Impossibility of Robustness and Recourse[(*)]

A formal result in Explainable AI

Hidde Fokkema[(1)], Rianne de Heide[(2)] and Tim van Erven[(1)].

(1): University of Amsterdam, Korteweg-de Vries Institute for Mathematics, (2): Vrije Universiteit Amsterdam, Department of Mathematics

University of Amsterdam
Korteweg-De Vries instituut for Mathematics
VU Vrije Universiteit Amsterdam

Abstract

For any way of measuring utility, there exists a (continuous) machine learning model f for which no attribution method φ_f can provide explanations that are both recourse sensitive and continuous.

Utility

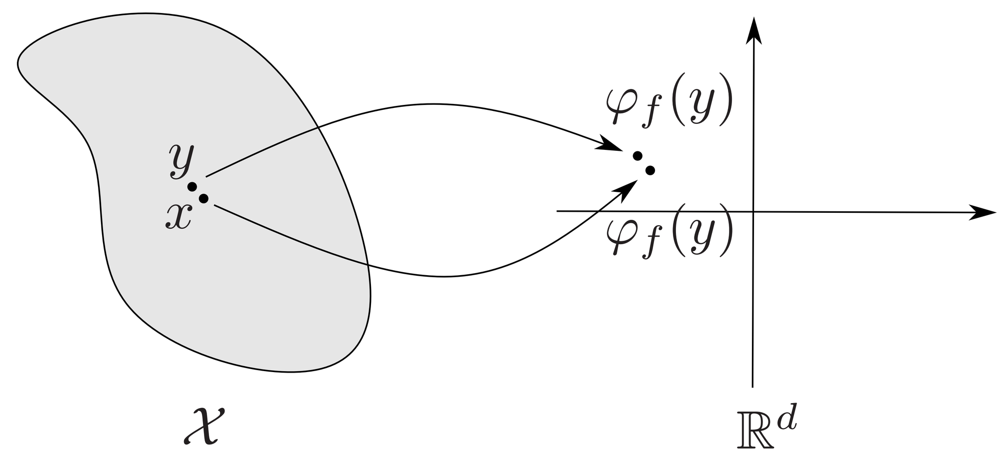
A *Utility function*, $u_f(x, y)$, measures the utility experienced by a user when changing x to y.
A user is satisfied if $u_f(x, y) \geq \tau$ for some *threshold* τ.
Some examples,

- **Flip the class label:**
  $u_f(x, y) = |\text{sign}(f(x)) - \text{sign}(f(y))| \geq 2.$
- **Increase score by amount** τ:
  $u_f(x, y) = f(y) - f(x) \geq \tau.$
- **Increase probability by** $p \times 100\%$:
  $u_f(x, y) = \dfrac{f(y)}{f(x)} \geq 1 + p.$

Robustness

An attribution function φ_f for f is called *Robust* if it is continuous.

Similar users require similar explanations



Attribution Methods

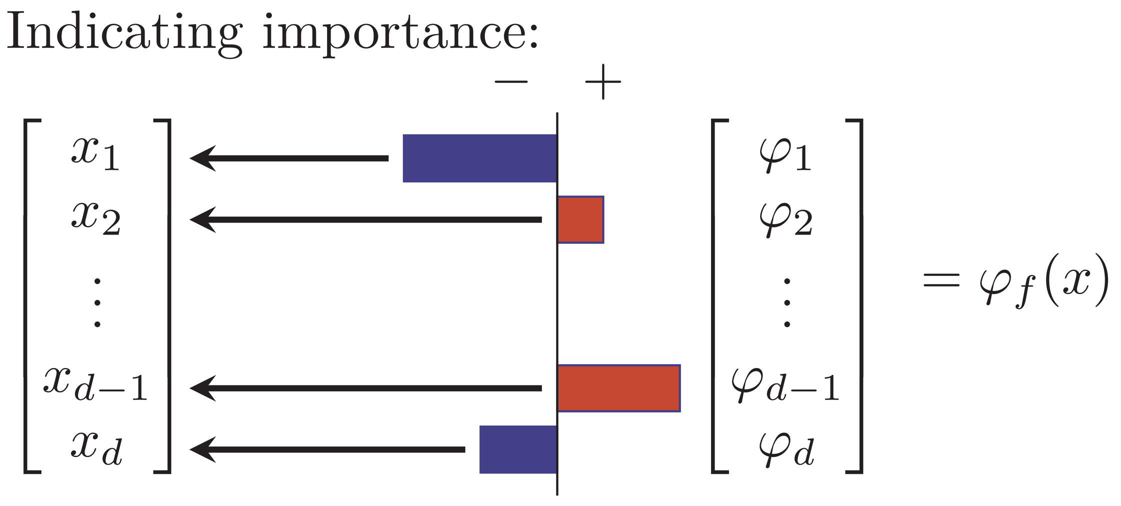
Machine learning model, e.g. a classifier:

$$f \colon \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}, \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mapsto y.$$

An *Attribution function* for f is a mapping

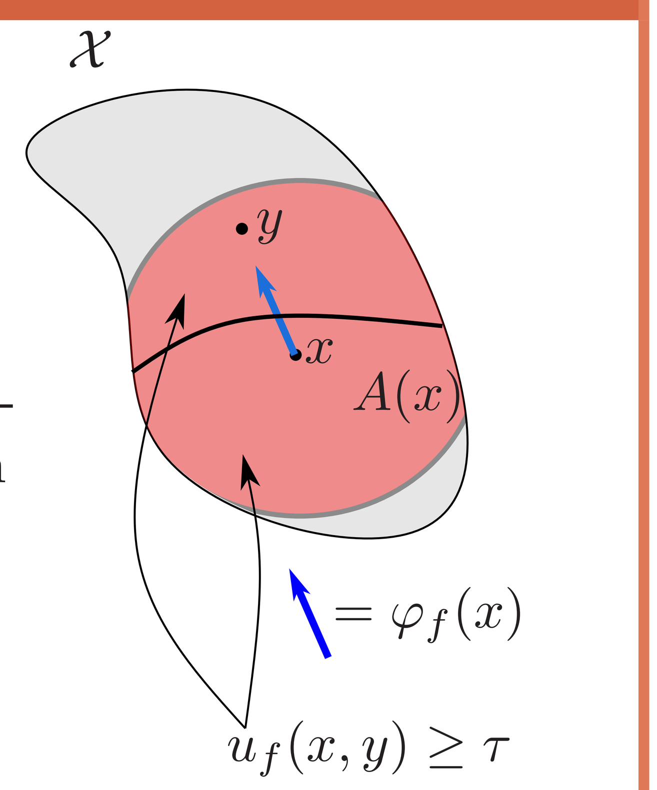
$$\varphi_f \colon \mathcal{X} \to \mathbb{R}^d.$$

Indicating importance:

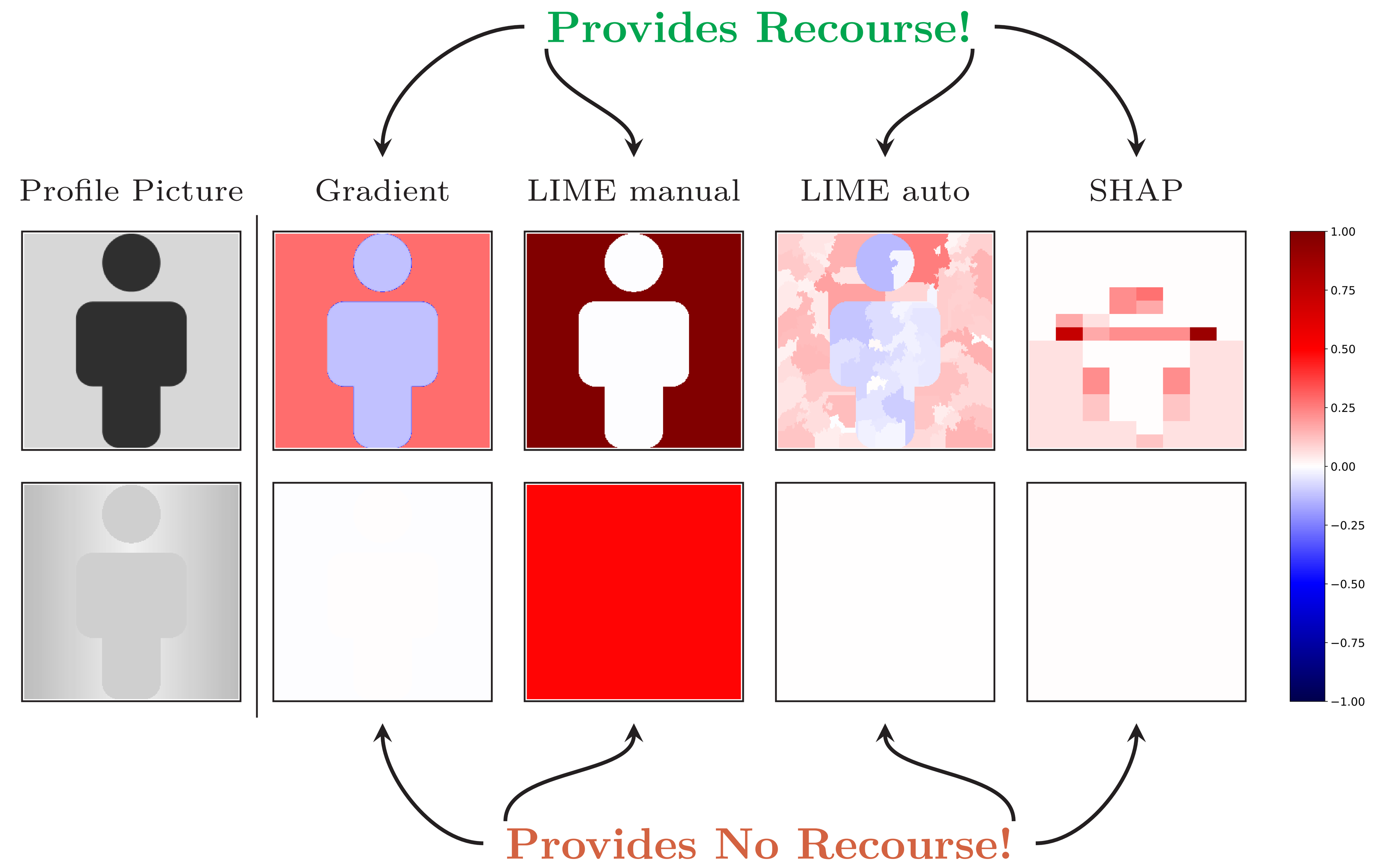


Recourse Sensitivity

Informally, an attribution method is called *Recourse Sensitive* if the user can achieve a sufficient utility increase when moving in the direction of $\varphi_f(x)$.



Profile picture example



Impossibility Result

Theorem 1. *Suppose $\mathcal{X} = \mathbb{R}^d$, $u_f(x, y) = |\text{sign}(f(x)) - \text{sign}(f(y))|$, $\tau = 2$ and $\delta > 0$. Then, there exists a continuous classifier $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and continuous.*

Recourse Sensitivity Expanded

Formally, define set of attainable points from x

$$A(x) = \{y \in \mathcal{X} \mid \|x - y\| \leq \delta, y \in C(x)\}.$$

The set $C(x)$ will impose some constraints. Examples:

- $C(x) = \mathcal{X}$, the unrestricted case.
- $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq k\}$, sparse change.
- $C(x) = \{y \in \mathcal{X} \mid y = x + \alpha z, \alpha \geq 0, z \in D\}$, certain directions D.

The points around x that are both attainable and achieve sufficient utility are given by

$$T(x) = \{y \in A(x) \mid u_f(x, y) \geq \tau\}.$$

An attribution φ_f is called *Recourse Sensitive* if $\varphi_f(x) = \alpha(y - x)$ for some $\alpha > 0$ and $y \in T(x)$, for all $x \in \mathcal{X}$ for which $T(x)$ is non-empty.

General Results

Impossibility

Theorem 2. *If u_f is of the form $u_f(x, y) = \widetilde{u}(f(x), f(y))$ and if there exist $z_1, z_2 \in \mathbb{R}^d$ such that $\widetilde{u}(z_1, z_2) \geq \tau$ and $\widetilde{u}(z_1, z_1) < \tau$. Then there exists a continuous $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and robust.*

Exact characterization, One feature version

Setting:

- $\mathcal{X} \subseteq \mathbb{R}^d$,
- $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq 1\}$,
- $\delta, \tau > 0, \alpha \in [0, \delta]$ and $u_f(x, x) < \tau$ for all $x \in \mathcal{X}$,
- $L^i = \{x \in \mathcal{X} \mid u_f(x, y) \geq \tau, y = x - \alpha e_i\}$,
- $R^i = \{x \in \mathcal{X} \mid u_f(x, y) \geq \tau, y = x + \alpha e_i\}$.

Theorem 3. *A continuous recourse sensitive attribution function φ_f for f exists if and only if there exist $\widetilde{L}^i \subseteq L^i$ and $\widetilde{R}^i \subseteq R^i$ for all $i = 1, \ldots, d$ such that $\bigcup_{i=1}^d \widetilde{L}^i \cup \widetilde{R}^i = \bigcup_{i=1}^d L^i \cup R^i$ and, $\widetilde{L}^i$ and $\widetilde{R}^i$ are all pairwise separated.*

Footnotes

(*) 'Full title':
Attribution-based Explanations that Provide Recourse Cannot be Robust.
arXiv:2205.15834