CWI Machine Learning Theory Bootcamp Tutorial, February 14–15, 2023

Formal Results in Explainable Machine Learning

Tim van Erven

University of Amsterdam

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

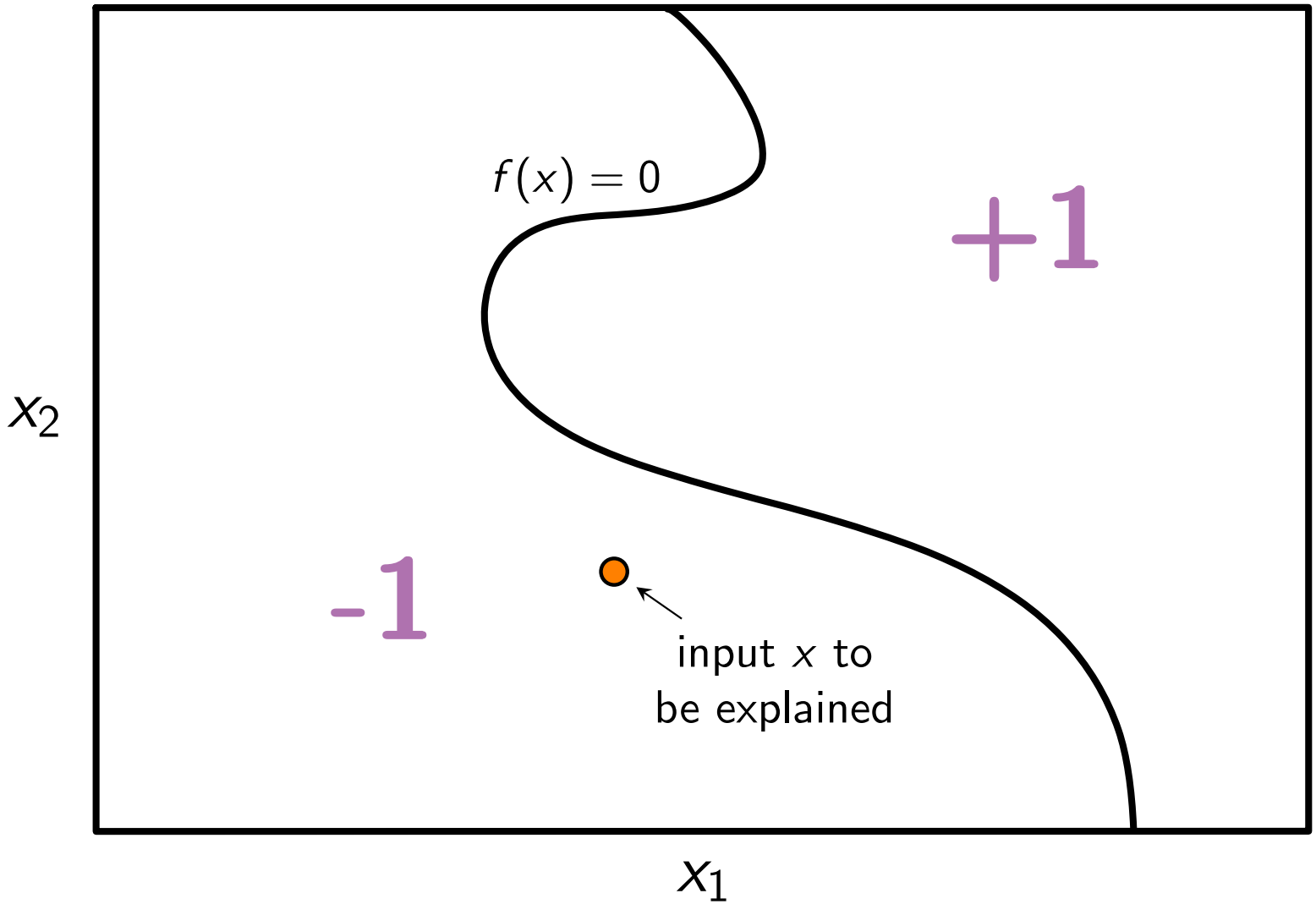
- Classify my company as high risk for money laundering?
- Reject my bank loan?
- Predict this patient can safely leave the intensive care?
- Mistake a picture of a husky for a wolf?
- Reject the profile picture I uploaded to get a public transport card?[1]
- ...

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

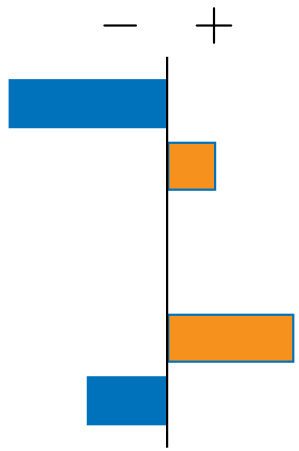
Local Post-hoc Explanations



- **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

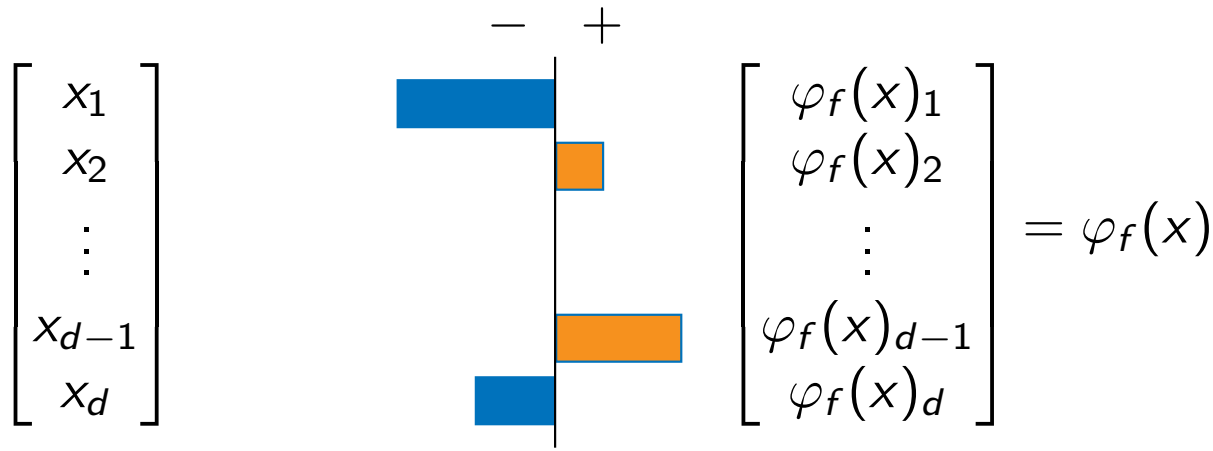
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d-1} \\ x_d \end{bmatrix}$$



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

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

Local Explanations via Attributions



$\phi_f(x) \in \mathbb{R}^d$ 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

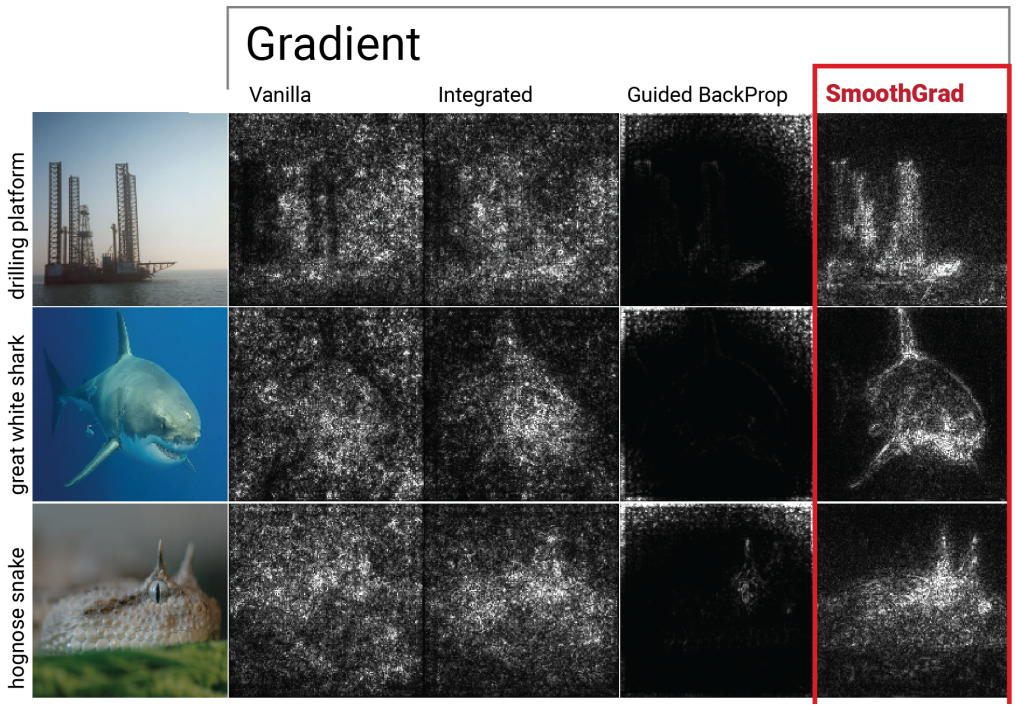
$$f(x) = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

$\phi_f(x)_i = \theta_i$ 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]



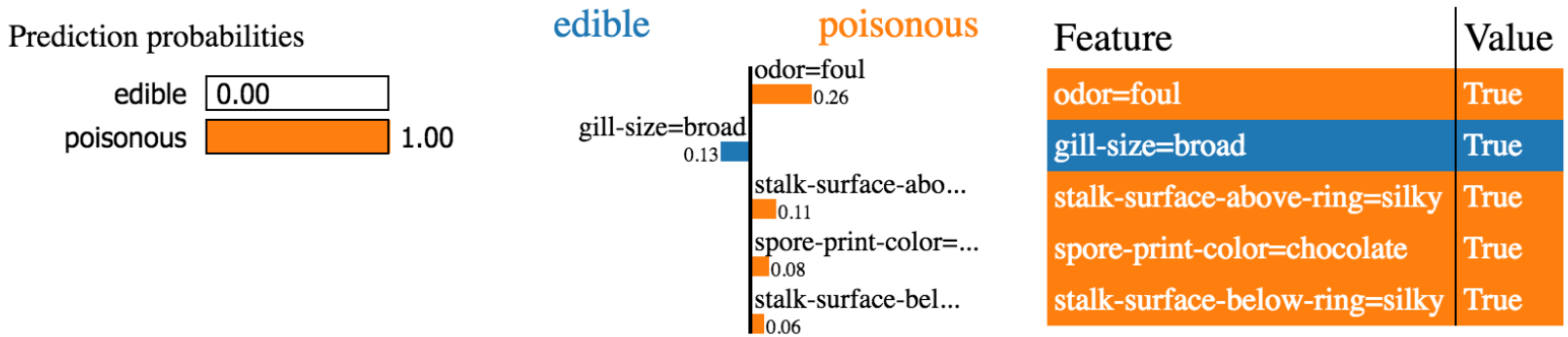
- Vanilla gradient: $\phi_f(x) = \nabla f(x)$
- SmoothGrad: $\phi_f(x) = \mathbb{E}_{Z \sim \mathcal{N}(x,\Sigma)}[\nabla f(Z)]$
- ...

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

Example Attribution Method: LIME

LIME: 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: Explaining Text

LIME for text:[4]

[4] Image source: https://towardsdatascience.com/what-makes-your-question-insincere-in-quora-26ee7658b010

Example: Explaining Text

LIME for text:[4]

Current development process in the literature:

- Specify method with intuitively reasonable properties
- Show examples where it does something intuitively reasonable
- Follow-up studies find that method fails for application X

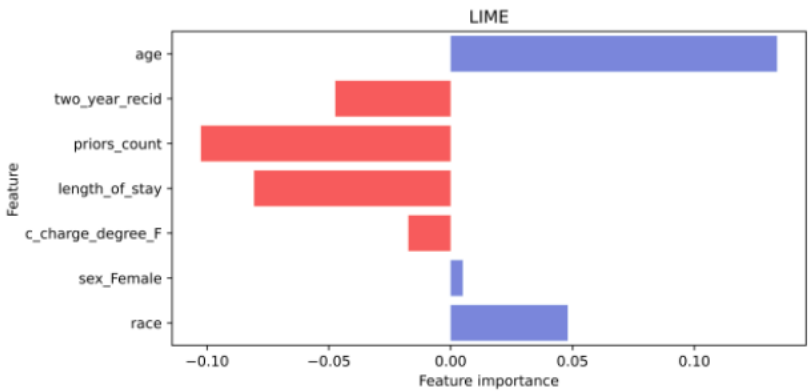
[4] Image source: `https://towardsdatascience.com/what-makes-your-question-insincere-in-quora-26ee7658b010`

Example: What is the Right Explanation?

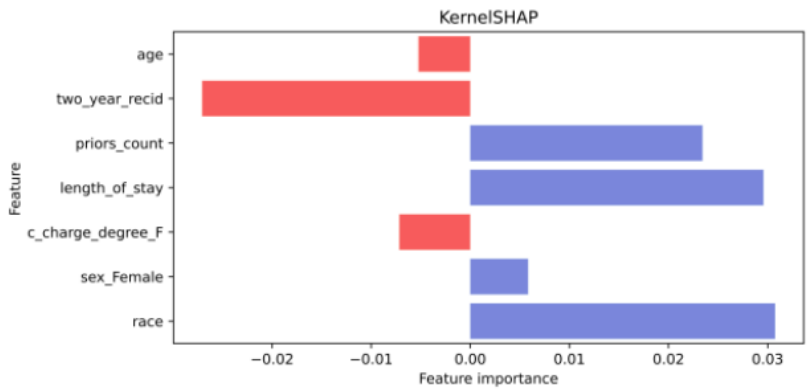
COMPAS data[5]:

- Data collected by Propublica reporters to show that commercial recidivism prediction algorithm used by judges in the USA is biased against black defendants compared to white defendants.

LIME Method



SHAP Method



[5] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Images from [Krishna et al., 2022]

Asymptotic Analysis of LIME for Images

Garreau, Mardaoui

What Does LIME Really See in Images?

ICML, 2021

LIME for Images

1. Decompose image into d superpixels (small, homogeneous patches)[6]
2. Can sample perturbed image $\tilde{X}$ by
   - Sample d Bernoulli$(1/2)$ variables $Z = (Z^1, \ldots, Z^d)$
   - If $Z^j = 1$, then keep j-th superpixel from original image
   - If $Z^j = 0$, then replace j-th superpixel by its average pixel value.

predicted: trailer_truck (35.2%)

LIME explanation

[6] Image courtesy of Damien Garreau

LIME for Images

1. Decompose image into d superpixels (small, homogeneous patches)
2. Can sample perturbed image $\tilde{X}$ by
   - Sample d Bernoulli$(1/2)$ variables $Z = (Z^1, \ldots, Z^d)$
   - If $Z^j = 1$, then keep j-th superpixel from original image
   - If $Z^j = 0$, then replace j-th superpixel by its average pixel value.
3. Query response $\tilde{Y} = f(\tilde{X})$
4. Weight image $\tilde{X}$ by distance to original[6]

$$\pi = \exp\left(-\frac{\mathrm{d}_{\cos}(Z, \mathbb{1})^2}{2\nu^2}\right) \qquad \text{for hyperparameter } \nu > 0$$

5. Sample n times and fit weighted ridge regression[7]

$$\hat{\theta}_n = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \min_{\theta_0 \in \mathbb{R}} \sum_{i=1}^{n} \pi_i (\tilde{Y}_i - Z_i^\intercal \theta + \theta_0)^2 + \lambda \|\theta\|^2$$

[6] $\mathrm{d}_{\cos}(u, v) = 1 - \frac{u^\intercal v}{\|u\| \|v\|}$ is the cosine distance between vectors

[7] In practice $\lambda = 1$ is tiny; in analysis take $\lambda = 0$ for simplicity.

Asymptotic Analysis of LIME for Images

- Recall that $Z = (Z^1, \ldots, Z^d)$ i.i.d. Bernoulli(1/2)
- Induces distribution on weight π and perturbed image $\tilde{X}$

Theorem (Garreau, Mardaoui, 2021)

Suppose f bounded and $\lambda = 0$. Then

$$\hat{\theta}_n \to \theta \qquad \text{in probability,}$$

where

$$\theta_j = c_1 \mathop{\mathbb{E}}_{Z}[\pi f(\tilde{X})] + c_2 \mathop{\mathbb{E}}_{Z}[\pi Z^j f(\tilde{X})] + c_3 \sum_{\substack{k \in \{1,\ldots,d\} \\ k \neq j}} \mathop{\mathbb{E}}_{Z}[\pi Z^k f(\tilde{X})]$$

for some constants c_1, c_2, c_3 that do not depend on f, and which can be computed in closed form.

Consequences

$$\theta_j = c_1 \mathop{\mathbb{E}}_{Z}[\pi f(\tilde{X})] + c_2 \mathop{\mathbb{E}}_{Z}[\pi Z^j f(\tilde{X})] + c_3 \sum_{\substack{k \in \{1,\dots,d\} \\ k \neq j}} \mathop{\mathbb{E}}_{Z}[\pi Z^k f(\tilde{X})]$$

Consequence 1

- Apart from sampling noise, LIME explanations are linear in f:

$$\beta^{f+g} = \beta^f + \beta^g$$

Consequence 2: Large Bandwidth

- As $\nu \to \infty$: $c_1 \to -2$, $c_2 \to 4$, $c_3 \to 0$, and $\pi \to 1$ a.s.

$$\theta_j \to 2\Big(\mathop{\mathbb{E}}_{Z}[f(\tilde{X}) | Z^j = 1] - \mathop{\mathbb{E}}_{Z}[f(\tilde{X})] \Big)$$

- Compares value of f with and without fixing the j-th superpixel to be as in the model.

An Impossibility Result

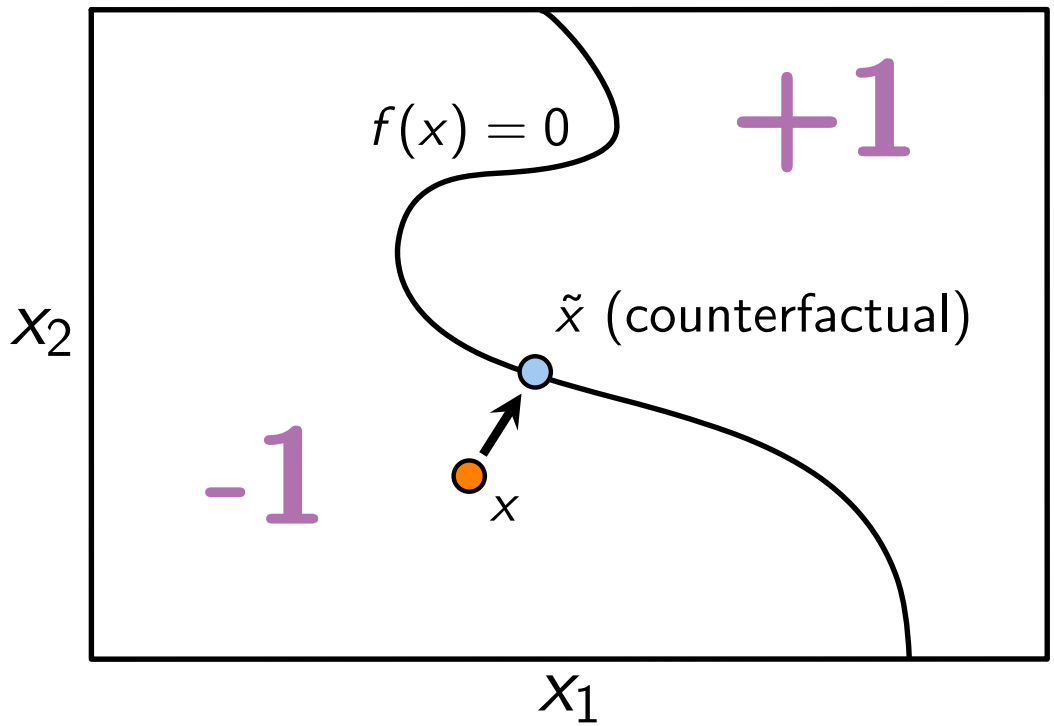
Fokkema, De Heide, Van Erven

Attribution-based Explanations that Provide Recourse Cannot be Robust

ArXiv:2205.15834 preprint, 2023

Example: Counterfactual Explanations

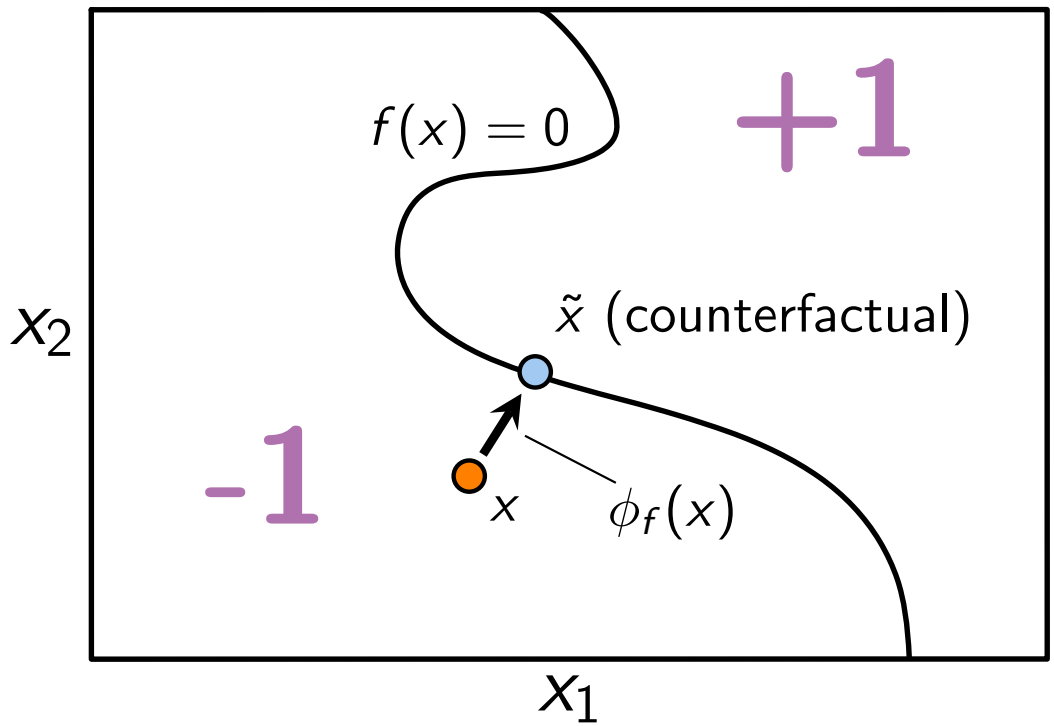
"If you would have had an income of **€40 000** instead of **€35 000**, your loan request would have been approved."



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

Example: Counterfactual Explanations

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

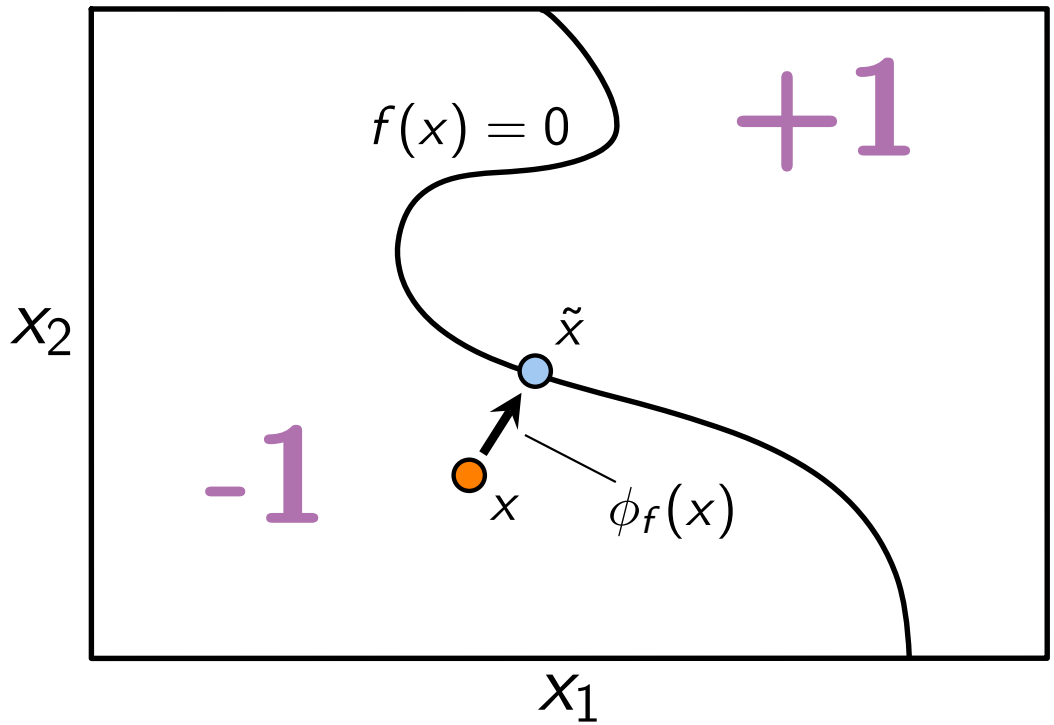


Counterfactual explanation: $\tilde{x} = \underset{x' : \text{sign}(f(x'))=+1}{\arg\min} \text{dist}(x', x)$

Viewed as attribution method: $\phi_f(x) = \tilde{x} - x$

Explanations with Recourse as their Goal

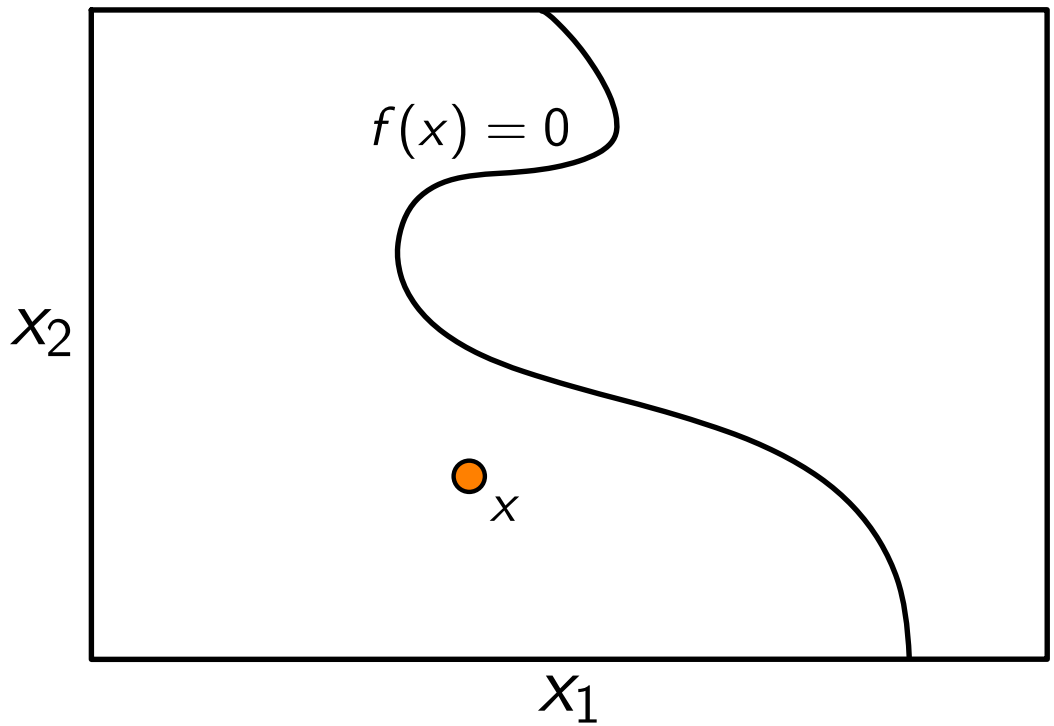
"If you change your current income of **€35 000** to **€40 000**, then your loan request will be approved."



- Attribution methods **provide recourse** if they tell the user how to **change their features** such that f **takes their desired value**.

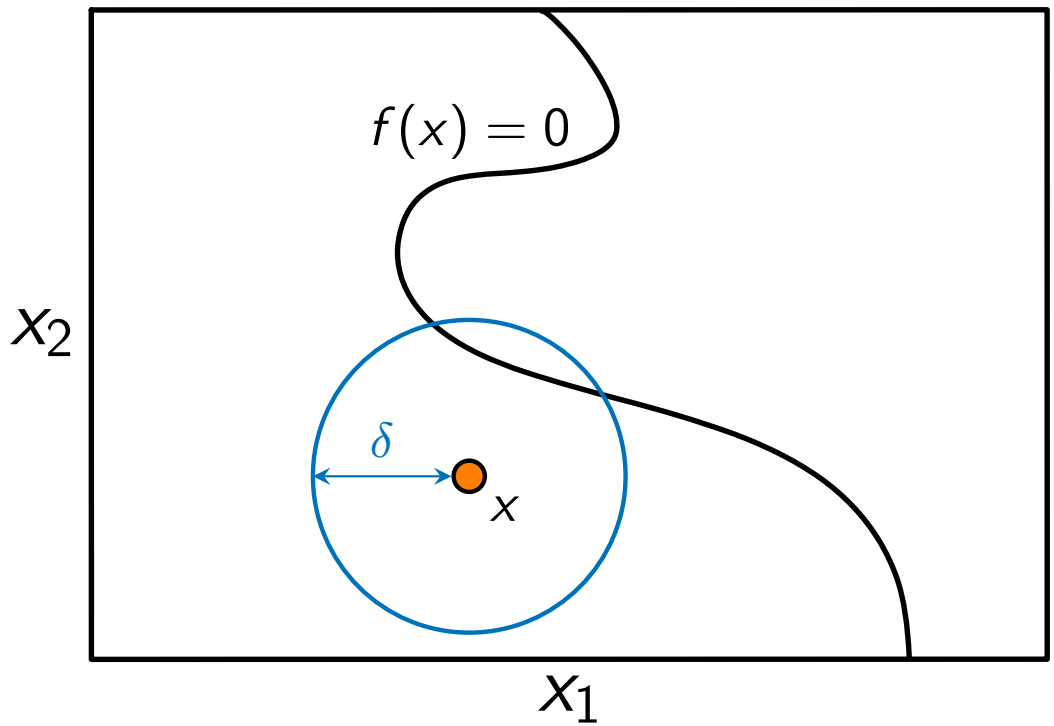
Recourse Sensitivity

- [Fokkema, De Heide, Van Erven, 2022]: our approach to define weakest possible requirement for providing recourse.



Recourse Sensitivity

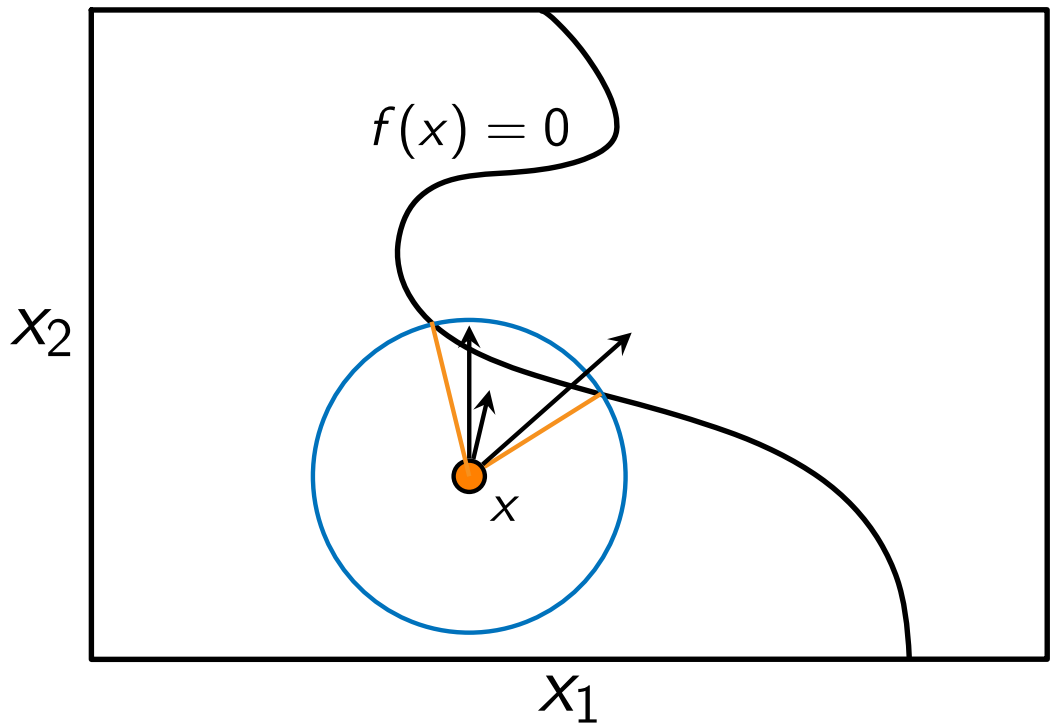
- [Fokkema, De Heide, Van Erven, 2022]: our approach to define weakest possible requirement for providing recourse.



1. Assume user can change their features by at most some $\delta > 0$

Recourse Sensitivity

- [Fokkema, De Heide, Van Erven, 2022]: our approach to define weakest possible requirement for providing recourse.



1. Assume user can change their features by at most some $\delta > 0$
2. $\phi_f(x)$ can point in **any direction that provides recourse** within distance δ, and length does not matter as long as it is > 0.
3. If no direction provides recourse, then $\phi_f(x)$ can be arbitrary.

Robustness of Explanations

Compare:

1. "If you change your current income of **€35 000** to **€40 000**, then your loan request will be approved."
2. "If you change your current income of **€35 001** to **€45 000**, then your loan request will be approved."

Minor changes in x should not cause big changes in explanations!

Robustness of Explanations

Compare:

1. "If you change your current income of **€35 000** to **€40 000**, then your loan request will be approved."
2. "If you change your current income of **€35 001** to **€45 000**, then your loan request will be approved."

Minor changes in x should not cause big changes in explanations!

Robustness: If f is continuous, then ϕ_f should also be **continuous**.
(e.g. survey of recourse by [Karimi et al., 2021])

Robustness of Explanations

Compare:

1. "If you change your current income of **€35 000** to **€40 000**, then your loan request will be approved."
2. "If you change your current income of **€35 001** to **€45 000**, then your loan request will be approved."

Minor changes in x should not cause big changes in explanations!

Robustness: If f is continuous, then ϕ_f should also be **continuous**.
(e.g. survey of recourse by [Karimi et al., 2021])

On the **robustness** of **interpretability methods**

D **Alvarez-Melis**, TS **Jaakkola** - arXiv preprint arXiv:1806.08049, **2018** - arxiv.org

We argue that robustness of explanations---ie, that similar inputs should give rise to similar explanations---is a key desideratum for interpretability. We introduce metrics to quantify robustness and demonstrate that current methods do not perform well according to these metrics. Finally, we propose ways that robustness can be enforced on existing interpretability approaches.

☆ Save 99 Cite Cited by 389 Related articles All 4 versions »

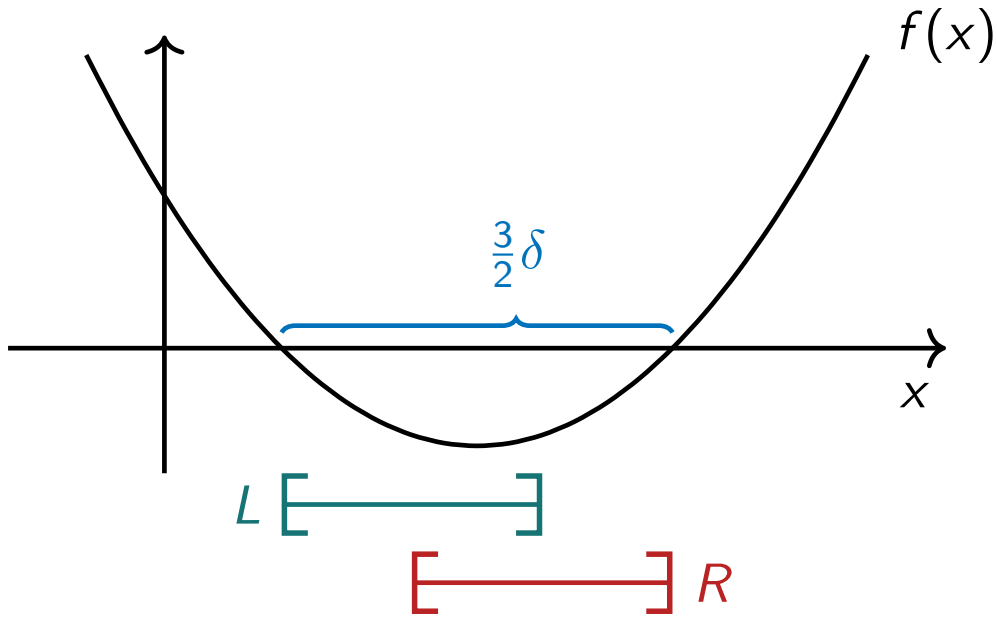
Impossibility in Binary Classification

Theorem (Fokkema, De Heide, Van Erven, 2022)

For any $\delta > 0$ there exists a continuous function f such that no attribution method ϕ_f can be both recourse sensitive and continuous.

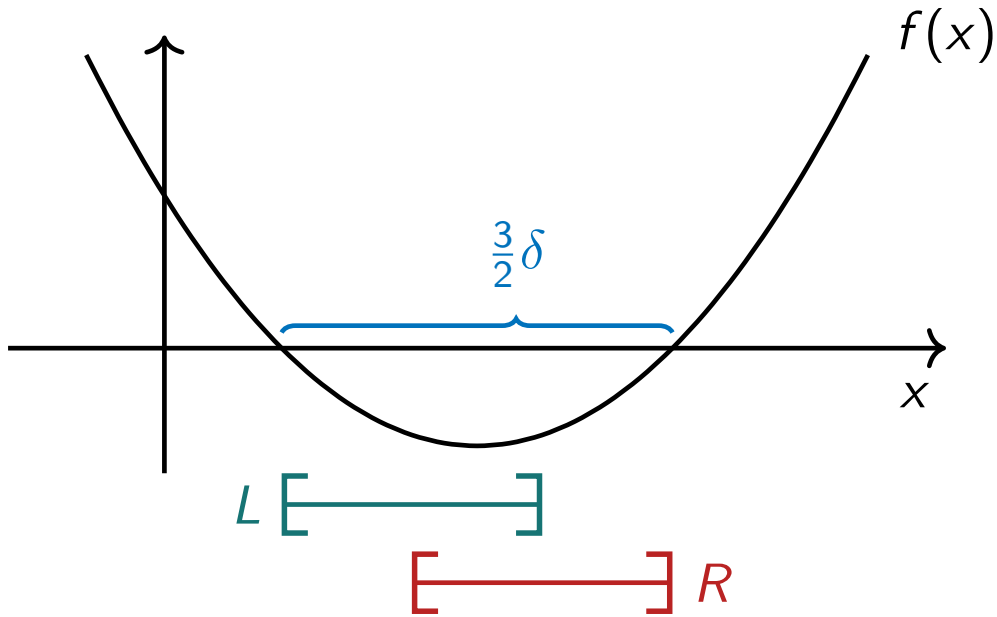
- Power of math: can reason about all explanation methods that could possibly exist

Proof Sketch



$$L = \{x : \text{recourse possible by moving at most } \delta \text{ left}\}$$
$$R = \{x : \text{recourse possible by moving at most } \delta \text{ right}\}$$

Proof Sketch

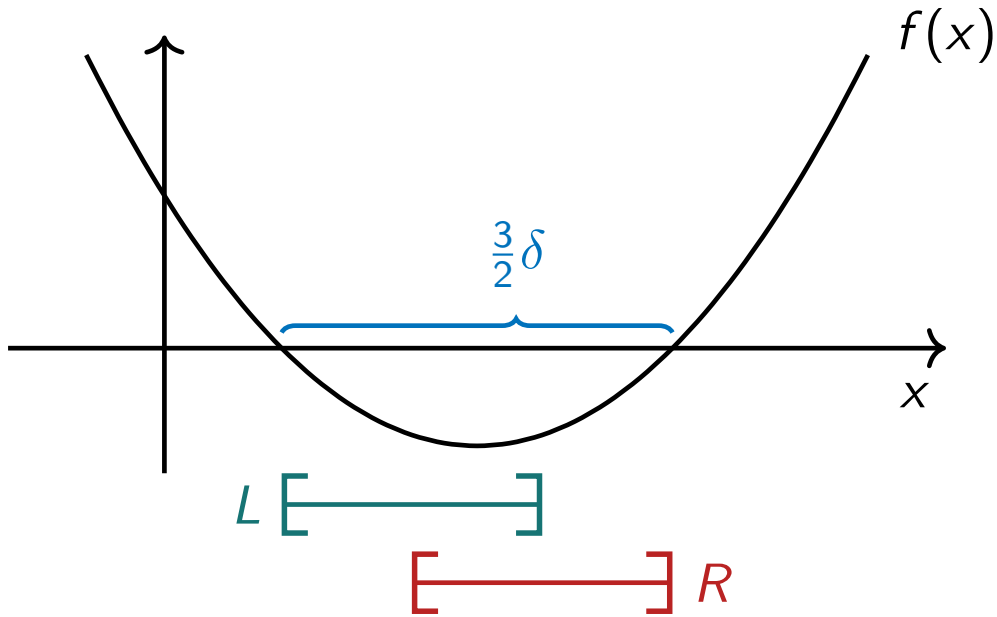


$$L = \{x : \text{recourse possible by moving at most } \delta \text{ left}\}$$
$$R = \{x : \text{recourse possible by moving at most } \delta \text{ right}\}$$

Recourse sensitivity implies:

$$\phi_f(x) \begin{cases} < 0 & \text{for } x \in L \setminus R \\ > 0 & \text{for } x \in R \setminus L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

Proof Sketch



$$L = \{x : \text{recourse possible by moving at most } \delta \text{ left}\}$$
$$R = \{x : \text{recourse possible by moving at most } \delta \text{ right}\}$$

Recourse sensitivity implies:

$$\phi_f(x) \begin{cases} < 0 & \text{for } x \in L \setminus R \\ > 0 & \text{for } x \in R \setminus L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

But this **contradicts continuity**!
(by the mean-value theorem)

Can embed 1D example in higher dimensions as well.

Conclusion

Zooming Out

- Most work on explainability is empirical
- Empirical approach has been very successful in deep learning, but struggles to find proper foundations for explainability
- Formal analysis is slow and leads to more modest claims, but builds up **solid foundations**

Other Noteworthy Formal Results (non-exhaustive list)

- Formal analyses of LIME for other modalities, Anchors, SHAP [Garreau and Luxburg, 2020, Mardaoui and Garreau, 2021, Lopardo et al., 2022, Bordt and von Luxburg, 2022]
- No-free-lunch theorem: no local post-hoc method can perform optimally across all neighborhoods [Han et al., 2022]

References

- Garreau, Mardaoui. **What Does LIME Really See in Images?**, ICML, 2021.
- H. Fokkema, R. de Heide and T. van Erven. **Attribution-based Explanations that Provide Recourse Cannot be Robust**, ArXiv:2205.15834 preprint, 2023.

Other references:

S. Bordt and U. von Luxburg. From Shapley values to generalized additive models and back. *ArXiv:2209.04012 preprint*, 2022.

D. Garreau and U. Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, 2020.

T. Han, S. Srinivas, and H. Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. In *Advances in Neural Information Processing Systems*, 2022.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2021.

G. Lopardo, D. Garreau, and F. Precioso. A sea of words: An in-depth analysis of anchors for text data. *ArXiv:2205.13789 preprint*, 2022.

D. Mardaoui and D. Garreau. An analysis of lime for text data. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3501. PMLR, 2021.

D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *ArXiv:1706.03825*, 2017.