Axiomatic Foundations of Explainability

Leila Amgoud , Jonathan Ben-Naim
CNRS – IRIT, France
{amgoud, bennaim}@irit.fr

Abstract

Improving trust in decisions made by classification models is becoming crucial for the acceptance of automated systems, and an important way of doing that is by providing explanations for the behaviour of the models. Different explainers have been proposed in the recent literature for that purpose, however their formal properties are under-studied.

This paper investigates theoretically explainers that provide reasons behind decisions independently of instances. Its contributions are fourfold. The first is to lay the foundations of such explainers by proposing key axioms, i.e., desirable properties they would satisfy. Two axioms are incompatible leading to two subsets. The second contribution consists of demonstrating that the first subset of axioms characterizes a family of explainers that return *sufficient* reasons while the second characterizes a family that provides *necessary* reasons. This sheds light on the axioms which distinguish the two types of reasons. As a third contribution, the paper introduces various explainers of both families, and fully characterizes some of them. Those explainers make use of the whole feature space. The fourth contribution is a family of explainers that generate explanations from finite datasets (subsets of the feature space). This family, seen as an abstraction of Anchors and LIME, violates some axioms including one which prevents incorrect explanations.

1 Introduction

Recent progress in data-driven AI has been largely due to machine learning and in particular deep learning models. However, the predictions of these models resist analysis due to their inherent non-linear behaviour and their vast amount of interacting parameters. This opacity impedes the relevance of those models from a theoretical point of view, since their properties are difficult to investigate, and from a practical point of view, as many applications, such as healthcare or embedded systems need guarantees to be deployed, and others, e.g in the legal, or financial domain require transparency to be accepted. Explanations help human users understand why a decision was reached. Explaining the functionality of classification systems and their rationale thus becomes a vital need. This has generated a lot of effort, see [Cyras *et al.*, 2021; Guidotti *et al.*, 2019; Miller, 2019; Biran and Cotton, 2017] for surveys on explainers of machine learning models. Existing explainers can be classified in two different ways. The first way distinguishes explainers that provide *local* explanations for individual instances (eg. [Ribeiro *et al.*, 2016; Ribeiro *et al.*, 2018; Dhurandhar *et al.*, 2018; Ignatiev *et al.*, 2019; Darwiche and Hirth, 2020]) and explainers that provide *global* explanations for classes independently of instances (eg. [Ignatiev *et al.*, 2019; Amgoud, 2021a]). The second way for classifying exiting explainers is based on the information used for generating explanations. Explainers, like Anchors and LIME [Ribeiro *et al.*, 2016; Ribeiro *et al.*, 2018; Amgoud, 2021b] use datasets while others, like those studied in [Ignatiev *et al.*, 2019; Ignatiev *et al.*, 2020; Darwiche and Hirth, 2020], use the whole set of instances.

Despite the popularity of existing explainers, their formal properties are under-studied. This makes their comparison difficult. Some explainers have been analysed against a set of metrics and have been shown to be efficient. However, some counter-intuitive results have been detected in [Narodytska *et al.*, 2019] for Anchors and LIME. This shows that the existing metrics are not sufficient for analysing the quality of an explainer and guiding the definition of novel ones. They are also not sufficient for an accurate comparison of explainers.

The present paper bridges this gap by investigating the theoretical foundations of explainers that provide global explanations (i.e. reasons behind assigning classes independently of instances). Foundations are important not only for a better understanding of the explanation process in general, but also for clarifying the basic assumptions underlying every explainer, and for comparing different (families of) explainers.

The paper contains four contributions. The first is to lay the foundations of explainers by proposing key axioms, i.e., desirable properties, they would satisfy. Two axioms are shown to be incompatible, leading to two subsets. The second contribution consists of demonstrating that the first subset of axioms characterizes the family of explainers that are based on *abductive* reasoning, hence producing *sufficient reasons*, and the second subset of axioms characterizes the family of explainers that are based on *counterfactual* reasoning, i.e., returning *necessary reasons*. These characterisations shed

light on the properties that distinguish the two types of reasons. As a third contribution, the paper introduces various explainers of both families, each of them generating explanations under complete information, i.e., using the whole feature space. It fully characterizes some of them including the one which retunrs the so-called Prime Implicants and studied in [Ignatiev *et al.*, 2019; Darwiche and Hirth, 2020; Audemard *et al.*, 2020]. The fourth contribution is a family of explainers that generate reasons from finite datasets (subsets of the feature space). This family, seen as an abstraction of Anchors [Ribeiro *et al.*, 2018] and LIME [Ribeiro *et al.*, 2016], violates some axioms including one which prevents incorrect explanations.

2 Classification

We start by introducing the initial material needed to classify, i.e., *classes* as well as *attributes* and their *domains*.

Definition 1 (Theory)**.** A *classification theory* is a triple $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ such that the following holds:

- $\mathcal{A}$ is a non-empty finite set of attributes (or features);
- d is a function on $\mathcal{A}$ such that, for every $a \in \mathcal{A}$, $d(a)$ is countable (discrete domains) with $|d(a)| > 1$;
- $\mathcal{C}$ is a finite set of classes such that $|\mathcal{C}| > 1$.

Next, we need to define the notion of *literal*, i.e., an assignment of a value to an attribute:

Definition 2 (Literal)**.** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory. A *literal* on $\mathbf{T}$ is a couple $\langle a, v \rangle$ such that $a \in \mathcal{A}$ and $v \in d(a)$. We denote by $\mathtt{Lit}_{\mathbf{T}}$ the set of all literals on $\mathbf{T}$. A subset L of $\mathtt{Lit}_{\mathbf{T}}$ is *consistent* iff, for any two elements $l = \langle a, v \rangle$ and $l' = \langle a', v' \rangle$ of L, if $a = a'$, then $v = v'$.

We turn to the notion of *instance*, i.e., an assignment of values to all attributes:

Definition 3 (Instance)**.** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory. An *instance* on $\mathbf{T}$ is a subset I of $\mathtt{Lit}_{\mathbf{T}}$ such that every attribute $a \in \mathcal{A}$ appears exactly once in I. We denote by $\mathtt{Inst}_{\mathbf{T}}$ the set of all instances on $\mathbf{T}$.

Notice that every instance is consistent, and every proper subset of an instance is also consistent.

Property 1. Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory and $I \in \mathtt{Inst}_{\mathbf{T}}$. I is consistent; for any $I' \subset I$, I' is consistent.

We are ready to define the notion of *classifier*. It is a function which assigns a single class to every instance. Furthermore, every class is assigned to at least one instance.

Definition 4 (Classifier)**.** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory. A *classifier* on $\mathbf{T}$ is a surjective function $\mathbf{R}$ from $\mathtt{Inst}_{\mathbf{T}}$ to $\mathcal{C}$.

Notation ($\mathtt{Inst}_{\mathbf{TR}}(.)$)**:** We denote by $\mathtt{Inst}_{\mathbf{TR}}(x)$ the set of all instances of a class x in $\mathbf{T}$ and $\mathbf{R}$, i.e., $\mathtt{Inst}_{\mathbf{TR}}(x) = \{I \in \mathtt{Inst}_{\mathbf{T}} : \mathbf{R}(I) = x\}$.

We show that every class is assigned to at least one instance and not assigned to at least one other instance.

Property 2. Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory and $\mathbf{R}$ a classifier on $\mathbf{T}$. For any $x \in \mathcal{C}$, the following holds: $\mathtt{Inst}_{\mathbf{TR}}(x) \neq \emptyset$ and $\mathtt{Inst}_{\mathbf{TR}}(x) \neq \mathtt{Inst}_{\mathbf{T}}$.

Let us now analyse the relation of a literal with a class. It may be *irrelevant* to the class, i.e., it has no impact on the class, or *relevant* to the class and thus its absence may prevent the class from being assigned to an instance, or *core* to the class, i.e. its absence automatically discards the class.

Notation ($\mathtt{Dif}_{\mathbf{T}}(.)$)**:** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory, $I \in \mathtt{Inst}_{\mathbf{T}}$, and $a \in \mathcal{A}$. We denote by $\mathtt{Dif}_{\mathbf{T}}(I, a)$ the set of all instances on $\mathbf{T}$ that *differs* from I with regard to a, i.e., $\mathtt{Dif}_{\mathbf{T}}(I, a)$ is the set of every $J \in \mathtt{Inst}_{\mathbf{T}} \setminus \{I\}$ such that, $\forall\, b \in \mathcal{A} \setminus \{a\}$, $\forall\, v \in d(b)$, if $\langle b, v \rangle \in I$, then $\langle b, v \rangle \in J$.

A literal $\langle a, v \rangle$ is *relevant* to a class x under a theory $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ and a classifier $\mathbf{R}$ iff there exists another value $v' \in d(a)$ which leads to another class than x. It is *core* to the class if the class is not proposed by $\mathbf{R}$ when the literal is absent.

Definition 5 (Relevance/Coreness)**.** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory, $\mathbf{R}$ a classifier on $\mathbf{T}$, $x \in \mathcal{C}$, and $l = \langle a, v \rangle \in \mathtt{Lit}_{\mathbf{T}}$. We say that l is *relevant* to x in $\mathbf{T}$ and $\mathbf{R}$ iff $\exists\, I \in \mathtt{Inst}_{\mathbf{TR}}(x)$ such that the following holds:

- $l \in I$;
- $\exists\, I' \in \mathtt{Dif}_{\mathbf{T}}(I, a)$, $I' \notin \mathtt{Inst}_{\mathbf{TR}}(x)$.

l is *core* to x in $\mathbf{T}$ and $\mathbf{R}$ iff $\forall\, I \in \mathtt{Inst}_{\mathbf{TR}}(x)$, $l \in I$.

Note that relevant literals exist since $\mathtt{Inst}_{\mathbf{T}}$ contains all the possible instances that can be built from a theory, i.e., all instances are assumed to be reasonable cases.

Let us illustrate the above notions with a classical example borrowed from [Darwiche and Hirth, 2020].

Example 1. *Consider the task of college admission. There are four binary attributes: Entrance exam (E), First time entrance (F), Work experience (W) and GPA. The decision is binary: a candidate is either admitted or denied. Consider a binary classifier, represented by the following rules:*

- *If* $E = 1$ *and* $F = 0$*, then Admit*
- *If* $E = 1$*,* $F = 1$*,* $W = 1$*, then Admit*
- *If* $E = 1$*,* $F = 1$ *and* $W = 0$ *and GPA* $= 1$*, then Admit*
- *If* $E = 1$*,* $F = 1$ *and* $W = 0$ *and GPA* $= 0$*, then Deny*
- *If* $E = 0$*, then Deny*

Note that $\langle E, 1 \rangle$ *is core to the class Admit while* $\langle GPA, 1 \rangle$ *is only relevant to Admit. However, there is no core literal to the class Deny.*

Obviously, if a literal is core to a class, then it is also relevant to that class. The converse does not hold.

Proposition 1. Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory and $\mathbf{R}$ a classifier on $\mathbf{T}$. For any $x \in \mathcal{C}$, for any $l \in \mathtt{Lit}_{\mathbf{T}}$, if l is core to x, then l is relevant to x.

3 Explanation Functions and Axioms

Explaining a classifier amounts either to describing its global behaviour, namely how it affects classes independently of instances, or to locally justifying its prediction for an instance. However, the latter is generally based on the former. Indeed, an explanation of an instance describes why the classifier assigned the class of the instance. Hence, in this paper we focus on explaining classes. An explanation answers the question:

why a class x is assigned by $\mathbf{R}$? There are different categories of explanations as reviewed in [Schneider and Handali, 2019]. However, in this paper we focus exclusively on explanations that are *literals* since they are easy to interpret by humans. Indeed, research in cognitive science revealed that in practice, humans expect an explanation for the key factors that caused the given output. Furthermore, most of existing explanation functions (rule-based explanations, prime implicants, examples) are based on literals. Other categories (eg. conversation-based) are beyond the scope of this paper. Note that there may be several reasons for assigning a class.

Definition 6. A *class question* is a triple $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ such that $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ is a classification theory, $\mathbf{R}$ is a classifier on $\mathbf{T}$, and x is an element of $\mathcal{C}$.

Formally, an explanation for a class is a set of subsets of literals. Every subset of literals, which may be the emptyset, is one reason behind predicting the class. Hence, what we call *class explanation* is the complete set of reasons.

Definition 7. Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory. A *class explanation* on $\mathbf{T}$ is a set of subsets of $\mathtt{Lit}_{\mathbf{T}}$. Every such subset is called a *reason*.

A class explainer, or explanation function, is a function which assigns to every class question a class explanation.

Definition 8. A *class explainer* is a function $\mathbf{F}$ mapping every question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ into a class explanation on $\mathbf{T}$.

We provide below some formal properties that a reasonable class explainer could satisfy. Such properties are important for assessing the quality of an explanation function and for comparing pairs of functions.

The first property states that an explainer should always provide explanations. It is important to provide explanations for humans (eg., customer for whom a loan has been refused).

Axiom 1 (Success)**.** A class explainer $\mathbf{F}$ satisfies *success* iff for any class question $\mathbf{Q}$, $\mathbf{F}(\mathbf{Q}) \neq \emptyset$.

The second property states that an explainer should provide informative explanations, and thus an empty explanation is not recommended.

Axiom 2 (Explainability)**.** A class explainer $\mathbf{F}$ satisfies *explainability* iff for any class question $\mathbf{Q}$, $\forall\, L \in \mathbf{F}(\mathbf{Q})$, $L \neq \emptyset$.

The next property states that reasons in an explanation should not contain *unnecessary* information.

Axiom 3 (Irreducibility)**.** A class explainer $\mathbf{F}$ satisfies *irreducibility* iff for any class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall L \in \mathbf{F}(\mathbf{Q}), \forall l \in L, \exists\, I \in \mathtt{Inst}_{\mathbf{T}} \setminus \mathtt{Inst}_{\mathbf{TR}}(x)$ s.t. $L \setminus \{l\} \subseteq I$.

The next property states that every reason is a subset of at least one instance. This ensures the *feasability* of reasons. Recall that the latter represent causes; when they occur, the classes they explain are suggested for instances.

Axiom 4 (Feasibility)**.** A class explainer $\mathbf{F}$ satisfies *feasibility* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall L \in \mathbf{F}(\mathbf{Q})$, $\exists\, I \in \mathtt{Inst}_{\mathbf{TR}}(x)$ s.t. $L \subseteq I$.

Class explanations are the basis for explaining individual instances. Indeed, explaining an instance amounts to justify its class. The next axiom states that class explanations should be not only sufficient for explaining instances but also for reproducing the predictions of the classifier. The second property makes it possible to use explanations on unseen data.

Axiom 5 (Representativity)**.** A class explainer $\mathbf{F}$ satisfies *representativity* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall I \in \mathtt{Inst}_{\mathbf{TR}}(x), \exists L \in \mathbf{F}(\mathbf{Q})$ s.t. $L \subseteq I$.

The following property states that an explanation should only contain information that impacts a prediction.

Axiom 6 (Relevance)**.** A class explainer $\mathbf{F}$ satisfies *relevance* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall L \in \mathbf{F}(\mathbf{Q})$, $\forall l \in L$, l is relevant to x.

We saw previously that some literals can be more than relevant for a class. They are core as their absence in an instance prevents a class from being assigned by a classifier. The next axiom is more demanding than the previous one, and requires that an explanation contains only core literals.

Axiom 7 (Coreness)**.** A class explainer $\mathbf{F}$ satisfies *coreness* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall L \in \mathbf{F}(\mathbf{Q})$, $\forall l \in L$, l is core to x.

The next property ensures that information that is not part of reasons of a class is irrelevant to the class. This ensures exhaustivity of the explanation provided for the class.

Axiom 8 (Exhaustivity)**.** A class explainer $\mathbf{F}$ satisfies *Exhaustivity* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall l \in \mathtt{Lit}_{\mathbf{T}}$, if l is relevant to x, then $\exists\, L \in \mathbf{F}(\mathbf{Q})$ s.t. $l \in L$.

The following property ensures that every core literal to a class should appear in the explanation of that class.

Axiom 9 (Completeness)**.** A class explainer $\mathbf{F}$ satisfies *completeness* iff for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\forall l \in \mathtt{Lit}_{\mathbf{T}}$, if l is core to x, then $\exists\, L \in \mathbf{F}(\mathbf{Q})$ s.t. $l \in L$.

The previous axioms describe properties of one class explanation. The last axiom is about the set of all such explanations that can be generated from a theory. It ensures their compatibility, avoiding thus erroneous explanations. The axiom states that the union of two reasons supporting different classes should be inconsistent. To illustrate the idea, consider an explainer that provides respectively $L = \{(a, v)\}$ and $L' = \{(b, v')\}$ for the classes x and y. Note that $L \cup L'$ is consistent, then there exists an instance I that contains $L \cup L'$. The two explanations support contradictory predictions for I.

Axiom 10 (Coherence)**.** A class explainer $\mathbf{F}$ satisfies *coherence* iff for any two class questions $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ and $\mathbf{Q}' = \langle \mathbf{T}', \mathbf{R}', x' \rangle$ s.t. $\mathbf{T} = \mathbf{T}'$, $\mathbf{R} = \mathbf{R}'$, and $x \neq x'$, $\forall L \in \mathbf{F}(\mathbf{Q}), \forall L' \in \mathbf{F}(\mathbf{Q}')$, $L \cup L'$ is inconsistent.

Feasibility guarantees the consistency of every reason.

Property 3. Let $\mathbf{F}$ be a class explainer that satisfies Feasibility, $\mathbf{Q}$ a class question. For any $L \in \mathbf{F}(\mathbf{Q})$, L is consistent.

From a couple of axioms, it follows that a reason causes the class it explains. Indeed, its appearance in any instance leads the classifier to assign that class to it.

Proposition 2. Let $\mathbf{F}$ be a class explainer that satisfies Feasibility, Representativity and Coherence, $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ a class question. The following holds:

$$\forall L \in \mathbf{F}(\mathbf{Q}), \forall I \in \mathtt{Inst}_{\mathbf{T}} \text{ s.t. } L \subseteq I, I \in \mathtt{Inst}_{\mathbf{TR}}(x).$$

Exhaustivity and Relevance ensure that the literals used in the explanation of a class are exactly *all* those that are relevant to the class. Likewise, Completeness and Coreness ensure that explanations are based on all and only core literals.

Theorem 1. *Let* $\mathbf{F}$ *be a class explainer and* $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ *a class question. The following two points hold:*

- $\mathbf{F}$ *satisfies Exhaustivity and Relevance iff*

$$\bigcup_{L \in \mathbf{F}(\mathbf{Q})} L = \{l \in \mathtt{Lit}_{\mathbf{T}} \ : l \textit{ is relevant to } x\};$$

- $\mathbf{F}$ *satisfies Completeness and Coreness iff*

$$\bigcup_{L \in \mathbf{F}(\mathbf{Q})} L = \{l \in \mathtt{Lit}_{\mathbf{T}} \ : l \textit{ is core to } x\}.$$

The above axioms are not all independent. Some of them follow from others. We considered them in the paper since they allow to discriminate between explainers. Some explainers may satisfy only an implied axiom while others may satisfy the one that does not follow from any other axiom.

Proposition 3. Let $\mathbf{F}$ be a class explainer.

- if $\mathbf{F}$ satisfies Representativity, then $\mathbf{F}$ satisfies Success;
- if $\mathbf{F}$ satisfies Coreness, then $\mathbf{F}$ satisfies Relevance;
- if $\mathbf{F}$ satisfies Exhaustivity, then $\mathbf{F}$ satisfies Completeness;
- if $\mathbf{F}$ satisfies Feasibility, Coherence and Representativity, then $\mathbf{F}$ satisfies Explainability, Exhaustivity.

Most of the axioms are compatible, i.e., there exists at least one explanation function that satisfies them all together (obviously for any classifier and any theory). It is no surprise that Coreness and Exhaustivity are incompatible since they express diverging strategies that may be followed by explainers. Finally, since core literals may not exist, the three axioms (Success, Explainability, Coreness) are incompatible.

Proposition 4. The following holds:

- Success, Explainability, Irreducibility, Feasibility, Representativity, Relevance, Exhaustivity, Completeness, and Coherence are compatible;
- Success, Irreducibility, Feasibility, Representativity, Relevance, Coreness, and Completeness are compatible;
- Explainability, Irreducibility, Feasibility, Relevance, Coreness, and Completeness are compatible;
- Coreness and Exhaustivity are incompatible.
- Success, Explainability and Coreness are incompatible.

4 Explainers Based on Abductive Reasoning

One of the most studied explainers is based on abductive reasoning. It looks for sets of literals that are sufficient for assigning a class to a given instance. It thus explains instances instead of classes. Its explanations are called minimal sufficient subsets in [Camburu *et al.*, 2020], prime implicants in [Shih *et al.*, 2018; Darwiche and Hirth, 2020] or abductive explanations in [Ignatiev, 2020]. In [Amgoud, 2021a], abductive reasoning is used for explaining classes. The idea is to highlight *factors that caused a class*.

In that spirit, we investigate a family of class explainers based on the abductive reasoning. We call them the *sufficiency explainers*. Such explainers generate explanations under complete information (i.e., the whole set of instances is available, which is reasonable for explaining some quite simple classifiers like decision trees) and adopt the following abductive principle: if a class x is assigned whenever a literal l is observed, then we extrapolate that l is a reason for x.

Let us formally define the sufficiency explainers. As a preliminary, we need a notation for the set of all those subsets of literals that are sufficient to force a certain class:

Definition 9. Let $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ be a class question. We denote by $\mathtt{Suff}_{\mathbf{Q}}$ the set of every $L \subseteq \mathtt{Lit}_{\mathbf{T}}$ such that:

- L is consistent;
- $\forall\, I \in \mathtt{Inst}_{\mathbf{T}}$, if $L \subseteq I$, then $I \in \mathtt{Inst}_{\mathbf{TR}}(x)$.

We are ready to define our family of explainers based on complete information and the abductive reasoning:

Definition 10 (Sufficiency)**.** A *sufficiency* class explainer is a class explainer $\mathbf{F}$ such that, for every class question $\mathbf{Q}$,

- $\mathbf{F}(\mathbf{Q}) \subseteq \mathtt{Suff}_{\mathbf{Q}}$,
- $\forall\, I \in \mathtt{Inst}_{\mathbf{TR}}(x), \exists\, L \in \mathbf{F}(\mathbf{Q}), L \subseteq I$.

Next, we characterize the sufficiency explainers with three axioms, namely Feasibility, Representativity, and Coherence. As a preliminary, we first show that every class explainer satisfying the three aforementioned axioms returns explanations which are subsets of those generated by $\mathtt{Suff}_{\mathbf{Q}}$:

Theorem 2. *If a class explainer* $\mathbf{F}$ *satisfies Feasibility, Representativity and Coherence, then, for any class question* $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$*, the inclusion* $\mathbf{F}(\mathbf{Q}) \subseteq \mathtt{Suff}_{\mathbf{Q}}$ *holds.*

We are ready for the characterization:

Theorem 3. *A class explainer* $\mathbf{F}$ *satisfies Feasibility, Representativity, Coherence iff* $\mathbf{F}$ *is a sufficiency class explainer.*

It is worth mentioning that a sufficiency class explainer violates Relevance, Coreness and Irreducibility (see Table 1). In what follows, we provide two specific explainers of this family. The first one, called the *all-abductive explainer* (`aAbd`), returns all sufficient reasons for a class.

Definition 11 (`aAbd`)**.** We denote by `aAbd` the class explainer transforming every class question $\mathbf{Q}$ into $\mathtt{Suff}_{\mathbf{Q}}$.

Example 1 (Cont.) Examples of reason for Admit are: $\{(E,1),(F,0)\}$, $\{(E,1),(F,0),(GPA,1)\}$.

The following result shows that the class explainer `aAbd` satisfies most of the axioms.

Theorem 4. *The following properties hold:*

- `aAbd` *satisfies Success, Explainability, Feasibility, Representativity, Exhaustivity, Completeness, Coherence;*
- `aAbd` *violates Irreducibility, Relevance and Coreness.*

We turn to a second specific sufficiency explainer, called the *min-abductive explainer* (`mAbd`). The latter returns the minimal sufficient reasons for a class.

	Sufficiency	aAbd	mAbd	aCtf	mCtf	xCtf	f-rAbd
Success	•	•	•	•		•	•
Explainability	•	•	•		•		•
Irreducibility			•	•	•	•	•
Feasibility	•	•	•	•	•	•	•
Representativity	•	•	•	•		•	
Relevance			•	•	•	•	
Coreness				•	•	•	
Exhaustivity	•	•	•				
Completeness	•	•	•	•	•	•	
Coherence	•	•	•				

Table 1: The symbol • stands for the axiom is satisfied by the explainer.

Definition 12 (mAbd)**.** The *min-abductive class explainer* (mAbd) is a class explainer transforming every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ into the set of every $L \subseteq \mathtt{Lit}_{\mathbf{T}}$ such that:

- L is consistent;
- $\forall\, I \in \mathtt{Inst}_{\mathbf{T}}$ such that $L \subseteq I$, $I \in \mathtt{Inst}_{\mathbf{TR}}(x)$;
- $\forall\, L' \subset L$, L' does not satisfy the above point.

Example 1 (Cont.) The class Admit has three reasons, which correspond to the three preconditions of the rules. The same holds for Deny.

The explainer mAbd refines aAbd by keeping only the minimal (for set-inclusion) explanations.

Proposition 5. For any class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, $\mathtt{mAbd}(\mathbf{Q}) = \{L \in \mathtt{aAbd}(\mathbf{Q}) \,:\, \forall L' \subset L, L' \notin \mathtt{aAbd}(\mathbf{Q})\}$.

The min-abductive explainer satisfies all our axioms except Coreness. Due to the minimality condition, mAbd ensures that every literal in an explanation is relevant to the explained class. Furthermore, it keeps only the minimally sufficient subset of literals for causing a class.

Theorem 5. mAbd *satisfies Success, Explainability, Irreducibility, Feasibility, Representativity, Relevance, Exhaustivity, Completeness, and Coherence, but violates Coreness.*

We now present below a representation theorem which characterizes the abductive explainer mAbd. We show that mAbd is the *only* explainer satisfying all axioms except coreness (recall that some axioms imply others).

Theorem 6. *A class explainer* $\mathbf{F}$ *satisfies Irreducibility, Feasibility, Representativity, and Coherence iff* $\mathbf{F} = \mathtt{mAbd}$.

5 Explainers Based on Counterfactual Reasoning

We turn to a second family of explainers, called the *necessity explainers*. It is based on complete information and the following counterfactual principle: if a literal l is observed whenever a class x is assigned, then we extrapolate that l is a reason for assigning x. Put differently, if l was not observed, then x would not have been assigned, hence the word counterfactual. As a preliminary to define the necessity explainers, we need a notation for those subsets of literals that are necessary to a certain class:

Definition 13. Let $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ be a class question. We denote by $\mathtt{Nec}_{\mathbf{Q}}$ the set of every $L \subseteq \mathtt{Lit}_{\mathbf{T}}$ such that:

- L is consistent;
- $\forall\, I \in \mathtt{Inst}_{\mathbf{T}}$, if $L \not\subseteq I$, then $I \notin \mathtt{Inst}_{\mathbf{TR}}(x)$.

Note that the necessary subsets of literals for a class x constitute the power set of the intersection of all instances of x.

Proposition 6. Let $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ be a class question. Then, $\mathtt{Nec}_{\mathbf{Q}} = \mathtt{Pow}[\bigcap \mathtt{Inst}_{\mathbf{TR}}(x)]$.

We are ready to define our family of explainers based on complete information and the counterfactual reasoning.

Definition 14 (Necessity)**.** A *necessity* class explainer is a class explainer $\mathbf{F}$ such that, for every class question $\mathbf{Q}$, $\mathbf{F}(\mathbf{Q}) \subseteq \mathtt{Nec}_{\mathbf{Q}}$.

Let us investigate a specific member of the family, which returns all necessary subsets of literals:

Definition 15. The *all-counterfactual explainer* (aCtf) is a function transforming every class question $\mathbf{Q}$ into $\mathtt{Nec}_{\mathbf{Q}}$.

Example 1 (Cont.) Let $\mathbf{Q}$ be the question centered on Admit and $\mathbf{Q}'$ the question centered on Deny. We have $\bigcap \mathtt{Inst}_{\mathbf{TR}}(\text{Admit}) = \{\langle \mathtt{E}, 1 \rangle\}$. Thus, $\mathtt{Nec}_{\mathbf{Q}} = \mathtt{Pow}(\{\langle \mathtt{E}, 1 \rangle\}) = \{\emptyset, \{\langle \mathtt{E}, 1 \rangle\}\}$. Thus, $\mathtt{aCtf}(\mathbf{T}, \mathbf{R}, \text{Admit}) = \{\emptyset, \{\langle \mathtt{E}, 1 \rangle\}\}$. Similarly, $\bigcap \mathtt{Inst}_{\mathbf{TR}}(\text{Deny}) = \emptyset$. Thus, $\mathtt{Nec}_{\mathbf{Q}'} = \mathtt{Pow}(\emptyset) = \{\emptyset\}$. Thus, $\mathtt{aCtf}(\mathbf{T}, \mathbf{R}, \text{Deny}) = \{\emptyset\}$.

We axiomatically analyse aCtf:

Theorem 7. aCtf *satisfies Success, Irreducibility, Feasibility, Representativity, Relevance, Coreness, Completeness. It violates Explainability, Exhaustivity, and Coherence.*

We turn to a second specific explainer, which minimizes the necessary subsets:

Definition 16 (mCtf)**.** The *min-counterfactual explainer* (mCtf) is the function transforming every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ into the set of every subset L of $\mathtt{Lit}_{\mathbf{T}}$ such that:

- $L \neq \emptyset$; L is consistent;
- $\forall\, I \in \mathtt{Inst}_{\mathbf{T}}$, if $L \not\subseteq I$, then $I \notin \mathtt{Inst}_{\mathbf{TR}}(x)$;
- $\forall\, L' \subset L$, L' does not satisfy the above two points.

Example 1 (Cont.) We have $\mathtt{mCtf}(\mathbf{T}, \mathbf{R}, \text{Admit}) = \{\{\langle \mathtt{E}, 1 \rangle\}\}$ and $\mathtt{mCtf}(\mathbf{T}, \mathbf{R}, \text{Deny}) = \emptyset$.

We axiomatically analyze mCtf. Note that we lose Success and Representativity, but we gain Explainability.

Theorem 8. mCtf *satisfies Explainability, Irreducibility, Feasibility, Relevance, Coreness, and Completeness. It violates Success, Representativity, Exhaustivity, and Coherence.*

Finally, we introduce a third specific explainer, which maximizes the necessary subsets.

Definition 17 (xCtf)**.** The *max-counterfactual explainer* (xCtf) is the function transforming every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ into the set of every subset L of $\mathtt{Lit}_{\mathbf{T}}$ such that:

- L is consistent;
- $\forall\, I \in \mathtt{Inst}_{\mathbf{T}}$, if $L \not\subseteq I$, then $I \notin \mathtt{Inst}_{\mathbf{TR}}(x)$;
- $\forall\, L' \supset L$, L' does not satisfy the above two points.

Notice that xCtf returns only one reason, namely the intersection of all instances of the class in question:

Proposition 7. Let $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ be a class question. Then, $\mathtt{xCtf}(\mathbf{Q}) = \{\bigcap \mathtt{Inst}_{\mathbf{TR}}(x)\}$.

Example 1 (Cont.) We have $\mathtt{xCtf}(\mathbf{T}, \mathbf{R}, \text{Admit}) = \{\{\langle \mathtt{E}, 1 \rangle\}\}$ and $\mathtt{xCtf}(\mathbf{T}, \mathbf{R}, \text{Deny}) = \{\emptyset\}$.

We axiomatically analyze xCtf and observe that it satisfies exactly the same axioms as aCtf. So, returning all necessary subsets or the largest one (i.e., the intersection of the instances of the class in question) lead to the same axioms.

Theorem 9. xCtf *satisfies Success, Irreducibility, Feasibility, Representativity, Relevance, Coreness, and Completeness. It violates Explainability, Exhaustivity, and Coherence.*

6 Explaining Under Incomplete Information

In this section, we investigate explanations under incomplete information (i.e., not all instances are available, which is typically the case with the dataset a classifier has been trained on, or the dataset generated for existing explainers like Anchors and LIME). Working with incomplete information makes sense, in particular, for complex classifiers whose querying may not be reasonable for all instances. Note that our abductive and counterfactual explainers (defined in the previous sections) work with the whole set of instances. However, in practice only a subset of instances (dataset) is available. The question is: does our previous results still hold if reasons are generated from a proper subset of $\mathtt{Inst}_{\mathbf{T}}$? The answer is unfortunately negative. We define a parameterized family of explainers that provide minimally sufficient reasons from a dataset. The parameter is a function which selects the dataset to be used. Such a definition abstracts Anchors and LIME since they both use datasets generated in different ways.

Definition 18 (Fragments)**.** Let $\mathbf{T} = \langle \mathcal{A}, d, \mathcal{C} \rangle$ be a classification theory, $\mathbf{R}$ a classifier on $\mathbf{T}$, and $S \subseteq \mathtt{Inst}_{\mathbf{T}}$. We say that S is a *fragment* in $\mathbf{T}$ and $\mathbf{R}$ iff, for every $x \in \mathcal{C}$, we have that $\mathtt{Inst}_{\mathbf{TR}}(x) \cap S \neq \emptyset$.

Definition 19. A *fragment selector* is a function f transforming every couple $(\mathbf{T}, \mathbf{R})$ such that $\mathbf{T}$ is a classification theory and $\mathbf{R}$ a classifier on $\mathbf{T}$ into a fragment in $\mathbf{T}$ and $\mathbf{R}$.

We are now ready to introduce the novel family.

Definition 20. Let f be a fragment selector. The *f-relaxed abductive explainer* (f-rAbd) is the function transforming every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ into the set of every subset L of $\mathtt{Lit}_{\mathbf{T}}$ such that:

- $\exists\, I \in f(\mathbf{T}, \mathbf{R})$ such that $L \subseteq I$;
- $\forall\, I \in f(\mathbf{T}, \mathbf{R})$ such that $L \subseteq I$, $I \in \mathtt{Inst}_{\mathbf{TR}}(x)$;
- $\forall\, L' \subset L$, L' does not satisfy the above point.

Property 4. Let f be a fragment selector and $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$ a class question. For any $L \in f\text{-rAbd}(\mathbf{Q})$, L is consistent.

We show that, for every fragment selector f, f-rAbd satisfies Success, Explainability, Feasibility and Irreducibility, and it violates the remaining axioms. This is not surprising since it generates explanations from a subset of instances.

Theorem 10. *Let f be a fragment selector. f-rAbd satisfies Success, Explainability, Feasibility and Irreducibility. It violates Coreness, Relevance, Completeness, Exhaustivity, Representativity and Coherence.*

The following result shows that f-rAbd satisfies a weak version of Representativity. Indeed, every instance of the set $f(\mathbf{T}, \mathbf{R})$ is a superset of at least one reason of its class.

Proposition 8. Let f be a fragment selector. f-rAbd satisfies *Weak Representativity*, i.e., for every class question $\mathbf{Q} = \langle \mathbf{T}, \mathbf{R}, x \rangle$, for every $I \in f(\mathbf{T}, \mathbf{R}) \cap \mathtt{Inst}_{\mathbf{TR}}(x)$, there exists $L \in f\text{-rAbd}(\mathbf{Q})$ such that $L \subseteq I$.

Existing heuristics explanation functions like Anchor and LIME violate Coherence, leading to incorrect outcomes in some cases. Recall that both Anchors and LIME are *not* class explainers, they are instance explainers, i.e., they provide reasons for assigning $\mathbf{R}(I)$ to an instance I.

7 Related Work

There haven't been a lot of axiomatic approaches to explainability. Most of existing works propose instances of explainers and analyse them either experimentally (eg. [Ignatiev *et al.*, 2019]) or formally (eg. [Darwiche and Hirth, 2020]). None of these works have discussed axioms. In [Wolf *et al.*, 2019], some axioms have been proposed for *instance* explainers. Our axioms concern class explainers.

Contrastive explanations are widely studied. They describe what should be modified in order to avoid a class. It has been shown in [Amgoud, 2021a] that they are dual to the reasons generated by mAbd. Hence, they represent the same concept. That's why in this paper, we investigated only one of them.

8 Conclusion

This paper studied foundations of explainers that justify classes. It provided key axioms that an explainer would satisfy and characterised various explainers that satisfy them. It highlighted the key axioms that separate sufficient reasons from necessary ones (i.e., counterfactuals). Another important result of the paper concerns the family of explainers that generate reasons from a subset of instances. We showed that they violate relevance, leading to erroneous explanations.

As a future work, we plan to extend our axioms for dealing with other types of explanations like the conversational ones.

Acknowledgments

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

References

[Amgoud, 2021a] Leila Amgoud. Explaining black-box classification models with arguments. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI*, pages 791–795, 2021.

[Amgoud, 2021b] Leila Amgoud. Non-monotonic explanation functions. In *Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU*, volume 12897 of *Lecture Notes in Computer Science*, pages 19–31, 2021.

[Audemard *et al.*, 2020] Gilles Audemard, Frédéric Koriche, and Pierre Marquis. On tractable XAI queries based on compiled representations. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR*, pages 838–849, 2020.

[Biran and Cotton, 2017] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, pages 1–6, 2017.

[Camburu *et al.*, 2020] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob N. Foerster, Thomas Lukasiewicz, and Phil Blunsom. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. *ArXiv*, abs/2009.11023, 2020.

[Cyras *et al.*, 2021] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4392–4399, 2021.

[Darwiche and Hirth, 2020] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *24th European Conference on Artificial Intelligence ECAI*, volume 325, pages 712–720. IOS Press, 2020.

[Dhurandhar *et al.*, 2018] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 590–601, 2018.

[Guidotti *et al.*, 2019] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2019.

[Ignatiev *et al.*, 2019] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, pages 15857–15867, 2019.

[Ignatiev *et al.*, 2020] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. From contrastive to abductive explanations and back again. In *XIXth International Conference of the Italian Association for Artificial Intelligence*, volume 12414 of *Lecture Notes in Computer Science*, pages 335–355, 2020.

[Ignatiev, 2020] Alexey Ignatiev. Towards trustable explainable AI. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pages 5154–5158, 2020.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Narodytska *et al.*, 2019] Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and João Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *Proceedings of the 22nd International Conference Theory and Applications of Satisfiability Testing - SAT*, pages 267–278, 2019.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, page 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1527–1535, 2018.

[Schneider and Handali, 2019] Johannes Schneider and Joshua Peter Handali. Personalized explanation for machine learning: a conceptualization. In *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS*, 2019.

[Shih *et al.*, 2018] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 5103–5111, 2018.

[Wolf *et al.*, 2019] Lior Wolf, Tomer Galanti, and Tamir Hazan. A formal approach to explainability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society AIES*, pages 255–261, 2019.