# Explaining Black-Box Classifiers: Properties and Functions

Leila Amgoud

CNRS – IRIT
Toulouse – France

November 24, 2021

# Introduction

- Noteworthy advances in data-driven AI
  - Voice recognition, image recognition, . . .
- Broad range of machine learning models
  - Interpretable models (eg. decision trees)
  - Non-interpretable models (eg. deep neural networks)

# Black-box Classifiers

Features: $\mathcal{F} = \{f_1, \ldots, f_n\}$

Domains: $\mathcal{D}_1, \ldots, \mathcal{D}_n$

Literals: $\mathcal{U}$ is the set of all pairs $(f_i, v)$ where $f_i \in \mathcal{F}$, $v \in \mathcal{D}_i$

Features space: $\mathcal{X}$ is the set of $n-$tuples of literals of the form

$$\{(f_1, v_{1i}), \ldots, (f_n, v_{nl})\}$$

Classes: $\mathcal{C} = \{c_1, \ldots, c_k\}$, $k \geq 2$

Classifier is a function $\mathtt{f} : \mathcal{X} \to \mathcal{C}$

Consistency A set $H \subseteq \mathcal{U}$ is consistent if $\nexists (f, v), (f', v') \in H$ s.t. $f = f'$ and $v \neq v'$

# Example

| Instances | *Vacation* | *Concert* | *Meeting* | *Exhibition* | *Hiking* |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 | 1 |
| $x_5$ | 0 | 1 | 1 | 0 | 0 |
| $x_6$ | 0 | 1 | 1 | 1 | 0 |
| $x_7$ | 1 | 1 | 0 | 1 | 1 |

# Explainability of Classifiers

Goal: to explain the classifier's outcomes. Useful for

- giving feedback for users
- improving trust in decisions made by the model
- improving the model's outcomes

| Instances | *Vacation* | *Concert* | *Meeting* | *Exhibition* | *Hiking* |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 | 1 |
| $x_5$ | 0 | 1 | 1 | 0 | 0 |
| $x_6$ | 0 | 1 | 1 | 1 | 0 |
| $x_7$ | 1 | 1 | 0 | 1 | 1 |

Why does the model predict class 1 for instance $x_7$?

# Approaches for Explaining Classifiers

- Trace-based approach retraces the *internal decision-making process* of the model
  - ✓ **Real**/**Certain** explanations
  - × Not easy to grasp for non experts
  - × Not feasible for **non-interpretable** models

- Input-Output approach looks for correlations between input data and predictions
  - ✓ Feasible for any model
  - × **Plausible** explanations

# Gloabal vs. Local Explanation Functions

Let $\texttt{f} : \mathcal{X} \rightarrow \mathcal{C}$ be a classifier

- Global explanations describe global behaviour of $\texttt{f}$ $(\texttt{g} : \mathcal{C} \rightarrow \mathcal{G})$
  *Eg.* $\texttt{f}$ *predicts hike whenever a person is on vacation*

- Local explanations focus on individual instances $(\texttt{s} : \mathcal{X} \rightarrow \mathcal{G}')$
  - *Eg.* $\texttt{f}$ *predicts not hike for instance* $x_1$ *because the person has a meeting*

# Research Questions

- What are the properties of reasonable explanation functions?
- What are the types of explanations?
- How to define reasonable explanation functions that generate each type of explanation?

# Properties

$$\texttt{f} : \mathcal{X} \rightarrow \mathcal{C} \qquad s : \mathcal{C} \rightarrow \mathcal{G} \qquad g : \mathcal{X} \rightarrow \mathcal{G}'$$

Non-emptyness: Every instance should have an explanation ($\forall x \in \mathcal{X}$, $g(x) \neq \emptyset$)

Non-Triviality Explanations should be informative ($\forall x \in \mathcal{X}$, $\emptyset \notin g(x)$)

Consistency: $\forall x \in \mathcal{X}$, $g(x) \subseteq s(f(x))$

Soundness: An explanation should contain only information that is **relevant** to a prediction.

Compleneness: Information that is not part of explanations is **irrelevant** to the predicted class.

Representativity: there exists $t : \mathcal{G}' \rightarrow \mathcal{C}$ such that for any $x \in \mathcal{X}$, $t(g(x)) = \texttt{f}(x)$

Coherence: Compatible explanations should concern compatible predictions.

# Properties (cont.)

| $\mathcal{Y}$ | *Vacation* | *Concert* | *Meeting* | *Exhibition* | *Hiking* |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 | 1 |
| $x_5$ | 0 | 1 | 1 | 0 | 0 |
| $x_6$ | 0 | 1 | 1 | 1 | 0 |
| $x_7$ | 1 | 1 | 0 | 1 | 1 |

- $\mathtt{g}(x_1) = \{U_1\}$ $\quad U_1 = \{(V, 0)\}$
- $\mathtt{g}(x_2) = \{U_2\}$ $\quad U_2 = \{(M, 0)\}$

**Incoherence**

$U_1 \cup U_2 = \{(V, 0), (M, 0)\}$ is consistent while $\mathtt{f}(x_1) \neq \mathtt{f}(x_2)$

# Types of Explanations: Global Explanations

Abductive explanations[1] = Key factors that cause a given class

*Class c is suggested because* $(f_i, v_i), \ldots, (f_k, v_k)$

Examples

- Not hike because there is a meeting
- Reject a loan because annual income is $30K$

## Argument Pro

*An argument pro a class* $c \in \mathcal{C}$ *is a pair* $\langle H, c \rangle$ *s.t.*

- $H \subseteq \mathcal{U}$
- *H is consistent*
- $\forall x \in \mathcal{X}$ *s.t.* $H \subseteq x$, $\mathtt{f}(x) = c$
- $\nexists H' \subset H$ *such that* $H'$ *satisfies the third condition.*

$\mathtt{Pros}(c)$ *denotes the set of all arguments pro c.*

[1] Other terminology: Prime Implicants, Miminimal Sufficient Subsets, Pertinent Positives.

# Example

| $\mathcal{X}$ | $f_1$ | $f_2$ | $\mathtt{f}(.)$ |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

- $\mathtt{Pros}(c_1) = \{a_1\}$
- $\mathtt{Pros}(c_2) = \{a_2\}$
- $\mathtt{Pros}(c_3) = \{a_3\}$

$$a_1 = \langle\{(f_1, 0), (f_2, 0)\}, c_1\rangle$$
$$a_2 = \langle\{(f_1, 0), (f_2, 1)\}, c_2\rangle$$
$$a_3 = \langle\{(f_1, 1)\}, c_3\rangle$$

**Proposition**

*Let $c \in \mathcal{C}$.*

- $(\texttt{Pros}(c) = \emptyset) \iff (\forall x \in \mathcal{X}, \texttt{f}(x) \neq c)$.
- $(\texttt{Pros}(c) = \{\langle \emptyset, c \rangle\}) \iff (\forall x \in \mathcal{X}, \texttt{f}(x) = c)$
- *If* $\exists x \in \mathcal{X}$ *s.t.* $\texttt{f}(x) = c$, *then* $\exists \langle H, c \rangle \in \texttt{Pros}(c)$. *Furthermore,* $H \subseteq x$.
- *If* $\exists \langle H, c \rangle \in \texttt{Pros}(c)$, *then* $\exists x \in \mathcal{X}$ *s.t.* $\texttt{f}(x) = c$.
- *Let* $c, c' \in \mathcal{C}$ *with* $c \neq c'$. $\forall \langle H, c \rangle \in \texttt{Pros}(c), \forall \langle H', c' \rangle \in \texttt{Pros}(c')$,

  $$H \cup H' \text{ is inconsistent.}$$

- *The function* `Pros` *satisfies all the properties.*

# Types of Explanations: Global Explanations (cont.)

Counterfactuals[2] = Changes that result in *another* outcome

- *If* $(f_i, v_i), \ldots, (f_k, v_k)$, the class would not have been $c$

Example

- If the annual income has been $45K$, the loan would have been offered

**Argument Con**

*Let $c \in \mathcal{C}$. An argument con $c$ is a pair $\langle H, \overline{c} \rangle$ s.t.*

- $H \subseteq \mathcal{U}$
- *$H$ is consistent*
- $\forall x \in \mathcal{X}$ *s.t.* $H \subseteq x$, $\mathtt{f}(x) \neq c$
- $\nexists H' \subset H$ *such that $H'$ satisfies the third condition.*

$\mathtt{Cons}(c)$ *denotes the set of all arguments con $c$.*

[2] Other terminology: Contrastive, Pertinent Negatives, Adversarial Examples.

# Example (cont.)

| $\mathcal{X}$ | $f_1$ | $f_2$ | $\mathtt{f}(.)$ |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

- $\mathtt{Cons}(c_1) = \{b_1, b_2\}$

$$b_1 = \langle\{(f_1, 1)\}, \overline{c_1}\rangle$$
$$b_2 = \langle\{(f_2, 1)\}, \overline{c_1}\rangle$$

- $\mathtt{Cons}(c_2) = \{b_3, b_4\}$

$$b_3 = \langle\{(f_1, 1)\}, \overline{c_2}\rangle$$
$$b_4 = \langle\{(f_2, 0)\}, \overline{c_2}\rangle$$

- $\mathtt{Cons}(c_3) = \{b_5\}$

$$b_5 = \langle\{(f_1, 0)\}, \overline{c_3}\rangle$$

# Global Explanations: Arguments Con Classes (cont.

**Proposition**

*Let $c \in \mathcal{C}$.*

- $(\texttt{Pros}(c) = \emptyset) \iff (\texttt{Cons}(c) = \{\langle \emptyset, \overline{c} \rangle\})$
- $(\texttt{Cons}(c) = \emptyset) \iff (\texttt{Pros}(c) = \{\langle \emptyset, c \rangle\})$
- *If $\mathcal{C} = \{c, c'\}$, then*
  - $\texttt{Pros}(c) = \{\langle H, c \rangle \mid \langle H, \overline{c'} \rangle \in \texttt{Cons}(c')\}$
  - $\texttt{Cons}(c) = \{\langle H, \overline{c} \rangle \mid \langle H, c' \rangle \in \texttt{Pros}(c')\}$
- *For all $\langle H, c \rangle \in \texttt{Pros}(c)$, $\langle H', \overline{c} \rangle \in \texttt{Cons}(c)$, the set $H \cup H'$ is inconsistent.*
- *The function* $\texttt{Cons}$ *satisfies all the properties.*

# Duality of Pros and Cons

**Supp**

*Let* $c \in \mathcal{C}$ *and* $\texttt{Supp}(c) = \{H_1, \ldots, H_k\}$ *s.t for every* $i = 1, \ldots, k$,

- $H_i \subseteq \mathcal{U}$
- $H_i$ *is consistent*
- $\forall \langle H, \overline{c} \rangle \in \texttt{Cons}(c)$, $H \cup H_i$ *is inconsistent*
- $\nexists H' \subset H_i$ *s.t.* $H'$ *satisfies the third condition.*

**Theorem**

*Let* $c \in \mathcal{C}$.

$$\texttt{Pros}(c) = \{\langle H, c \rangle \mid H \in \texttt{Supp}(c)\}$$

# Duality of Pros and Cons (cont.)

**Att**

*Let* $c \in \mathcal{C}$ *and* $\texttt{Att}(c) = \{H_1, \ldots, H_k\}$ *s.t for every* $i = 1, \ldots, k$,

- $H_i \subseteq \mathcal{U}$
- $H_i$ *is consistent*
- $\forall \langle H, \overline{c} \rangle \in \texttt{Pros}(c)$, $H \cup H_i$ *is inconsistent*
- $\nexists H' \subset H_i$ *s.t.* $H'$ *satisfies the third condition.*

**Theorem**

*Let* $c \in \mathcal{C}$.

$$\texttt{Cons}(c) = \{\langle H, c \rangle \mid H \in \texttt{Att}(c)\}$$

# Local Explanations: Abductive Explanations

Why $\mathtt{f}(x) = c$?

**Abductive Explanation**

*Let $x \in \mathcal{X}$. An abductive explanation of $x$ is any member of the set:*

$$\mathtt{AE}(x) = \{H \subseteq \mathcal{U} \mid H \in \mathtt{Supp}(\mathtt{f}(x)) \textit{ and } H \subseteq x\}.$$

# Example

| $\mathcal{X}$ | $f_1$ | $f_2$ | f(.) |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

- $\texttt{Pros}(c_1) = \{a_1\}$
- $\texttt{Pros}(c_2) = \{a_2\}$
- $\texttt{Pros}(c_3) = \{a_3\}$

$a_1 = \langle\{(f_1, 0), (f_2, 0)\}, c_1\rangle$

$a_2 = \langle\{(f_1, 0), (f_2, 1)\}, c_2\rangle$

$a_3 = \langle\{(f_1, 1)\}, c_3\rangle$

- $\texttt{AE}(x_1) = \{\{(f_1, 0), (f_2, 0)\}\}$
- $\texttt{AE}(x_2) = \{\{(f_1, 0), (f_2, 1)\}\}$
- $\texttt{AE}(x_3) = \{\{(f_1, 1)\}\}$
- $\texttt{AE}(x_4) = \{\{(f_1, 1)\}\}$

# Local Explanations: Abductive Explanations (cont.)

Why $\mathtt{f}(x) = c$?

## Abductive Explanation

*Let $x \in \mathcal{X}$. An abductive explanation of $x$ is any member of the set:*

$$\mathtt{AE}(x) = \{H \subseteq \mathcal{U} \mid H \in \mathtt{Supp}(\mathtt{f}(x)) \text{ and } H \subseteq x\}.$$

## Proposition

- *Let $x \in \mathcal{X}$.*
  - $\mathtt{AE}(x) \neq \emptyset$
  - $\mathtt{AE}(x) = \{\emptyset\} \iff \forall y \in \mathcal{X}, \mathtt{f}(y) = \mathtt{f}(x)$
  - $\mathtt{AE}(x) \subseteq \{H \subseteq \mathcal{U} \mid \langle H, \mathtt{f}(x)\rangle \in \mathtt{Pros}(\mathtt{f}(x))\}$
- *The function* $\mathtt{AE}$ *satisfies all the properties.*

# Local Explanations: General Counterfactuals

Why $x$ is not labelled with any other class than $\mathtt{f}(x)$?

**General Counterfactual**

*Let $x \in \mathcal{X}$. A general counterfactual of $x$ is any member of the set:*

$$\mathtt{CF}(x) = \{H \setminus x \mid \langle H, \overline{\mathtt{f}(x)} \rangle \in \mathtt{Cons}(\mathtt{f}(x))\}.$$

# Example (cont.)

| $\mathcal{X}$ | $f_1$ | $f_2$ | $\mathtt{f}(.)$ |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

- $\mathtt{Cons}(c_3) = \{b_5\}$

$b_5 = \langle\{(f_1, 0)\}, \overline{c_3}\rangle$

- $\mathtt{CF}(x_4) = \{\{(f_1, 0)\}\}$

# Local Explanations: General Counterfactuals (cont.)

Why $x$ is not labelled by any other class than $\mathtt{f}(x)$?

## General Counterfactual

*Let $x \in \mathcal{X}$. A general counterfactual of $x$ is any member of the set:*

$$\mathtt{CF}(x) = \{H \setminus x \mid \langle H, \overline{\mathtt{f}(x)} \rangle \in \mathtt{Cons}(\mathtt{f}(x))\}.$$

## Theorem

*Let $x \in \mathcal{X}$. $H \in \mathtt{CF}(x, \mathtt{f}(x)) \iff H$ satisfies the conditions below:*

- $H \subseteq \mathcal{U}$
- *$H$ is consistent*
- $\mathtt{f}(x_{\downarrow H})$[a] $\neq \mathtt{f}(x)$
- *$\nexists H' \subset H$ s.t. $H'$ satisfies the above conditions.*

[a] $\mathtt{f}(x_{\downarrow H})$ denotes the set of literals obtained by replacing the values of features in $x$ by those in $h$ and keeping the remaining ones unchanged.

# Local Explanations: Specific Counterfactuals

Why $x$ is not labelled with $c'$ instead of $\texttt{f}(x)$?

# Example (cont.)

Why $x_4$ is not labelled with $c_1$ instead of $c_3$?

| $\mathcal{X}$ | $f_1$ | $f_2$ | $\mathtt{f}(.)$ |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

- $\mathtt{CF}(x_4) = \{H\}$ $\quad H = \{(f_1, 0)\}$
- But, $x_{4\downarrow H} = x_2$ and $\mathtt{f}(x_2) \neq c_1$

# Local Explanations: Specific Counterfactuals (cont.)

Why $x$ is not labelled $c'$ instead of $\mathtt{f}(x)$?

## Specific Counterfactual

*Let $x \in \mathcal{X}$, $c \in \mathcal{C}$ s.t. $\mathtt{f}(x) \neq c$. A specific counterfactual of $(x, c)$ is a set $H \subseteq \mathcal{U}$ s.t.*

- $\exists y \in \mathcal{X}$ *s.t.* $\mathtt{f}(y) = c$ *and* $y = x_{\downarrow H}$
- $\nexists H' \subset H$ *s.t.* $H'$ *satisfies the above conditions.*

| $\mathcal{X}$ | $f_1$ | $f_2$ | $\mathtt{f}(.)$ |
|---|---|---|---|
| $x_1$ | 0 | 0 | $c_1$ |
| $x_2$ | 0 | 1 | $c_2$ |
| $x_3$ | 1 | 0 | $c_3$ |
| $x_4$ | 1 | 1 | $c_3$ |

The specific counterfactual of $(x_4, c_1)$ is $x_1$

# Limits

Arguments pro/con classes are built from the **whole feature space** $\mathcal{X}$

- ✓ **Correct** explanations
- × **Not feasible** in practice

Solution: To define the same type of explanations from $\mathcal{Y} \subseteq \mathcal{X}$

- × **Plausible** explanations

# Plausible Abductive Explanations

## Plausible Abductive Explanation

*Let $\mathtt{g}_p$ be an explanation function of a classification model $\mathtt{f}$ applied to theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ s.t. for $\mathcal{Y} \subseteq \mathcal{X}$, $x \in \mathcal{Y}$, $H \subseteq \mathcal{U}$, $H \in \mathtt{g}_p^{\mathcal{Y}}(x)$ iff:*

- $H \subseteq x$
- $\forall y \in \mathcal{Y}$ *s.t.* $H \subseteq y$, $\mathtt{f}(y) = \mathtt{f}(x)$
- $\nexists H' \subset H$ *such that* $H'$ *satisfies the above conditions.*

*$H$ is called plausible abductive explanation of $x$.*

## Proposition

*The function $\mathtt{g}_p$ is incoherent and non-monotonic.*

# Plausible Abductive Explanations

| $\mathcal{Y}$ | *Vacation* | *Concert* | *Meeting* | *Exhibition* | *Hiking* |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 | 1 |
| $x_5$ | 0 | 1 | 1 | 0 | 0 |
| $x_6$ | 0 | 1 | 1 | 1 | 0 |
| $x_7$ | 1 | 1 | 0 | 1 | 1 |

- $\mathtt{g}_p(x_1) = \{U_1, U_2\}$
- $\mathtt{g}_p(x_2) = \{U_4, U_5\}$
- $\mathtt{g}_p(x_5) = \{U_1, U_2, U_3\}$

$U_1 = \{(V, 0)\}$
$U_2 = \{(M, 1)\}$
$U_3 = \{(C, 1), (E, 0)\}$
$U_4 = \{(V, 1)\}$
$U_5 = \{(M, 0)\}$

## Incoherence

$U_1 \cup U_5 = \{(V, 0), (M, 0)\}$ is consistent while $\mathtt{f}(x_1) \neq \mathtt{f}(x_2)$

# Plausible Abductive Explanations

| $\mathcal{Y}$ | *Vacation* | *Concert* | *Meeting* | *Exhibition* | *Hiking* |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 | 1 |
| $x_5$ | 0 | 1 | 1 | 0 | 0 |
| $x_6$ | 0 | 1 | 1 | 1 | 0 |
| $x_7$ | 1 | 1 | 0 | 1 | 1 |
| $x_8$ | 1 | 1 | 0 | 0 | 1 |

- $\mathrm{g}_p^{\mathcal{Y}}(x_5) = \{U_1, U_2, U_3\}$

$$U_1 = \{(V, 0)\}$$
$$U_2 = \{(M, 1)\}$$
$$U_3 = \{(C, 1), (E, 0)\}$$

## Non-monotonicity

$U_3 \notin \mathrm{g}_p^{\mathcal{Z}}(x_5)$ where $\mathcal{Z} = \mathcal{Y} \cup \{x_8\}$

# Argument-based Explanation Functions

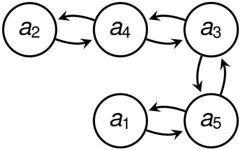
- An argument pro a class $c \in \mathcal{C}$ is a pair $\langle H, c \rangle$ s.t.
  - $H \subseteq \mathcal{U}$ (set of literals)
  - $H$ is consistent
  - $\forall x \in \mathcal{Y}$ s.t. $H \subseteq x$, $f(x) = c$
  - $\nexists H' \subset H$ such that $H'$ satisfies the third condition.

  $\mathsf{arg}(\mathcal{Y})$: the set of all arguments built from $\mathcal{Y}$

- Example
  - $a_1 = \langle U_1, 0 \rangle$ $\quad U_1 = \{(V, 0)\}$
  - $a_2 = \langle U_2, 0 \rangle$ $\quad U_2 = \{(M, 1)\}$
  - $a_3 = \langle U_3, 0 \rangle$ $\quad U_3 = \{(C, 1), (E, 0)\}$
  - $a_4 = \langle U_4, 1 \rangle$ $\quad U_4 = \{(V, 1)\}$
  - $a_5 = \langle U_5, 1 \rangle$ $\quad U_5 = \{(M, 0)\}$

# Attacks

- Let $\langle H, c\rangle$, $\langle H', c'\rangle$ be arguments. $\langle H, c\rangle$ attacks $\langle H', c'\rangle$ iff:
  - $H \cup H'$ is consistent, and
  - $c \neq c'$.




- A set $\mathcal{E}$ of arguments is a naive extension iff:
  - $\nexists a, b \in \mathcal{E}$ s.t. $a$ attacks $b$, and
  - $\nexists \mathcal{E}' \subseteq \texttt{arg}(\mathcal{Y})$ s.t. $\mathcal{E} \subset \mathcal{E}'$ and $\mathcal{E}'$ satisfies the first condition.

- Example
  - $\mathcal{E}_1 = \{a_1, a_2, a_3\}$
  - $\mathcal{E}_2 = \{a_1, a_4\}$
  - $\mathcal{E}_3 = \{a_2, a_5\}$
  - $\mathcal{E}_4 = \{a_4, a_5\}$

# Argument-based Explanation Functions

## Definition

Let $\mathtt{g}_*$ be an explanation function of a classification model $\mathtt{f}$ applied to theory a $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ s.t. for $\mathcal{Y} \subseteq \mathcal{X}_\mathcal{T}$, for $x \in \mathcal{Y}$,

$$\mathtt{g}_*^{\mathcal{Y}}(x) = \{H \mid \exists \langle H, \mathtt{f}(x) \rangle \in \bigcap \mathcal{E}_i \text{ and } H \subseteq x\}$$

where $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are naive extensions.

## Proposition

*The function $\mathtt{g}_*$ is coherent and non-monotonic.*

## Example

$$\bigcap_{i=1}^{4} \mathcal{E}_i = \emptyset \Rightarrow \forall x \in \mathcal{Y}, \mathtt{g}_*(x) = \emptyset$$

# Other Non-Monotonic Functions

| Extensions | Covered instances | Covered classes |
|---|---|---|
| $\mathcal{E}_1 = \{a_1, a_2, a_3\}$ | $x_1, x_3, x_5, x_6$ | 0 |
| $\mathcal{E}_2 = \{a_1, a_4\}$ | $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ | 0, 1 |
| $\mathcal{E}_3 = \{a_2, a_5\}$ | $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ | 0, 1 |
| $\mathcal{E}_4 = \{a_4, a_5\}$ | $x_2, x_4, x_7, x_8$ | 1 |

- Select $\mathcal{E}_2$
  - $\{(V, 1)\}$ is the reason for predicting the class 1
  - $\{(V, 0)\}$ is the reason for predicting the class 0
- Select $\mathcal{E}_3$
  - $\{(M, 0)\}$ is the reason for predicting the class 1
  - $\{(M, 1)\}$ is the reason for predicting the class 0

# Summary

## Conclusions

- Explanations of non-interpretable models are generated under incomplete information
  - they are only **plausible**
- Trade-off to be found between properties

## Challenges

- Novel non-monotonic explanation functions that:
  - guarantee existence of explanations
  - approximate ”real” explanations
  - satisfy desirable properties
- More properties of explanation functions
- Investigate suitability of non-monotonic functions for explaining interpretable models