Activation maximization/optimization [38]
+ Activation maximization on DBN (Deep Belief Network) [76]
+ Activation maximization, multifaceted feature visualization [77]
Visualization via regularized optimization [78]
Semantic dictionary [39]
Network dissection [36], [37]
Decision trees
Propositional logic, rule-based [82]
Sparse decision list [83]
Decision sets, rule sets [84], [85]
Encoder-generator framework [86]
Filter Attribute Probability Density Function [87]
MUSE (Model Understanding through Subspace Explanations) [85]

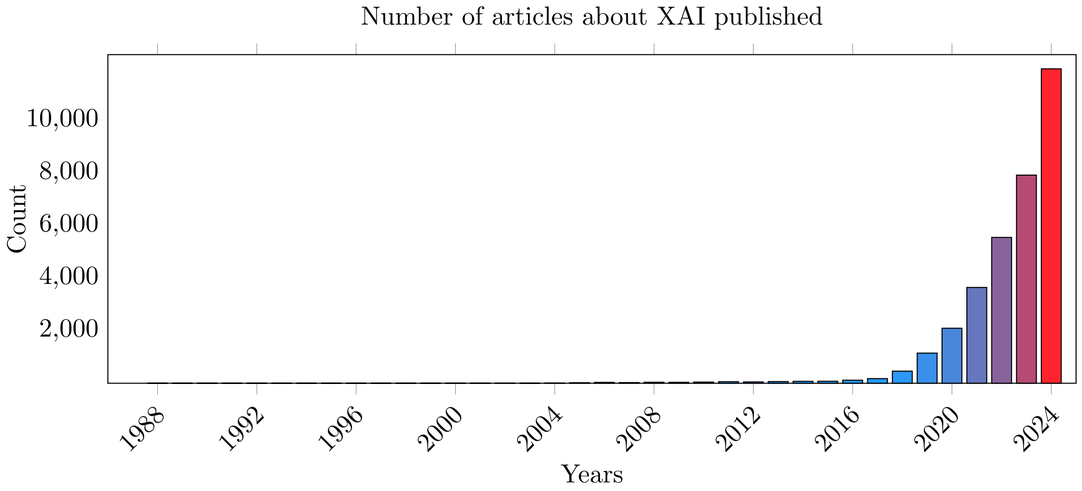
(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

(2014.03) SEDC [129]
(2015.08) OAE [51]
(2016.05) HCLS [110, 112]
(2017.06) Feature Tweaking [186]
(2017.11) CF Expl. [196]
(2017.12) Growing Spheres [114]
(2018.02) CEM [55]
(2018.02) POLARIS [209]
(2018.05) LORE [80]
(2018.06) Local Foil Trees [190]
(2018.09) Actionable Recourse [189]
(2018.11) Weighted CFs [77]
(2019.01) Efficient Search [175]
(2019.04) CF Visual Expl. [76]
(2019.05) MACE [99]
(2019.05) DiCE [145]
(2019.05) CERTIFAI [179]
(2019.06) MACEM [56]
(2019.06) Expl. using SHAP [165]
(2019.07) Nearest Observable [201]
(2019.07) Guided Prototypes [191]
(2019.07) REVISE [95]
(2019.08) CLEAR [202]
(2019.08) MC-BRP [123]
(2019.09) FACE [162]
(2019.09) Equalizing Recourse [83]
(2019.10) Action Sequences [163]
(2019.10) C-CHVAE [156]
(2019.11) FOCUS [124]
(2019.12) Model-based CFs [127]
(2019.12) LIME-C/SHAP-C [164]
(2019.12) EMAP [41]
(2019.12) PRINCE [71]
(2019.12) LowProFool [18]
(2020.01) ABELE [79]
(2020.01) SHAP-based CFs [66]
(2020.02) CEML [11–13]
(2020.02) MINT [100]
(2020.03) ViCE [74]
(2020.03) Plausible CFs [22]
(2020.04) SEDC-T [193]
(2020.04) MOC [52]
(2020.04) SCOUT [199]
(2020.04) ASP-based CFs [28]
(2020.05) CBR-based CFs [103]
(2020.06) Survival Model CFs [106]
(2020.06) Probabilistic Recourse [101]
(2020.06) C-CHVAE [155]
(2020.07) FRACE [210]
(2020.07) DACE [96]
(2020.07) CRUDS [60]
(2020.07) Gradient Boosted CFs [5]
(2020.08) Gradual Construction [97]
(2020.08) DECE [44]
(2020.08) Time Series CFs [16]
(2020.08) PermuteAttack [87]
(2020.10) Fair Causal Recourse [195]
(2020.10) Recourse Summaries [167]
(2020.10) Strategic Recourse [43]
(2020.11) PARE [172]

(2020) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

Methods	H
Linear probe [101]	
Regression based on CNN [106]	
Backwards model for interpretability of linear models [107]	
GDM (Generative Discriminative Models): ridge regression + least square [100]	
GAM, GA^{2}M (Generative Additive Model) [82], [102], [103]	
ProtoAttend [105]	
Other content-subject-specific models:	N
+ Kinetic model for CBF (cerebral blood flow) [131]	N
+ CNN for PK (Pharmacokinetic) modelling [132]	N
+ CNN for brain midline shift detection [133]	N
+ Group-driven RL (reinforcement learning) on personalized healthcare [134]	N
+ Also see [108]–[112]	N
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N
CCA (Canonical Correlation Analysis) [113]	
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD	
F-SVD (Frame Singular Value Decomposition) [114] on electromyography data	
DWT (Discrete Wavelet Transform) + Neural Network [135]	
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [136]	
GAN-based Multi-stage PCA [118]	
Estimating probability density with deep feature embedding [119]	
t-SNE (t-Distributed Stochastic Neighbour Embedding) [77]	
+ t-SNE on CNN [120]	
+ t-SNE, activation atlas on GoogleNet [121]	
+ t-SNE on latent space in meta-material design [122]	
+ t-SNE on genetic data [137]	
+ mm-t-SNE on phenotype grouping [138]	
Laplacian Eigenmaps visualization for Deep Generative Model [124]	
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [125]	
KNN with triplet loss and *query-result activation map pair* [139]	
Group-based Interpretable NN with RW-based Graph Convolutional Layer [123]	
TCAV (Testing with Concept Activation Vectors) [96]	
+ RCV (Regression Concept Vectors) uses TCAV with Br score [140]	
+ Concept Vectors with UBS [141]	
+ ACE (Automatic Concept-based Explanations) [56] uses TCAV	
Influence function [129] helps understand adversarial training points	
Representer theorem [130]	
SocRat (Structured-output Causual Rationalizer) [127]	
Meta-predictors [126]	
Explanation vector [128]	
# Also listed elsewhere: [14], [43], [85], [94]	N
# Also listed elsewhere: [14], [60], [85] etc	N
CNN with separable model [142]	

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI



Explainability methods

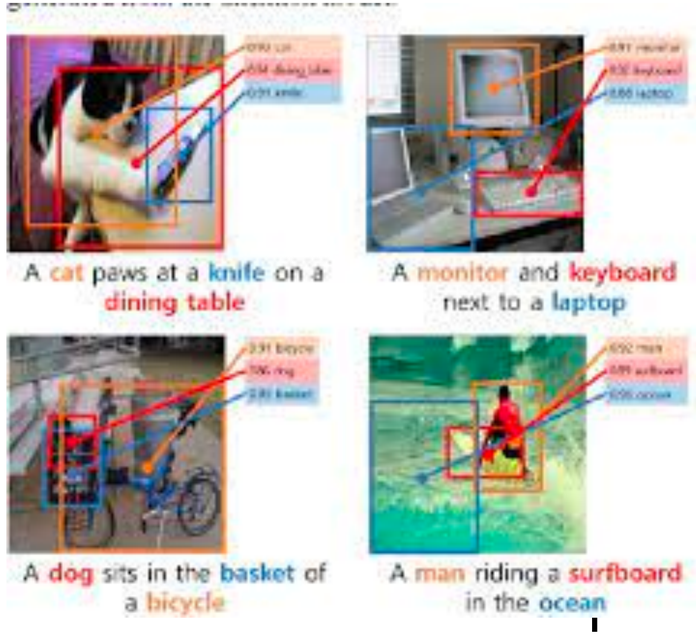
Examples

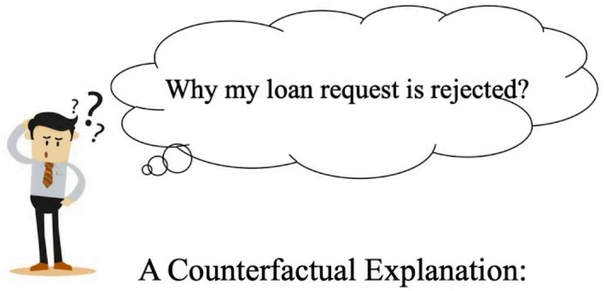
- Highlight important features
- Counterfactuals

- Caption generation
- Prototype examples
- Network probing

Text with highlighted words

Why does the older generation think that just because they don't understand video games and technology, they feel like they have to hate them and blame every bad thing on them?





A Counterfactual Explanation:

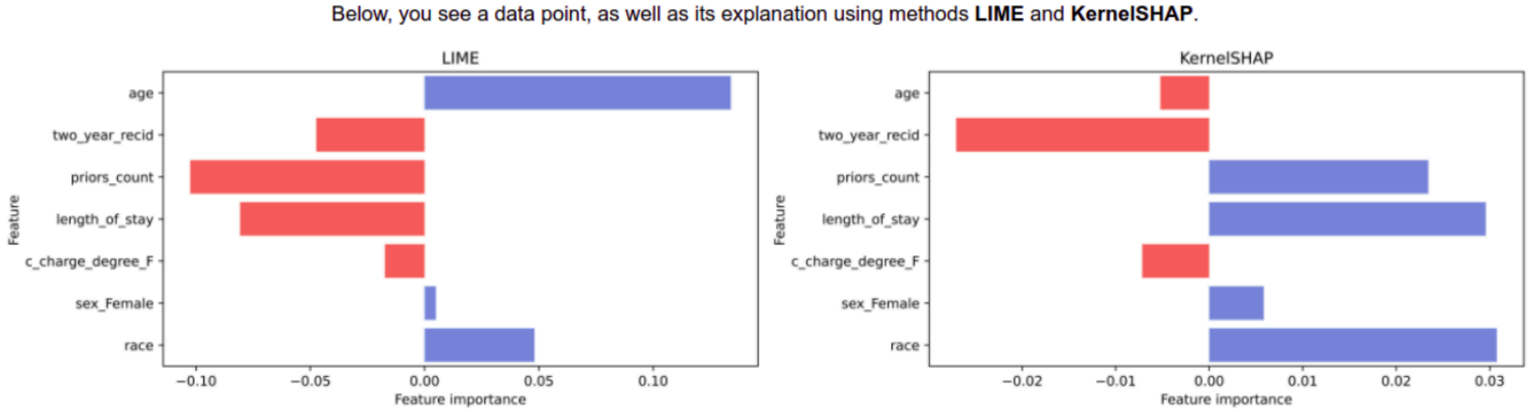
If you had an income of $40,000 rather than $30,000, your loan request would have been approved.

the minimal change made to alter the decision

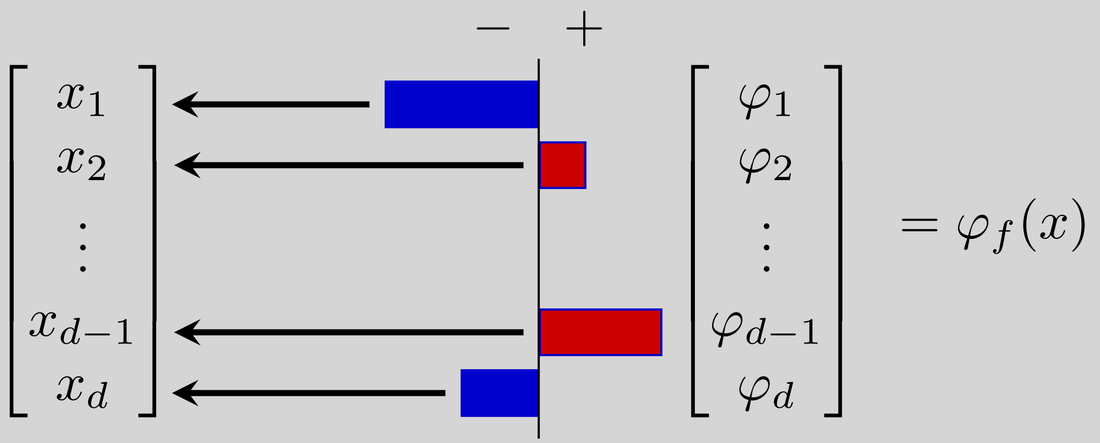
Explainability methods

Some problems

- Easy manipulation
- Methods can disagree with each other
- Plausible explanations may not be faithful to the real model



Attribution method



Highlight important features:

- Positive values mean a positive influence of that feature on the outcome
- Negative values mean a negative influence of that feature on the outcome

Not always the correct interpretation!

"Your income of € 40 000 contributed positively.

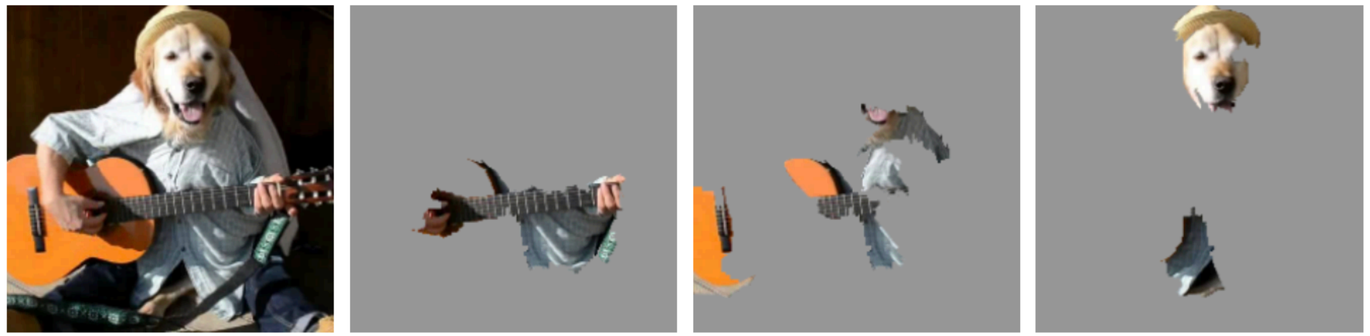
However, your 5 different credit cards contributed negatively to our decision."

Attribution methods

Examples

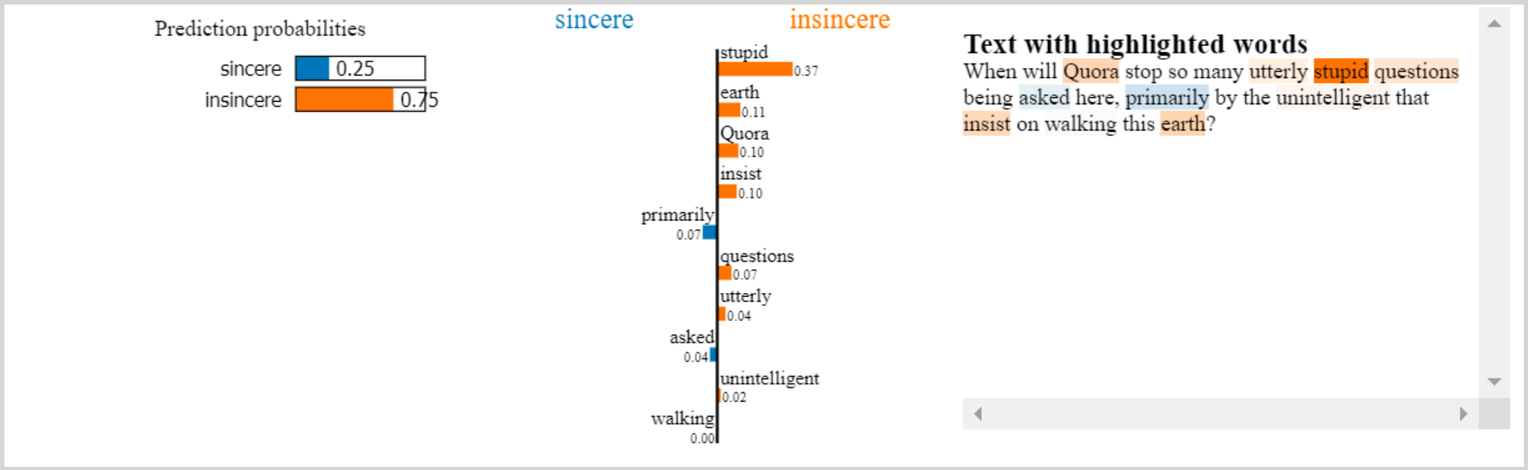
- LIME
- SHAP
- SmoothGrad

Pictures



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

Text



Leading example

2 parties:

- Credit Loan Applicant (A)


- Credit Loan Provider (P)

Loan application process:

- (A) provides (P) with a set of features:

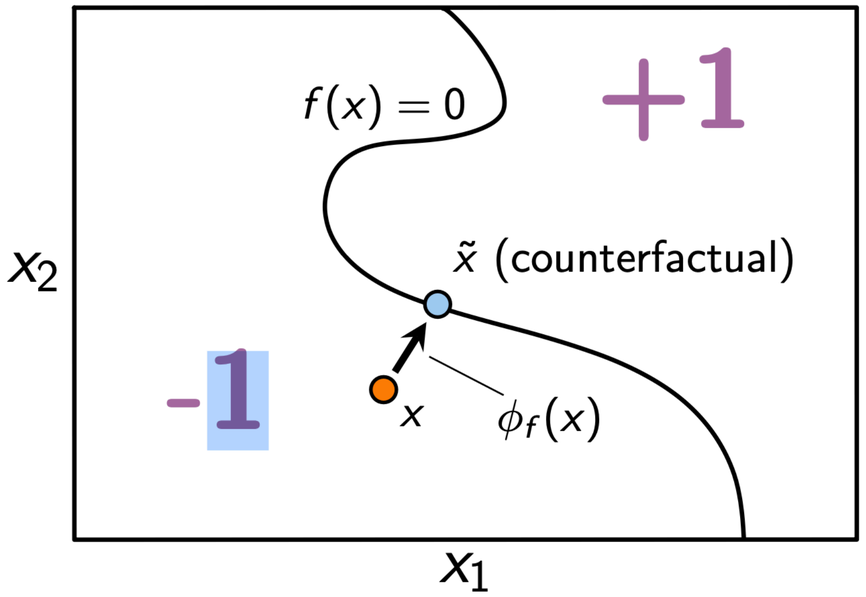
$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- (P) has an automated decision system f

$$f(x) = +1 \quad \text{Accepted}$$
$$f(x) = -1 \quad \text{Rejected}$$

- (A) can ask for a counterfactual explanation

Counterfactual method



- Tell (A) how the decision can be changed from -1 to +1
- Minimal effort for (A)
- Provides "***Recourse***"

"If you would have had an income of € 40 000 instead of €35 000, then we would have accepted your credit application."

Leading example

2 parties:

- Credit Loan Applicant (A)

- Credit Loan Provider (P)

Loan application process:

- (A) provides (P) with a set of features:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- (P) has an automated decision system f

$f(x) = 1$ if accepted
$f(x) = -1$ if not

- (A) can ask for a counterfactual explanation

This example is seen as a:

Counterfactual literature

Positive example

Strategic classification

Negative example

Modelling the setting: A statistical point of view

Model

Learning theoretic setting for classification

$$f\colon \mathcal{X} \subseteq \mathbb{R}^d \to \{-1, +1\}$$

We assume that $(X_0, Y) \sim P$

Loss is measured by 0/1-loss

$$\ell(f(x), y) = 1\{f(x) \neq y\}$$

Goal is to minimize ***expected loss (Risk/Accuracy)***

$$R = \mathbb{E}_P[\ell(f(X_0), Y)] = P(f(X_0) \neq Y).$$

The optimal classifier is the ***Bayes Classifier***

$$f_P^* = \arg\min \mathbb{E}_P[\ell(f(X_0), Y)]$$

Model

Adding recourse

Counterfactual point is defined as

$$\varphi(X_0) = X^{\mathsf{CF}} \in \underset{z:\, f(z)=+1}{\arg\min}\ c(X_0, z)$$

For simplicity, we assume that every negative X_0 accepts recourse

By adding recourse in the mix,

$$X_0 \to X,$$

where X is either X_0 or X^{CF}, we induce a new distribution

$$(X_0, X, Y) \sim Q.$$

Risk with **Recourse** is defined as

$$R_Q(f) = \mathbb{E}_Q[\ell(f(X), Y)] = Q(f(X) \neq Y)$$

Note that Q depends on f in general

Model

When is Recourse accepted

The function $r(X_0)$ **models** how likely X_0 is to accept recourse

In general X_0 does not need to change to X^{CF},

This is modelled by setting

$$X = BX^{\text{CF}} + (1 - B)X_0, \quad B \sim \text{Ber}(r(X_0)).$$

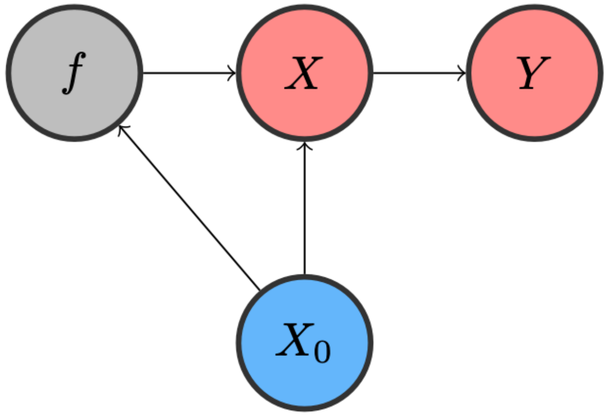
For the rest of the talk we will assume $r(X_0) = 1$

Modelling Q

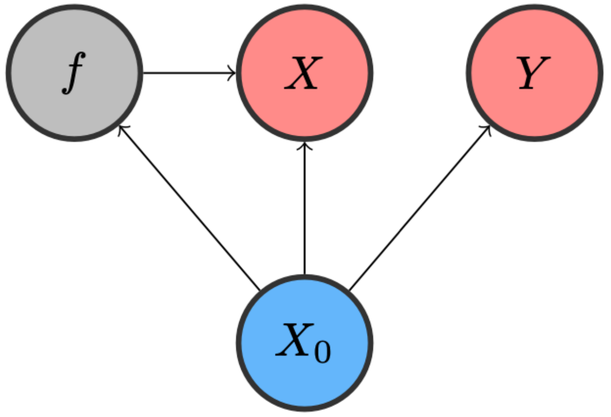
Statistical choice in dependency structure

We define 2 extreme cases:

- ***Compliant:*** $Q(Y \mid X_0, X) = P(Y \mid X)$
- ***Defiant:*** $Q(Y \mid X_0, X) = P(Y \mid X_0)$



Compliant case



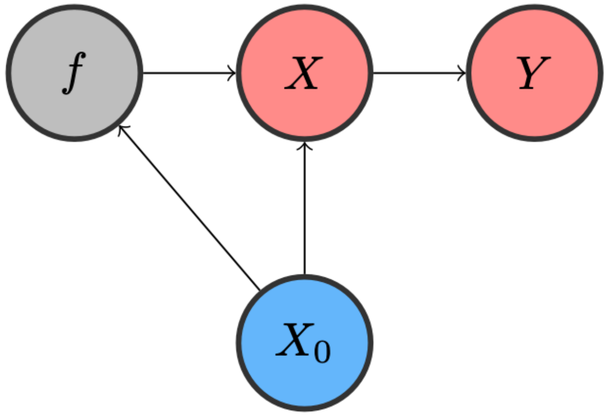
Defiant case

Modelling Q

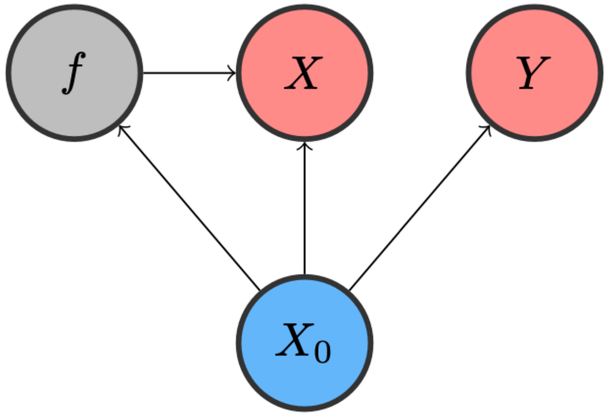
Examples

Some 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 case



Defiant case

Optimal classifier

Optimal Classifier

Example (Compliant)

We assume that

$$X \mid Y = +1 \sim N(\mu, \Sigma)$$

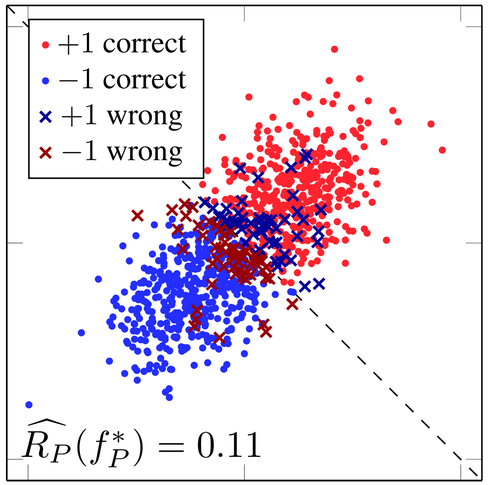
$$X \mid Y = -1 \sim N(\nu, \Sigma)$$

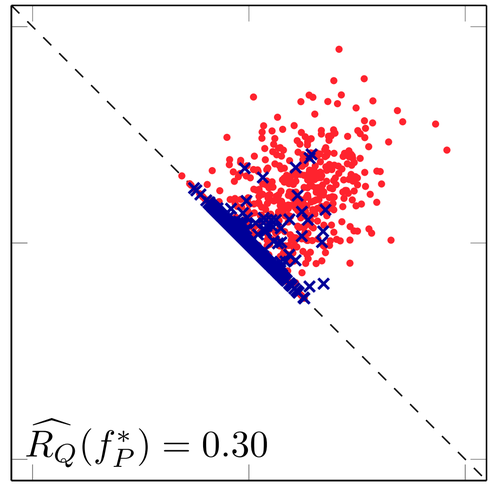
$$P(Y = +1) = P(Y = -1) = \frac{1}{2}$$

Optimal classifier is (assuming $\|\mu\|_{\Sigma^{-1}} = \|\nu\|_{\Sigma^{-1}}$)

$$f_P^*(x) = \text{sign}(\theta^\top x),$$

$$\theta = \Sigma^{-1}(\mu - \nu)$$





- $R_P(f_P^*) = \Phi(\|\mu - \nu\|_{\Sigma^{-1}})$
- $R_Q(f_P^*) = \frac{1}{4} + \frac{1}{2}\Phi(\|\mu - \nu\|_{\Sigma^{-1}})$

$$R_Q(f_P^*) > R_P(f_P^*) \text{ if } R_P(f_P^*) < \frac{1}{2}$$

Optimal Classifier

Formal result

Theorem

Let ℓ be the $0/1$ loss, f_P^* is the Bayes classifier for the distribution P, and suppose that $P(Y = 1 \mid X_0 = x) = \frac{1}{2}$ for all x on the decision boundary of f_P^*, then:

A. For the Compiant case,

$$R_Q(f_P^*) = \frac{1}{2}P(f_P(X_0) = -1) + P(f_P(X_0) = 1, Y = -1) > R_P(f_P^*)$$

B. For the Defiant case,

$$R_Q(f_P^*) = P(Y = -1) > R_P(f_P^*)$$

Optimal Classifier

Proof sketch (Compliant)

$$R_Q(f_P^*) = \frac{1}{2}P(f_P(X_0) = -1) + P(f_P(X_0) = 1, Y = -1) > R_P(f_P^*)$$

- Every point is now classified as $+1$
- The mistakes you make are
  - Original $f_P^*(X_0) = +1$ but $Y = -1$,
  - Half of the original $f_P^*(X_0) = -1$,
    - because $P(Y = +1 \mid X) = P(Y = -1 \mid X) = \frac{1}{2}$ on the decision boundary

Optimal Classifier

Proof sketch (Defiant)

$$R_Q(f_P^*) = P(Y = -1) > R_P(f_P^*)$$

- Every point is now classified as $+1$
- The mistakes you make are
  - Original $f_P^*(X_0) = +1$ but $Y = -1$,
  - Original $f_P^*(X_0) = -1$, but $Y = -1$, because the label does not change in this case

Modelling the setting: A Causal point of view

Leading example

Software developer looking for a job:

Master's degree

Qualified

Github Activity

Software company:

- ▶ Invites for an interview based on:
  - Having a Master's degree
  - Having many Github commits

Potential candidates:

- ▶ Use tools to increase Github activity
- ▶ Feature becomes unpredictive

Model

We now take a causal point of view

We assume that there is a Structural Causal Model (SCM), inducing the data:

$$\mathbb{M} = \langle \mathbb{X}, \mathbb{U}, f \rangle$$

Where,

- $\mathbb{X}$ are all observed variables;
- $\mathbb{U}$ the exogenous noise;
- f the structural equations.

The features are denoted by

$$X_0 = (X_0^1, \ldots, X_0^d) \, .$$

A target variable Y_0, from which a label

$$L_0 = 1[Y_0 \geq 0]$$

is derived.

We again assume access to the Bayes classifier:

$$h_0(x) = P(L_0 = 1 \mid X_0 = x)$$

$$\widehat{L_0}(x) = 1\left[h_0(x) \geq \frac{1}{2}\right]$$

Model

Recourse as interventions

Recourse Explanations are framed as interventions

$$a^{\text{CF}} = \text{do}\left(\left\{X^i = \theta^i\right\}_{i \in I}\right)$$

Important variables are:

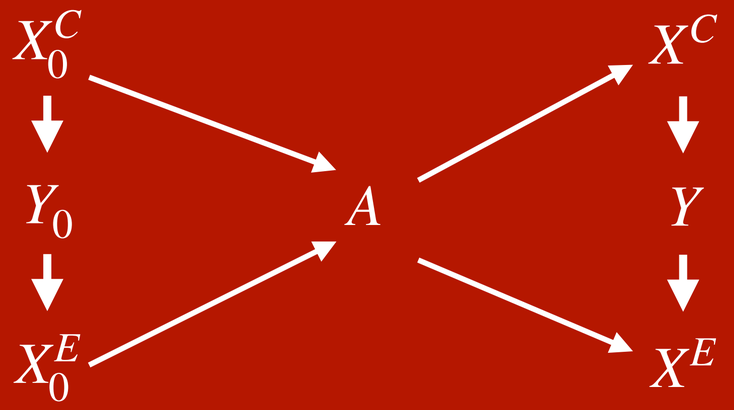
- X_0^C, the direct causes of Y_0 :

$$X_0^C \rightarrow Y_0$$

- X_0^E, the direct effects of Y_0:

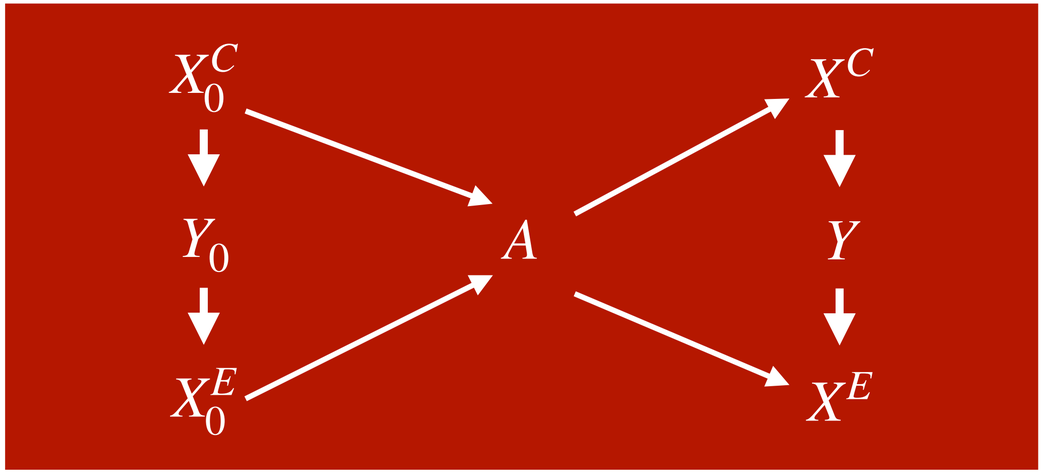
$$Y_0 \rightarrow X_0^E$$

Results in a graph of the form:



Performative validity and how to ensure it

Performative validity



Classifier after giving recourse is:

$$\widehat{L}(x) = 1\left[P(L = 1 \mid X = x) \geq \frac{1}{2}\right]$$

Def. Performative validity, for all x

$$\widehat{L}(x) \geq \widehat{L_0}(x)$$

Captures the idea that everyone should **improve**

Can be ensured when

$$P(L = 1 \mid X = x) = P(L_0 = 1 \mid X_0 = x)$$

Problematic Explanations

When are interventions not problematic?

$$1\left[A = \text{do}(\emptyset)\right] \perp\!\!\!\perp L \mid X$$

In words: Knowing if an intervention was performed, does not give any extra information about the real label

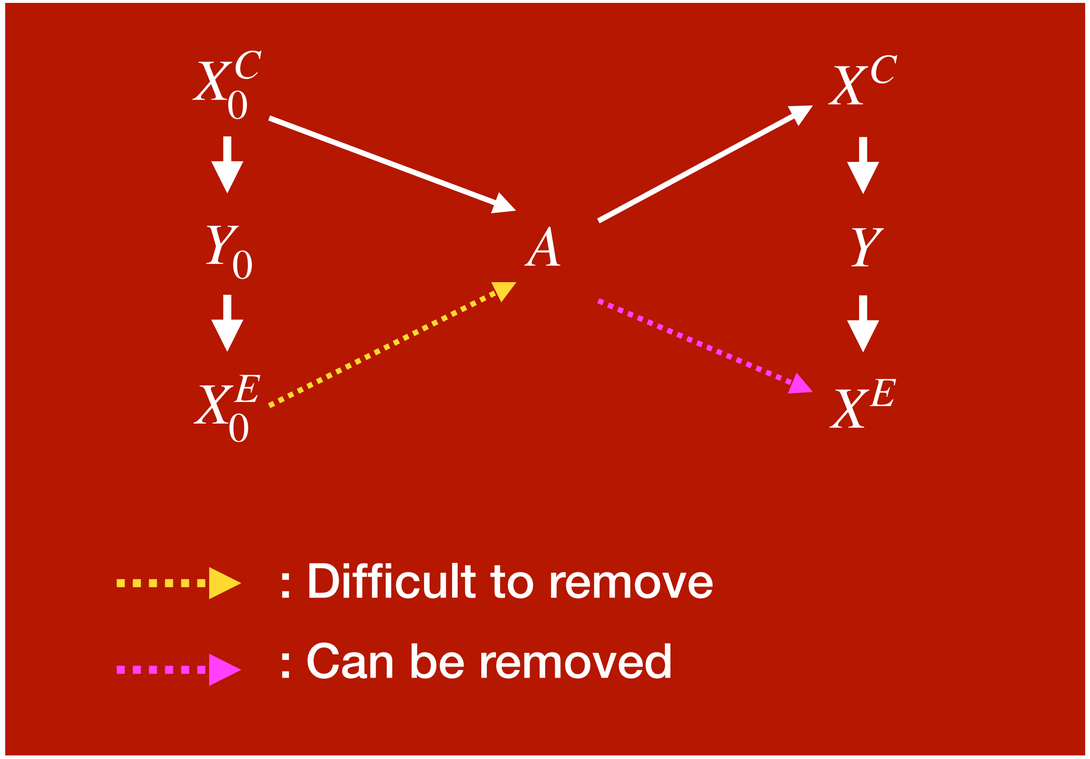
We identify 2 problematic situations:

1. Recourse explanations influenced by effect variables
2. Recourse explanations intervening on effect variables

Problematic Explanations

We identify 2 problematic situations:

1. Recourse explanations influenced by effect variables
2. Recourse explanations intervening on effect variables



Problematic Explanations

How can validity be ensured?

There are 3 ways recourse explanations are found:

1. **Counterfactual Explanations [Wachter et al.]:**

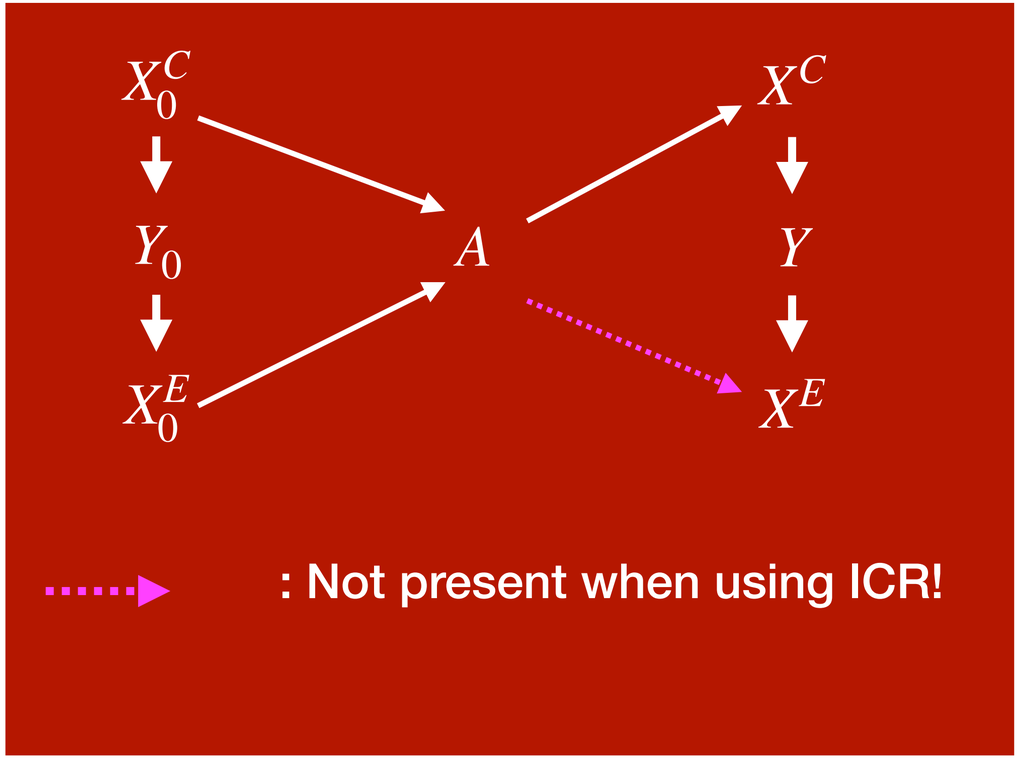
$$a^{\text{CF}}(x) = \arg\min c(x, x') \;\; \text{s.t} \;\; \hat{L}(x') = 1$$

2. **Causal Recourse [Karimi et al.]:**

$$a^{\text{CR}}(x) = \arg\min c(x, a) \;\; \text{s.t} \;\; P(\hat{L}(x') = 1 \mid X = x, \text{do}(a))$$

3. **Improvement-Focused Recourse [König et al.]:**

$$a^{\text{ICR}}(x) = \arg\min c(x, a) \;\; \text{s.t} \;\; P(L = 1 \mid X = x, \text{do}(a))$$



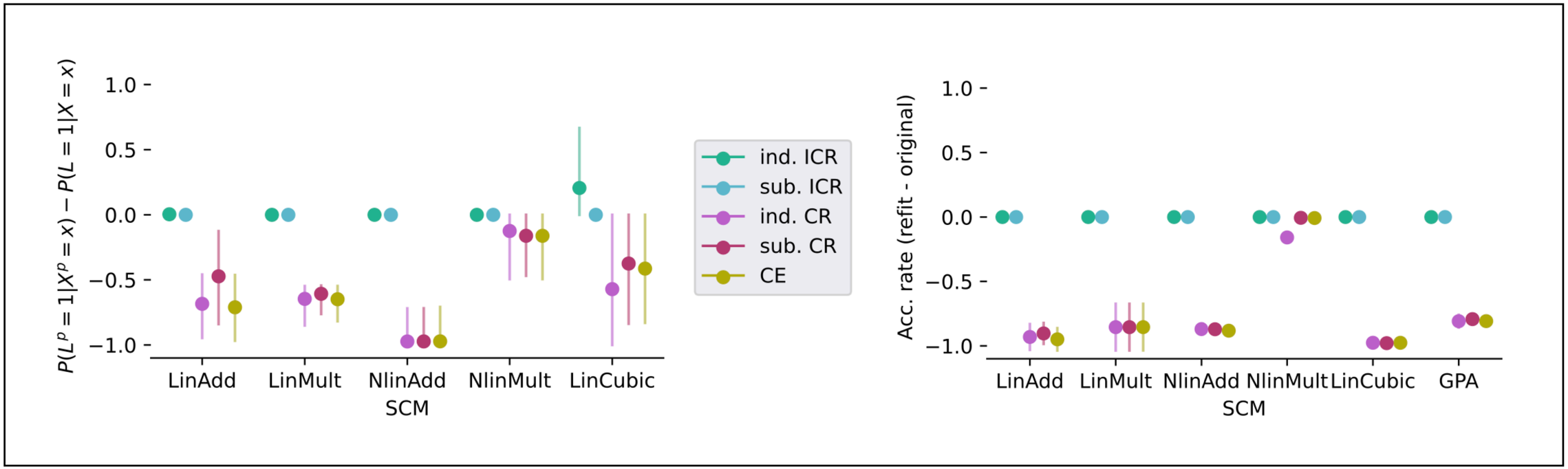
Ensuring Performative Validity

Proof sketch

- ➡ We need to identify when $1\left[A = \text{do}(\emptyset)\right] \perp\!\!\!\perp L \mid X$
- ➡ The SCM $\mathbb{M}$ induces a causal graph over all variables
- ➡ All conditional independencies can be read off graphically using d-separation
  - ➡ Problem reduces to finding all d-open paths and closing them
  - ➡ Realise there are only 2 types of paths that are d-open
    - ➡ The first path will always exist, but the conditional dependence can be removed by an assumption on the structural equations (Faithfulness violation)
    - ➡ The second type can be removed by using ICR recommendations, they only target causes

Ensuring Performative Validity

Experiments



Some perspectives

Some perspectives

When can Recourse still be beneficial?

Some possible answers:

- In situations where Accuracy is not the most important metric
- When counterfactuals have other explanatory properties
- Should Recourse be replaced by Contestability?

Thank you for your attention!

Strategising

Strategising

Example

Can (P) strategise against this accuracy drop?

- Need to assume that not everyone gets an explanation, i.e.

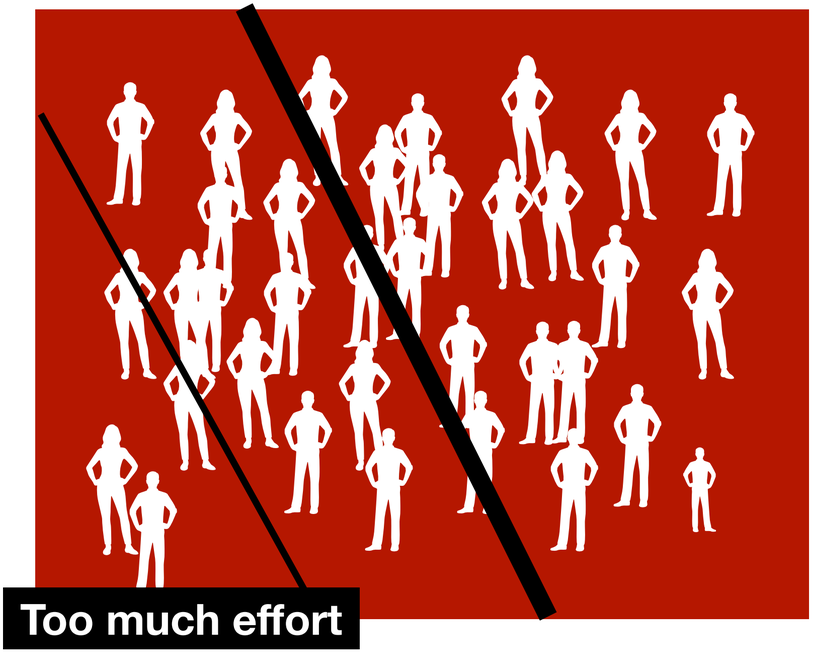
$$r(x_0) = 1\{\|\varphi(x_0) - x_0\| < D\} \text{ for some } D > 0$$

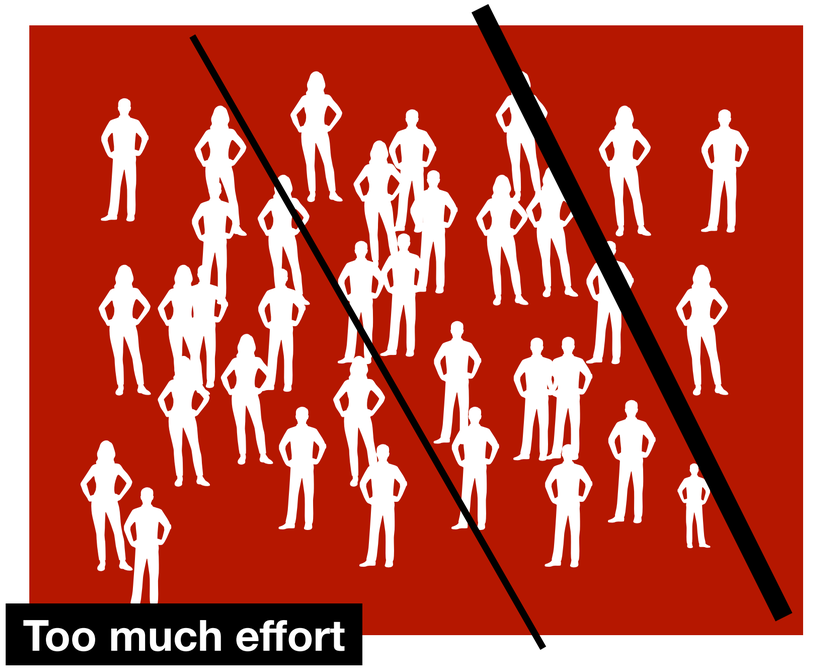
Too much effort

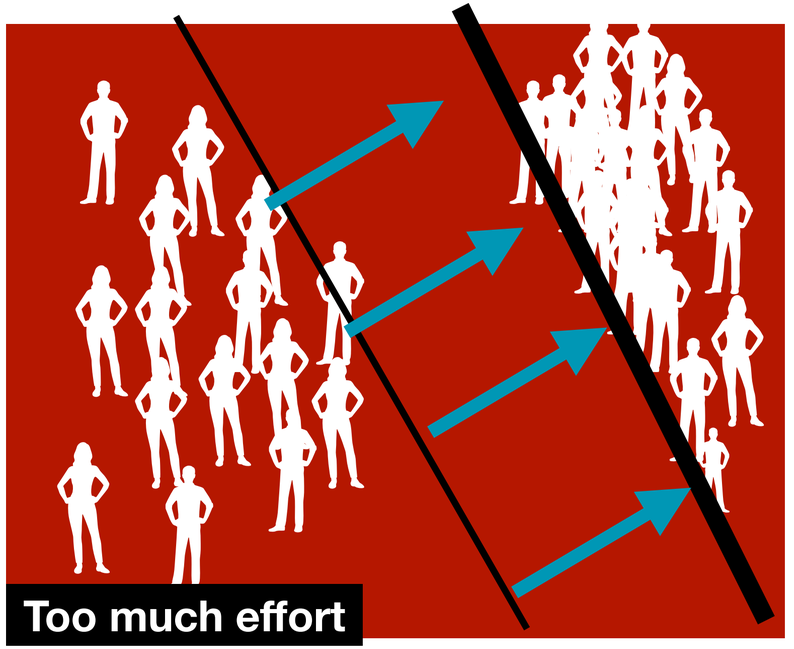
Yes and No

Strategising

Example







Exactly the same people get a loan

More effort

Strategising

- Can generalise this result
  - Defiant case: Accuracy can only be as good as before
  - Compliant case: In principle, Accuracy could have arbitrary improvement, but would increase the cost for every applicant substantially

Strategising

Formal result, a new definitions

Definition

Let $\mathcal{F}$ be some model class, $r(x_0) \in \{0,1\}$ and $\varphi^r(x_0) = r(x_0)\varphi(x_0) + (1 - r(x_0))\varphi(x_0)$, define

$$\mathcal{F}_\varphi^r := \{f' : x_0 \mapsto f(\varphi^r(x_0)) \mid f \in \mathcal{F}\}.$$

The set of functions induced by the recourse map.

If $\mathcal{F}_\varphi^r = \mathcal{F}$, we call $\mathcal{F}$ ***invariant under recourse***

For example,

$$\mathcal{F} = \{f(x) = \text{sign}(a^\top x + b) \mid a, b \in \mathbb{R}^d\} \text{ and } r(x_0) = 1\{\|\varphi(x_0) - x_0\| < D\}.$$

Then $\mathcal{F} = \mathcal{F}_\varphi$

Strategising

Formal result (Defiant)

Theorem

If $\mathscr{F}$ is a recourse invariant model class, then

$$\min_{f \in \mathscr{F}} R_P(f) = \min_{f \in \mathscr{F}} R_Q(f).$$

You can only do as well as before

Strategising

Formal result (Compliant)

Theorem

If $\mathscr{F}$ is a recourse invariant model class, then

$$\min_{f \in \mathscr{F}} R_Q(f) \leq \min_{f \in \mathscr{F}} R_P(f) - \gamma.$$

Where $\gamma \in \mathbb{R}$ depends on P and f.

Improvement is possible when $\gamma > 0$!

For example, in the Gaussian example

However, there will be a cost for every individual

Non-Optimal classifier

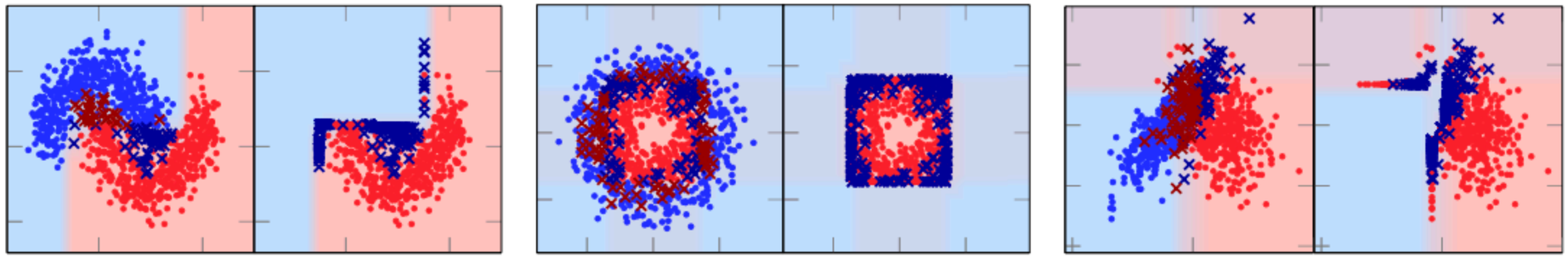
Non-Optimal Classifier

More general

What if we consider non-optimal classifiers?

Need to make some extra assumptions:

- Assume $f(x) = \text{sign}(g(x) - \frac{1}{2})$ for some probabilistic classifier $g(x) \colon \mathcal{X} \to [0,1]$
- The function g is "ϵ-close" to $P(Y = 1 \,|\, X = x)$ along the decision boundary



ϵ-close assumption

More general

What is the assumption?

Informally:

- The function g is "ϵ-close" to $P(Y = 1 \,|\, X = x)$ along the decision boundary

Formally:

$$\int_{\{x_0 :\, g(x_0) < \frac{1}{2}\}} \left| \frac{1}{2} - P(Y = 1 \,|\, X = \varphi(x_0)) \right| P(\mathrm{d}x_0) < \epsilon$$

Implied by uniform bound,
$\left| \frac{1}{2} - P(Y = 1 \,|\, X_0 = x) \right| < \epsilon$ for all x such that $g(x) = \frac{1}{2}$

Non-Optimal Classifier

Formal result

Theorem

Let ℓ be the $0/1$ loss, $g: \mathcal{X} \to [0,1]$ a continuous probabilistic classifier and assume the ϵ -condition:

A. For the Compliant case, $R_Q(f)$ is lower and upper bounded by

$$(\frac{1}{2} \pm \epsilon) P(f(X_0) = -1) + P(f(X_0) = +1, Y = -1)$$

B. For the Defiant case,

$$R_Q(f) = P(Y = -1)$$

Non-Optimal Classifier

Implications

Theorem

Let ℓ be the $0/1$ loss, $g: \mathcal{X} \to [0,1]$ a continuous probabilistic classifier and assume the ϵ -condition:

A. For the Compliant case, $R_Q(f) \geq R_P(f)$ if

$$P(Y = -1 \,|\, f(X_0) = -1) \geq \frac{1}{2} + \epsilon$$

B. For the Defiant case, $R_Q(f) \geq R_P(f)$ if and only if

$$P(Y = -1 \,|\, f(X_0) = -1) \geq \frac{1}{2}$$

Non-Optimal Classifier

Interpretation

Recourse will harm the risk if

A. For the Compliant case, if f approximates the true conditional distribution and f performs ϵ better on the negative class

B. For the Defiant case, if f performs better than random on the negative class

Non-Optimal Classifier

Experimental results

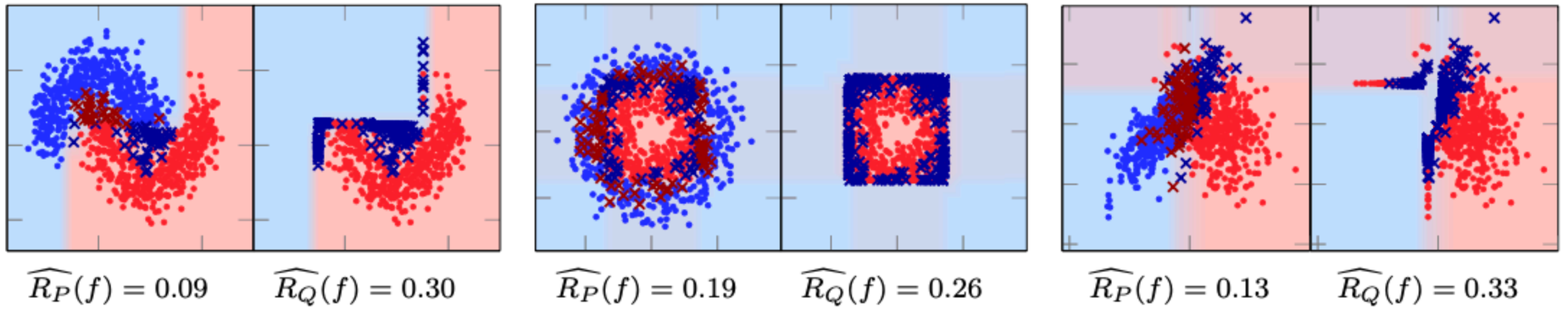


Table 1: Estimated risks on synthetic data sets. Lower risk is bold.

	Moons data		Circles data		Gaussians data	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
Logistic Regression (LR)	**0.13**	0.33	0.51	**0.34**	**0.14**	0.32
GradientBoostedTrees (GBT)	**0.08**	0.30	**0.19**	0.26	**0.13**	0.33
Decision Tree (DT)	**0.08**	0.29	**0.19**	0.23	**0.13**	0.34
Naive Bayes (NB)	**0.13**	0.33	**0.17**	**0.16**	**0.15**	0.28
QuadraticDiscriminantAnalysis (QDA)	**0.13**	0.33	**0.17**	**0.16**	**0.12**	0.33
Neural Network(4)	**0.12**	0.32	**0.23**	0.30	**0.13**	0.36
Neural Network(4, 4)	**0.04**	0.26	**0.17**	0.22	**0.12**	0.40
Neural Network(8)	**0.04**	0.23	**0.16**	0.20	**0.12**	0.36
Neural Network(8, 16)	**0.04**	0.26	**0.16**	0.18	**0.11**	0.35
Neural Netowrk(8, 16, 8)	**0.04**	0.26	**0.16**	0.18	**0.11**	0.35

Table 2: Estimated risks on real data sets. Lower risk is bold.

	Credit data						Census data						HELOC data					
	Wachter		GS		CoGS		Wachter		GS		CoGS		Wachter		GS		CoGS	
	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q
LR	0.17	**0.05**	0.17	**0.05**	0.17	**0.04**	**0.21**	0.29	**0.21**	0.33	**0.21**	0.32	**0.29**	0.41	**0.29**	0.41	**0.29**	0.44
GBT	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	0.15	**0.04**	**0.15**	0.23	**0.15**	0.33	**0.20**	**0.21**	**0.20**	0.25	**0.20**	0.37
DT	0.29	**0.12**	0.29	**0.05**	0.29	**0.05**	**0.23**	**0.21**	**0.23**	0.43	**0.23**	0.45	**0.19**	0.25	**0.19**	0.21	**0.19**	0.31
NB	0.11	**0.06**	0.11	**0.06**	0.11	**0.07**	**0.19**	0.78	**0.19**	0.76	**0.19**	0.81	**0.29**	0.44	**0.29**	0.43	**0.29**	0.48
QDA	0.12	**0.06**	0.12	**0.06**	0.12	**0.07**	**0.20**	0.78	**0.20**	0.75	**0.20**	0.82	**0.32**	0.46	**0.32**	0.47	**0.32**	0.52
NN(4)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.06**	**0.16**	0.26	**0.16**	0.25	**0.16**	0.26	**0.29**	0.47	**0.29**	0.46	**0.29**	0.50
NN(4, 4)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.30	**0.15**	0.27	**0.15**	0.30	**0.29**	0.47	**0.29**	0.47	**0.29**	0.51
NN(8)	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**	**0.07**	**0.16**	0.34	**0.16**	0.33	**0.16**	0.33	**0.28**	0.44	**0.28**	0.46	**0.28**	0.51
NN(8, 16)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.36	**0.15**	0.34	**0.15**	0.36	**0.27**	0.42	**0.27**	0.45	**0.27**	0.46
NN(8, 16, 8)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.36	**0.15**	0.34	**0.15**	0.36	**0.27**	0.42	**0.27**	0.45	**0.27**	0.46

Non-Optimal Classifier

Proof sketch (Compliant)

Upper/lower: $(\frac{1}{2} \pm \epsilon) P(f(X_0) = -1) + P(f(X_0) = +1, Y = -1)$

- ➡ Every point is now classified as $+1$
- ➡ The mistakes you make are
  - ➡ Original $f(X_0) = +1$ but $Y = -1$,
  - ➡ Within ϵ-distance of half the original $f(X_0) = -1$,
- ➡ Simplify $R_P(f) \leq (\frac{1}{2} - \epsilon) P(f(X_0) = -1)$

Non-Optimal Classifier

Proof sketch (Defiant)

$$R_Q(f) = P(Y = -1) > R_P(f)$$

- Every point is now classified as $+1$
- The mistakes you make are
  - Original $f(X_0) = +1$ but $Y = -1$,
  - Original $f(X_0) = -1$, but $Y = -1$, because the label does not change in this case

Non-Optimal Classifier

Some more examples

What if $r(x)$ is not constant:

- $r(x) = p \in [0,1]$
- $r(x) = e^{-\frac{1}{2\sigma}\|x-\varphi(x)\|}$

- On y-axis: $R_Q - R_P$
- From L to R: Moons, Circles, Gaussians