Learning Symbolic Models for Interpretability in Healthcare Applications

By Jennifer J. Schnur

Abstract

Recent trends in data science research have gravitated toward "black box" modeling approaches, in which the meaningful pathway from input to output is not explicitly understood by human users. Despite impressive predictive performance within a wide array of contexts, "black box" models raise difficulties for domain experts to extract the concrete patterns that machine learning algorithms are designed to uncover. In this work, we discuss the potential of symbolic modeling as a more interpretable machine learning approach, especially within high-stakes fields, such as healthcare. We suggest the first step toward achieving such a model is discovering the useful building blocks, or structural relationships, that live within a particular dataset. Using simple association measures, we find that ground-truth building blocks can be discovered on synthetic machine learning problems.

Why the Healthcare Sector Should Care

- Symbolic models are represented by explicit mathematical notation that users can work with and analyze using procedures that are typically learned by the end of high school. As a result, the learned models are highly accessible for interpretation and analysis in the clinical setting.
- A concrete understanding of risk patterns will allow clinical staff to design more tailored interventions and procedures to improve patient outcomes.
- Symbolic model representation is highly portable, and extracted patterns can therefore seamlessly transfer to other health applications.
- Interpretable modeling facilitates transparency and accountability to patients and stakeholders due to the traceable nature of predictions.

Introduction

In recent years, there has been explosive growth and utilization of machine learning models across nearly every field for predictive and analytical tasks. Machine learning refers to the algorithmic process of leveraging patterns to estimate some target variable or group structure in a dataset. In most cases, the final product is a model that defines the rules for arriving at a prediction, given some input data. Modeling approaches vary widely depending on context, data

characteristics, and the particular problem at hand, often reflecting business needs or scientific questions.

Of particular interest, with respect to machine learning ethics, is the concept of interpretability. An interpretable model allows the user to explicitly understand how predictions are made, which is most often measured by a human's ability to consistently predict the model's result (Molnar, 2020). This sits in contrast with "black box" models that obfuscate the pathway from model inputs to output, hiding the learned patterns. Recurring evidence has shown that "black box" models often outperform standard techniques in terms of predictive accuracy at the expense of human interpretability, which is commonly labeled the accuracy-interpretability tradeoff (Adadi et al., 2018). Data scientists theorize that this tradeoff stems from the fact that common interpretable modeling approaches miss out on potential non-linear associations or feature interactions that may benefit a particular problem; therefore, accounting for these components might reduce or eliminate the tradeoff entirely (Rudin, 2019). In order to create high-performing interpretable models, we must capture these patterns in ways that humans can understand them. Symbolic (mathematical) notation may provide a useful path forward in this research, since it is a human-created grammar that has become universal across different languages and cultures, enabling a concise representation of highly sophisticated concepts (Venezia, 2016).

Interpretable models are important, especially in high-stakes domains such as healthcare, where patients' lives hang in the balance. From a healthcare provider's perspective, it is crucial to understand specifically how certain patient features, such as demographics, socioeconomic circumstances, and clinical information contribute to disease diagnostics or risk for adverse outcomes. Clinical staff cannot design better processes or justify interventions without an in-depth grasp of how their care and treatments, in combination with the patient's attributes, might affect the health trajectory.

Method

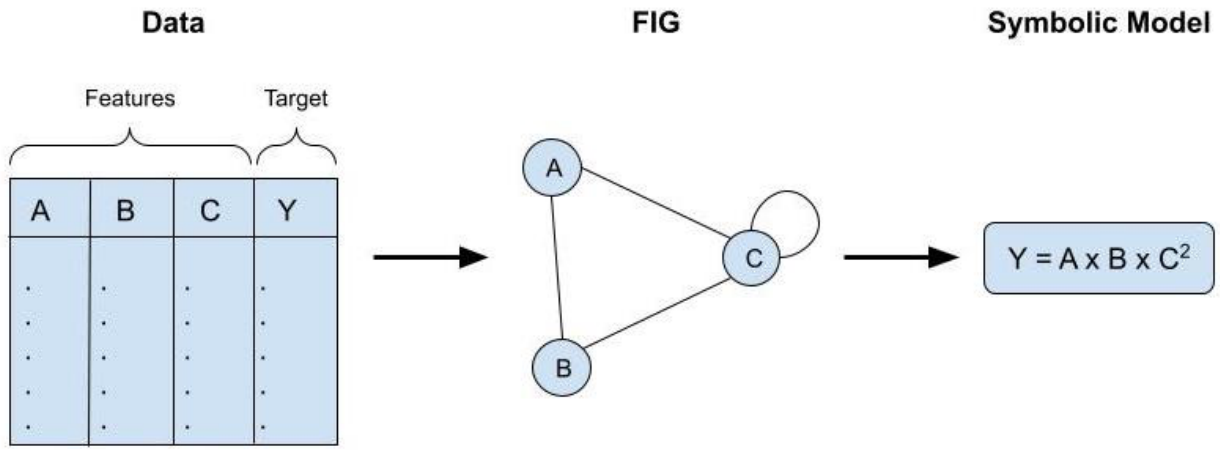
Symbolic Regression (SR) refers to the task of learning a free-form mathematical expression to model some target variable in a dataset. The model consists of a set of symbols (i.e. features and mathematical operators) that fuse together to resemble a formula, which can be interpreted and analyzed using standard mathematical procedures, such as factorization, expansion, derivation, integration, substitution, etc. Due to their interpretability, symbolic models are desirable and have been learned in a variety of ways, ranging from genetic

algorithms that evolve a population of expressions through "breeding" to reinforcement learning which promotes the most successful symbolic structures over a series of learning episodes (La Cava et al., 2021).

SR is a difficult task because an optimal expression can be any length and consist of any combination of features and operators. Therefore, the most common approaches tend to be computationally expensive and rely heavily on random choices during the learning process. We envision a more deterministic approach that begins with the discovery of meaningful building blocks, or mathematical substructures, that prove useful in solving the problem at hand.

In order to quantify the difficulty of this first step in the SR approach, we explored a baseline method for predicting feature interactions on synthetic machine learning problems[1]. Specifically, we tested if simple association measures (e.g. Pearson correlation, distance correlation, etc.) can help create graph structures that link informative features together, where links represent useful interactions. For each problem, these predicted interactions form a feature interaction graph (FIG) that may be used for building symbolic models downstream, as illustrated in Figure 1. While typical interpretable feature engineering methods suffer from the sheer complexity of combining and testing features for model development, the FIG may allow us to extract composite features directly from the connected components in the graph structure.

Figure 1.



These experiments are not without their limitations: (1) this work solely considers multiplicative interactions (the product of 2 features) and does not account for any

[1] The Feynman Symbolic Regression Database: https://space.mit.edu/home/tegmark/aifeynman.html

mathematical transformations (e.g. sin, cos, tan, square root, log, exponential), which limits the types of expressions we can construct (although we will explore this idea in future work) and (2) the datasets are synthetic and do not represent any real world phenomena that would typically include noise, diversity of feature distributions, and collinearities.

Findings

The initial experiments reveal two important aspects of the link prediction problem that will need further investigation, specifically (1) interaction evaluation and (2) heuristic choices.

Problem 1: Interaction Evaluation

Surprisingly, it is not always clear how to evaluate whether a particular interaction is contained in a symbolic expression. Let's consider Problem 8 from Table 1. The target variable is modeled by ground-truth expression,

$$Y = \frac{Gm_1m_2(r_1-r_2)}{r_1r_2}.$$

We can confidently say there are multiplicative interactions between variables G and m_1 or m_1 and m_2, but what about r_1 and r_2? It appears so in the simplified expression, but if we expand the expression to be

$$Y = \frac{Gm_1m_2}{r_2} - \frac{Gm_1m_2}{r_1},$$

the interaction seems to disappear. Due to this flexibility of representation, we need a solid standard for evaluation. For the purposes of this study, we only consider interactions contained in the *expanded* version of each expression, but acknowledge that this remains an area for investigation. For the interaction (link) prediction task, we measure success via F1 score, which ranges from 0 to 1 and is calculated using the correct and incorrect link predictions within the learned FIG.

Problem 2: Heuristic Choice

We tested our approach on 8 relatively simple problems using 2 different measures of association, Pearson correlation and distance correlation. The results (Table 1) show that distance correlation performs just as well or better than Pearson correlation as the chosen heuristic for link prediction in most cases, which makes sense since distance correlation is a

more robust measure that can detect nonlinear dependencies between variables. However, distance correlation still fails considerably on more complex problems (e.g. Problems 8). These experiments illustrate the difficulty in creating FIGs from data for moderate-to-large symbolic expression sizes and show that more work is needed to develop a consistent framework.

Table 1.

Problem No.	Target Variable's (Y's) Ground-Truth Symbolic Expression	Link Prediction F1 Score	
		Heuristic = Pearson Correlation	Heuristic = Distance Correlation
1	$Y = n\mu$	1.0	1.0
2	$Y = gmz$	1.0	1.0
3	$Y = \frac{Bqv}{p}$	0.92	1.0
4	$Y = \frac{q_1}{4\pi\epsilon r^2}$	0.80	1.0
5	$Y = \frac{\epsilon h^2}{\pi m q^2}$	0.89	0.93
6	$Y = \frac{a^2 q^2}{6\pi\epsilon c^3}$	0.90	0.80
7	$Y = \frac{mx^2(w^2 + w_0^2)}{4}$	0.94	0.88
8	$Y = \frac{Gm_1m_2(r_1 - r_2)}{r_1 r_2}$	0.15	0.43

Learning Symbolic Models for Interpretability in Healthcare Applications

By Jennifer J. Schnur

Practical Insights / Applications

The mission of most healthcare practitioners revolves around constant improvement because "what makes us better makes you better," [2] as Northwestern Medicine's slogan goes. For clinical staff focused on enhancing their processes, whether it's diagnostics, treatment management, or triage, the first step is to understand the abundant relationships held within the electronic medical record that facilitate prediction on a variety of patient-centered problems. This can only be accomplished if models remain interpretable, where possible, in the clinical setting.

In this work, we explored a simple approach to learn the useful feature interactions that contribute to prediction, so that an optimal symbolic model can be created downstream. Symbolic models provide readily available, clear patterns for clinical staff to work with and analyze (Schnur et al., 2022). In particular, they can be used to generate "what if?" scenarios (e.g. how will the patient's risk for infection change if we modify the patient's medication dosage?). Similarly, healthcare practitioners can examine counterfactuals (e.g. would the patient have experienced bleeding if we had missed a step in our post-surgery safety process?) Healthcare professionals are most concerned about how their interventions and behaviors might impact patient outcomes, and therefore modeling approaches must cater to those needs.

Using symbolic models, healthcare practitioners can easily trace or simulate predictions and share this information with patients and stakeholders, which improves accountability and transparency. Disclosing the concrete information that leads to a prediction increases an individual's ability to make decisions, relational trust, and loyalty (Felzmann et al., 2020), which are crucial aspects of successful healthcare. Furthermore, the patterns held within symbolic models are modular, meaning that they can be split up into meaningful chunks to be independently analyzed or transferred by health practitioners from one application to another. This is useful given the highly connected nature of health risk (Goh et al., 2007).

Conclusions and Next Steps

The overuse of "black box" models has pervaded many fields, leaving domain experts at loss for concretely understanding the patterns that lead to predictions. However, the results from this study show that the creation of high-performance interpretable models is an attainable goal.

[2] https://www.nm.org/healthbeat/medical-advances/better

Learning Symbolic Models for Interpretability in Healthcare Applications

By Jennifer J. Schnur

We show that simple association measures (distance correlation, in particular) are able to capture the majority of the important feature interactions to estimate a target variable from synthetic data. However, research into other measures and more robust heuristics may be required to achieve the same results on more complex symbolic regression problems. In future work, we will continue to explore the problem of discovering meaningful building blocks from data, as the first step in the symbolic regression pipeline. Then we will focus on the later stages of symbolic regression, in which we utilize the learned FIG structure to create the optimal symbolic model. Finally, we will apply these methods to real-world healthcare data.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333-3361.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685-8690.

La Cava, W., Orzechowski, P., Burlacu, B., de França, F. O., Virgolin, M., Jin, Y., ... & Moore, J. H. (2021). Contemporary symbolic regression methods and their relative performance. *arXiv preprint arXiv:2107.14351*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Schnur, J. J., & Chawla, N. V. (2022). Information fusion via symbolic regression: A tutorial in the context of human health. *Information Fusion*.

Venezia, A. (2016). The Development of Notation in Mathematical Analysis. https://digitalcommons.lmu.edu/honors-thesis/107/