UNIVERSITY OF AMSTERDAM
Korteweg de Vries Institute for Mathematics

The Risks of Recourse in Binary Classification

Hidde Fokkema*, Damien Garreau† and Tim van Erven*.

*: University of Amsterdam, Korteweg-de Vries Institute for Mathematics, †:University of Würzburg, CAIDAS

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

Summary

Algorithmic recourse provides explanations that help users **overturn an unfavorable decision** by a machine learning system, e.g. a rejected bank loan. But we prove (mathematically) that this is often **harmful when applied at scale to a whole population**. Results:

1. For optimal deterministic classifiers, providing recourse **always increases** the risk = average population loss.
2. For (near-)optimal probabilistic classifiers, the risk will also increase.
3. The party deploying the classifier has a **strategic incentive** to **undo the effect of recourse** to prevent risk increase.

Setting

We are given data $(X_0, Y) \sim P$ and a classifier:

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

which should have small **risk** measured by:

$$R_P(f) := \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0), Y)] = P(f(X_0) \neq Y).$$

For $\ell = 0/1$-loss

A **counterfactual** for X_0 is

$$\varphi(X_0) = X^{\mathrm{CF}} \in \mathop{\mathrm{argmin}}_{z\colon f(z)=+1} c(X_0, z).$$

The point $\varphi(X_0)$ gives **recourse** to the user. The user accepts or rejects the counterfactual with some probability $r(X_0)$:

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

This induces a joint distribution Q on (X, X_0).

How Do Users Implement Recourse?

For the conditional distribution of $Y \mid X, X_0$ we are faced with a choice. We define 2 extreme cases:

- **Compliant:** $Q(Y \mid X_0, X) = P(Y \mid X)$,
  The new features faithfully represent a change.
- **Defiant:** $Q(Y \mid X_0, X) = P(Y \mid X_0)$
  The change in features is only cosmetic.

Main Results

Let $f_P^* = \mathrm{argmin}_f\, R_P(f)$ be the Bayes-optimal classifier, and let $R_Q(f) = \mathbb{E}_{(X,Y)\sim Q}[\ell(f(X), Y)]$ denote the risk under recourse.

Theorem 1 (Bayes-Optimal Classifier Risk Increase). *Suppose that $P(Y = 1 \mid X_0 = x) = \frac{1}{2}$ for all x on the decision boundary of f_P^*. Then, in both the **defiant** and **compliant** settings, **recourse always increases the risk**:*

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

The inequality is strict if recourse happens with positive probability: $P(B = 1, f_P^*(X_0) = -1) > 0$.

Alternatively, suppose we have a continuous probabilistic classifier $g\colon \mathcal{X} \to [0, 1]$.

Theorem 2 (Probabilistic Classifier Risk Increase). *Let $f(x) = \mathrm{sign}(g(x) - \frac{1}{2})$. Then,*

*(a) For the **defiant** case: the **risk increases if the classifier is better than random guessing**:*

$$R_Q(f) \geq R_P(f) \qquad \text{if and only if} \qquad P(Y = -1 \mid B = 1, f(X_0) = -1) \geq \tfrac{1}{2}.$$

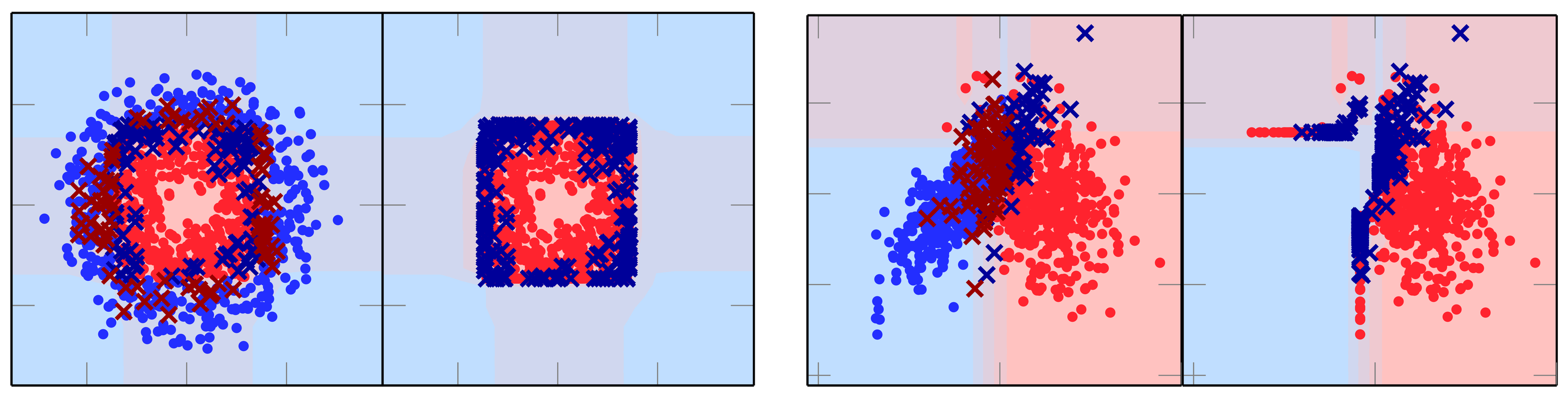
And, suppose that the decision boundary of g is close to the optimal decision boundary: $|P(Y = 1 \mid X_0 = x) - \frac{1}{2}| \leq \varepsilon$ for all x such that $g(x) = \frac{1}{2}$. Then

*(b) For the **compliant** case: the **risk increases if the classifier is ε-better than random guessing**:*

$$R_Q(f) \geq R_P(f) \qquad \text{if} \qquad P(Y = -1 \mid B = 1, f(X_0) = -1) \geq \tfrac{1}{2} + \varepsilon.$$

Experiments

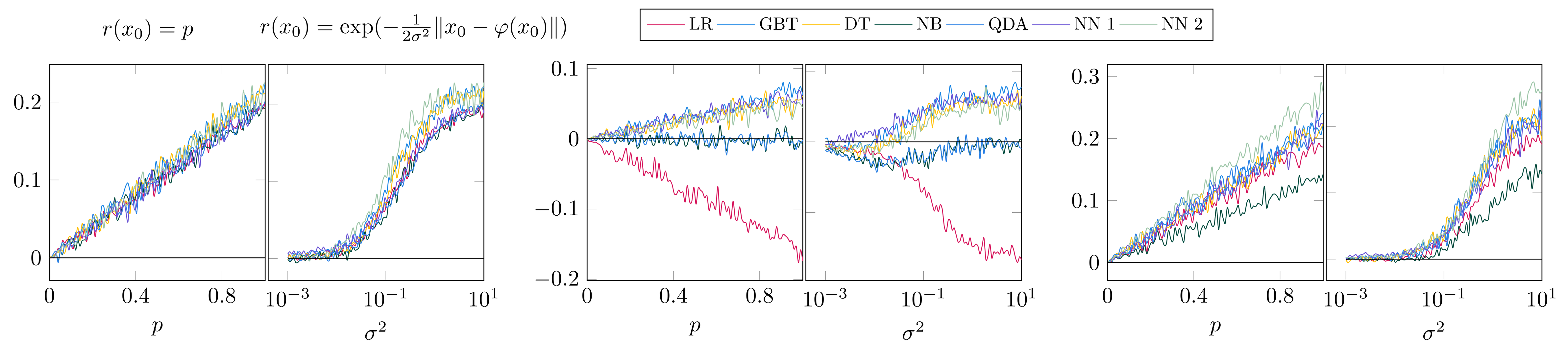
The increase in risk can be observed in synthetic examples as well as on real data:



$\widehat{R_P}(f) = 0.19$ $\widehat{R_Q}(f) = 0.26$ $\widehat{R_P}(f) = 0.13$ $\widehat{R_Q}(f) = 0.33$

	Census data	
	R_P	R_Q
LR	**0.21** ± **0.03**	0.30 ± 0.02
GBT	0.15 ± 0.01	**0.05** ± **0.01**
DT	**0.25** ± **0.04**	**0.23** ± **0.05**
NB	**0.18** ± **0.02**	0.76 ± 0.02
QDA	**0.21** ± **0.02**	0.76 ± 0.03
NN 1	**0.16** ± **0.01**	**0.27** ± **0.11**
NN 2	**0.16** ± **0.02**	0.32 ± 0.05

Here, we plot the dependence of the risk difference on $r(x_0)$ with different synthetic examples. On the y-axis, the difference $R_Q(f) - R_P(f)$ is plotted.



Strategizing

Call a function class $\mathcal{F}$ **invariant under recourse** if any $f \in \mathcal{F}$ has a unique f' such that $f = f' \circ \varphi$. NB the effect of recourse, Q_f, depends on f.

Theorem 4 (Defiant Case). *If $r(x_0) \in \{0, 1\}$ and $\mathcal{F}$ is invariant under recourse. Then,*

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

In the Compliant case, the situation is a bit more difficult.

Theorem 5 (Compliant Case). *If $r(x_0) \in \{0, 1\}$ and $\mathcal{F}$ is invariant under recourse. Then,*

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

where Δ is given by

$$\Delta := \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0), Y)] - \mathop{\mathbb{E}}_{(X_0,Y)\sim Q_{f'}}[\ell(f(X_0), Y)].$$

Is Algorithmic Recourse a Dead End?

Yes:

- Worse classification accuracy is bad: For instance, **many more customers cannot repay their bank loan after receiving recourse.**
- Counterfactuals are still useful explanations for other purposes
- Provide 'contestability' instead of recourse?

No: Recourse might still be valuable in situations where

- Accuracy (for the party deploying the classifier) is not that important.

The field must **change** its **motivating examples**.

Rest of the Paper

Read our paper online!

- **Surrogate losses:** we show conditions for when risk increases when a surrogate loss is used.
- **More Experiments**