From Contrastive to Abductive Explanations and Back Again

Alexey Ignatiev[1] , **Nina Narodytska**[2] , **Nicholas Asher**[3] , **Joao Marques-Silva**[3]
[1] Monash University, Melbourne, Australia
[2] VMware Research, CA, USA
[3] IRIT, CNRS, Toulouse, France
alexey.ignatiev@monash.edu, nnarodytska@vmware.com,
{nicholas.asher, joao.marques-silva}@irit.fr

Abstract

Explanations of Machine Learning models often address a 'Why?' question. These can are related with selecting feature-value pairs sufficient for the prediction. Recent work has investigated explanations that address a 'Why Not?' question, i.e. finding a change of feature values that guarantees a change of prediction. These two forms of explanations of ML models appear to be mostly unrelated. However, this paper demonstrates otherwise and establishes a rigorous formal relationship between 'Why?' and 'Why Not?' explanations. The paper proves that for any given instance, 'Why?' explanations are minimal hitting sets of 'Why Not?' explanations and vice-versa. Moreover, the paper devises novel algorithms for extracting and enumerating both forms of explanations.

1 Background

Recent work (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019a; Darwiche and Hirth 2020; Marques-Silva et al. 2020; Huang et al. 2021) proposed model-based approaches to computing *rigorous* explanations of machine learning (ML) models, offering the strongest *formal* guarantees with respect to the underlying ML model. They contrast with the majority of *heuristic* approaches to explainability represented by model-agnostic explanations (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2018). Most work on ML explainability aims to answer a 'Why prediction π?' question. The answer to this question has been referred to as PI-explanations (Shih, Choi, and Darwiche 2018), *abductive* explanations (Ignatiev, Narodytska, and Marques-Silva 2019a), and (minimal) sufficient reasons (Darwiche and Hirth 2020). Another kind of explanations aims to answer a 'Why Not?' question (Miller 2019); these are known by the name of *contrastive* explanations.

This paper focuses on the relationship between *local* (i.e. applied to a concrete data instance) abductive and contrastive explanations. First, it formally defines contrastive explanations. Second, it demonstrates that local abductive and contrastive explanations are related by a minimal hitting set relationship, which builds on the seminal work of R. Reiter (Reiter 1987). Crucially, this novel hitting set relationship reveals a wealth of algorithms for computing and for enumerating contrastive and abductive explanations.

Explainability in Machine Learning. We assume classification problems with a set of classes $\mathbb{K}$ and an ML model $\mathbb{M}$, which is represented by a finite set of first-order logic (FOL) sentences $\mathcal{M}$. Consider a set of features $\mathcal{F}$; an *instance* is an assignment of values to features. The space of instances is defined by $\mathbb{F}$. A prediction $\pi \in \mathbb{K}$ is associated with each instance $X \in \mathbb{F}$. Let us define a predicate $\mathcal{M}_\pi \subseteq \mathbb{F}$, s.t. $\mathcal{M}_\pi(X)$ is true iff the input X is consistent with prediction π given the ML model $\mathbb{M}$. A consistent conjunction of feature literals τ is an *implicant* of $\mathcal{M}_\pi$, denoted by $\tau \models \mathcal{M}_\pi$, if

$$\forall (X \in \mathbb{F}).\tau(X) \rightarrow \mathcal{M}_\pi(X) \tag{1}$$

is true. An implicant is called *prime* if none of its proper subsets is an implicant. Local abductive explanations for an instance X are *prime implicants* of $\mathcal{M}_\pi$ consistent with X.

Analysis of Inconsistent Formulas. Consider *inconsistent* formulas $\mathcal{F}$, i.e. $\mathcal{F} \models \bot$, represented as conjunctions of clauses. Also, let $\mathcal{F} = \mathcal{B} \cup \mathcal{R}$, where $\mathcal{R}$ contains *relaxable* clauses (allowed not to be satisfied to restore consistency), and $\mathcal{B}$ contains the *non-relaxable* clauses (these must be satisfied). The following definition characterizes two dual notions used in the analysis of inconsistent formulas.

Definition 1. *Let $\mathcal{F} = \mathcal{B} \cup \mathcal{R}$ s.t. $\mathcal{F} \models \bot$. $\mathcal{U} \subseteq \mathcal{R}$ is a* Minimal Unsatisfiable Subset *(MUS) iff $\mathcal{B} \cup \mathcal{U} \models \bot$ and $\forall_{\mathcal{U}' \subsetneq \mathcal{U}}$, $\mathcal{B} \cup \mathcal{U}' \not\models \bot$. $\mathcal{T} \subseteq \mathcal{R}$ is a* Minimal Correction Subset *(MCS) iff $\mathcal{B} \cup \mathcal{R} \setminus \mathcal{T} \not\models \bot$ and $\forall_{\mathcal{T}' \subsetneq \mathcal{T}}$, $\mathcal{B} \cup \mathcal{R} \setminus \mathcal{T}' \models \bot$.*

A fundamental result in reasoning about inconsistent clause sets is the minimal hitting set (MHS) duality relationship between MUSes and MCSes (Reiter 1987; Birnbaum and Lozinskii 2003): *MCSes are MHSes of MUSes and vice-versa.* This result has been extensively used in the development of algorithms for MUSes and MCSes (Bailey and Stuckey 2005; Liffiton and Sakallah 2008; Liffiton et al. 2016), and also applied in a number of various settings.

2 Contributions: Explanations and Duality

As mentioned above, recent work (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019a; Darwiche and Hirth 2020) proposed to relate model-based abductive explanations explanations with prime implicants.

Definition 2 (Abductive Explanation)**.** *Given an instance τ, with a prediction π, and an ML model represented with a predicate $\mathcal{M}_\pi$, i.e. $\tau \models \mathcal{M}_\pi$, an* abductive explanation *(AXp)*

is a minimal subset of literals $\sigma \subseteq \tau$ such that $\sigma \models \mathcal{M}_\pi$.

We propose the following definition of a (local) contrastive explanation, which captures the intuitive notion of the contrastive explanation discussed in (Miller 2019).

Definition 3 (Contrastive Explanation)**.** *Given an instance τ, with a prediction π, and an ML model represented by a predicate $\mathcal{M}_\pi$, i.e. $\tau \models \mathcal{M}_\pi$, a* contrastive explanation *(CXp) is a minimal subset of literals $\rho \subseteq \tau$ such that $\tau \setminus \rho \not\models \mathcal{M}_\pi$.*

Explainability and Inconsistent Formulas. Consider a set of feature values τ with a prediction is π, i.e. $\tau \models \mathcal{M}_\pi$. Equivalently, $\tau \land \neg\mathcal{M}_\pi \models \bot$. Thus,

$$\tau \land \neg\mathcal{M}_\pi \tag{2}$$

is inconsistent, with the background knowledge being $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$ and the relaxable clauses being $\mathcal{R} \triangleq \tau$. As proposed in (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019a), a (local abductive) explanation is a subset-minimal set σ of the literals in τ, such that, $\sigma \land \neg\mathcal{M}_\pi \models \bot$. Therefore, the following holds:

Proposition 1. *Local abductive explanations are MUSes of the pair $(\mathcal{B}, \mathcal{R})$, $\tau \land \neg\mathcal{M}_\pi$, where $\mathcal{R} \triangleq \tau$ and $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$.*

Now, consider an MCS $\rho \subseteq \tau$ of equation 2. As a result, $\bigwedge_{l \in \tau \setminus \rho}(l) \land \neg\mathcal{M}_\pi \not\models \bot$. Hence, given Definition 3, observe that the following holds:

Proposition 2. *Local contrastive explanations are MCSes of the pair $(\mathcal{B}, \mathcal{R})$, $\tau \land \neg\mathcal{M}_\pi$, where $\mathcal{R} \triangleq \tau$ and $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$.*

Duality. Given the results above, and the hitting set duality between MUSes and MCSes (Reiter 1987; Birnbaum and Lozinskii 2003), we have the following.

Theorem 1. *AXps are MHSes of CXps and vice-versa.*

Propositions 1–2, and Theorem 1 can serve to exploit the vast body of work on the analysis of inconsistent formulas for computing both CXps and AXps and, arguably more importantly, to enumerate explanations (Ignatiev et al. 2020; Marques-Silva et al. 2021; Huang et al. 2021).

Note that earlier work (Ignatiev, Narodytska, and Marques-Silva 2019b) established a relation between prime implicants and implicates as a way to relate global AXps and so-called *counterexamples*. In contrast, this work delved into the fundamentals of reasoning about inconsistency, concretely the duality between MCSes and MUSes, and established a relation between *local* AXps and CXps.

3 Relevance to KR & Significance

The importance of our work (Ignatiev et al. 2020) has been underlined recently. Indeed, the minimal hitting set duality between AXps and CXps served to prove that deciding membership of a feature in some explanation is in polynomial time in the case of decision tree ML models (Huang et al. 2021). Furthermore, the duality relationship between AXps and CXps is the central property driving the explanation enumeration algorithms proposed for graph-based ML models (Huang et al. 2021) as well as monotone classifiers (Marques-Silva et al. 2021). Finally, we believe the results of our work may be crucial in the context of classifiers represented in other knowledge compilation (KC) languages (Darwiche and Marquis 2002) and prove overall helpful at the intersection of KR and ML.

References

Bailey, J., and Stuckey, P. J. 2005. Discovery of minimal unsatisfiable subsets of constraints using hitting set dualization. In *PADL*, 174–186.

Birnbaum, E., and Lozinskii, E. L. 2003. Consistent subsets of inconsistent systems: structure and behaviour. *J. Exp. Theor. Artif. Intell.* 15(1):25–46.

Darwiche, A., and Hirth, A. 2020. On the reasons behind decisions. In *ECAI*, 712–720.

Darwiche, A., and Marquis, P. 2002. A knowledge compilation map. *J. Artif. Intell. Res.* 17:229–264.

Huang, X.; Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2021. On efficiently explaining graph-based classifiers. In *KR*, Accepted for publication.

Ignatiev, A.; Narodytska, N.; Asher, N.; and Marques-Silva, J. 2020. From contrastive to abductive explanations and back again. In *AI*IA*, 335–355.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019a. Abduction-based explanations for machine learning models. In *AAAI*, 1511–1519.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019b. On relating explanations and adversarial examples. In *NeurIPS*, 15857–15867.

Liffiton, M. H., and Sakallah, K. A. 2008. Algorithms for computing minimal unsatisfiable subsets of constraints. *J. Autom. Reasoning* 40(1):1–33.

Liffiton, M. H.; Previti, A.; Malik, A.; and Silva, J. M. 2016. Fast, flexible MUS enumeration. *Constraints* 21(2):223–250.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *NIPS*, 4765–4774.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2020. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *NeurIPS*.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2021. Explanations for monotonic classifiers. In *ICML*, Accepted for publication.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267:1–38.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artif. Intell.* 32(1):57–95.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*, 1527–1535.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, 5103–5111.