University of Tübingen, November 23, 2022

Impossibility in Explainable Machine Learning:

Attribution-based Explanations that Provide Recourse Cannot be Robust

Tim van Erven

UNIVERSITY OF AMSTERDAM

Joint work with:

Hidde Fokkema

Rianne de Heide

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?
- Give a certain medical diagnosis?
- Make a certain mistake?
- 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

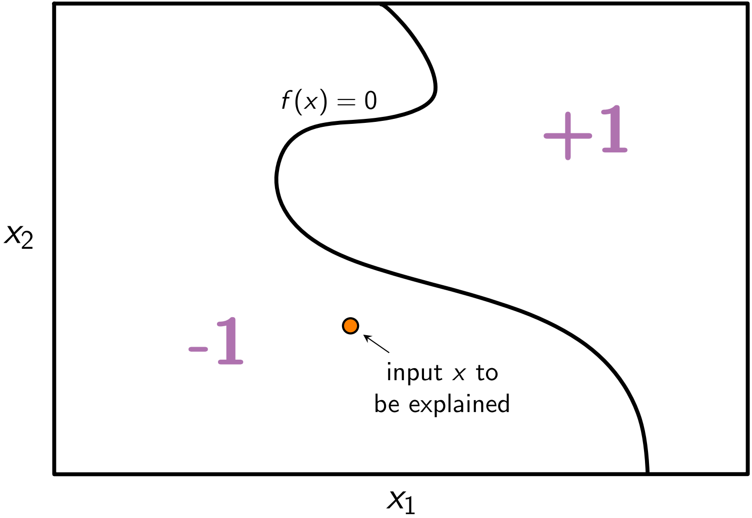
- Classify my company as high risk for money laundering?
- Reject my bank loan?
- Give a certain medical diagnosis?
- Make a certain mistake?
- Reject the profile picture I uploaded to get a public transport card?[1]
- ...

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

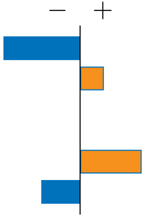
Local Post-hoc Explanations



- **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

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

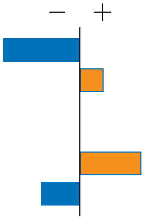


$$\begin{bmatrix} \varphi_f(x)_1 \\ \varphi_f(x)_2 \\ \vdots \\ \varphi_f(x)_{d-1} \\ \varphi_f(x)_d \end{bmatrix} = \varphi_f(x)$$

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

Local Explanations via Attributions

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



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

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

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

$$\phi_f(x)_i = \theta_i \qquad \text{could be } \textbf{coefficient} \text{ of } x_i$$

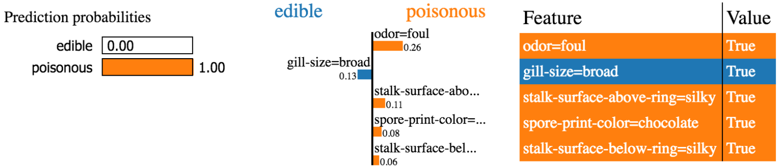
- NB This example is **too simple!** In general $\phi_f(x)$ will depend on x. But many methods can be viewed as local linearizations of f.

Examples of Local Attribution Methods

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:[2]

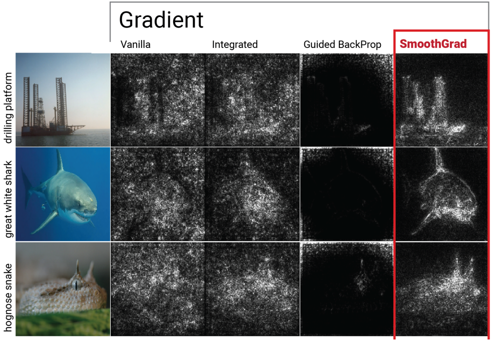


(classifying edibility of mushrooms)

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

Example: Gradient-based Explanations

Various gradient methods[3]

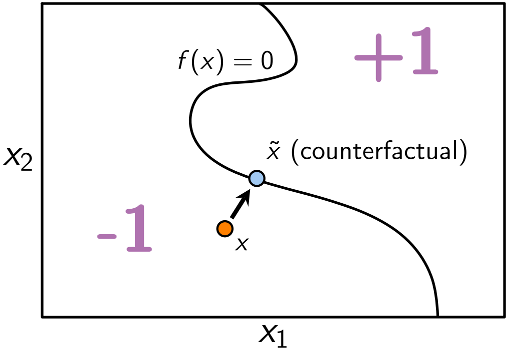


- Vanilla gradient: $\phi_f(x) = \nabla f(x)$
- SmoothGrad: $\phi_f(x) = \mathbb{E}_{Z \sim \mathcal{N}(x,\Sigma)}[\nabla f(Z)]$
- ...

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

Example: Counterfactual Explanations

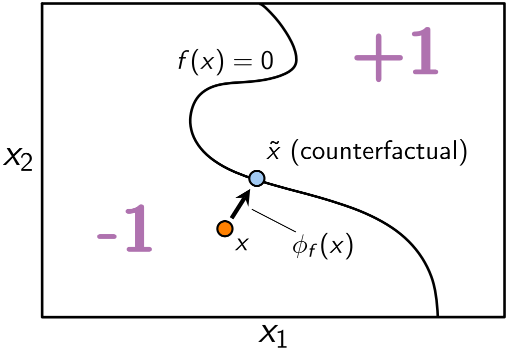
"If you would have had an income of **€40 000** instead of **€35 000**, your loan request would have been approved."



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

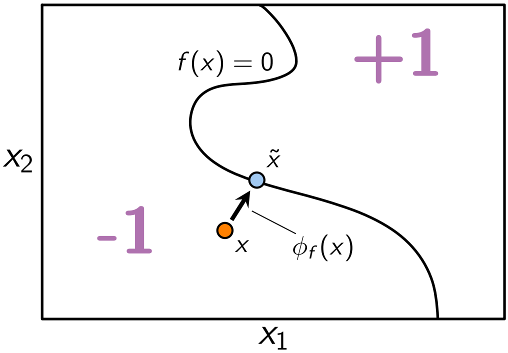
Viewed as attribution method: $\phi_f(x) = \tilde{x} - x$

How Do We Evaluate Explanations?

- When are they good? Are some better than others?
- What is even the **goal** they are trying to achieve?

Explanations with Recourse as their Goal

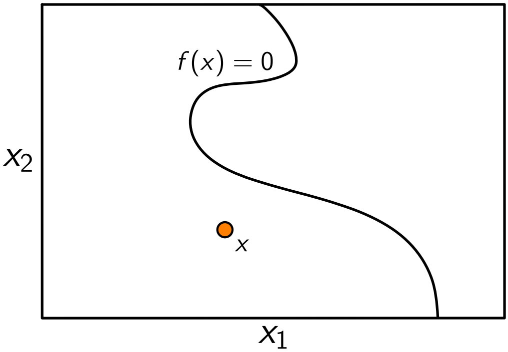
"If you change your current income of **€35 000** to **€40 000**, then your loan request will be approved."



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

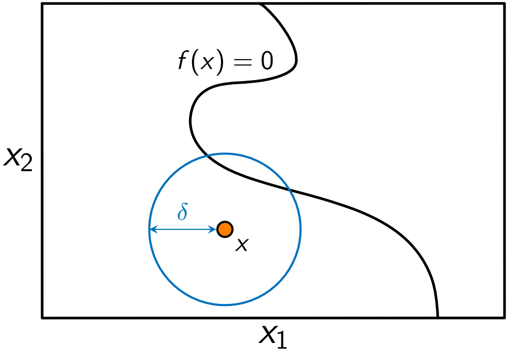
Recourse Sensitivity

- Our definition: weakest possible requirement for providing recourse.



Recourse Sensitivity

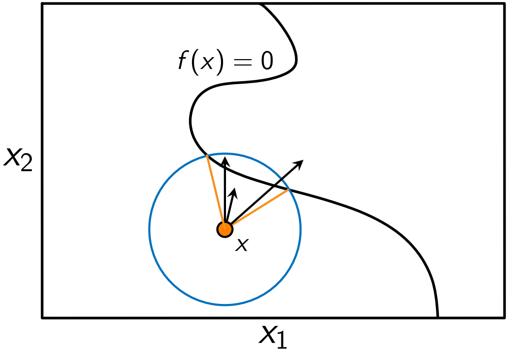
- Our definition: weakest possible requirement for providing recourse.



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

Recourse Sensitivity

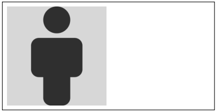
- Our definition: weakest possible requirement for providing recourse.



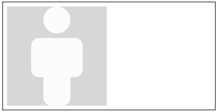
1. Assume user can change their features by at most some $\delta > 0$
2. $\phi_f(x)$ can point in **any direction that provides recourse** within distance δ, and length does not matter as long as it is > 0.
3. If no direction provides recourse, then $\phi_f(x)$ can be arbitrary.

Recourse Sensitivity: Example

Profile picture is accepted if contrast
between profile and background is large enough:



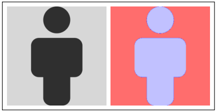
(a) Accepted profile picture



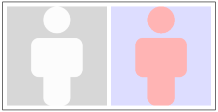
(b) Rejected profile picture

Recourse Sensitivity: Example

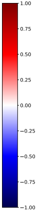
Profile picture is accepted if contrast
between profile and background is large enough:



(a) Accepted profile picture



(b) Rejected profile picture

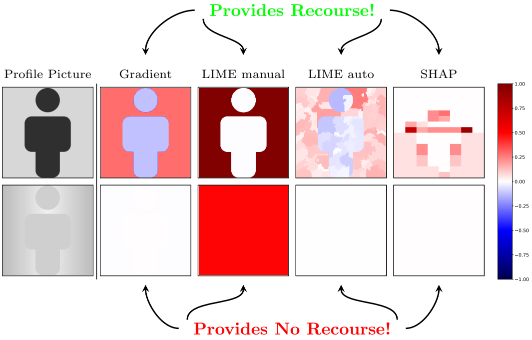


Recourse Sensitivity: Example

Profile picture is accepted if contrast between profile and background is large enough:

(a) Accepted profile picture

(b) Rejected profile picture



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])

Impossibility:

No Single Method Can Be Both Recourse Sensitive and Robust

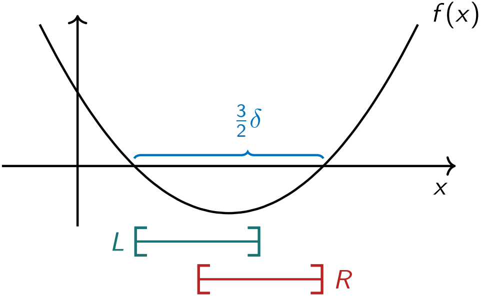
Impossibility in Binary Classification

Suppose the user wants to switch to the $+1$ class in a binary classification setting.

Theorem (For Binary Classification)

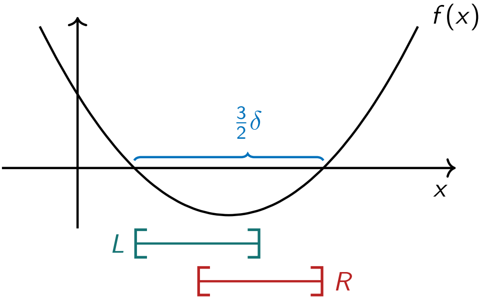
For any $\delta > 0$ there exists a continuous function f such that no attribution method ϕ_f can be both recourse sensitive and continuous.

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

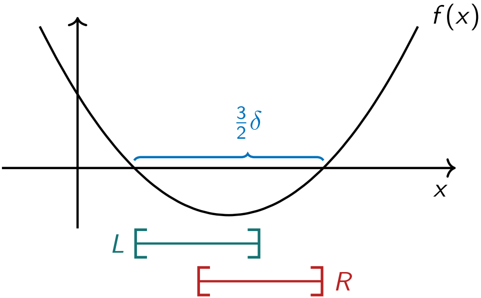


$$L = \{x : \text{recourse possible by moving at most } \delta \text{ left}\}$$
$$R = \{x : \text{recourse possible by moving at most } \delta \text{ right}\}$$

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.

Characterizing Impossible Functions in 1D

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

Theorem

Let $d = 1$, $\delta > 0$. Then there exists a **recourse sensitive** *and* **continuous** *attribution method ϕ_f for a function f if and only if there exist $\tilde{L} \subseteq L$ and $\tilde{R} \subseteq R$ such that*

1. *$\tilde{L} \cup \tilde{R} = L \cup R$ and*
2. *$\tilde{L}$ and $\tilde{R}$ are* **separated**.

Sets A and B are separated if $\mathrm{cl}(A) \cap B = \emptyset$ and $A \cap \mathrm{cl}(B) = \emptyset$.

Characterizing Impossible Functions in 1D

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

Theorem

Let $d = 1$, $\delta > 0$. Then there exists a **recourse sensitive** *and* **continuous** *attribution method ϕ_f for a function f if and only if there exist $\tilde{L} \subseteq L$ and $\tilde{R} \subseteq R$ such that*

1. $\tilde{L} \cup \tilde{R} = L \cup R$ *and*
2. $\tilde{L}$ *and* $\tilde{R}$ *are* **separated**.

Sets A and B are separated if $\text{cl}(A) \cap B = \emptyset$ and $A \cap \text{cl}(B) = \emptyset$.

Proof Ideas:

- $\tilde{L}$ and $\tilde{R}$ determine the sign of ϕ_f on $L \cup R$
- Separatedness gives just enough room for ϕ_f to cross through 0 in between $\tilde{L}$ and $\tilde{R}$

Recourse Beyond Classification

Utility Function:

User with input x is satisfied with point y if $u_f(x, y) \geq \tau$ for some $\tau \geq 0$.

Examples:

- Classification with desired class $+1$: $u_f(x, y) := f(y) \geq +1$
- Absolute increase: $u_f(x, y) := f(y) - f(x) \geq \tau$
- Relative increase by $p \times 100\%$: $u_f(x, y) := \frac{f(y)}{f(x)} \geq 1 + p$

Impossibility for General Utility Functions

Theorem (For General Utility Functions)

Let $\delta > 0, \tau \geq 0$. Assume that

1. $u_f(x, y) = \tilde{u}(f(x), f(y))$ *depends on* x, y *only via* f*;*
2. *There exist* $z_1, z_2 \in \mathbb{R}$ *for which* $\tilde{u}(z_1, z_2) \geq \tau$ *and* $\tilde{u}(z_1, z_1) < \tau$*.*

Then there exists a continuous function f *such that no attribution method* ϕ_f *can be both recourse sensitive and robust.*

Impossibility for General Utility Functions

Theorem (For General Utility Functions)

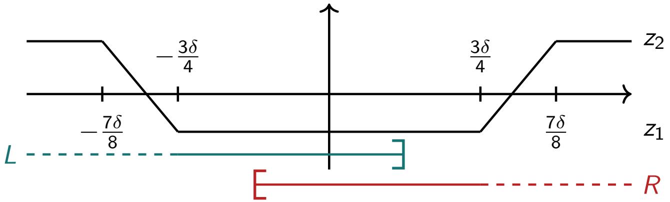
Let $\delta > 0, \tau \geq 0$. Assume that

1. $u_f(x, y) = \tilde{u}(f(x), f(y))$ *depends on* x, y *only via* f;
2. *There exist* $z_1, z_2 \in \mathbb{R}$ *for which* $\tilde{u}(z_1, z_2) \geq \tau$ *and* $\tilde{u}(z_1, z_1) < \tau$.

Then there exists a continuous function f *such that no attribution method* ϕ_f *can be both recourse sensitive and robust.*

Proof Idea:

- Like impossibility for binary classification with this f:



Conclusion

Summary:

- Exist f for which recourse sensitivity $+$ robustness is **impossible**, for classification and other utility functions
- Exact **characterisation** of impossible f, but only **for 1D**
- Further extensions in the paper:
  - Include constraints on user actions
  - Characterisation in arbitrary dimensions when user can only change a single feature
  - Sufficient conditions on f under which impossibility is avoided

Conclusion

Summary:

- Exist f for which recourse sensitivity + robustness is **impossible**, for classification and other utility functions
- Exact **characterisation** of impossible f, but only **for 1D**
- Further extensions in the paper:
  - Include constraints on user actions
  - Characterisation in arbitrary dimensions when user can only change a single feature
  - Sufficient conditions on f under which impossibility is avoided

Discussion:

Is impossibility a really bad problem?
Not, but need to **refine formal goals** of explainability for recourse. E.g.:

- Accept that robustness sometimes fails
- Set-valued explanations
- Randomized explanations
- ...

References

- H. Fokkema, R. de Heide and T. van Erven. **Attribution-based Explanations that Provide Recourse Cannot be Robust**, ArXiv:2205.15834 preprint, 2022.

Other references:

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.

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