



# Descriptive Accuracy in Explanations: The Case of Probabilistic Classifiers

Emanuele Albini<sup>1</sup> , Antonio Rago<sup>1</sup> <sup>(✉)</sup>, Pietro Baroni<sup>2</sup> ,  
and Francesca Toni<sup>1</sup>

<sup>1</sup> Department of Computing, Imperial College London, London, UK  
{emanuele,antonio,ft}@imperial.ac.uk

<sup>2</sup> Dip.to di Ingegneria dell'Informazione, Università degli Studi di Brescia,  
Brescia, Italy  
pietro.baroni@unibs.it

**Abstract.** A user receiving an explanation for outcomes produced by an artificially intelligent system expects that it satisfies the key property of *descriptive accuracy* (DA), i.e. that the explanation contents are in correspondence with the internal working of the system. Crucial as this property appears to be, it has been somehow overlooked in the XAI literature to date. To address this problem, we consider the questions of formalising DA and of analysing its satisfaction by explanation methods. We provide formal definitions of *naive*, *structural* and *dialectical* DA, using the family of probabilistic classifiers as the context for our analysis. We evaluate the satisfaction of our given notions of DA by several explanation methods, amounting to two popular feature-attribution methods from the literature and a novel form of explanation that we propose and complement our analysis with experiments carried out on a varied selection of concrete probabilistic classifiers.

**Keywords:** Explainable AI · Probabilistic classifiers · Descriptive accuracy

## 1 Introduction

Equipping automated decision systems with explanation capabilities is a compelling need in order to achieve user *trust*, a prerequisite for acceptance, especially if the systems are based on some model or technique that is not directly understandable by the users. This need lies, in particular, at the basis of the rapid growth of the research field of explainable AI (XAI) in recent years. As pointed out in [10], trust, which is an attitude of the trustors (in our case, the systems' users), is distinguished from trustworthiness, which is a property of the trustees (in our case the explained systems). This makes the goal of achieving trust, and the role of explanations there for, a rather tricky issue. On the one hand, there can be situations where trust is achieved by explanations which are convincing but somehow deceptive. On the other hand, there can be situations

where an otherwise trustworthy system loses users’ trust due to problems in its explanations’ capabilities.

These considerations call for the need of identifying some basic requirements that explanations should satisfy in order to lead to (deservingly) trustworthy AI systems. In this paper, for the specific setting of *probabilistic classifiers*, we focus on the property of *descriptive accuracy* (DA) described in [19], for machine learning in general, as “the degree to which an interpretation method objectively captures the relationships learned by machine learning models”. DA appears to be a crucial requirement for any explanation: its absence would lead to the risk of misleading (if not deceptive) indications for the user.

In this work we consider the issue of providing formal counterparts to the general notion of DA and then of assessing their satisfaction by both existing and suitably defined explanation approaches. Specifically, we make the following contributions.

- We introduce three formal notions of DA (Sect. 4): *naïve* DA, as a precursor to *dialectical* DA, both applicable to any probabilistic classifier, and *structural* DA, applicable to probabilistic classifiers that are equipped with a *structural description*, as is the case for *Bayesian network classifiers* (BCs) and *Chain Classifiers* (CCs).
- We study whether concrete explanation methods satisfy our notions of DA (Sect. 5). We focus our analysis on existing *feature attribution methods* from the literature (LIME [24] and SHAP [15]) as well as a novel method we define.
- We evaluate our forms of DA empirically (when they cannot be guaranteed formally) on a variety of BCs and CCs (Sect. 6) showing that they are often violated in practice by LIME and SHAP.

## 2 Related Work

Numerous methods for providing explanations have been proposed (e.g. see the survey by [8]) and their desirable properties have been considered from a variety of perspectives (e.g. see the survey by [26]). We draw inspiration from [19] and focus, in particular, on their notion of *descriptive accuracy* (DA) for (model-based or post-hoc) interpretable machine learning. As mentioned in the introduction, DA concerns the degree to which an interpretation (in our setting, explanation) method objectively captures the relationships learned by the machine-learned models.

DA is seen, in [19], as a crucial property for achieving interpretable machine learning, alongside, in particular, *predictive accuracy*, wrt (test) data, of the predictions produced by the interpretations/explanations. Whereas DA is concerned with the inner workings of models, predictive accuracy is concerned with the input-output behaviour thereof. Predictive accuracy is thus closely linked with properties of *fidelity* or *faithfulness* which have been considered by several works (see e.g. [8, 13]). In the case of explanations concerning a single instance, *local fidelity* has been defined as a measure of how well an explanation model

approximates the original model in a neighbourhood of the instance in need of explaining [1, 24].

Overall, whereas formal counterparts of predictive accuracy/fairness/fidelity have been extensively studied in the XAI literature, to the best of our knowledge, formal counterparts of DA appear to be lacking up to now. This gap is particularly significant for the classes of post-hoc explanations methods which, *per se*, have no relations with the underlying operation of the explained model and therefore cannot rely on any implicit assumption that DA is guaranteed, in a sense, by construction. This applies, in particular, to the family of model-agnostic local explanation methods, namely methods which are designed to be applicable to any model (and hence need to treat the model itself purely as a black-box) and whose explanations are restricted to illustrate individually a single outcome of the model without aiming to describe its behaviour in more general terms. This family includes the well-known class of *additive feature attribution* methods, such as LIME [24] and SHAP [15], where the explanation for the outcome of a model basically consists in ascribing to each input feature a numerical weight. We will provide three *formal* characterisations for DA, allowing evaluation of explanation methods for satisfaction of DA in precise terms and we will study our notions of DA in the context of both LIME and SHAP, showing that they are not able to satisfy them.

### 3 Preliminaries

As DA is inherently related to the internal operation of a model, rather than just to its input/output behaviour, any formal notion of DA cannot be completely model-agnostic. It follows that an investigation of DA needs to find a balance between the obvious need for wide applicability and the potential advantages of model-tailored definitions. For this reason we will focus on the broad family of *probabilistic classifiers*.

We consider (discrete) probabilistic classifiers with *feature variables*  $\mathbf{X} = \{X_1, \dots, X_m\}$  ( $m > 1$ ) and *class variables*  $\mathbf{C} = \{C_1, \dots, C_n\}$  ( $n \geq 1$ ). Each (random) variable  $V_i \in \mathbf{V} = \mathbf{X} \cup \mathbf{C}$  is equipped with a discrete set of possible *values*  $\Omega_{V_i}$ : we define the *feature space* as  $\mathcal{X} = \Omega_{X_1} \times \dots \times \Omega_{X_m}$  and the *class space* as  $\mathcal{C} = \Omega_{C_1} \times \dots \times \Omega_{C_n}$ . From now on, we call any vector  $\mathbf{x} \in \mathcal{X}$  an *input* and denote as  $\mathbf{x}(X_i)$  the value of feature  $X_i$  in  $\mathbf{x}$ . Given input  $\mathbf{x}$ , a *probabilistic classifier*  $\mathcal{PC}$  computes, for each class variable  $C_i$  and value  $\omega \in \Omega_{C_i}$ , the probability  $P(C_i = \omega | \mathbf{x})$  that  $C_i$  takes value  $\omega$ , given  $\mathbf{x}$ . We then refer to the *resulting value* for a class variable  $C_i \in \mathbf{C}$  given input  $\mathbf{x}$  as  $\mathcal{PC}(C_i | \mathbf{x}) = \argmax_{\omega \in \Omega_{C_i}} P(C_i = \omega | \mathbf{x})$ . Table 1 gives a probabilistic classifier for a (toy) financial setting where the values of class variables *problematic external event* and *drop in consumer confidence* are determined based on the feature variables *company share price trend*, *devaluation of currency*, *healthy housing market* and *negative breaking news cycle*. Here, for any variable  $V_i \in \mathbf{V}$ ,  $\Omega_{V_i} = \{+, -\}$ .

For  $X_i \in \mathbf{X}$ , we will abuse notation as follows, to simplify some of the formal definitions later in the paper:  $\mathcal{PC}(X_i | \mathbf{x}) = \mathbf{x}(X_i)$  (basically, the “resulting value”

**Table 1.** An example of *probabilistic classifier* with  $\mathbf{X} = \{s, d, h, n\}$  and  $\mathbf{C} = \{c, e\}$ . Here, e.g. for  $\mathbf{x}$  (highlighted in bold) such that  $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$ ,  $\mathcal{PC}(c|\mathbf{x}) = +$  (as  $P(c = +|\mathbf{x}) = .60$ ), and  $\mathcal{PC}(e|\mathbf{x}) = +$  (as  $P(e = +|\mathbf{x}) = .60$ )).

|          |     |     |     |     |     |   |     |     |     |     |   |     |     |     |   |     |
|----------|-----|-----|-----|-----|-----|---|-----|-----|-----|-----|---|-----|-----|-----|---|-----|
| <i>s</i> | +   | +   | +   | +   | +   | + | +   | +   | -   | -   | - | -   | -   | -   | - | -   |
| <i>d</i> | +   | +   | +   | +   | -   | - | -   | -   | +   | +   | + | +   | -   | -   | - | -   |
| <i>h</i> | +   | +   | -   | -   | +   | + | -   | -   | +   | +   | - | -   | +   | +   | - | -   |
| <i>n</i> | +   | -   | +   | -   | +   | - | +   | -   | +   | -   | + | -   | +   | -   | + | -   |
| <i>c</i> | +   | -   | +   | +   | +   | - | +   | -   | +   | -   | + | +   | +   | +   | - | +   |
| <i>P</i> | .60 | .65 | 1   | .60 | .60 | 1 | 1   | .65 | .60 | .65 | 1 | .60 | .60 | 1   | 1 | .65 |
| <i>e</i> | +   | -   | +   | +   | +   | - | +   | -   | +   | -   | + | +   | +   | +   | - | +   |
| <i>P</i> | .60 | 1   | .60 | .60 | .60 | 1 | .60 | 1   | 1   | .65 | 1 | 1   | 1   | .65 | 1 | .65 |

for a feature variable, given an input, is the value assigned to that variable in the input) and  $P(X_i = \mathbf{x}(X_i)) = 1$  (basically, the probability of a feature variable being assigned its value, in the given input, is 1). We will also use notation:

$$P(V = v|\mathbf{x}, \text{set}(V_i = v_i)) = \begin{cases} P(V = v|\mathbf{x}'), & \text{if } V_i \in \mathbf{X}, \\ P(V = v|\mathbf{x}, V_i = v_i), & \text{if } V_i \in \mathbf{C}, \end{cases}$$

where, in the first case,  $\mathbf{x}'(V_i) = v_i$  and  $\mathbf{x}'(V_j) = \mathbf{x}(V_j)$  for all  $V_j \in \mathbf{X} \setminus \{V_i\}$ . Basically, this notation allows to gauge the effects of changes in value for (input or class) variables on the probabilities computed by the classifiers (for assignments of values to any variables).

Various types of probabilistic classifiers exist. In Sect. 6 we will experiment with (explanations for) a variety of (discrete) *Bayesian Classifiers* (BCs, see [2] for an overview), where the variables in  $\mathbf{V}$  constitute the nodes in a Bayesian network, i.e. a directed acyclic graph whose edges indicate probabilistic dependencies amongst the variables. We will also experiment with (explanations for) *chained probabilistic classifiers* (CCs, e.g. as defined by [23] for the case of BCs). These CCs result from the combination of simpler probabilistic classifiers (possibly, but not necessarily, BCs), using an ordering  $>_C$  over  $\mathbf{C}$  such that the value of any  $C_i \in \mathbf{C}$  is treated as a feature value for determining the value of any  $C_j \in \mathbf{C}$  with  $C_j >_C C_i$ , and thus a classifier computing values for  $C_i$  can be chained with one for computing values for  $C_j$ . For illustration, in Table 2 we re-interpret the classifier from Table 1 as a CC amounting to a chain of two classifiers, using  $e >_C c$ : the classifier (a) determines the value of  $c$  as an additional input for the classifier (b). Then, the overall classifier determines the value of  $c$  first based on the feature variables  $d, h$  and  $n$ , and then  $e$  based on  $s$  and  $c$  (treated as a feature variable in the chaining). Note that, in Table 2 and throughout the paper, we abuse notation and use inputs for overall (chained) classifiers ( $\mathbf{x}$  in the caption of the table) as inputs of all simpler classifiers forming them (rather than the inputs' restriction to the specific input variables of the simpler classifiers).

For some families of probabilistic classifiers (e.g. for BCs) it is possible to provide a graphical representation which gives a synthetic view of the depen-

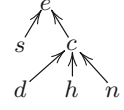
**Table 2.** An example of *chained probabilistic classifier* (CC) with (a) the first probabilistic classifier  $\mathcal{PC}_1$  with  $\mathbf{X}_1 = \{d, h, n\}$ ,  $\mathbf{C}_1 = \{c\}$ , and (b) the second probabilistic classifier  $\mathcal{PC}_2$  with  $\mathbf{X}_2 = \{s, c\}$ ,  $\mathbf{C}_2 = \{e\}$  (both inputs highlighted in bold). Here, e.g. for  $\mathbf{x}$  as in the caption of Table 1,  $\mathcal{PC}(c|\mathbf{x}) = \mathcal{PC}_1(c|\mathbf{x}) = +$  and  $\mathcal{PC}(e|\mathbf{x}) = \mathcal{PC}_2(e|\mathbf{x}, \text{set}(c = \mathcal{PC}_1(c|\mathbf{x}))) = +$ . (c) A *structural description* for the CC in (a–b), shown as a graph.

|     |     |     |   |     |     |   |     |
|-----|-----|-----|---|-----|-----|---|-----|
| $d$ | +   | +   | + | +   | -   | - | -   |
| $h$ | +   | +   | - | -   | +   | + | -   |
| $n$ | +   | -   | + | -   | +   | - | +   |
| $c$ | +   | -   | + | +   | +   | + | -   |
| $P$ | .60 | .65 | 1 | .60 | .60 | 1 | .65 |

(a)

|     |     |   |   |     |
|-----|-----|---|---|-----|
| $s$ | +   | + | - | -   |
| $c$ | +   | - | + | -   |
| $e$ | +   | - | + | -   |
| $P$ | .60 | 1 | 1 | .65 |

(b)



(c)

dence and independence relations between the variables. In these cases, we will assume that the classifier is accompanied by a *structural description*, namely a set  $\mathcal{SD} \subseteq \mathbf{V} \times \mathbf{V}$ . The structural description identifies for each variable  $V_j \in \mathbf{V}$  a (possibly empty) set of *parents*  $\mathcal{PA}(V_j) = \{V_i \mid (V_i, V_j) \in \mathcal{SD}\}$  with the meaning that the evaluation of  $V_j$  is completely determined by the evaluations of  $\mathcal{PA}(V_j)$  in the classifier. In the case of BCs, the parents of each (class) variable correspond to the variables in its unique *Markov boundary* [20, 21]  $\mathcal{M} : \mathbf{V} \rightarrow 2^{\mathbf{V}}$ , where, for any  $V_i \in \mathbf{V}$ ,  $\mathcal{M}(V_i)$  is the  $\subseteq$ -minimal set of variables such that  $V_i$  is conditionally independent of all the other variables  $(\mathbf{V} \setminus \mathcal{M}(V_i))$ , given  $\mathcal{M}(V_i)$ . In the case of CCs, even when no information is available about the internal structure of the individual classifiers being chained, a structural description may be extracted to reflect the connections between features and classes. For illustration, for the CC in Table 2(a–b), the structural description is  $\mathcal{SD} = \{(d, c), (h, c), (n, c), (s, e), (c, e)\}$ , given in Table 2(c) as a graph.

We remark that similar notions have been considered earlier in the literature. For instance in [27] a notion of support graph derived from a Bayesian network has been considered. This support graph however is built with reference to a given variable of interest and is meant to support the construction of arguments which provide a sort of representation of the reasoning inside the network. In our case we provide a structural description which does not refer to a single variable of interest and is not used for building explanations but rather to verify whether they satisfy structural DA, as it will be described later. A deeper analysis of the possible connections between our proposal and argumentation-based approaches for explaining Bayesian networks is an interesting subject for future work.

In the remainder, unless specified otherwise, we assume as given a probabilistic classifier  $\mathcal{PC}$  with feature variables  $\mathbf{X}$  and class variables  $\mathbf{C}$ , without making any assumptions.

## 4 Formalising Descriptive Accuracy

We aim to define DA formally, in such a way that DA, independently of any specific explanation method (but with a focus on the broad class of local expla-

nations, and specifically feature attribution methods to obtain them). To do so we will consider different abstractions of explanation, with the capability to encompass a broad range of existing notions in the literature as instances. The abstractions are based on the combinations of alternative choices along two dimensions. First, we consider two basic elements that an explanation may refer to: (1) *input features*; (2) pairs of variables representing *relations* between variables. Second, we assume that the basic elements inside an explanation can be: (a) regarded as an undifferentiated set (we call these elements *unsigned*, in contrast with (b)); (b) partitioned into two sets according to their *positive or negative* role in the explanation. The combinations (1)-(a) and (2)-(a) will correspond respectively to the abstract notions of *unipolar* and *relational unipolar* explanations while the combinations (1)-(b) and (2)-(b) will correspond respectively to the notions of *bipolar* and *relational bipolar* explanations. We will introduce a notion of *naive* DA for all the kinds of abstract explanations we consider and a notion of *dialectical* DA tailored to the two cases of relational explanations. We see naive DA as a very weak pre-requisite for explanations (it can be regarded as related to the basic notion of relevance), and prove that it is implied by dialectical DA for both bipolar and relational bipolar explanations (Propositions 1 and 2, resp.): thus, naive DA can be seen as a step towards defining dialectical DA. (Naive and) Dialectical DA are applicable to *any* probabilistic classifiers. In the specific setting of classifiers with underlying graph structures, such as BCs and CCs, we will also define a notion of *structural* DA for relational unipolar/bipolar explanations.

#### 4.1 Unipolar Explanations and Naive DA

We begin with a very general notion of *unipolar explanation*: we only assume that, whatever the nature and structure of the explanation, it can be regarded at an abstract level as a *set of features*:

**Definition 1.** *Given an input  $\mathbf{x} \in \mathcal{X}$  and the resulting value  $\omega = \mathcal{PC}(C|\mathbf{x})$  for class  $C \in \mathbf{C}$  given  $\mathbf{x}$ , a unipolar explanation (for  $C = \omega$ , given  $\mathbf{x}$ ) is a triple  $\langle \mathbf{F}, C, \mathbf{x} \rangle$  where  $\mathbf{F} \subseteq \mathbf{X}$ .*

Intuitively, the features in a unipolar explanation are those deemed “relevant” for explaining the resulting value assigned by the classifier to a class variable, for the input under consideration. It is easy to see that it is straightforward to derive unipolar explanations from the outcomes produced by existing explanation methods when they return features accompanied by additional information (e.g. feature importance as in the case of the attribution methods LIME and SHAP): basically, in these settings the unipolar explanations disregard the additional information, and amount to (a subset of) the set of features alone (e.g. the  $k$  most important features).

The simplest form of DA, i.e. *naive DA*, matches the intuition that the inclusion of features in a unipolar explanation should play a role in the underlying model, and is formally defined as follows:

*Property 1.* A unipolar explanation  $\langle \mathbf{F}, C, \mathbf{x} \rangle$  satisfies *naive descriptive accuracy* iff for every  $X_i \in \mathbf{F}$  there exists an input  $\mathbf{x}' \in \mathcal{X}$  with  $\mathbf{x}'(X_j) = \mathbf{x}(X_j)$  for every  $X_j \neq X_i$  and with  $\mathbf{x}'(X_i) \neq \mathbf{x}(X_i)$ , such that, letting  $\omega = \mathcal{PC}(C|\mathbf{x})$ , it holds that  $P(C = \omega|\mathbf{x}) \neq P(C = \omega|\mathbf{x}')$ .

Naive DA holds when, for each feature, there is at least one case (i.e. an alternative input  $\mathbf{x}'$  to the input  $\mathbf{x}$  being explained) where a change in the value of the feature has an effect on the probability of the value of the class variable. It is a rather weak requirement as it excludes individually “irrelevant” features from playing a role in the explanation.

For illustration, given the probabilistic classifier in Table 1 and  $\mathbf{x}$  as in the table’s caption, the unipolar explanation  $\langle \{s, d, h, n\}, c, \mathbf{x} \rangle$  does not satisfy naive DA, given that both  $s$  and  $d$  are “irrelevant” here: changing the value of either does not affect the probability of  $c$ . Meanwhile, it is easy to see that  $\langle \{h, n\}, c, \mathbf{x} \rangle$  satisfies naive DA.

## 4.2 Bipolar Explanations and Dialectical DA

Unipolar explanations consist of “minimal” information, i.e. just the features playing a role in explanations. At a finer level of granularity, we consider *bipolar explanations*, where the features are partitioned into two sets: those having a positive effect on the resulting value and those having a negative effect. The notions of positive and negative effect may admit different specific interpretations in different contexts, the general underlying intuition being that the corresponding features provide, resp., reasons for and against the resulting value being explained. Whatever the interpretation, we assume that positive and negative features are disjoint, as, in an explanation, a feature with a twofold role would be confusing for the user.

**Definition 2.** Given an input  $\mathbf{x} \in \mathcal{X}$  and the resulting value  $\omega = \mathcal{PC}(C|\mathbf{x})$  for class  $C \in \mathbf{C}$  given  $\mathbf{x}$ , a bipolar explanation (for  $C = \omega$ , given  $\mathbf{x}$ ) is a quadruple  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  where  $\mathbf{F}_+ \subseteq \mathbf{X}$ ,  $\mathbf{F}_- \subseteq \mathbf{X}$ , and  $\mathbf{F}_+ \cap \mathbf{F}_- = \emptyset$ ; we refer to features in  $\mathbf{F}_+$  and  $\mathbf{F}_-$  resp. as positive and negative reasons.

It is easy to see that existing explanation methods can be regarded as producing bipolar explanations when those methods return features accompanied by additional positive or negative information: in these settings, as in the case of unipolar explanations, bipolar explanations disregard the additional information, and amount to (a subset of) the set of features with their polarity (e.g. the  $k$  features with the highest positive importance as positive features and the  $k$  features with the lowest negative importance as negative features).

Taking into account the distinction between positive and negative reasons, we introduce a property requiring that the role assigned to features is justified:

*Property 2.* A bipolar explanation  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  satisfies *dialectical descriptive accuracy* iff for every  $X_i \in \mathbf{F}_+ \cup \mathbf{F}_-$ , for every  $\mathbf{x}' \in \mathcal{X}$  with  $\mathbf{x}'(X_j) = \mathbf{x}(X_j)$  for all  $X_j \neq X_i$  and  $\mathbf{x}'(X_i) \neq \mathbf{x}(X_i)$ , letting  $\omega = \mathcal{PC}(C|\mathbf{x})$ , it holds that

- if  $X_i \in \mathbf{F}_+$  then  $P(C = \omega|\mathbf{x}) > P(C = \omega|\mathbf{x}')$ ;
- if  $X_i \in \mathbf{F}_-$  then  $P(C = \omega|\mathbf{x}) < P(C = \omega|\mathbf{x}')$ .

In words, if a feature is identified as a positive (negative) reason for the resulting value for a class variable, given the input, the feature variable's value leads to increasing (decreasing, resp.) the posterior probability of the class variable's resulting value (with all other feature values unchanged).

For illustration, in the running example with  $\mathcal{PC}$  in Table 1, the bipolar explanation  $\langle \{d, n\}, \{h\}, c, \mathbf{x} \rangle$ , given input  $\mathbf{x}$  as in the table's caption does not satisfy dialectical DA. Indeed,  $d$  is a positive reason in the explanation but, for  $\mathbf{x}'$  agreeing with  $\mathbf{x}$  on all features other than  $d$  (with  $\mathbf{x}'(d) = -$ ), we obtain  $P(c = +|\mathbf{x}) = .60 \not< P(c = +|\mathbf{x}') = .60$ . Instead, it is easy to see that the bipolar explanation  $\langle \{n\}, \{h\}, c, \mathbf{x} \rangle$ , satisfies dialectical DA.

In general, unipolar explanations can be directly obtained from bipolar explanations by ignoring the distinction between positive and negative reasons, and the property of naive DA can be lifted:

**Definition 3.** A bipolar explanation  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  satisfies naive descriptive accuracy iff the unipolar explanation  $\langle \mathbf{F}_+ \cup \mathbf{F}_-, C, \mathbf{x} \rangle$  satisfies naive descriptive accuracy.

It is then easy to see that dialectical DA strengthens naive DA:<sup>1</sup>

**Proposition 1.** If a bipolar explanation  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  satisfies dialectical DA then it satisfies naive DA.

### 4.3 Relational Unipolar Explanations and Naive DA

Moving towards a richer explanation notion, we pursue the idea of providing a detailed view of the relations between variables of a probabilistic classifier, reflecting influences possibly occurring amongst the variables. To this purpose we first introduce *relational unipolar explanations* as follows.

**Definition 4.** Given  $\mathbf{x} \in \mathcal{X}$  and the resulting value  $\omega = \mathcal{PC}(C|\mathbf{x})$  for  $C \in \mathbf{C}$  given  $\mathbf{x}$ , a relational unipolar explanation (for  $C = \omega$ , given  $\mathbf{x}$ ) is a triple  $\langle \mathcal{R}, C, \mathbf{x} \rangle$  where  $\mathcal{R} \subseteq \mathbf{V} \times \mathbf{V}$ .

In words, a relational unipolar explanation includes a set  $\mathcal{R}$  of pairs of variables (i.e. a relation between variables) where  $(V_i, V_j) \in \mathcal{R}$  indicates that the value of  $V_i$  has a role in determining the value of  $V_j$ , given the input.

For illustration, for  $\mathcal{PC}$  in Table 1,  $\langle \{(s, e), (c, e)\}, e, \mathbf{x} \rangle$  may be a relational unipolar explanation for  $\mathbf{x}$  in the table's caption, indicating that  $s$  and  $c$  both influence (the value of)  $e$ . Note that relational unipolar explanations admit unipolar explanations as special instances: given a unipolar explanation  $\langle \mathbf{F}, C, \mathbf{x} \rangle$ , it is straightforward to see that  $\langle \mathbf{F} \times \{C\}, C, \mathbf{x} \rangle$  is a relational unipolar explanation. However, as demonstrated in the illustration, relational unipolar explanations may include relations besides those between feature and class variables found in unipolar explanations.

The notion of naive DA can be naturally extended to relational unipolar explanations:

<sup>1</sup> All proofs are omitted due to space limitations.

*Property 3.* A relational unipolar explanation  $\langle \mathcal{R}, C, \mathbf{x} \rangle$  satisfies *naive descriptive accuracy* iff for every  $(V_i, V_j) \in \mathcal{R}$ , letting  $v_i = \mathcal{PC}(V_i | \mathbf{x})$  and  $v_j = \mathcal{PC}(V_j | \mathbf{x})$ , there exists  $v'_i \in \Omega_{V_i}$ ,  $v'_i \neq v_i$ , such that  $P(V_j = v_j | \mathbf{x}) \neq P(V_j = v_j | \mathbf{x}, \text{set}(V_i = v'_i))$ .

For illustration, for  $\mathcal{PC}$  in Table 1,  $\langle \{(s, e), (n, e)\}, e, \mathbf{x} \rangle$  satisfies naive DA for  $\mathbf{x}$  in the table's caption, but  $\langle \{(s, e), (d, e)\}, e, \mathbf{x} \rangle$  does not, as changing the value of  $d$  to  $-$  (the only alternative value to  $+$ ), the probability of  $e = +$  remains unchanged.

It is easy to see that, for relational unipolar explanations  $\langle \mathbf{F} \times \{C\}, C, \mathbf{x} \rangle$ , corresponding to unipolar explanations  $\langle \mathbf{F}, C, \mathbf{x} \rangle$ , Property 3 implies Property 1.

#### 4.4 Relational Bipolar Explanations and Dialectical DA

Bipolar explanations and dialectical DA can also be naturally extended to accommodate relations, as follows:

**Definition 5.** Given an input  $\mathbf{x} \in \mathcal{X}$  and the resulting value  $\omega = \mathcal{PC}(C | \mathbf{x})$  for class  $C \in \mathbf{C}$  given  $\mathbf{x}$ , a relational bipolar explanation (RX) is a quadruple  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  where:

$\mathcal{R}_+ \subseteq \mathbf{V} \times \mathbf{V}$ , referred to as the set of positive reasons;

$\mathcal{R}_- \subseteq \mathbf{V} \times \mathbf{V}$ , referred to as the set of negative reasons;

$\mathcal{R}_+ \cap \mathcal{R}_- = \emptyset$ .

*Property 4.* An RX  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  satisfies *dialectical descriptive accuracy* iff for every  $(V_i, V_j) \in \mathcal{R}_+ \cup \mathcal{R}_-$ , letting  $v_i = \mathcal{PC}(V_i | \mathbf{x})$ ,  $v_j = \mathcal{PC}(V_j | \mathbf{x})$ , it holds that, for every  $v'_i \in \Omega_{V_i} \setminus \{v_i\}$ :

if  $(V_i, V_j) \in \mathcal{R}_+$  then  $P(V_j = v_j | \mathbf{x}) > P(V_j = v_j | \mathbf{x}, \text{set}(V_i = v'_i))$ ;

if  $(V_i, V_j) \in \mathcal{R}_-$  then  $P(V_j = v_j | \mathbf{x}) < P(V_j = v_j | \mathbf{x}, \text{set}(V_i = v'_i))$ .

An RX can be seen as a graph of variables connected by edges identifying positive or negative reasons. Examples of RXs for the running example are shown as graphs in Fig. 1 (where the nodes also indicate the values ascribed to the feature variables in the input  $\mathbf{x}$  and to the class variables by any of the toy classifiers in Tables 1 and 2). Here, (iii) satisfies dialectical DA, since setting to  $-$  the value of any variable with a positive (negative) reason to another variable will reduce (increase, resp.) the probability of the latter's value being  $+$ , whereas (ii) does not, since setting  $d$  to  $-$  does not affect the probability of  $c$ 's value being  $+$  and (i) does not since setting  $d$  to  $-$  does not affect the probability of  $e$ 's value being  $+$ .

Similarly to the case of unipolar/bipolar explanations, relational unipolar explanations can be obtained from RXs by ignoring the distinction between positive and negative reasons, and the property of dialectical DA can be lifted:

**Definition 6.** An RX  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  satisfies naive DA iff the relational unipolar explanation  $\langle \mathcal{R}_+ \cup \mathcal{R}_-, C, \mathbf{x} \rangle$  satisfies naive DA.

It is then easy to see that dialectical DA strengthens naive DA:

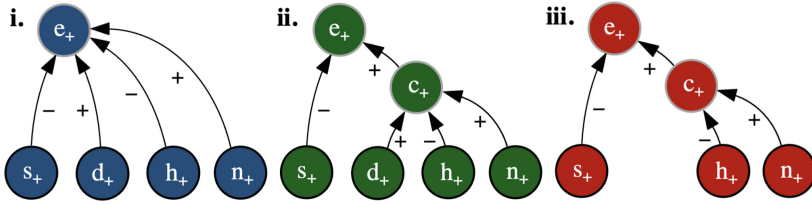


**Fig. 1.** Example RXs (shown as graphs, with positive and negative reasons given by edges labelled  $+$  and  $-$ , resp.) with input  $\mathbf{x}$  such that  $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$  (represented as  $s_+$ ,  $d_+$ ,  $h_+$ ,  $n_+$ ) and (resulting) class values  $c = +$  (represented as  $c_+$ ) and  $e = +$  (represented as  $e_+$ ).

**Proposition 2.** *If an RX  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  satisfies dialectical DA then it satisfies naive DA.*

Note that bipolar explanations  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  can be regarded as special cases of RXs, i.e.  $\langle \{(X, C) | X \in \mathbf{F}_+\}, \{(X, C) | X \in \mathbf{F}_-\}, C, \mathbf{x} \rangle$  (indeed, the RX in Fig. 1(i) is a bipolar explanation). Thus, from now on we will often refer to all forms of bipolar explanation as RXs.

#### 4.5 Relational Bipolar Explanations and Structural DA

When a classifier is equipped with a structural description, one can require that the relations used for explanation purposes in RXs are subsets of those specified by the structural description, so that the RXs correspond directly to (parts of) the inner working of the model. This leads to the following additional form of DA:

*Property 5.* Given a probabilistic classifier  $\mathcal{PC}$  with structural description  $\mathcal{SD}$ :

- a relational unipolar explanation  $\langle \mathcal{R}, C, \mathbf{x} \rangle$  satisfies *structural descriptive accuracy* iff  $\mathcal{R} \subseteq \mathcal{SD}$ ; and
- an RX  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  satisfies *structural descriptive accuracy* iff  $\mathcal{R}_+ \cup \mathcal{R}_- \subseteq \mathcal{SD}$ .

For instance, suppose that  $\mathcal{SD}$  is the structural description in Table 2(c). Then, the RXs in Fig. 1(ii-iii) satisfy structural DA, while the RX in Fig. 1(i) does not, since the relations from  $d$ ,  $h$  and  $n$  to  $e$  are not present in the structural description.

## 5 Achieving DA in Practice

Here, we study the satisfaction of the proposed properties by explanation methods. We focus in particular on two existing methods in the literature (LIME [24] and SHAP [15]). After showing that these methods do not satisfy the properties introduced in Sect. 4, we introduce a novel form of explanation guaranteed to satisfy them. Thus, this novel form of explanation can be seen as a “champion” for our proposed forms of DA, showing that they can be satisfied in practice.

We start with LIME and SHAP. The explanations they produce (given an input  $\mathbf{x}$  and a classifier, computing  $C=\omega$ , given  $\mathbf{x}$ ) basically consist in computing, for each feature  $X_i \in \mathbf{X}$ , a real number  $w(\mathbf{x}, X_i, C)$  indicating the importance of  $X_i$ , which has assigned value  $\mathbf{x}(X_i)$  in the given input  $\mathbf{x}$ , towards the probability of the class variable  $C$  being assigned value  $\omega = \mathcal{PC}(C|\mathbf{x})$  by the classifier, in the context of  $\mathbf{x}$ .<sup>2</sup> The absolute value of this number can be interpreted as a measure of the feature importance in the explanation, while its sign, in the context of explaining probabilistic classifiers, indicates whether the feature has a positive or negative role wrt the classifier’s resulting value for the explained instance. Features which are assigned a value of zero can be regarded as irrelevant. Clearly, such explanations correspond to bipolar explanations  $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$  as in Definition 2, with

- $\mathbf{F}_+ = \{X_i \in \mathbf{X} \mid w(\mathbf{x}, X_i, C) > 0\}$  and
- $\mathbf{F}_- = \{X_i \in \mathbf{X} \mid w(\mathbf{x}, X_i, C) < 0\}$ .

In the remainder, with an abuse of terminology, we call these bipolar explanations *LIME/SHAP explanations*, depending on whether  $w$  is calculated using, resp., the method of LIME/SHAP. For illustration, consider the classifier in Table 1 and  $\mathbf{x}$  such that  $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$ , as in the caption of Fig. 1, for which the classifier computes  $e = +$ . In this simple setting, SHAP computes  $w(\mathbf{x}, s_+, e_+) = -0.20$ ,  $w(\mathbf{x}, d_+, e_+) = 0.03$ ,  $w(\mathbf{x}, h_+, e_+) = -0.05$ , and  $w(\mathbf{x}, n_+, e_+) = 0.25$  (adopting here the same conventions on variable assignments as in the caption of the figure). This results in the SHAP explanation in Fig. 1(i). Thus features  $d$  and  $n$  (with their current values) are ascribed positive roles and  $s$  and  $h$  are ascribed negative roles in determining the outcome  $\mathcal{PC}(e|\mathbf{x}) = +$ . However, as stated earlier, for feature  $d$  this is in contrast with the property of naive DA. In fact, by inspection of Table 1, it can be noted that in changing the value of this variable individually we would still have  $P(e = +|\mathbf{x}) = 1$ . To put it in intuitive terms, assigning a positive importance to this variable suggests to the user that its current value (namely  $+$ ) has a role in determining the outcome  $e = +$  (though minor), which is misleading. The following proposition generalizes these considerations:

**Proposition 3.** *In general, LIME and SHAP explanations are not guaranteed to satisfy naive nor dialectical DA.*

The illustration above proves this result for SHAP explanations, by providing a counterexample to naive (and hence dialectical) DA in the context of the classifier in Table 1. As to LIME, we built counterexamples by introducing spurious features within trained probabilistic classifiers and showing that they are assigned non-null importance.

Concerning structural DA, LIME and SHAP explanations may in general satisfy it only if  $\mathbf{X} \times \mathbf{C} \subseteq \mathcal{SD}$ , i.e. if the structural description includes all the

<sup>2</sup> We omit the formal definitions of how these well-known methods compute  $w$ , as not essential for the considerations on property satisfaction drawn in this paper.

possible relations from feature variables to class variables. This is for instance the case for naive BCs [16], but not for more general BCs or CCs.

To guarantee the missing properties too, we define novel *dialectically accurate relational explanations* (*DARXs*):

**Definition 7.** *Given a probabilistic classifier with structural description  $\mathcal{SD}$ , a dialectically accurate relational explanation (DARX) is a relational bipolar explanation  $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$  where, letting  $v_x = \mathcal{PC}(V_x | \mathbf{x})$  for any  $V_x \in \mathbf{V}$ :*

- $\mathcal{R}_+ = \{(V_i, V_j) \in \mathcal{SD} | \forall v'_i \in \Omega_{V_i} \setminus \{v_i\} \text{ it holds that } P(V_j = v_j | \mathbf{x}) > P(V_j = v_j | \mathbf{x}, \text{set}(V_i = v'_i))\};$
- $\mathcal{R}_- = \{(V_i, V_j) \in \mathcal{SD} | \forall v'_i \in \Omega_{V_i} \setminus \{v_i\} \text{ it holds that } P(V_j = v_j | \mathbf{x}) < P(V_j = v_j | \mathbf{x}, \text{set}(V_i = v'_i))\}.$

**Proposition 4.** *DARXs are guaranteed to satisfy naive, structural and dialectical DA.*

For illustration, suppose  $\mathcal{SD}$  corresponds exactly to the links in Fig. 1(iii). Then, this figure shows the DARX for  $e$  given the input in the figure’s caption and the classifier in Table 1 (or Table 2). Here, the satisfaction of naive DA ensures that no spurious reasons, i.e. where the corresponding variables do not, in fact, influence one another, are included in the DARX. Note that, when explaining  $e$  with the same input, SHAP may draw a positive reason from  $d$  to  $e$  (as in Fig. 1(i)) when, according to  $\mathcal{SD}$ ,  $d$  does not directly affect  $e$ . Further, the satisfaction of dialectical DA means that each of the reasons in the DARX in Fig. 1(iii) is guaranteed to have the desired dialectical effect. Note that DARXs are local explanations, meaning that they are meant to explain the behaviour of the classifier given a specific input, not the behaviour of the classifier in general. In other words, they assign a positive or negative role to variables with reference to the specific input considered and it may of course be the case that, given a different input, the same variable has a different role.

## 6 Empirical Evaluation

As mentioned in Sect. 3, we experiment with (chains of) BCs as well as chains (in the form of trees) of tree-based classifiers (referred to as *C-DTs* below). As far as BCs are concerned, we experiment with different types, corresponding to different restrictions on the structure of the underlying Bayesian network and conditional dependencies: naive BCs (*NBC*) [16]; tree-augmented naive BCs (*TAN*) [7]; and *chains* of BCs [29], specifically in the form of chains of NBCs (*CNBC*) and of the unrestricted BCs suggested in [22] (*CUBC*). We choose C-DTs and (chains of) BCs as they are endowed with a natural structural description, allowing us to evaluate structural DA, while they are popular methods with tabular data, e.g. in the case of BCs, for medical diagnosis [14, 17].<sup>3</sup>

<sup>3</sup> We use several datasets or (pre-computed) Bayesian networks, and deploy the best-performing type of the chosen type of classifier for each. As structural descriptions, we use those described in Sect. 3. When training classifiers on datasets, we split them into train/test sets (with 75/25% ratio) and optimise the hyper-parameters.

**Table 3.** Average percentages of reasons (over 100 samples) *violating* DA (i.e.  $|\{(V_i, V_j) \in \mathcal{R}_- \cup \mathcal{R}_+ \text{ such that } (V_i, V_j) \text{ violates DA}\}|/|\mathcal{R}_- \cup \mathcal{R}_+|$ ) for several instantiated RXs. (\*) NBC (Naive BC), TAN (Tree-Augmented NBC), CUBC (Chain of Unrestricted BCs), C-DTs (Chain of Decision Trees); (†) results must be 0.0% due to the BC type.

| Dataset     | Source | Classifier* | SHAP   |                 |             | LIME   |                 |             |
|-------------|--------|-------------|--------|-----------------|-------------|--------|-----------------|-------------|
|             |        |             | Naive  | Structural      | Dialectical | Naive  | Structural      | Dialectical |
| Shuttle     | [28]   | NBC         | 0%     | 0% <sup>†</sup> | 16.43%      | 0%     | 0% <sup>†</sup> | 17.14%      |
| German      | [28]   | NBC         | 0%     | 0% <sup>†</sup> | 54.56%      | 0%     | 0% <sup>†</sup> | 49.55       |
| California  | [11]   | TAN         | 0%     | 0% <sup>†</sup> | 16.75%      | 0%     | 0% <sup>†</sup> | 16.75%      |
| Child       | [3]    | CUBC        | 0%     | 70.97%          | 75.35%      | 63.74% | 89.59%          | 91.16%      |
| Insurance   | [3]    | CUBC        | 0%     | 67.07%          | 78.77%      | 59.56% | 89.26%          | 93.07%      |
| HELOC       | [5]    | C-DTs       | 51.77% | 100%            | 94.42%      | 62.21% | 100%            | 97.87%      |
| LendingClub | [4]    | C-DTs       | 16.19% | 100%            | 94.47%      | 72.95% | 100%            | 98.57%      |

Our experiments aim to evaluate the satisfaction/violation of structural and dialectical DA empirically for various concrete RXs (i.e. LIME, SHAP and their structural variants) when they are not guaranteed to satisfy the properties, as shown in Sect. 5. The main questions we aim to address concern *actual DA* and *efficiency*, as follows.

**Actual DA.** While some approaches may not be guaranteed to satisfy DA in general, they may for the most part in practice. *How much DA is achieved in the concrete settings of SHAP, and LIME explanations?* We checked the average percentages of reasons in LIME and SHAP explanations which do not satisfy our notions of descriptive accuracy. The results are in Table 3. We note that: (1) LIME often violate *naive descriptive accuracy*, e.g. in the *Child* and *Insurance* BCs, whereas SHAP does not; (2) LIME and SHAP systematically violate *structural descriptive accuracy*; (3) LIME and SHAP often violate *dialectical descriptive accuracy*.

**Efficiency.** We have defined DARXs so that they are guaranteed to satisfy structural and dialectical DA. *Is the enforcement of these properties viable in practice, i.e. how expensive is it to compute DARXs?* Formally, the computational cost for DARXs can be obtained as follows. Let  $t_p$  be the time to compute a prediction and its associated posterior probabilities.<sup>4</sup> The upper bound of the time complexity to compute a DARX is  $T_{DARX}(\Omega) = O(t_p \cdot \sum_{V_i \in \mathbf{V}} |\Omega_{V_i}|)$ , which is *linear* with respect to the sum of the number of variables’ values, making DARXs *efficient*, while the time required by LIME is exponential with respect to the number of variables’ values for discrete variables, while it requires a large (by default 5000) number of sample in the case of continuous variables.

<sup>4</sup> In our experiments, using a machine with *Intel i9-9900X* at 3.5 Ghz and 32 GB of RAM (no GPU acceleration)  $t_p$  ranges from  $3\mu s$  to  $436ms$  depending on the classifier.

## 7 Discussion and Conclusions

We have introduced a three-fold notion of DA for explanations of probabilistic classifiers, which, despite its intuitiveness, is often not satisfied by prominent explanation methods, and shown that it can be satisfied, by design, by the novel explanation concept of DARXs.

A variety of approaches devoted in particular to the explanation of Bayesian networks exist in the literature [12, 18]. At a high level these approaches can be partitioned into three main families [12]: explanation of evidence (which concerns explaining observations by abducting the value of some unobserved variables), explanation of model (which aims at presenting the entire underlying model to the user), and explanation of reasoning. Explanation of reasoning is the closest to our approach and, according to [12] is in turn divided into: (i) explanation of the results obtained by the system and the reasoning process that produced them; (ii) explanation of the results not obtained by the system, despite the user's expectations; (iii) hypothetical reasoning, i.e. what results the system would have returned if one or more given variables had taken on different values from those observed. Our approach is mainly related to point (i), even if it may support some form of hypothetical reasoning too. We remark anyway that the spirit of DARX is not advancing the state of the art in explanations for Bayesian networks but rather providing a concrete example of a method satisfying the DA properties we introduced and showing that even with this baseline approach we can get improvements with respect to popular model-agnostic methods.

Our work opens several avenues for future work. It would be interesting to experiment with other forms of probabilistic classifiers, including (chained) neural networks, possibly in combination with methods for extracting *causal models* from these classifiers to provide structural descriptions for satisfying structural DA. It would also be interesting to study the satisfaction of (suitable variants of) DA by other forms of explanations, including *minimum cardinality* explanations [25] and set-based explanations [6, 9].

**Acknowledgements.** Toni was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors listed.

## References

1. Alvarez-Melis, D., Jaakkola, T.S.: A causal framework for explaining the predictions of black-box sequence-to-sequence models. In: EMNLP (2017)
2. Bielza, C., Larrañaga, P.: Discrete Bayesian network classifiers: a survey. *ACM Comput. Surv.* **47**(1), 5:1–5:43 (2014)
3. BNlearn: Bayesian network repository - an R package for Bayesian network learning and inference (2020). <https://www.bnlearn.com/bnrepository>

4. Club, L.: Lending Club Loans (2019). <https://kaggle.com/wordsforthewise/lending-club>
5. FICO Community: Explainable Machine Learning Challenge (2019). <https://community.fico.com/s/explainable-machine-learning-challenge>
6. Dhurandhar, A., et al.: Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: NeurIPS (2018)
7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. *Mach. Learn.* **29**(2–3), 131–163 (1997). <https://doi.org/10.1023/A:1007465528199>
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5), 93:1–93:42 (2019)
9. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: AAAI (2019)
10. Jacovi, A., Marasovic, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: FAccT (2021)
11. Kelley Pace, R., Barry, R.: Sparse spatial autoregressions. *Stat. Probab. Lett.* **33**(3), 291–297 (1997)
12. Lacave, C., Díez, F.J.: A review of explanation methods for Bayesian networks. *Knowl. Eng. Rev.* **17**(2), 107–127 (2002). <https://doi.org/10.1017/S026988890200019X>
13. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: AIES (2019)
14. Lipovetsky, S.: Let the evidence speak - using Bayesian thinking in law, medicine, ecology and other areas. *Technometrics* **62**(1), 137–138 (2020)
15. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NeurIPS (2017)
16. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. *J. ACM* **7**(3), 216–244 (1960)
17. McLachlan, S., Dube, K., Hitman, G.A., Fenton, N.E., Kyrimi, E.: Bayesian networks in healthcare: distribution by medical condition. *Artif. Intell. Medicine* **107**, 101912 (2020)
18. Mihaljevic, B., Bielza, C., Larrañaga, P.: Bayesian networks for interpretable machine learning and optimization. *Neurocomputing* **456**, 648–665 (2021)
19. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**(44), 22071–22080 (2019)
20. Neapolitan, R.E., Jiang, X.: Probabilistic Methods for Financial and Marketing Informatics. Elsevier, Amsterdam (2010)
21. Pearl, J.: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference. Morgan Kaufmann, Burlington (1989)
22. Provan, G.M., Singh, M.: Learning Bayesian networks using feature selection. In: AISTATS (1995)
23. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5782, pp. 254–269. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04174-7\\_17](https://doi.org/10.1007/978-3-642-04174-7_17)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should i trust you?”: explaining the predictions of any classifier. In: ACM SIGKDD (2016)
25. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining Bayesian network classifiers. In: IJCAI (2018)

26. Sokol, K., Flach, P.A.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: FAT\* (2020)
27. Timmer, S.T., Meyer, J.-J.C., Prakken, H., Renooij, S., Verheij, B.: Explaining Bayesian networks using argumentation. In: Destercke, S., Denoeux, T. (eds.) ECSQARU 2015. LNCS (LNAI), vol. 9161, pp. 83–92. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-20807-7\\_8](https://doi.org/10.1007/978-3-319-20807-7_8)
28. UCI, C.: Machine Learning Repository (2020). <https://archive.ics.uci.edu/ml/datasets.php>
29. Zaragoza, J.H., Sucar, L.E., Morales, E.F., Bielza, C., Larrañaga, P.: Bayesian chain classifiers for multidimensional classification. In: IJCAI (2011)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

