Non-Monotonic Explanation Functions

Leila Amgoud

Non-Monotonic Explanation Functions

Leila Amgoud

CNRS – IRIT, France
amgoud@irit.fr

Abstract. Explaining black-box classification models is a hot topic in AI, it has the overall goal of improving trust in decisions made by such models. Several works have been done and diverse explanation functions have been proposed. The most prominent ones, like Anchor and LIME, return abductive explanations which highlight key factors that cause predictions. Despite their popularity, the two functions may return inaccurate and sometimes incorrect explanations.
In this paper, we study abductive explanations and identify the origin of this shortcoming. We start by defining two kinds of explanations: *absolute explanations* that are generated from the whole feature space, and *plausible explanations* (like those provided by Anchors and LIME) that are constructed from a proper subset of the feature space. We show that the former are coherent in that two compatible sets of features cannot explain distinct classes while the latter may however be incoherent, leading thus to incorrect explanations. Then, we show that explanations are provided by non-monotonic functions. Indeed, an explanation may no longer be valid if new instances are received. Finally, we provide a novel function that is based on argumentation and that returns plausible explanations. We show that the function is non monotonic and its explanations are coherent.

Keywords: Classification · Explainability · Argumentation.

1 Introduction

In the last few years, the discussion around artificial intelligence is gaining more and more interest. This is mainly due to noteworthy advances made in data-driven AI in general, and deep learning in particular. In this sub-field of AI, the idea is to learn a targeted objective (like the class of an object) from a vast quantity of data. However, the predictions of existing models can hardly be explained in a transparent way. This opacity is seen as a great limitation, which impedes the relevance of those models in practical applications like healthcare but also in embedded systems for mobility despite their successes. Explanations are essential for increasing societal trust, and thus acceptance of those models. They help users understand why a decision was reached.

Explaining the functionality of complex classification models and their rationale becomes a vital need. Consequently, several works have been done in the literature (see [1–3] for recent surveys on explaining machine learning models). They consider as input a classifier, and provide explanations of its predictions. Those works can be divided into two families: The first family opens somehow the classifier to provide insight into the internal decision-making process (eg. [4, 5]), its explanations describe thus the classifier's algorithm. However, the use of increasingly complex algorithms (e.g. deep

neural networks) has made them more difficult to explain and the inner workings of an algorithm may be inexplicable even to the developers of those classifiers [6]. Furthermore, a predominant finding from research in philosophy of sciences and social sciences is that a full causal explanation of a prediction is undesired for humans, as they do not need to understand the algorithm followed by a model. In other words, an explanation does not necessarily hinge on the general public understanding of how algorithmic systems function. Consequently, the second family provides explanations without opening the black-box (eg. [5, 7–11]) and pays particular attention to what constitutes a *good* explanation for human users. Several notions have been defined in this second family including *abductive* explanations (eg. [12]), *counterfactuals* (eg. [5]), *semifactuals* [13], *contrastive* explanations (eg. [8, 9]), explanations based on *pertinent positives* and *pertinent negatives* (eg. [9]), *adversarial examples* (eg. [14]), examples (eg. [15]), and *counter-examples* (eg. [12]).

In this paper, we focus on complex classifiers whose internal processes are inexplicable, and we follow the second approach for explaining their predictions. We study the abductive explanations, which highlight key factors that cause predictions. Such explanations are provided by the two prominent explanation functions: Anchors and LIME [16, 11]. Despite their popularity, the two functions may return inaccurate and sometimes incorrect explanations, and the reason behind this shortcoming remains unclear.

In this paper, we elucidate the origin of the shortcoming. For that purpose, we start by defining two explanations functions. The first function generates *absolute* explanations from the whole feature space. This approach is followed for instance in [17, 12]. We show that this function is coherent, i.e., its explanations are compatible and do not justify distinct classes. However, in practice the whole feature space is not available. Consequently, functions like Anchors and LIME generate (abductive) explanations from a proper subset of the feature space. Indeed, the explanations of a given instance are constructed from some instances in its close neighbourhood.

Our second function generates thus what we call *plausible* explanations from a subset of instances, which may be the neighbourhood of instances, or a set of instances on which the classifier returns a good confidence, or simply a dataset on which it was trained. Such explanations are only plausible (compared to absolute) since they are generated from incomplete information. We show that they are non monotonic in that an explanation may no longer be valid if the subset of instances is extended with further ones. We show also that this function is incoherent, leading to incorrect explanations (as in Anchors and LIME).

To sum up, generating explanations is a non-monotonic reasoning problem, and defining a non-monotonic function that is coherent remains a challenge in the XAI literature. In this paper we provide such a function that is based on argumentation.

The paper is structured as follows: Section 2 presents the background on classification. Section 3 defines the two first functions of explanation and Section 4 investigates their properties and links. Section 5 defines the argument-based explanation function and the last section concludes.

2 Classification Problem

Throughout the paper, we assume a finite and non-empty set $\mathcal{F} = \{f_1, \ldots, f_n\}$ of *features* (called also *attributes*) that take respectively their values from *finite* domains $\mathcal{D}_1, \ldots, \mathcal{D}_n$. Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$. For every feature f and every possible value v of f, the pair (f, v) is called *literal feature*, or *literal* for short. Let $\mathcal{U}$ denote the universe of all possible literal features. The set $\mathcal{X}$ contains all possible $n-$tuples of literal features or sets of the form $\{(f_1, v_{1i}), \ldots, (f_n, v_{nl})\}$, i.e., $\mathcal{X}$ contains all the possible *instantiations* of the n features of $\mathcal{F}$. $\mathcal{X}$ and its elements are called respectively *feature space* and *instances*. Note that $\mathcal{X}$ is finite since $\mathcal{F}$ and the n domains in $\mathcal{D}$ are finite. Let $\mathcal{C} = \{c_1, \ldots, c_m\}$, with $m > 1$, be a finite and non-empty set of possible distinct *classes*.

Definition 1 *A* theory *is a tuple* $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$.

We denote by $\mathcal{X}_\mathcal{T}$ the feature space of theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$. When it is clear from the context, we write $\mathcal{X}$ for short.

A classification model is a function that assigns to every instance $x \in \mathcal{X}_\mathcal{T}$ of a theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ a single prediction, which is a class from the set $\mathcal{C}$.

Definition 2 *Let* $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ *be a theory. A* classification model *is a function* $\mathtt{f}$ *s.t.* $\mathtt{f} : \mathcal{X}_\mathcal{T} \to \mathcal{C}$.

Let us illustrate the above notions with a simple example.

Example 1. Assume a classification problem of deciding whether to hike or not. Hence, $\mathcal{C} = \{c_0, c_1\}$ where c_0 stands for not hiking and c_1 for hiking. The decision is based on four attributes: Being in vacation (V), having a concert (C), having a meeting (M) and having an exhibition (E), thus $\mathcal{F} = \{V, C, M, E\}$. Each attribute is binary, i.e., takes two possible values from $\mathcal{D}_i = \{0, 1\}$ ($i = 1, \ldots, 4$). Assume a classification model $\mathtt{f}$ that assigns classes to instances of $\mathcal{Y} \subset \mathcal{X}$ as shown in the table below.

$\mathcal{Y}$	V	C	M	E	H
x_1	0	0	1	0	c_0
x_2	1	0	0	0	c_1
x_3	0	0	1	1	c_0
x_4	1	0	0	1	c_1
x_5	0	1	1	0	c_0
x_6	0	1	1	1	c_0
x_7	1	1	0	1	c_1

A set of literal features is consistent if it does not contain two literals having the same feature but distinct values.

Definition 3 *A set* $H \subseteq \mathcal{U}$ *is* consistent *iff* $\nexists (f, v), (f', v') \in H$ *such that* $f = f'$ *and* $v \neq v'$. *Otherwise,* H *is said to be* inconsistent.

3 Abductive Explanation Functions

In the machine learning literature, there is a decent amount of work on explaining outcomes of classifiers. Several types of explanations have then been identified, and the most prominent one is the so-called *abductive explanation*. The latter amounts at answering the following question:

Why does an outcome hold?

In other words, why a particular class is assigned to an instance? The answer consists in highlighting factors (i.e., literal features) that caused the given class. Research in cognitive science revealed that in practice, humans provide partial explanations instead of full ones. Instead of a full account, they expect an explanation for the key factors that caused the given output instead of another. Consequently, the two popular explanation functions Anchors and LIME [16, 11] as well as the one that generates the so-called *pertinent positives* [9] generate sets of literals that cause a prediction. Below are two examples of such explanations:

(E_1) *I won't hike (c_0) because I am not on vacation* $(V, 0)$.
(E_2) *You were denied your loan because* your annual income was £30K.

In the first example, the decision of not hiking is due to the attribute (Vacation) which takes the value 0. The second example is on decision about credit worthiness, hence there are again two possible classes (0 and 1). The explanation E_2 means that when the feature *Salary* has the value £30K, the outcome is 0.

In what follows, we define the notion of explanation function. The latter explains the predictions of a classification model in the context of a fixed theory. In other words, it takes as input a classification model $\mathtt{f}$ and a theory $\mathcal{T}$, and returns for every instance $x \in \mathcal{X}_{\mathcal{T}}$, the set of reasons that cause $\mathtt{f}(x)$. The same function is thus applicable to any model $\mathtt{f}$ and any theory. Generally, a function generates explanations from a subset of instances which may be the dataset on which the classifier has been trained, or the set of instances on which it returns a good confidence, or the neighbourhood of some instances, etc.

Definition 4 *An* explanation function *is a function* $\mathtt{g} : \mathcal{X}_{\mathcal{T}} \rightarrow 2^{2^{\mathcal{U}}}$ *that generates from* $\mathcal{Y} \subseteq \mathcal{X}_{\mathcal{T}}$ *the reasons behind the predictions of a classifier* $\mathtt{f}$ *in theory* $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ *. For* $x \in \mathcal{X}_{\mathcal{T}}$, $\mathtt{g}^{\mathcal{Y}}(x)$ *denotes the set of explanations of* x*, or reasons of assigning the class* $\mathtt{f}(x)$ *to* x*.*

Abductive explanations have been studied in the AI literature a long time ago (eg. [18]). More recently, they have been used for interpreting blackbox classification models (eg. [17, 4, 19]). The basic idea behind these works is to look for regular behavior of a model in the whole feature space. More precisely, an abductive explanation is defined as a minimal (for set inclusion) set of literals (or features in [4]) that is sufficient for predicting a class. It is worth mentioning that a given prediction may have several abductive explanations. In what follows, we call such explanations "absolute" since they are defined by exploring the whole feature space, thus under *complete information*, and consequently they cannot be erroneous as we will see later.

Definition 5 *Let* $\mathtt{g}_a$ *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* $\mathcal{Y} \subseteq \mathcal{X}_\mathcal{T}$, $x \in \mathcal{Y}$, $H \subseteq \mathcal{U}$, $H \in \mathtt{g}_a^\mathcal{Y}(x)$ *iff:*

- $H \subseteq x$
- $\forall y \in \mathcal{X}_\mathcal{T}$ *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* absolute abductive explanation *of* $(x, \mathtt{f}(x))$.

Example 2. Consider the theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ such that $\mathcal{F} = \{f_1, f_2\}$, $\mathcal{D}_1 = \mathcal{D}_2 = \{0, 1\}$, and $\mathcal{C} = \{c_1, c_2, c_3\}$. Assume a model $\mathtt{f}$ that assigns classes to instances as shown in the table below.

$\mathcal{Y}$	f_1	f_2	c
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 absolute explanations of x_1, x_2, x_3, x_4 are as follows:

- $\mathtt{g}_a^\mathcal{Y}(x_1) = \{H_1\}$
- $\mathtt{g}_a^\mathcal{Y}(x_2) = \{H_2\}$
- $\mathtt{g}_a^\mathcal{Y}(x_3) = \{H_3\}$
- $\mathtt{g}_a^\mathcal{Y}(x_4) = \{H_3\}$

$H_1 = \{(f_1, 0), (f_2, 0)\}$
$H_2 = \{(f_1, 0), (f_2, 1)\}$
$H_3 = \{(f_1, 1)\}$

The definition of absolute explanations requires exploring the whole feature space (see the second condition of Def. 5), which is quite difficult and maybe not feasible in practice. Consequently, functions like Anchor and LIME [16, 11] generate explanations from a proper subset of the feature space. They focus on the closest instances of the instance to be explained. In what follows, we define an explanation function whose outcomes are called *plausible*. It somehow generalises and improves Anchor and LIME since it also provides abductive explanations, but those that are minimal (for set inclusion). Minimality of an explanation is important since it discards irrelevant information from being part of the explanation. This condition is violated by Anchors and LIME.

Definition 6 *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* $\mathcal{Y} \subseteq \mathcal{X}$, *and* $x \in \mathcal{Y}$, $H \in \mathtt{g}_p^\mathcal{Y}(x)$ *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.

Note that the function $\mathtt{g}_p$ can be applied to different subsets of $\mathcal{X}$, and as we will see later the results may not be the same.

Example 1 (Cont.) Consider the following sets:

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

It can be checked that:

- $\mathsf{g}_p^{\mathcal{Y}}(x_1) = \{U_1, U_2\}$
- $\mathsf{g}_p^{\mathcal{Y}}(x_5) = \{U_1, U_2, U_3\}$
- $\mathsf{g}_p^{\mathcal{Y}}(x_2) = \mathsf{g}_p^{\mathcal{Y}}(x_4) = \mathsf{g}_p^{\mathcal{Y}}(x_7) = \{U_4, U_5\}$

Every (absolute, plausible) abductive explanation is consistent. Furthermore, an instance may have one or several (absolute, plausible) abductive explanations.

Proposition 1 *Let $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ be a theory, $\mathcal{Y} \subseteq \mathcal{X}$ and $x \in \mathcal{Y}$.*

- *Every (absolute, plausible) abductive explanation of x is consistent.*
- *If $\mathcal{Y} = \mathcal{X}$, then $\mathsf{g}_a^{\mathcal{Y}}(x) = \mathsf{g}_p^{\mathcal{Y}}(x)$*
- *$\mathsf{g}_a^{\mathcal{Y}}(x) \neq \emptyset$ and $\mathsf{g}_p^{\mathcal{Y}}(x) \neq \emptyset$*
- *$\mathsf{g}_p^{\mathcal{Y}}(x) = \{\emptyset\}$ iff $\forall y \in \mathcal{Y} \setminus \{x\}$, $\mathtt{f}(y) = \mathtt{f}(x)$.*

Proof. An (absolute, plausible) explanation is a part of an instance. Every instance is consistent, then its subparts are all consistent. The second property is straightforward. Let us show the third property. Since x is consistent, then $\exists H \subseteq x$ such that H is minimal (for set inclusion) such that $\forall y \in \mathcal{Y}$ s.t. $H \subseteq y$, $\mathtt{f}(y) = \mathtt{f}(x)$. The last property is straightforward.

4 Properties of Explanation Functions

We discuss below two formal properties of explanation functions. Such properties are important for assessing the quality of an explanation function and for comparing pairs of functions. The first property ensures *coherence* of the set of explanations provided by a function. More precisely, it states that a set of literals cannot cause two distinct predictions. Let us illustrate the idea with Example 1.

Example 1 (Cont.) Recall that $\{(V, 0)\} \in \mathsf{g}_p^{\mathcal{Y}}(x_1)$ and $\{(M, 0)\} \in \mathsf{g}_p^{\mathcal{Y}}(x_2)$. Note that the set $\{(V, 0), (M, 0)\}$ is consistent, then there exists an instance $y \in \mathcal{X} \setminus \mathcal{Y}$ such that $\{(V, 0), (M, 0)\} \subseteq y$. By definition of an abductive explanation, $\mathtt{f}(y) = c_0$ (due to $\{(V, 0)\}$) and $\mathtt{f}(y) = c_1$ (due to $\{(M, 0)\}$) which is impossible since every instance has one class. This example shows that the function g_p is not coherent even locally, i.e. when working with $\mathcal{Y} \subset \mathcal{X}$.

Below is a formal definition of the notion of coherence of an explanation function.

Definition 7 (Coherence) *An explanation function g is coherent iff for any model $\mathtt{f}$, any theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$, any $\mathcal{Y} \subseteq \mathcal{X}_{\mathcal{T}}$, the following holds:*
$\forall x, x' \in \mathcal{X}_{\mathcal{T}}$, $\forall H \in \mathsf{g}^{\mathcal{Y}}(x)$, $\forall H' \in \mathsf{g}^{\mathcal{Y}}(x')$, if $H \cup H'$ is consistent, then $\mathtt{f}(x) = \mathtt{f}(x')$.

We show that the absolute abductive explanation function is coherent while this property is violated by the plausible function as shown above.

Proposition 2 *The function* $\mathtt{g}_a$ *is coherent and* $\mathtt{g}_p$ *is incoherent.*

Proof. Example 1 shows a counter-example for the coherence of $\mathtt{g}_p$. Let us now show that $\mathtt{g}_a$ is coherent. Let $x, x' \in \mathcal{X}$, $c, c' \in \mathcal{C}$ and $H \in \mathtt{g}_a(x)$, $H' \in \mathtt{g}_a(x')$. Assume that $H \cup H'$ is consistent. Then, $\exists z \in \mathcal{X}$ s.t. $H \cup H' \subseteq z$. By Def. 5, $\mathtt{f}(z) = c$ (due to H) and $\mathtt{f}(z) = c'$ (due to H'). Since (z) is unique, then $c = c'$.

Remark: As said before, Anchors and LIME explanation functions are somehow instances of $\mathtt{g}_p$ as they generate abductive explanations from a proper subset of the feature space. They even do not use a whole dataset but only instances that are in the neighbourhood of the instance being explained. Hence, the two functions violate coherence.

Let us now investigate another property of explanation functions, that of monotony. The idea is to check whether an instance generated from a dataset remains plausible when the dataset is expanded with new instances.

Definition 8 (Monotony) *An explanation function* $\mathtt{g}$ *is* monotonic *iff for any model* $\mathtt{f}$*, any theory* $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ *, any* $x \in \mathcal{X}_{\mathcal{T}}$*, the following property holds:*
$\mathtt{g}^{\mathcal{Y}}(x) \subseteq \mathtt{g}^{\mathcal{Z}}(x)$ *whenever* $\mathcal{Y} \subseteq \mathcal{Z} \subseteq \mathcal{X}_{\mathcal{T}}$*. It is non-monotonic otherwise.*

The absolute function $\mathtt{g}_a$ is clearly monotonic since it explores the whole feature space, thus complete information. However, the plausible function is not monotonic.

Proposition 3 *The function* $\mathtt{g}_a$ *is monotonic and* $\mathtt{g}_p$ *is non-monotonic.*

Let us illustrate the non-monotonicity of $\mathtt{g}_p$ with an example.

Example 1 (Cont.) Consider the instance x_5. Recall that $U_3 = \{(C, 1), (E, 0)\}$ is a plausible explanation of x_5. Assume we receive the new instance x_8 below:

$\mathcal{Y}$	V	C	M	E	H
x_8	1	1	0	0	c_1

Note that $\{(C, 1), (E, 0)\}$ is no longer a plausible explanation of x_5 that can generated from the set $\mathcal{Y} \cup \{x_5\}$. Indeed, the second condition of Def. 6 is not satisfied.

The above example shows that a plausible explanation of an instance is not necessarily an absolute one, while, ideally $\mathtt{g}_p$ should approximate $\mathtt{g}_a$.

Property 1. Let $\mathcal{Y} \subset \mathcal{X}_{\mathcal{T}}$ and $x \in \mathcal{Y}$. $\mathtt{g}_p^{\mathcal{Y}}(x) \not\subseteq \mathtt{g}_a^{\mathcal{Y}}(x)$.

To sum up, we have shown that in practice explanations are constructed from a dataset, which is a subset of the feature space of a theory. However, due to incompleteness of information in the dataset, some explanations may be incorrect, i.e., they are not absolute. Furthermore, we have seen that the actual functions (like Anchors and LIME) that generate plausible explanations suffer from another weakness which is incoherence. The latter leads also to incorrect explanations. Thus, defining a nonmonotonic explanation function that is coherent remains a challenge in the literature. In the next section, we define such a function.

5 Argument-based Explanation Function

Throughout this section we consider an arbitrary theory $\mathcal{T} = \langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$ and a subset $\mathcal{Y} \subseteq \mathcal{X}_{\mathcal{T}}$ of instances. We define a novel explanation function which is based on arguments. The latter support classes, in the sense they provide the minimal sets of literals that are causing a class. They are thus independent from instances. An advantage of not considering instances is to reduce the number of arguments that can be built. Furthermore, explanations of an instance are in explanations of its predicted class.

Definition 9 *Let $c \in \mathcal{C}$. An* argument *in favor of c is a pair $\langle H, c \rangle$ s.t.*

- $H \subseteq \mathcal{U}$
- *H is consistent*
- *$\forall y \in \mathcal{Y}$ s.t. $H \subseteq y$, $\mathtt{f}(y) = c$*
- *$\nexists H' \subset H$ that verifies the above conditions.*

H and c are called respectively support *and* conclusion *of an argument. Let* $\arg(\mathcal{Y})$ *denote the set of arguments built from $\mathcal{Y}$.*

Note that the set $\arg(\mathcal{Y})$ is finite since $\mathcal{Y}$ is finite. Furthermore, its elements are closely related to the plausible explanations of the function $\mathtt{g}_p$.

Proposition 4 *Let $c \in \mathcal{C}$. $\langle H, c \rangle \in \arg(\mathcal{Y}) \iff \exists x \in \mathcal{Y}$ s.t. $H \in \mathtt{g}_p^{\mathcal{Y}}(x, c)$.*

Consider the initial version of our running example, which contains seven instances.

Example 1 (Cont.) There are two classes in the theory: c_0, c_1. Their arguments are given below:

- $a_1 = \langle U_1, c_0 \rangle$ $\quad U_1 = \{(V, 0)\}$
- $a_2 = \langle U_2, c_0 \rangle$ $\quad U_2 = \{(M, 1)\}$
- $a_3 = \langle U_3, c_0 \rangle$ $\quad U_3 = \{(C, 1), (E, 0)\}$
- $a_4 = \langle U_4, c_1 \rangle$ $\quad U_4 = \{(V, 1)\}$
- $a_5 = \langle U_5, c_1 \rangle$ $\quad U_5 = \{(M, 0)\}$

These arguments may be conflicting. This is particularly the case when they violate the coherence property, namely when their supports are consistent but their conclusions are different.

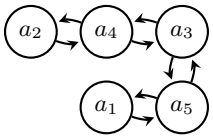
Definition 10 *Let $\langle H, c \rangle$, $\langle H', c' \rangle \in \arg(\mathcal{Y})$. We say that $\langle H, c \rangle$ attacks $\langle H', c' \rangle$ iff:*

- *$H \cup H'$ is consistent, and*
- *$c \neq c'$.*

Obviously, the above attack relation is symmetric.

Property 2. Let $a, b \in \arg(\mathcal{Y})$. If a attacks b, then b attacks a.

Example 1 (Cont.) The attacks between the arguments are depicted below:



Note that every argument in favor of a class attacks at least one argument in favor of the other class. This shows that the plausible explanations generated by the function $\mathtt{g}_p$ are incoherent, and cannot all be correct.

Arguments and their attack relation form an argumentation system as follows.

Definition 11 *An* argumentation system *built from $\mathcal{Y} \subseteq \mathcal{X}$ is a pair $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$ where $\mathcal{R} \subseteq \arg(\mathcal{Y}) \times \arg(\mathcal{Y})$ such that for $a, b \in \arg(\mathcal{Y})$, $(a, b) \in \mathcal{R}$ iff a attacks b (in the sense of Def. 10).*

Since arguments are conflicting, they should be evaluated using a semantics. There are different types of semantics in the literature. In this paper, we consider extension-based ones that have been introduced by Dung in [20]. They compute sets of arguments that can be jointly accepted. Each set is called an extension and represents a coherent position. Since the attack relation is symmetric, then it has been shown that *stable and preferred semantics coincide with the naive*, which returns maximal (for set $\subseteq$) sets that do not contain conflicting arguments. So, we focus here on naive semantics.

Definition 12 *Let $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$ be an argumentation system and $\mathcal{E} \subseteq \arg(\mathcal{Y})$. The set $\mathcal{E}$ is a* naive extension *iff:*

- *$\nexists a, b \in \mathcal{E}$ s.t. $(a, b) \in \mathcal{R}$, and*
- *$\nexists \mathcal{E}' \subseteq \arg(\mathcal{Y})$ s.t. $\mathcal{E} \subset \mathcal{E}'$ and $\mathcal{E}'$ satisfies the first condition.*

Let $\sigma(AS)$ denote the set of all naive extensions of AS.

Example 1 (Cont.) The argumentation system has four naive extensions:

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

Each naive extension refers to a possible set of explanations. Note that $\mathcal{E}_1$ and E_4 promote respectively the arguments in favor of c_0 and those in favor of c_1.

We are now ready to define the new explanation function. For a given instance x, it returns the support of any argument in favour of $\mathtt{f}(x)$ that is in every naive extension and the support should be part of x. The intuition is the following: when two explanations cannot hold together (coherence being violated), both are discarded since at least one of them is incorrect. Our approach is thus very cautious.

Definition 13 *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* $x \in \mathcal{Y}$*,*

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

where $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$.

Example 1 (Cont.) In the example, $\bigcap_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i = \emptyset$. Hence, $\forall x \in \mathcal{Y}$, $\mathtt{g}_*(x, \mathtt{f}(x)) = \emptyset$.

The above example shows that this function may return an emptyset of explanations, meaning with the available information, it is not possible to generate reasonable abductive explanations. Note that generation of argument is a nonmonotonic process.

Example 2 (Cont.) Assume a set $\mathcal{Y} = \{x_1, x_2\}$. It can be checked that $\arg(\mathcal{Y}) = \{b_1, b_2\}$ where $b_1 = \{(f_2, 0)\}$ and $b_2 = \{(f_2, 1)\}$. The two arguments are not conflicting, thus $\mathcal{R} = \emptyset$ and there is a single naive extension which contains the two. Hence, $\mathtt{g}_*(x_1, c_1) = \{\{(f_2, 0)\}\}$ and $\mathtt{g}_*(x_2, c_2) = \{\{(f_2, 1)\}\}$.

We show that the new function is non-monotonic and coherent.

Proposition 5 *The function* $\mathtt{g}_*$ *is non-monotonic and coherent.*

Finally, when $\mathtt{g}_*$ is applied on the whole feature space, the attack relation of the corresponding argumentation system would be empty, and the generated explanations coincide with the absolute ones.

Proposition 6 *If* $\mathcal{Y} = \mathcal{X}$*, then* $\mathtt{g}_* = \mathtt{g}_a$.

6 Related Work

Most work on finding explanations in the ML literature is experimental, focusing on specific models, exposing their internal representations to find correlations *post hoc* between these representations and the predictions. There haven't been a lot of formal characterizations of explanations in AI, with the exception of [12], which defines abductive explanations and adversarial examples in a fragment of first order logic, and [17], who focused on semi-factuals, that they consider a specific form of counterfactuals. Both works considered binary classifiers with binary features. Furthermore, they generate explanations from the whole feature space, which in practice is not reasonable since a classification model is trained on a dataset. In our work, we focused on abductive explanations for general classifiers, and discussed two particular properties: monotony and coherence.

Unlike our work, which explains existing Black-box models, [21, 22] proposed novel classification models that are based on arguments. Their explanations are defined in dialectical way as fictitious dialogues between a proponent (supporting an output) and an opponent (attacking the output) following [20]. The authors in [23–26] followed the same approach for defining explainable multiple decision systems, recommendation systems, or scheduling systems. In the above papers an argument is simply an instance and its label while our arguments pro/con are much richer. This shows that they are proposed for different purposes.

7 Conclusion

This paper presented a preliminary investigation on functions that would explain predictions of black-box classifiers. It focused on one type of explanations, those that identify the key features that caused predictions. Such explanations are popular in the XAI literature, however there are a few formal attempts at formalizing them and investigating their properties. Existing definitions consider the whole feature space, and this is not reasonable in practice.

In this paper, we argue that generating explanations consists of reasoning under incomplete information, and reasonable functions are thus nonmonotonic. We have shown that existing (nonmonotonic) functions like Anchors and LIME may return incoherent results, which means incorrect explanations. Finally, we provided the first function that satisfies coherence while generating explanations from datasets. The function is based on a well-known nonmonotonic reasoning (NMR) approach, which is argumentation. This makes thus a connection between XAI and NMR.

This work can be extended in different ways. First, we have seen that the novel function g^* may return an emptyset for an instance. While cautious reasoning is suitable when dealing with conflicting information by NMR models, it may be a great weakness in XAI since a user would always expect an explanation for the outcome provided by a classifier. Hence, a future work consists of exploring other functions that guarantee outputs. Another line of research consists of using weighted semantics for evaluating arguments. Such semantics would then lead to weighted explanations.

Acknowledgements

Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged.

References

1. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI Workshop on Explainable Artificial Intelligence (XAI). (2017) 1–6
2. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys **51**(5) (2019) 93:1–93:42
3. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267** (2019) 1–38
4. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: The Thirty-Third Conference on Artificial Intelligence, AAAI. (2019) 1511–1519
5. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR **abs/1711.00399** (2017)
6. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: Proceedings of the 24th International Conference on Intelligent User Interfaces IUI. (2018) 211–223

7. Biran, O., McKeown, K.R.: Human-centric justification of machine learning predictions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI. (2017) 1461–1467
8. Luss, R., Chen, P., Dhurandhar, A., Sattigeri, P., Shanmugam, K., Tu, C.: Generating contrastive explanations with monotonic attribute functions. CoRR (2019)
9. Dhurandhar, A., Chen, P., Luss, R., Tu, C., Ting, P., Shanmugam, K., Das, P.: Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Annual Conference on Neural Information Processing Systems, NeurIPS. (2018) 590–601
10. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. (2019) 279–288
11. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18). (2018) 1527–1535
12. Ignatiev, A., Narodytska, N., Marques-Silva, J.: On relating explanations and adversarial examples. In: Thirty-third Conference on Neural Information Processing Systems, NeurIPS. (2019) 15857–15867
13. Byrne, R.: Semifactual "even if" thinking. Thinking and reasoning **8**(1) (2002) 41–67
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR. (2014)
15. Cai, C.J., Jongejan, J., Holbrook, J.: The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces IUI. (2019) 258–262
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should itrust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016) 1135–1144
17. Darwiche, A., Hirth, A.: On the reasons behind decisions. In: 24th European Conference on Artificial Intelligence ECAI. Volume 325., IOS Press (2020) 712–720
18. Dimopoulos, Y., Dzeroski, S., Kakas, A.: Integrating explanatory and descriptive learning in ILP. In: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI. (1997) 900–907
19. Kakas, A., Riguzzi, F.: Abductive concept learning. New Generation Computing **18**(3) (2000) 243–294
20. Dung, P.M.: On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence **77** (1995) 321–357
21. Amgoud, L., Serrurier, M.: Agents that argue and explain classifications. Autonomous Agents and Multi-Agent Systems **16**(2) (2008) 187–209
22. Cocarascu, O., Stylianou, A., Cyras, K., Toni, F.: Data-empowered argumentation for dialectically explainable predictions. In: 24th European Conference on Artificial Intelligence ECAI. Volume 325., IOS Press (2020) 2449–2456
23. Zhong, Q., Fan, X., Luo, X., Toni, F.: An explainable multi-attribute decision model based on argumentation. Expert Systems with Applications **117** (2019) 42–61
24. Cyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., Hapuarachchi, T.: Explanations by arbitrated argumentative dispute. Expert Systems with Applications **127** (2019) 141–156
25. Cyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: The Thirty-Third Conference on Artificial Intelligence, AAAI. (2019) 2752–2759
26. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: Fantastic explanations and how to find them. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI. (2018) 1949–1955