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

Brett Mullins
University of Massachusetts, Amherst

Introduction

Rule-based explanations provide simple reasons explaining the behavior of machine learning classifiers at given points in the feature space. 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 explanations based on existing explanation algorithms.

Contributions

- We present a novel framework of explainability for 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.

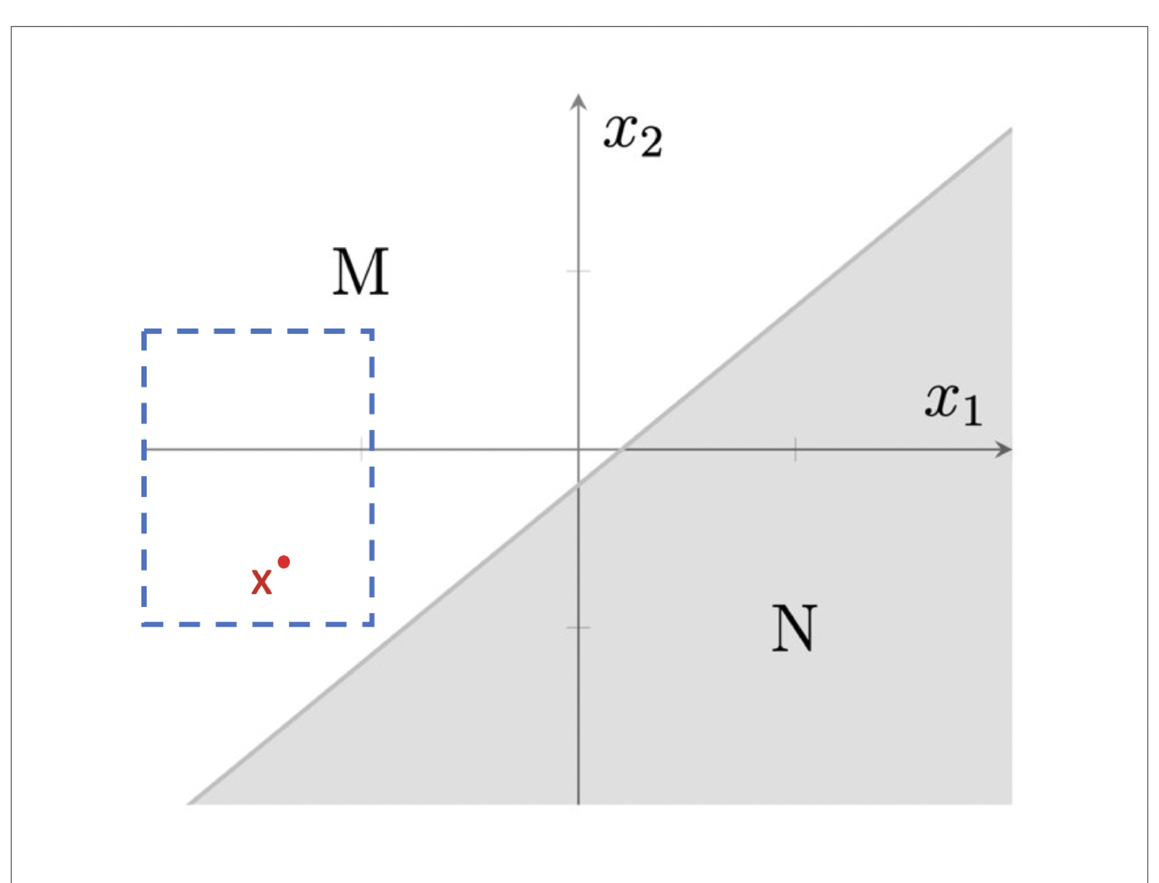

Figure 1: Example rule-based explanation of x for a linear classifier

Rule-Based Explanations

Given classifier $f : X \to Y$, 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 $f(x)$. These explanations have the following properties:

- Local
- Post-Hoc
- Perturbation-Resistant

Figure 1 illustrates a rule-based explanation that is an open rectangle on a continuous feature space. There are several existing algorithms for generating rule-based explanations including Anchors [1] and LORE [2]. We say that a classifier is *explainable* if there exists explanations for all of the feature space except for a set of edge cases.

Explanation Topology

Rule-based explanations are regions of the feature space that satisfy some predicate or definable property φ. We restrict to predicates satisfying the following properties:

- If two explanations for a given point overlap, then there exists an explanation in their intersection covering the point.
- Each point is covered by an explanation.

Then collection of sets satisfying φ is a topological basis [3]. Closing this collection under countable union and finite intersection, we obtain explanation topology $\mathcal{T}_\varphi$.

Sets belonging to $\mathcal{T}_\varphi$ are called open. The main result of this paper characterizes explainability in terms of open sets and small sets.

Main Result (Explainability is a simple topological property)

Theorem: 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.

Explanation Schemes

An explanation scheme (X, φ, μ) is a reference frame for analyzing explainability. It consists of the following:

- X - feature space
- φ - rule generating the explanation topology $\mathcal{T}_\varphi$
- μ - coverage measure defined on $\mathcal{T}_\varphi$

Coverage measures the size of an explanation. For instance, μ is often a probability measure if one is known.

A set of edge cases must be small with respect to both topology $\mathcal{T}_\varphi$ and measure μ. The corresponding notions of smallness are $\mathcal{T}_\varphi$-meagre and μ-null.

Application: Ensembles

Ensembles aggregate predictions from a collection of classifiers and are commmonly used in practice e.g. Random Forests and XGBoost. Ensembles are often viewed as complex, whereas their consitutent classifiers are weak or simple. We show that if the constitutent classifiers are explainable for a given scheme then their ensemble is explainable.

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

Implications

1. For continuous feature spaces, explanations can take nearly any desired shape.
2. If features are unbounded and a probability measure is not known, then the user should only consider explanations that are bounded.

Future Work

Extend Formal Framework

- Minimum Coverage Guarantee
- Fuzzy Explanations

Explore Connections

- Computational Complexity
- Synthetic Topology

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *ArXiv*, abs/1805.10820, 2018.

[3] James Munkres. *Topology*. Prentice Hall, 2nd edition, 2000.

Contact Information

- Website: bcmullins.github.io
- Email: bmullins@umass.edu

DREAM LAB
DATA SYSTEMS RESEARCH FOR EXPLORATION, ANALYTICS, AND MODELING

UMass Amherst