Logics for Binary-input Classifiers and Their Explanations

Emiliano Lorini, Xinghan Liu

Logics for Binary-input Classifiers and Their Explanations

Xinghan Liu[1,*,†], Emiliano Lorini[1,*,†]

[1]IRIT-CNRS, University of Toulouse, 118 Route de Narbonne, 31062, Toulouse, France

Abstract

We present our work [1, 2] on modal logics for binary-input classifiers and their explanations. They are able to represent classifiers that propositional logic cannot. In particular, black box classifier is understood as uncertainty among admissible classifiers which are coherent with an agent's partial knowledge, and represented in a product modal logic framework. We also briefly show the logics' application to XAI.

Keywords

Boolean function, black box classifier, product modal logic, explainable AI (XAI)

1. Introduction

The notions of explanation and explainability have been extensively investigated by philosophers [3, 4, 5] and are key aspects of AI-based systems. Classifier systems compute a given function in the context of a classification or prediction task. Explaining why the system has classified a given instance in a certain way is crucial for making the system intelligible and for finding biases in the classification process. Thus, a variety of notions of explanations have been discussed in the area of explainable AI (XAI) including abductive, contrastive and counterfactual explanations [6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

At the mathematical level, a Boolean classifier is nothing but a Boolean function f, and traditionally is represented by a propositional formula φ. Using modal logic we can model binary-input classifiers which have finite-valued output and are possibly partial. Moreover, it enables us to represent black box classifiers which are key research objects in XAI. A classifier is a white box, if it is determined and given in the model, while black box is understood as uncertainty (indeterminacy) among admissible classifiers which are coherent with an agent's partial knowledge about the classifier. In this paper we present four modal logics for binary-input classifiers in a unified framework: BCL (Binary-input Classifier Logic) and WBCL (Weak BCL); PLC (Product logic for binary-input Classifier) and WPLC (Weak PLC), according to whether the set of atomic propositions is finite or countably infinite, and whether the represented classifiers are white box or black box, see Table 1.

1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022

*Corresponding author.

†These authors contributed equally.

xinghan.liu@univ-toulouse.fr (X. Liu); Emiliano.Lorini@irit.fr (E. Lorini)

	BCL	WBCL	PLC	WPLC
Cardinality of language is	finite	infinite	finite	infinite
Classifiers are	white box	white box	black box	black box

Table 1
Four modal logics for classifiers

Furthermore, we exemplify how to apply them to XAI by a) expressing abductive explanation in our language; b) defining its counterpart in the case of black box classifiers; c) formalizing the explanation verification as a model checking problem.

2. Four Logics: BCL, WBCL, PLC and WPLC

Language Let Atm_0 be a countable set of atomic propositions with elements noted $p, p', \ldots$ to denote input variables of classifiers. We introduce a finite set Val to denote the *output values* (classifications, decisions) of the classifier. Elements of Val are noted $c, c', \ldots$ for classes. For any $c \in Val$, we call $\mathrm{t}(c)$ a *decision atom*, to be read as "the actual decision (or output) takes value c", and have $Dec = \{\mathrm{t}(c) : c \in Val\}$. Finally, let $Atm = Atm_0 \cup Dec$. The modal language $\mathscr{L}(Atm)$ is hence defined by the following grammar:

$$\varphi \quad ::= \quad p \mid \mathrm{t}(c) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_{\mathrm{I}}\varphi \mid \Box_{\mathrm{F}}\varphi,$$

where p ranges over Atm_0, c ranges over Val, and X is a finite subset of Atm_0 which we note $X \subseteq^{\mathrm{fin}} Atm_0$. Connectives $\vee, \rightarrow, \leftrightarrow, \Diamond_{\mathrm{I}}$ and $\Diamond_{\mathrm{F}}$ are defined in the normal way. Let $Atm(\varphi), Atm_0(\varphi), Dec(\varphi)$ denote the set of all atomic propositions, input variables, and decision atoms in the formula φ respectively. Finally, let $\mathscr{L}^{-\Box_{\mathrm{F}}}(Atm)$ denote the $\Box_{\mathrm{F}}$-free fragment of $\mathscr{L}(Atm)$.

Semantics The language is built to model (possibly partial) functions from 2^{Atm_0} to Val, and their interactions. Let us begin with the language $\mathscr{L}^{-\Box_{\mathrm{F}}}(Atm)$, which is interpreted relative to classifier models whose class is defined as follows.

Definition 1 (Classifier model). *A classifier model (CM) is a tuple $C = (S, f)$ where:*

- *$S \subseteq 2^{Atm_0}$ is the set of states, and*
- *$f : S \longrightarrow Val$ is a decision (or classification) function.*

*A pointed CM is a pair (C, s) where C is a CM and $s \in S$. We call $C = (S, f)$ finite if S is finite. The class of (finite) classifier models is noted **CM** (**finite-CM**).*

Hence, the classifier f has more than 2 outputs if $|ran(f)| > 2$; has countably infinite variables if Atm_0 is countably infinite; and is partial if $S \neq 2^{Atm_0}$. Essentially, one can view a CM just as an S5 model on Atm_0 with partition labelled by elements in Val, which is indicated by the satisfaction relation defined below, where $\Box_{\mathrm{I}}$ tentatively seems nothing but an S5 operator.

Definition 2 (Satisfaction relation 1). *Let $\varphi \in \mathscr{L}^{-\Box_F}(Atm)$, $C = (S, f)$ be a CM and $s \in S$:*

$$\begin{aligned}
(C, s) \models p &\iff p \in s,\\
(C, s) \models \mathsf{t}(c) &\iff f(s) = c,\\
(C, s) \models \neg\varphi &\iff (C, s) \not\models \varphi,\\
(C, s) \models \varphi \wedge \psi &\iff (C, s) \models \varphi \text{ and } (C, s) \models \psi,\\
(C, s) \models \Box_I\varphi &\iff \forall s' \in S, (C, s') \models \varphi.
\end{aligned}$$

As mentioned, we think of black box classifier as uncertainty over a set of admissible classifiers coherent with the agent's partial knowledge. This thought is formalized as the multi-classifier model defined below, which is nothing but a set of CMs with the same set of states.

Definition 3 (Multi-classifier model). *A multi-classifier model (MCM) is a pair $\Gamma = (S, \Phi)$ where $S \subseteq 2^{Atm_0}$ and $\Phi \subseteq Val^S$, where Val^S is the set of functions with domain S and codomain Val. A pointed MCM is a triple (Γ, s, f) where $\Gamma = (S, \Phi)$ is an MCM, $s \in S$ and $f \in \Phi$. We call $\Gamma = (S, \Phi)$ finite if S is finite. The class of all (finite) multi-classifier models is noted **MCM** (**finite-MCM**).*

Definition 4 (Satisfaction relation 2). *Let $\varphi \in \mathscr{L}(Atm)$, $\Gamma = (S, \Phi)$ an MCM, $s \in S$ and $f \in \Phi$:*

$$\begin{aligned}
(\Gamma, s, f) \models p &\iff p \in s,\\
(\Gamma, s, f) \models \mathsf{t}(c) &\iff f(s) = c,\\
(\Gamma, s, f) \models \neg\varphi &\iff (\Gamma, s, f) \not\models \varphi,\\
(\Gamma, s, f) \models \varphi \wedge \psi &\iff (\Gamma, s, f) \models \varphi \text{ and } (\Gamma, s, f) \models \psi,\\
(\Gamma, s, f) \models \Box_I\varphi &\iff \forall s' \in S : (\Gamma, s', f) \models \varphi,\\
(\Gamma, s, f) \models \Box_F\varphi &\iff \forall f' \in \Phi : (\Gamma, s, f') \models \varphi.
\end{aligned}$$

Both $\Box_I\varphi$ and $\Box_F\varphi$ have standard modal reading but range over different sets. $\Box_I\varphi$ has to be read "φ *necessarily* holds for the actual function, regardless of the input instance", while its dual $\Diamond_I\varphi =_{def} \neg\Box_I\neg\varphi$ has to be read "φ *possibly* holds for the actual function, regardless of the input instance". Similarly, $\Box_F\varphi$ has to be read "φ *necessarily* holds for the actual input instance, regardless of the function" and its dual $\Diamond_F\varphi$ has to be read "φ *possibly* holds for the actual input instance, regardless of the function". Therefore, the agent knows that the actual classification for s is c, if $(\Gamma, s, f) \models \Box_F\mathsf{t}(c)$, i.e. only classifiers outputting c for s are admissible; and $(\Gamma, s, f) \models \Diamond_F\mathsf{t}(c)$ means that classifying s as c is coherent with agent's partial knowledge. With these two modal dimensions, our framework subjects to the so-called *product modal logic*.

An important abbreviation is the following, where $X \subseteq^{\text{fin}} Atm_0$:

$$[X]\varphi =_{def} \bigwedge_{Y \subseteq X} \Big((\bigwedge_{p \in Y} p \wedge \bigwedge_{p \in X \setminus Y} \neg p) \to \Box_I((\bigwedge_{p \in Y} p \wedge \bigwedge_{p \in X \setminus Y} \neg p) \to \varphi)\Big).$$

Complex as it seems, $[X]\varphi$ means nothing but "φ necessarily holds, if the values of the input variables in X are kept fixed". It can be justified by checking that $(\Gamma, s, f) \models [X]\varphi$, if and only if $\forall s' \in S$, if $s \cap X = s' \cap X$ then $(\Gamma, s', f) \models \varphi$. Its dual $\langle X\rangle\varphi =_{def} \neg[X]\neg\varphi$ has to be read "φ *possibly* holds, if the values of the input variables in X are kept fixed". These modalities have a *ceteris paribus* reading and were first introduced in [16]. Notice when $X = \emptyset$, $[\emptyset]$ coincides with $\Box_I$.

Axiomatics We have to separate two cases, when Atm_0 is finite or countably infinite. The reason lays on the axiom **Funct** in Table 2, which intends to express the "functionality" property syntactically. We define $\mathsf{cn}_{X,Atm_0} =_{def} \bigwedge_{p\in X} p \wedge \bigwedge_{p\in Atm_0\setminus X} \neg p$. But when Atm_0 is infinite, cn_{X,Atm_0} is not a well-formed formula, and **Funct** has to be abandoned.

Definition 5 (Axiomatics). *We define* PLC *as the extension of classical propositional logic with all axioms and inference rules in Table 2;* WPLC *as* PLC *minus* **Funct***;* BCL *as all* $\Box_{\mathrm{F}}$*-free axioms and inference rule in Table 2; and* WBCL *as* BCL *minus* **Funct**.

$\big(\blacksquare\varphi \wedge \blacksquare(\varphi \to \psi)\big) \to \blacksquare\psi$	$(\mathbf{K}_{\blacksquare})$
$\blacksquare\varphi \to \varphi$	$(\mathbf{T}_{\blacksquare})$
$\blacksquare\varphi \to \blacksquare\blacksquare\varphi$	$(\mathbf{4}_{\blacksquare})$
$\neg\blacksquare\varphi \to \blacksquare\neg\blacksquare\varphi$	$(\mathbf{5}_{\blacksquare})$
$\Box_{\mathrm{F}}\Box_{\mathrm{I}}\varphi \leftrightarrow \Box_{\mathrm{I}}\Box_{\mathrm{F}}\phi$	(**Comm**)
$\bigvee_{c\in Val} \mathsf{t}(c)$	$(\mathbf{AtLeast}_{\mathsf{t}(c)})$
$\mathsf{t}(c) \to \neg\mathsf{t}(c')$ if $c \neq c'$	$(\mathbf{AtMost}_{\mathsf{t}(c)})$
$\big(\mathsf{cn}_{X,Atm_0} \wedge \mathsf{t}(c)\big) \to \Box_{\mathrm{I}}\big(\mathsf{cn}_{X,Atm_0} \to \mathsf{t}(c)\big)$	(**Funct**)
$p \to \Box_{\mathrm{F}} p$	$(\mathbf{Indep}_{\Box_{\mathrm{F}},p})$
$\neg p \to \Box_{\mathrm{F}}\neg p$	$(\mathbf{Indep}_{\Box_{\mathrm{F}},\neg p})$
$\dfrac{\varphi}{\blacksquare\varphi}$	$(\mathbf{Nec}_{\blacksquare})$

Table 2
Axioms and rules of inference, with $\blacksquare \in \{\Box_{\mathrm{I}}, \Box_{\mathrm{F}}\}$

We obtained the technical results in Theorem 1 and Table 3, whose proofs are in [1, 2].

Theorem 1. *Let* Atm_0 *be finite, then* BCL *and* PLC *are sound and complete with respect to* ***CM*** *and MCM respectively. Let* Atm_0 *be infinite, then* WBCL *and* WPLC *are sound and complete with respect to* ***CM*** *and* ***MCM*** *respectively.*

	Finite variables	**Infinite variables**
Fragment $\mathscr{L}^{-\Box_{\mathrm{F}}}(Atm)$	Polynomial	NP-complete
Full language $\mathscr{L}(Atm)$	Polynomial	in NEXPTIME

Table 3
Summary of complexity results

3. Classifier Explanations: Objective and Subjective

In the jargon of Boolean functions, a term or a property is a conjunction of literals (an atom or its negation), which we denote by λ. We use $Term(X)$ to denote all terms whose atoms are in X.

In the XAI literature recently people have focused on *local explanation*, namely to answer why the given instance is classified as such and so. A central notion is called *abductive explanation* [11] (or *sufficient reason* [17]). It is expressible in $\mathscr{L}(Atm)$ as follows:

$$\mathtt{AXp}(\lambda, c) =_{def} \lambda \wedge [Atm(\lambda)]\mathrm{t}(c) \wedge \bigvee_{p \in Atm(\lambda)} \langle Atm(\lambda) \setminus \{p\}\rangle \neg \mathrm{t}(c).$$

The three conjuncts mean that 1) λ is a "part" of the instance; 2) atoms in $Atm(\lambda)$ staying the same valuation as in s, $\mathrm{t}(c)$ necessarily holds regardless of other atoms; 3) λ is the "minimal" such property, in the sense that any its proper part $\lambda' \subset \lambda$ fails condition 2). Hence intuitively, it is sufficient and necessary to answer why the actual classification is c by saying "because the instance obtains property λ".

Let $C = (S, f)$ be a CM. We say that f is X-definite for $X \subseteq^{\text{fin}} Atm_0$, if $\forall s, s' \in S$, $s \cap X = s' \cap X$ then $f(s) = f'(s)$. And it is easy to see that f is X-definite, iff $(C, s) \models \mathtt{Def}(X)$ where $\mathtt{Def}(X) =_{def} \bigwedge_{c \in Val} \Box_{\mathrm{I}}(\langle X\rangle \mathrm{t}(c) \rightarrow [X]\mathrm{t}(c))$. When the classifier is X-definiteness for some $X \subseteq^{\text{fin}} Atm_0$, AXp always exists for the actual classification. We may call it the "principle of sufficient reason" (PSR) in term of Spinoza [Ethics, 1p11d2], and obtain the following validity.

$$\models_{\mathbf{CM}} (\mathrm{t}(c) \wedge \mathtt{Def}(X)) \rightarrow \bigvee_{\lambda \in Term(X)} \mathtt{AXp}(\lambda, c)$$

However, a sufficient reason may not be known to the agent when the classifier is a black box. We define λ as a *subjective* abductive explanation of the actual classification c, noted $\mathtt{SubAXp}(\lambda, c)$, if the agent knows that λ is an abductive explanation of the actual classification c, that is:

$$\mathtt{SubAXp}(\lambda, c) =_{def} \Box_{\mathrm{F}} \mathtt{AXp}(\lambda, c).$$

To see how SubAXp fails PSR, consider the following example. Suppose a classifier trained for deciding whether a paper is acceptable for a conference which has four input features: *si*gnificance, *or*iginality, *cl*arity and *an*onymity. Let 1 and 0 denote acceptance and rejection respectively.

Example 1 (Fail of PSR in black box). *Let $\Gamma = (S, \Phi)$ be an MCM of this black box, where $S = 2^{\{si,or,cl,an\}}$ and $s_1 = \{si, or, an\} \in S$. Consider $f_1, f_2 \in \Phi$ whose syntactic expressions are $\Box_{\mathrm{I}}(\mathrm{t}(1) \leftrightarrow ((or \wedge an) \vee (cl \wedge an))$, and $\Box_{\mathrm{I}}(\mathrm{t}(1) \leftrightarrow (si \wedge an))$ resp.. Then,*

$$(\Gamma, s_1, f_1) \models \mathtt{AXp}(or \wedge an, 1) \wedge \bigwedge_{\lambda \in Term(\{si,or,cl,an\})} \neg \mathtt{SubAXp}(\lambda, 1).$$

Therefore, it is of particular interest to determine how much knowledge is needed to verify subjective AXps. This problem can be studied in form of model checking. Let $\Gamma^{\Sigma,S,s_0} = (S, \Phi^{\Sigma,S,s_0})$ denote an MCM induced by Σ a finite subset of $\mathscr{L}^{-\Box_{\mathrm{F}}}(Atm)$, $S \subseteq 2^{Atm_0}$, $s_0 \in S$, where $\Phi^{\Sigma,S,s_0} =_{def} \{f \in Val^S : \forall \psi \in \Sigma, (S, f, s_0) \models \psi\}$. We can formalize the following model checking problem.

Subjective AXp existence
Given: finite $\Sigma \subset \mathscr{L}^{-\Box_{\mathrm{F}}}(Atm), S \subseteq 2^{Atm_0}, s_0 \in S$.
Question: Does it exist a term λ s.t. $(\Gamma^{\Sigma,S,s_0}, s_0, f) \models \mathtt{AXp}(\lambda, f(s_0))$ for all $f \in \Phi^{\Sigma,S,s_0}$?

There are many other explanation notions, and logical extensions discussed in [1, 2]. And we are working on applying this family of modal logics to more topics in XAI.

References

[1] X. Liu, E. Lorini, A logic for binary classifiers and their explanation, in: P. Baroni, C. Benzmüller, Y. N. Wáng (Eds.), Logic and Argumentation - 4th International Conference, CLAR 2021, Hangzhou, China, 2021, Proceedings, Springer, 2021, pp. 302–321.

[2] X. Liu, E. Lorini, A logic of "black box" classifier systems, in: Logic, Language, Information, and Computation: 28th International Workshop, WoLLIC 2022, Iași, Romania, 2022, Proceedings, Springer Nature, 2022, pp. 158–174.

[3] C. G. Hempel, P. Oppenheim, Studies in the logic of explanation, Philosophy of science 15 (1948) 135–175.

[4] B. Kment, Counterfactuals and explanation, Mind 115 (2006) 261–310.

[5] J. Woodward, Explanation and invariance in the special sciences, The British journal for the philosophy of science 51 (2000) 197–254.

[6] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[7] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), volume 8(1), 2017, pp. 8–13.

[8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[9] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Advances in neural information processing systems, 2018, pp. 592–603.

[10] A. Shih, A. Choi, A. Darwiche, A symbolic approach to explaining bayesian network classifiers, arXiv preprint arXiv:1805.03364 (2018).

[11] A. Ignatiev, N. Narodytska, J. Marques-Silva, Abduction-based explanations for machine learning models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 1511–1519.

[12] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[13] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).

[14] T. Miller, Contrastive explanation: A structural-model approach, The Knowledge Engineering Review 36 (2021).

[15] X. Huang, Y. Izza, A. Ignatiev, M. Cooper, N. Asher, J. Marques-Silva, Tractable explanations for d-dnnf classifiers, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5719–5728.

[16] D. Grossi, E. Lorini, F. Schwarzentruber, The ceteris paribus structure of logics of game forms, Journal of Artificial Intelligence Research 53 (2015) 91–126.

[17] A. Darwiche, A. Hirth, On the reasons behind decisions, in: ECAI 2020 - 24th European Conference on Artificial Intelligence, volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 712–720.