The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^2HCAI)

The Shape of Explanations: A Topological Account of Rule-Based Explanations in Machine Learning

Brett Mullins[1]

[1] University of Massachusetts Amherst
bmullins@umass.edu

Abstract

Rule-based explanations provide simple reasons explaining the behavior of machine learning classifiers at given points in the feature space. Several recent methods (Anchors, LORE, etc.) purport to generate rule-based explanations for arbitrary or black-box classifiers. But what makes these methods work in general? We introduce a topological framework for rule-based explanation methods and provide a characterization of explainability in terms of the definability of a classifier relative to an explanation scheme. We employ this framework to consider various explanation schemes and argue that the preferred scheme depends on how much the user knows about the domain and the probability measure over the feature space.

1 Introduction

Explanations for predictions of machine learning models act as reasons for a predictive model's behavior and are desirable for trustworthy and transparent machine learning (Molnar 2022). With this being said, there is not much agreement in the machine learning community on exactly what counts as an explanation (Doshi-Velez and Kim 2017; Burkart and Huber 2021). Machine learning practitioners have developed a large set of domain-specific explanation methods in recent years. These methods are tailored either to the specific task the model is seeking to perform, e.g. regression, classification, object detection, etc., or to the type of inputs and outputs of the model, e.g., tabular features, images, sentences, etc. (Islam et al. 2022).

A promising technique for explaining the predictions of structured or tabular classifiers is rule-based explanations. A rule-based explanation is a predicate defining a simple region in the feature space that is sufficient for classifying a given point. In this paper, we take advantage of the connection between the inherent definability of rule-based explanations and definability in topology to develop a general framework to represent varieties of explanations based on existing explanation algorithms.

To summarize this paper, we make the following contributions:

- We present a novel framework of explainability for rule-based classifiers based on existing explanation algorithms.
- We characterize explainability as a topological property relative to an explanation scheme i.e. relative to a choice of explanation shape and a measure of explanation size. We conjecture that all classifiers "in the wild" satisfy this notion of explainability.
- Employing our framework, we identify two principles for explanation algorithms that apply both theoretically and in practice. The first is that rule-based explanations can take nearly any desired shape. The second holds that if no probability measure is known over the feature space and at least one feature is not bounded, then explanations should be bounded, i.e. include all unbounded features.

This paper proceeds as follows. In Section 2, we discuss various existing explanation algorithms and provide a brief introduction to topology. We introduce explanation schemes as a framework for explainability and characterize explainability as a topological property in Sections 3, 4, respectively. In Section 5, we derive principles for both formal and practical explanation algorithms. In Sections 6, 7, we conclude by discussing limitations and open problems and surveying related work.

2 Background

Rule-Based Explanations

In this section, we introduce rule-based explanation algorithms and consider their representative properties. Given a classifier and a point in the feature space, a rule-based explanation algorithm generates a rule defined in terms of the features of the classifier that both covers the given point and is sufficient for its classification. Rule-based explanations are perturbation resistant in the sense that the explanation applies to a neighborhood about a given point. Additionally, this sort of explanation is often called post-hoc, since it occurs after the model is constructed, and local, since the explanation is specific to the given point (Guidotti et al. 2018b). Let us consider four such explanation algorithms: Anchors (Ribeiro, Singh, and Guestrin 2018), PALEX (Jia et al. 2020), LoRE (Guidotti et al. 2018a), and LoRMIkA (Rajapaksha, Bergmeir, and Buntine 2020).

Given black-box access to a classifier and a point, these algorithms evaluate the classifier on a collection of points sampled in a neighborhood of the given point and return a rule such that the points satisfying the rule (or at least some proportion of points above a given threshold) evaluate to the same label as the given point. Anchors and PALEX each return rules that are the conjunction of predicates involving individual features which only use $<$, $=$, $>$ relations and $\wedge$ connective. By contrast, LoRE and LoRMIkA return a rule defined by a simple decision tree as well as counterfactuals, though we ignore the latter in the present analysis. Notice that all four algorithms return rules that define rectangles in the feature space.

Rule-based explanation algorithms often claim to be model agnostic i.e. the explanation algorithm makes no assumption on the given classifier's functional form to generate post-hoc explanations (Ignatiev, Narodytska, and Marques-Silva 2019a). This suggests that such algorithms will generate explanations for any functional form.

Topology

Let X be a set. A *topology* $\mathcal{T}$ is a collection of subsets of X with the following properties:

1. $X, \emptyset \in \mathcal{T}$;
2. $\mathcal{T}$ is closed under arbitrary union;
3. $\mathcal{T}$ is closed under finite intersection.

We refer to $(X, \mathcal{T})$ as a *topological space*. In a topological space $(X, \mathcal{T})$, a subset $\mathcal{O} \subseteq X$ is called *open* in X if $\mathcal{O} \in \mathcal{T}$. A subset $C \subseteq X$ is called *closed* in X if $X \setminus C \in \mathcal{T}$. The *interior* of a set $A \subseteq X$, $\text{Int}(A)$, is the largest open set contained in A, and the *closure* of a A, $\overline{A}$, is the smallest closed set containing A.

A *basis* $\mathcal{B}$ for $\mathcal{T}$ is a collection of open sets of $\mathcal{T}$ such that every open set of $\mathcal{T}$ is the union of elements of $\mathcal{B}$. We refer to sets in the basis as *basic open sets*, and, if $\mathcal{B}$ is a basis for $\mathcal{T}$, then we say that $\mathcal{T}$ is generated by $\mathcal{B}$. If $A \in \mathcal{T}$ and $\mathcal{T}$ has basis $\mathcal{B}$, then we say that A can be defined in terms of basic open sets from $\mathcal{B}$.

We say that a set $A \subseteq X$ is *dense* in X if $\mathcal{O} \cap A \neq \emptyset$ for every non-empty open set $\mathcal{O} \subseteq X$. A set $B \subseteq X$ is said to be *nowhere dense* in X if $\text{Int}(\overline{B}) = \emptyset$. Nowhere dense sets fail to cover any part of X with respect to $\mathcal{T}$. A set $C \subseteq X$ is called *meagre* if C is the countable union of nowhere dense sets. If C is meagre with respect to topology $\mathcal{T}$, then C is said to be $\mathcal{T}$-meagre. The notion of meagre is what we mean by topologically small.

3 Explanation Schemes

Let $f : X \to Y$ be a classifier. A rule-based explanation for $x \in X$ is a well-defined region of the feature space containing x whose classification is invariant within the region, i.e. belonging to the region is sufficient to be classified as $f(x)$. The intuition of a rule-based explanation is that the label assigned by the classifier is unaffected by perturbations so long as the perturbed point remains within the region (Guidotti et al. 2018a). What properties, then, do these regions have, and what shape do they take? In this section, we develop a general topological framework to represent the properties of rule-based explanations. We explore how to represent explainability within this framework in Section 4 below.

Rule-based explanations are definable subsets of the feature space that belong to a common class, i.e. they satisfy some predicate or definable property φ. For instance, φ may be the predicate open rectangle or open ball. More generally, the candidate subsets for explanations belong to the class of subsets of the feature space satisfying φ: $\{A \subseteq X \mid \varphi(A)\}$. Not all predicates, however, are suitable to be explanations. Let us restrict attention to the following set of rules.

Definition 1. *A scalable rule φ for X is a predicate φ such that $X_\varphi = \{A \subseteq X \mid \varphi(A)\}$ satisfies the following two conditions:*

1. $\bigcup X_\varphi = X$
2. *If $x \in A_1 \cap A_2$ for $A_1, A_2 \in X_\varphi$, there exists $A_3 \in X_\varphi$ such that $x \in A_3$ and $A_3 \subseteq A_1 \cap A_2$.*

Condition 1 says that each point in the feature space is covered by a rule i.e. a potential explanation. Condition 2 says that rules can be defined as small as needed. While not all scalable rules make for desirable explanations for the user, we hold the converse is true. Scalable rules provide the necessary minimal structure for this analysis.

Observe that the collection of subsets satisfying a scalable rule φ meets the conditions for being a topological basis (Munkres 2000). By closing this collection under the operations of countable union and finite intersection, we obtain the topology $\mathcal{T}_\varphi$ We say that $\mathcal{T}_\varphi$ is the *explanation topology* generated by scalable rule φ. The explanation topology consists of sets that can be defined in terms of rules of the form φ.

Though there may be many rules covering a given point in the feature space, we assign a notion of size or robustness to a rule called coverage (Ribeiro, Singh, and Guestrin 2018). We define coverage as a measure μ in the measure-theoretic sense. Typically, if a probability measure p is known over the feature space, then coverage of a given rule A is the measure of A with respect to p. Similarly, if a probability measure is unknown, then the counting measure is typical for discrete feature spaces, Lebesgue measure is typical for continuous feature spaces, and a product measure is typical for spaces with both discrete and continuous features. If an explanation has zero coverage with respect to μ, we say that explanation is μ-null. Usually, the user will prefer an explanation with greater coverage to an explanation with lesser coverage; however, other factors may influence a user's preferences. For instance, a user may prefer explanations that involve fewer features to more features for a given amount of coverage (Molnar 2022).

Let us conclude by packaging together the terms introduced in this section:

Definition 2. *An explanation scheme is a tuple (X, φ, μ) where X is the feature space, φ is a scalable rule generating the explanation topology $\mathcal{T}_\varphi$ on X, and μ is a measure on X representing coverage.*

4 Explainability via Definability

In this section, we introduce a notion of explainability relative to an explanation scheme and demonstrate that explainability is equivalent to a simple topological property. In particular, we find that a classifier is explainable if the preimage of each label is the union of a low complexity set, i.e. an open set, and a set of points that is small with respect to both the scalable rule and coverage measure. Let us first introduce our notion of explainability.

Definition 3. *A classifier $f : X \to Y$ is explainable for scheme (X, φ, μ) if each $x \in X$ has an explanation except on a set of edge cases.*

Edge cases are typically taken to be a small number of instances in which some desired property does not hold. For instance, in the present setting, the prototypical edge case is a point on the decision boundary of a classifier for a continuous feature space. We formalize the notion of smallness as topologically small with respect to rules of the form φ and small in coverage with respect to measure μ.

Definition 4. *A set E is a set of edge cases for scheme (X, φ, μ) if E is $\mathcal{T}_\varphi$-meagre and μ-null.*

Being meagre and null are necessary for a set of edge cases since neither alone implies smallness in the sense of our prototypical edge case set. For instance, given the standard topology, it is known that $\mathbb{R}$ can be partitioned into a meagre set and a Lebesgue null set (Oxtoby 1980). This identification between the notion of edge cases and topologically and measure-theoretically small sets leads to our main result:

Theorem 1. *A classifier $f : X \to Y$ is explainable for scheme (X, φ, μ) if and only if, for $y \in Y$, there exists open set $\mathcal{O}_y \in \mathcal{T}_\varphi$ such that $f^{-1}(y) = \mathcal{O}_y \cup E_y$ and E_y is $\mathcal{T}_\varphi$-meagre, μ-null.*

Proof. $(\to)$. Suppose $f : X \to Y$ is explainable for (X, φ, μ). Let $\mathcal{O} \subseteq X$ be the set of points in the feature space with explanations and $E = X \setminus \mathcal{O}$. Then E is $\mathcal{T}_\varphi$-meagre, μ-null. Let $y \in Y$. Define $\mathcal{O}_y = f^{-1}(y) \cap \mathcal{O}$ and $E_y = f^{-1}(y) \cap E$. Since meagre and null sets are closed under subset, E_y is $\mathcal{T}_\varphi$-meagre, μ-null. Then, for $x \in \mathcal{O}_y$, there is explanation A_x covering x such that $A_x \subseteq \mathcal{O}_y$. Since $\mathcal{O}_y = \cup_{x \in \mathcal{O}_y} A_x$, $\mathcal{O}_y$ is open in $\mathcal{T}_\varphi$.
$(\leftarrow)$. Suppose, for $y \in Y$, $f^{-1}(y) = \mathcal{O}_y \cup E_y$ where $\mathcal{O}_y$ is open for $\mathcal{T}_\varphi$ and E_y is $\mathcal{T}_\varphi$-meagre, μ-null. Let $x \in X$, and, for some $y \in Y$, $f(x) = y$. Then $x \in \mathcal{O}_y \cup E_y$. WLOG, suppose $x \in \mathcal{O}_y$. Then there exists basic open set $A_x \subseteq \mathcal{O}_y$ covering x such that A_x satisfies φ. So A_x is an explanation for x. □

Though Theorem 1 is a simple characterization, this result allows us to determine explainability by only considering the geometry of the classifier with respect to a set of rules. As an example, consider a linear classifier on n continuous features with rules of the form of open squares. The explanation topology generated from open squares is the standard Euclidean topology on $\mathbb{R}^n$. In this topology, the preimage of one label is an open halfspace and the other is a closed halfspace which satisfies the condition for Theorem 1. So the linear classifier is explainable relative to this explanation scheme.

As a general application of Theorem 1, let us consider whether or not explainability is preserved by a voting ensemble classifier. We define a voting ensemble as follows (Dietterich 2000):

Definition 5. *If $f_1, \ldots, f_k$ are classifiers such that $f_i : X \to Y$, $1 \leq i \leq k$, then a voting ensemble of $f_1, \ldots, f_k$ is a classifier $f : X \to Y$ given by $f(x) = g(f_1(x), \ldots, f_k(x))$ where $g : Y^k \to Y$ where g returns the most common label.*

For instance, a random forest is a common voting ensemble (Breiman 2001). Below, we prove that voting ensembles preserve explainability with respect to a given explanation scheme by appealing only to topological properties.

Theorem 2. *If $f_1, \ldots, f_k$ are classifiers explainable for explanation scheme (X, φ, μ) and f is the voting ensemble of $f_1, \ldots, f_k$, then f is explainable for (X, φ, μ).*

Proof. Let $y \in Y$. Then $g^{-1}(y) = \{v \in Y^k \mid g(v) = y\}$. Suppose $v \in g^{-1}(y)$. For $1 \leq i \leq k$, $f_i^{-1}(v_i) = \mathcal{O}^i_{y,v} \cup E^i_{y,v}$. The set of points satisfying v is given by

$$\begin{aligned}\bigcap_{i=1}^{k} f_i^{-1}(v_i) &= \bigcap_{i=1}^{k} (\mathcal{O}^i_{y,v} \cup E^i_{y,v}) \\ &= \left(\bigcap_{i=1}^{k} \mathcal{O}^i_{y,v}\right) \cup E_{y,v} \\ &= \mathcal{O}_{y,v} \cup E_{y,v}\end{aligned}$$

where $E_{y,v} = \bigcap_{i=1}^{k} (\mathcal{O}^i_{y,v} \cup E^i_{y,v}) \setminus (\bigcap_{i=1}^{k} \mathcal{O}^i_{y,v})$ and $\mathcal{O}_{y,v} = \bigcap_{i=1}^{k} \mathcal{O}^i_{y,v}$. Distributing the intersection operator, $E_{y,v} = \bigcup_i \bigcap_{j=1}^{k} C_{ij}$ where for every i there is at least one j such that $C_{ij} = E^i_{y,v}$. Since meagre and null sets are closed under subset and countable union and $E_{y,v}$ is the union of $\mathcal{T}_\varphi$-meagre and μ-null sets, $E_{y,v}$ is $\mathcal{T}_\varphi$-meagre and μ-null. Correspondingly, $\mathcal{O}_{y,v}$ is open in $\mathcal{T}_\varphi$, since it is the finite intersection of $\mathcal{T}_\varphi$-open sets.

Then we obtain $f^{-1}(y)$ as follows:

$$\begin{aligned}f^{-1}(y) &= \bigcup_{v \in g^{-1}(y)} (\mathcal{O}_{y,v} \cup E_{y,v}) \\ &= \left(\bigcup_{v \in g^{-1}(y)} \mathcal{O}_{y,v}\right) \cup \left(\bigcup_{v \in g^{-1}(y)} E_{y,v}\right) \\ &= O_y \cup E_y\end{aligned}$$

where $\mathcal{O}_y = \bigcup_{v \in g^{-1}(y)} \mathcal{O}_{y,v}$ and $E_y = \bigcup_{v \in g^{-1}(y)} E_{y,v}$. Observe that $\mathcal{O}_y$ is open in $\mathcal{T}_\varphi$, since it is the union of $\mathcal{T}_\varphi$-open sets, and E_y is $\mathcal{T}_\varphi$-meagre and μ-null, since it is the union of $\mathcal{T}_\varphi$-meagre and μ-null sets. □

One may suspect that this notion of explainability is too permissive and designates all classifiers as explainable. In general, this is not the case. Consider the classifier with a single continuous feature that maps rational numbers to 1 and irrational numbers to 0. If φ is the predicate for open intervals, then for $x \in \mathbb{R}$ if x is rational (irrational) then every rule satisfying φ that covers x contains an irrational (rational) number. Since this holds for each point in the feature space, this classifier is not explainable.

Note though that the above example is qualitatively different from typical classifiers deployed in applications. This leads us to make the following imprecise conjecture:

Conjecture 1. *(In-the-wild conjecture) Classifiers deployed in applications satisfy explainability with respect to a typical explanation scheme.*

If the in-the-wild conjecture is true, then rule-based explanations and, thus, explanation algorithms based on this formal model are model agnostic.

5 Implications for Rule-Based Explanation Algorithms

In this section, we employ the framework developed above to argue for two principles for designing rule-based explanation algorithms.

Principle 1. *For continuous feature spaces, explanations can take nearly any desired shape.*

The candidate scalable rules discussed thus far - open balls and open rectangles - both generate the Euclidean or standard topology on continuous feature spaces i.e. both scalable rules generate the same explanation topology. However, the standard topology has many more bases. A typical result in topology is to prove that two bases generate the same topology (Munkres 2000). Scalable rules φ, ψ generate equivalent explanation topologies if for every potential explanation $A \subseteq X$ satisfying φ and $x \in A$, then there is some potential explanation $B \subseteq X$ satisfying ψ such that $x \in B$ and $B \subseteq A$, and vice versa.

If two explanation schemes share a coverage measure and their respective scalable rules each generate the same explanation topology, then they share the same class of explainable models. From this perspective, one can substitute any scalable rule that generates the standard topology in place of open balls or open rectangles in an explanation scheme without affecting which models are explainable for the new scheme. Recall, however, that there may be other reasons to prefer some sorts of explanations to others (Molnar 2022). For instance, users may prefer shorter explanations or explanations that include specific features they care about (Watson and Floridi 2021).

Principle 2. *If features are unbounded and a probability measure is not known, then the user should only consider scalable rules that are bounded.*

Let us say that a feature is bounded if the feature has a maximum value and a minimum value. Suppose the user does not know a probability measure over a continuous feature space, a single feature F is not bounded, and coverage is a monotone non-decreasing function of Lebesgue measure. Then all rules unbounded in F have equal coverage, namely ∞. This is to say that coverage cannot distinguish between the size of unbounded rules; although, some such rules are clearly *larger* than others. This is evident by considering the coverage of such an unbounded rule on the subspace excluding the feature F. Since this feature subspace is bounded, it is possible (though not necessary) for coverage to distinguish between rules of various sizes.

This issue can be resolved by restricting rules φ to the class of bounded scalable rules. A bounded rule is a rule that includes an upper bound predicate for features with no maximum value and a lower bound predicate for features with no minimum value. Practical examples include open balls with a rational center point and radius and open rectangles with rational corner points.

As mentioned above, one measure of the complexity of a rule is the number of predicates conjoined. While restricting to bounded rules solves the discrimination problem for Lebesgue measure, it introduces a lower bound on the complexity of potential explanations. For instance, each unbounded feature requires an upper bound predicate if unbounded above and a lower bound predicate if unbounded below. In the least, the length of explanations is the number of unbounded features.

6 Limitations and Open Problems

The topological notion of explainability developed in this paper relates the shape of a rule to whether or not a classifier is explainable. A limitation of this approach is that it only considers a single type of explanation, rule-based explanations, and, for rule-based explanations, may not capture a user's preferences over potential explanations. Below, we consider two properties of existing rule-based explanation algorithms and suggest directions for future research.

Coverage Guarantee

For a given explanation scheme (X, φ, μ), suppose a user is only interested in potential explanations with coverage at least α. In practice, explanations of sufficiently low coverage may either fail to serve as perturbation resistent explanations in the case where the rule is too narrowly defined or not be relevant to the data distribution in the case where the rule is μ-null. The inclusion of a coverage guarantee extends explainability in the sense of Definition 3 which, for an explainable model, guarantees than an explanation exists not that an explanation of a given coverage exists.

One approach to extending topological explainability is to modify the rule φ defining potential explanations to depend on μ. Let us define $\varphi_{\mu,\alpha}$ to restrict to potential explanations with coverage greater than α such that the set of potential explanations becomes $\{A \subseteq X | \varphi(A), \mu(A) \geq \alpha\}$. Observe, however, that $\varphi_{\mu,\alpha}$ is not a scalable rule by failing to satisfy Condition 2 from Definition 1 and so is not a basis for a topology. Finding a simple way to encode a coverage guarantee without compromising topological structure would be a significant improvement to the present formalism.

Fuzzy Explanations

Existing rule-based explanation algorithms such as Anchors return explanations where some but not necessarily all points satisfying the rule evaluate to the same label. This property results from searching for rules via sampling (Ribeiro, Singh, and Guestrin 2018). Likewise, a user may prefer a fuzzy explanation in cases where one's classifier is complex, perhaps due to overfitting. We term explanations of this sort fuzzy explanations and call the degree to which a fuzzy explanation shares the same label as the point to be explained its fidelity. Let us formalize fidelity as follows:

Definition 6. *For explanation scheme* (X, φ, μ)*, the fidelity of explanation* A *for classifier* $f : X \to Y$ *at* $x \in X$ *is given by* $\frac{\mu(A \cap X_{f(x)})}{\mu(A)}$ *where* $X_{f(x)} = \{x' \in X | f(x') = f(x)\}$ *if* $\mu(A) > 0$ *and* 0 *otherwise.*

Explainability in the sense of Definition 3 considers only maximum fidelity rules. Unlike the case with coverage guarantees, extending topological explainability to fuzzy explanations satisfying a given level of fidelity does not compromise the structure of the explanation topology $\mathcal{T}_\varphi$; rather, the characterization of an explainable classifier requires modification. This fuzzy explainability is a strict generalization of the former notion; there are simply more potential explanations for each point in the feature space. The difficulty here is that we can no longer represent explainability as the preimage of each label being an open set excluding a set of edge cases, since potential fuzzy explanations are not contained to the preimage of a single label. In particular, the subset of a potential fuzzy explanation contained in a label preimage can be of arbitrarily high topological complexity.

7 Related Work

Rule-based explanation methods are often based on work in rule induction (Grzymala-Busse 2005; Macha et al. 2022). Additional rule-based explanation algorithms include EXPLAN (Rasouli and Yu 2020), LIMREF (Rajapaksha and Bergmeir 2022), and XPlainer (Ignatiev, Narodytska, and Marques-Silva 2019b). There are several notions of explainability beyond rule-based explanations. Let us briefly mention three: counterfactuals, surrogate models, and feature importance.

Given a point in the feature space, counterfactuals provide a collection of nearby points that the model labels differently from the given point. Counterfactuals explain a prediction by describing which changes in the feature values would have yielded a different result. Guidotti 2022 provides an up-to-date comprehensive survey of counterfactual explanations.

Surrogate models approximate an arbitrary model with an interpretable model. If the approximation is sufficiently close, then interpreting the surrogate model explains the behavior of the arbitrary model. Global approaches seek to learn a single surrogate model, while local approaches learn a surrogate model at a given point in the feature space. For instance, Bastani, Kim, and Bastani 2019 proposes a method for learning a simple decision tree as a global surrogate, and LIME is a well-known technique learing a weighted linear model as a local surrogate (Ribeiro, Singh, and Guestrin 2016).

Feature importance explains model behavior by measuring how much each feature contributes to a prediction. A well-known approach is SHAP which is based on the game-theoretic concept of Shapley values (Lundberg and Lee 2017). Many local surrogate methods such as LIME will also yield some way of ranking features by importance (Ribeiro, Singh, and Guestrin 2016).

Several threads of research have sought to better understand explainability methods. Mullins 2019 is a prior attempt to express rule-based explainability in terms of topology. Adversarial attacks have been used to quantify the degree to which various methods can be manipulated (Wilking, Jakobs, and Morik 2022). Logic-based approaches encode models in a formal language and derive explanations (Ignatiev, Narodytska, and Marques-Silva 2019a). Watson and Floridi 2021 introduces an expansive framework to represent explanation methods as a communication game between two learners.

References

Bastani, O.; Kim, C.; and Bastani, H. 2019. Interpreting Blackbox Models via Model Extraction. *arXiv*, abs/1706.09773.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Burkart, N.; and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70: 245–317.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, abs/1702.08608.

Grzymala-Busse, J. W. 2005. Rule induction. In *Data mining and knowledge discovery handbook*, 277–294. Springer.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018a. Local Rule-Based Explanations of Black Box Decision Systems. *arXiv*, abs/1805.10820.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018b. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5).

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019a. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1511–1519.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019b. On validating, repairing and refining heuristic ML explanations. *arXiv*, abs/1907.02509.

Islam, M. R.; Ahmed, M. U.; Barua, S.; and Begum, S. 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3): 1353.

Jia, Y.; Bailey, J.; Ramamohanarao, K.; Leckie, C.; and Ma, X. 2020. Exploiting patterns to explain individual predictions. *Knowledge and Information Systems*, 62(3): 927–950.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Macha, D.; Kozielski, M.; Wróbel, Ł.; and Sikora, M. 2022. RuleXAI—A package for rule-based explanations of machine learning model. *SoftwareX*, 20: 101209.

Molnar, C. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.*

Mullins, B. 2019. Identifying the Most Explainable Classifier. *arXiv*, abs/1910.08595.

Munkres, J. R. 2000. *Topology*. Prentice Hall, Inc, 2nd ed. edition.

Oxtoby, J. C. 1980. *Measure and category: a survey of the analogies between topological and measure spaces*. Springer-Verlag.

Rajapaksha, D.; and Bergmeir, C. 2022. LIMREF: Local Interpretable Model Agnostic Rule-based Explanations for Forecasting, with an Application to Electricity Smart Meter Data. *arXiv*, abs/2202.07766.

Rajapaksha, D.; Bergmeir, C.; and Buntine, W. 2020. LoRMIkA: Local rule-based model interpretability with K-optimal associations. *Information Sciences*, 540: 221–241.

Rasouli, P.; and Yu, I. C. 2020. EXPLAN: explaining black-box classifiers using adaptive neighborhood generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, KDD '16, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Watson, D. S.; and Floridi, L. 2021. The explanation game: a formal framework for interpretable machine learning. In *Ethics, Governance, and Policies in Artificial Intelligence*, 185–219. Springer.

Wilking, R.; Jakobs, M.; and Morik, K. 2022. Fooling Perturbation-Based Explainability Methods. In *Workshop on Trustworthy Artificial Intelligence as a part of the ECML/PKDD 22 program*.