The Limits of Explainable Machine Learning: Some Things Are Simply Impossible

Tim van Erven

University of Amsterdam

Joint work with:

Hidde Fokkema 

Rianne de Heide
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 new OV chipcard?¹
- ...
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 new OV chipcard?¹
▶ . . .

A Communication Limit:
▶ Cannot communicate millions of parameters!
▶ Can communicate only some relevant aspects and/or need high-level concepts in common with user

¹Personal experience
Binary Classification

\[ f(x) = 0 \]

\( x_1 \) (credit score)

\( x_2 \) (age)

Loan

No Loan
Binary Classification

\[ f(x) = 0 \]

Input \( x \) to be explained.
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 \(f\).
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:**

<table>
<thead>
<tr>
<th>Prediction probabilities</th>
<th>edible</th>
<th>poisonous</th>
</tr>
</thead>
<tbody>
<tr>
<td>edible</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>poisonous</td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>odor=foul</td>
<td>True</td>
</tr>
<tr>
<td>gill-size=broad</td>
<td>True</td>
</tr>
<tr>
<td>stalk-surface-above-ring=silky</td>
<td>True</td>
</tr>
<tr>
<td>spore-print-color=chocolate</td>
<td>True</td>
</tr>
<tr>
<td>stalk-surface-below-ring=silky</td>
<td>True</td>
</tr>
</tbody>
</table>

(classifying edibility of mushrooms)

---

²Image source: https://github.com/marcotcr/lime
Example Attribution Method: LIME

**LIME:** Do local linear approximation of \( f \) near \( x \) (optionally in dimensionality reduced space), and report coefficients.

---

**LIME for images:**

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

---

\(^2\)Image source: [Ribeiro et al., 2016]
### Various gradient methods\(^3\)

<table>
<thead>
<tr>
<th>Gradient</th>
<th>Vanilla</th>
<th>Integrated</th>
<th>Guided BackProp</th>
<th>SmoothGrad</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>drilling platform</strong></td>
<td><img src="path/to/image1" alt="Image" /></td>
<td><img src="path/to/image2" alt="Image" /></td>
<td><img src="path/to/image3" alt="Image" /></td>
<td><img src="path/to/image4" alt="Image" /></td>
</tr>
<tr>
<td><strong>great white shark</strong></td>
<td><img src="path/to/image5" alt="Image" /></td>
<td><img src="path/to/image6" alt="Image" /></td>
<td><img src="path/to/image7" alt="Image" /></td>
<td><img src="path/to/image8" alt="Image" /></td>
</tr>
<tr>
<td><strong>hognose snake</strong></td>
<td><img src="path/to/image9" alt="Image" /></td>
<td><img src="path/to/image10" alt="Image" /></td>
<td><img src="path/to/image11" alt="Image" /></td>
<td><img src="path/to/image12" alt="Image" /></td>
</tr>
</tbody>
</table>

- **Vanilla gradient:** \( \phi_f(x) = \nabla f(x) \)
- **SmoothGrad:** \( \phi_f(x) = \mathbb{E}_{Z \sim \mathcal{N}(x, \Sigma)}[\nabla f(Z)] \)
- ...
**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: 
\[
\tilde{x} = \arg \min_{x': \text{sign}(f(x')) \neq \text{sign}(f(x))} \text{dist}(x', x)
\]
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: \( \tilde{x} = \text{arg min}_{x' : \text{sign}(f(x')) \neq \text{sign}(f(x))} \text{dist}(x', x) \)

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.
Impossibility:

No Single Method Can Be Both Recourse Sensitive and Robust

**Theorem**

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

- Our definition: weakest possible requirement for providing recourse.

\[ f(x) = 0 \]

Assume user can change their features by at most some \( \delta > 0 \).

\( \phi f(x) \) can point in any direction that provides recourse within distance \( \delta \), and length does not matter as long as it is >0.

If no direction provides recourse, then \( \phi f(x) \) can be arbitrary.
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 $\delta$, 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

Provides Recourse!

Profile Picture Gradient LIME manual LIME auto SHAP

Provides No Recourse!
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 $\phi_f$ should also be continuous. (e.g. survey of recourse by [Karimi et al., 2021])
Conclusion

Summary:
- In binary classification: exist $f$ for which recourse sensitivity + robustness is **impossible**
- Further extensions in the paper:
  - Generalization to multiclass and regression using utility functions
  - Include constraints on user actions
  - Exact characterization of impossible $f$ when user can only change a single feature

Discussion:
Is the field of explainable machine learning in trouble?
No, but need to refine goals of explainability for recourse. E.g.:
- Accept that robustness sometimes fails
- Set-valued explanations
- Randomized explanations
Conclusion

Summary:
- In binary classification: exist \( f \) for which recourse sensitivity + robustness is impossible
- Further extensions in the paper:
  - Generalization to multiclass and regression using utility functions
  - Include constraints on user actions
  - Exact characterization of impossible \( f \) when user can only change a single feature

Discussion:
Is the field of explainable machine learning in trouble?
Not, but need to refine goals of explainability for recourse. E.g.:
- Accept that robustness sometimes fails
- Set-valued explanations
- Randomized explanations
- ...
References


Other references:


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

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.

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