International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 1 ǁ January 2024 ǁ PP. 26-37

On Mathematics For Explainable Machine Learning: Method And Applications

YANG LIU

2

Abstract. Explainable Machine Learning (XML) aims to provide transparent and interpretable machine learning solutions that can be trusted and verified by humans. In this paper, we present various mathematical tools and techniques that can enhance the explainability of machine learning models, such as neural networks, probabilistic models, and symbolic systems. We demonstrate how mathematical concepts, such as logic, geometry, algebra, and calculus, can be used to analyze, visualize, and simplify the complex decision-making processes of machine learning models. We also propose a novel framework for measuring and comparing the explainability of different machine learning models, based on mathematical criteria and metrics. We evaluate our framework on several real-world datasets and applications, showing the benefits and challenges of XML.

Contents

Date of Submission: 22-12-2023 Date of acceptance: 03-01-2024

1. Introduction

In recent years, machine learning systems have seen tremendous growth in adoption across high-stakes domains like healthcare, finance, and transportation (see for example, [25] and [20]). However, the increasing use of complex, black-box machine learning models has given rise to issues of trust, ethics, and transparency (see for example, [23] and [1]). This has led to emerging emphasis on explainable machine learning (explainable machine learning) - an approach focused on imparting interpretability and meaning to machine learning model behaviors and predictions (see for example, [12] and [2]).

Mathematics is intrinsically linked to the goals of transparency and explainability in machine learning systems (see for example, [11] and [16]). The purpose of this theoretical research is to explore the role of mathematical concepts as key enablers for achieving model interpretability in the nascent field of explainable machine learning (see for example, [9] and [19]). Using available literature, we critically analyze established techniques like sensitivity analysis, optimization, type curves, etc. that facilitate embedding formal mathematical meaning and provable explanations within machine learning systems (see for example, [15] and [10]).

The significance of this research lies in formally delineating the mathematical underpinnings crucial for developing trustworthy and transparent machine learning models, a central motivation behind explainable machine learning research (see for example, [6] and [14]). This paper examines the limitations in current approaches and outlines recommendations to enrich integration of mathematical knowledge into explainable machine learning frameworks (see for example, [21] and [13]). Overall, we contend that mathematical rigor provides an inherent explanatory framework to elevate model interpretability, thereby addressing core objectives of explainable machine learning (see for example, [7] and [24]).

2. Literature Review

In recent years, machine learning systems have seen tremendous growth in adoption across high-stakes domains like healthcare, finance, and transportation (see for example, [20] and [5]). Explainable machine learning techniques aim to impart model transparency and interpretability, crucial for trustworthy systems deployed in high-stakes domains. Researchers have explored various mathematical concepts to enable inherent explainability (see for example, [1] and [22]).

2.1. **Mathematical Optimization.** Optimization functions allow encoding constraints and domain knowledge to produce optimized, provably sound predictions (see for example, [2] and [17]). Mathematical objectives serve as formalizations to elevate result reliability. Techniques like integer programming impart inherent transparency through factoring human-interpretable constraints (see for example, [16] and [8]). However, challenges exist in scaling such methods.

2.2. **Sensitivity Analysis.** Sensitivity analysis reveals the level of influence input variables have on model output (see for example, [19] and [18]). One approach assigns a sensitivity index to each feature revealing its explanatory power. While promising for explainability, thorough analysis on complex models can be computationally expensive (see for example, [10] and [4]).

2.3. **Formalized Mathematical Concepts.** Efforts are underway to create mathematical knowledge graphs and repositories that impart more structured conceptual information to models (see for example, [14] and [5]). However, progress is impeded by a shortage of formalized datasets and difficulties in translating informal concepts into mathematical representations (see for example, [13] and [3]).

In summary, mathematics is increasingly leveraged for inherent explainability through optimization, sensitivity analysis, and knowledge formalization. Key limitations identified include computational scalability issues and lack of structured mathematical knowledge resources. Further research must address these gaps to fully realize the promise

of mathematical explainable machine learning.II. Mathematics and explainable machine learning

Mathematics plays a central role in explainable machine learning, as it provides the theoretical foundations for many explainable machine learning methods. Optimization, graph theory, information theory, game theory, logic, and probability are some of the mathematical concepts and techniques used in explainable machine learning. For instance, optimization can be used to find the optimal explanation for a given model output, while graph theory can be used to represent the causal relationships among features and outcomes. Information theory can be used to measure the amount of information conveyed by an explanation, while game theory can be used to model the strategic interactions between the explainers and the explainees. Logic can be used to express the rules and constraints of an explanation, and probability can be used to quantify the uncertainty and confidence of an explanation.

One example of a mathematical equation related to explainable machine learning is the Shapley value, which is a game-theoretic concept that assigns a fair contribution to each feature in a model output. The Shapley value of feature i for a model f and an input x is defined as:

$$\phi_i(f,x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_x(S \cup \{i\}) - f_x(S)]$$

where F is the set of all features, S is a subset of features, and $f_x(S)$ is the model output when only the features in S are present and the rest are replaced by some baseline value. The Shapley value can be used to quantify the contribution of each feature to the model's output, enabling users to understand which features are most important for the model's predictions.

Another example of a mathematical equation related to explainable machine learning is the mutual information, which is an information-theoretic concept that measures the amount of information shared by

two random variables. The mutual information between a model output Y and an explanation E is defined as:

$$I(Y;E) = \sum_{y \in Y} \sum_{e \in E} p(y,e) \log \frac{p(y,e)}{p(y)p(e)}$$

where $p(y,e)$ is the joint probability distribution of Y and E, and $p(y)$ and $p(e)$ are the marginal probability distributions of Y and E, respectively. The mutual information can be used to quantify the informativeness of an explanation, as well as the trade-off between simplicity and accuracy.

In addition to these mathematical concepts, explainable machine learning also relies on various algorithms and techniques to generate explanations. Some common approaches include feature attribution, SHAP values, LIME, and treeExplainer. Feature attribution methods calculate the contribution of each feature to the model's output, while SHAP values assign a unique value to each feature for a specific instance, indicating its contribution to the predicted outcome. LIME generates explanations by approximating the behavior of the model using a simpler, interpretable model, such as a linear model. TreeExplainer uses decision trees to generate explanations, highlighting the most important splits and features that contributed to the prediction.

3. Trustworthiness and explainable machine learning

Trustworthiness is a critical aspect of machine learning, as it encompasses the reliability, dependability, and credibility of machine learning systems. explainable machine learning plays a vital role in enhancing the trustworthiness of machine learning by providing insights into the decision-making process and enabling users to understand why certain decisions were made. In this section, we will explore the relationship between trustworthiness and explainable machine learning, and discuss the principles and guidelines that can help ensure the trustworthiness of machine learning systems.

3.1. **Mathematical Definition of Trustworthiness.** Trustworthiness can be mathematically defined as the probability of an machine learning system behaving in a desired manner, given a specific input and context. This definition captures the idea that trustworthiness is a measure of how reliable and predictable an machine learning system is, and how well it can perform its intended tasks.

Mathematically, trustworthiness can be represented as follows:

$$T(S, I, C) = P(S(I, C) = D)$$

Where:

- $T(S, I, C)$ represents the trustworthiness of an machine learning system S, given input I and context C.
- $P(S(I, C) = D)$ represents the probability of the machine learning system producing the desired output D, given input I and context C.

3.2. **Principles of Trustworthiness and explainable machine learning.** There are several principles and guidelines related to trustworthiness and explainable machine learning. The OECD Principles on machine learning, for instance, emphasize the importance of transparency and accountability in machine learning systems. The principles state that machine learning systems should be designed in a way that respects the rule of law, human rights, democratic values, and diversity, and they should include appropriate safeguards to ensure a fair and just society.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems provides guidelines for ethical machine learning and autonomous systems. The guidelines emphasize the importance of transparency, accountability, and human oversight in machine learning systems, as well as the need to address potential biases and ensure that machine learning systems align with human values.

The explainable machine learning Principles, developed by the Explainable machine learning Foundation, provide a framework for evaluating the quality and effectiveness of explainable machine learning

methods. The principles emphasize the importance of explanation, accuracy, knowledge, fairness, privacy, and control in explainable machine learning systems. They also stress the need for explainable machine learning systems to be customizable, adaptable, and accessible to diverse users and stakeholders.

3.3. **Mathematical Theorems for Trustworthiness and explainable machine learning.** Several mathematical theorems can be used to evaluate the trustworthiness of machine learning systems and the effectiveness of explainable machine learning methods. One such theorem is the Bayesian Theorem, which states that the probability of a hypothesis (H) given evidence (E) is equal to the probability of the evidence given the hypothesis multiplied by the prior probability of the hypothesis, divided by the probability of the evidence.

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

This theorem can be applied to machine learning systems to determine the probability of a particular outcome given a specific input and context. For example, the probability of an machine learning system producing a correct diagnosis given a patient's symptoms can be calculated using Bayesian inference.

Another important theorem is the Vapnik-Chervonenkis (VC) theorem, which states that the expected generalization error of a learning algorithm is upper bounded by the sum of the training error and a term that depends on the complexity of the hypothesis space.

$$E(\text{Err}_{\text{gen}}) \leq E(\text{Err}_{\text{train}}) + O\left(\frac{\log(N)}{N}\right) \cdot H(h)$$

Where:

- $E(\text{Err}_{\text{gen}})$ represents the expected generalization error.
- $E(\text{Err}_{\text{train}})$ represents the training error.
- N represents the number of samples.
- $H(h)$ represents the entropy of the hypothesis space.

This theorem provides a theoretical bound on the generalization error of a learning algorithm, which can be useful in evaluating the trustworthiness of machine learning systems. It also highlights the tradeoff between the complexity of the hypothesis space and the generalization error.

3.4. **Evaluating Trustworthiness Using explainable machine learning.** explainable machine learning can be used to evaluate the trustworthiness of machine learning systems by providing insights into the decision-making process. One approach is to use feature attribution methods, which quantify the contribution of individual features to the model's predictions. This can help identify which features are most important for the model's decisions and whether there are any biases or errors in the decision-making process.

Another approach is to use techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to generate explanations for the model's predictions. These techniques assign a value to each feature for a specific prediction, indicating its contribution to the outcome. This can help identify which features are driving the model's decisions and whether there are any unexpected interactions between features.

In addition to these techniques, it is also important to consider the ethical and social implications of machine learning systems. This includes ensuring that the data used to train the model is representative of the population it will be applied to, and that the model is fair and unbiased. It also includes considering the potential consequences of the model's predictions, and whether they may have any negative impacts on individuals or society.

Overall, explaining machine learning systems and their decisions is a complex task that requires a combination of technical and ethical considerations. By using techniques such as feature attribution and explanation generation, and considering the ethical and social implications of machine learning systems, we can better understand how

machine learning systems make decisions and improve their trustworthiness.

4. Quantitative Aspects of Accountability and explainable machine learning

Accountability and explainable machine learning (explainable machine learning) are vital for the ethical deployment of machine learning systems. To ensure fairness and transparency, it is essential to implement mathematical frameworks that can quantify these concepts.

4.1. **Impact Assessments.** The impact of an machine learning system can be quantitatively assessed by evaluating its performance function, $P(\theta)$, where θ represents the system's parameters. This function measures the degree to which the system meets its intended objectives, taking into account factors such as accuracy, efficiency, and reliability.

$$P(\theta) = \alpha A(\theta) + \beta E(\theta) + \gamma R(\theta)$$

Here, $A(\theta)$, $E(\theta)$, and $R(\theta)$ denote the accuracy, efficiency, and reliability of the system respectively, while α, β, γ are weights representing their relative importance.

4.2. **Decision Explanations.** The explanatory power of an machine learning model can be measured by its interpretability score, $I(M)$, where M represents the model. This score quantifies how well a human user can understand the model's decisions.

$$I(M) = \delta C(M) + \varepsilon T(M)$$

Here, $C(M)$ is a measure of the model's complexity (lower complexity leads to higher interpretability), $T(M)$ is a measure of transparency (more transparency leads to higher interpretability), and δ, ε represent their relative importance.

4.3. **Fairness Principles.** Fairness in machine learning systems can be mathematically represented by a fairness function $F(D)$. Here, D

denotes decision outcomes across different demographic groups in a population.

$$F(D) = \sum |P(D_i|G = g_i) - P(D_i)|$$

This function calculates the absolute difference between decision outcomes for each group g_i compared to overall outcomes D_i. A lower value indicates greater fairness.

By integrating these mathematical frameworks into our discussion on accountability and explainable machine learning, we can better understand and quantify these complex concepts. These functions allow us to objectively assess how well an machine learning system aligns with ethical standards, thereby contributing significantly to our discussion on accountability in explainable machine learning.

In conclusion, through mathematical equations and principles such as performance functions $P(\theta)$, interpretability scores $I(M)$, and fairness functions $F(D)$, we can provide concrete metrics for assessing accountability in explainable machine learning systems. These quantitative measures enable us to make evidence-based assessments about whether an machine learning system is accountable or not a crucial step towards ensuring ethical machine learning practices.

Acknowledgment

This work is supported partly by a research grant from the Municipal Finance of Shenzhen.

Data Availability Statement

The author confirms that the data supporting the findings of this study are available within the article or its supplementary materials.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

[2] John T Baldwin. *Model theory and the philosophy of mathematical practice: Formalization without foundationalism*. Cambridge University Press, 2018.

[3] Michaela Benk and Andrea Ferrario. Explaining interpretable machine learning: Theory, methods and applications. *Methods and Applications (December 11, 2020)*, 2020.

[4] Andrea Bunt, Michael Terry, and Edward Lank. Challenges and opportunities for mathematics software in expert problem solving. *Human–Computer Interaction*, 28(3):222–264, 2013.

[5] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. Trustworthy ai. *Reflections on Artificial Intelligence for Humanity*, pages 13–39, 2021.

[6] Carlos Iván Chesnevar, Ana Gabriela Maguitman, and Ronald Prescott Loui. Logical models of argument. *ACM Computing Surveys (CSUR)*, 32(4):337–383, 2000.

[7] Changyu Deng, Xunbi Ji, Colton Rainey, Jianyu Zhang, and Wei Lu. Integrating machine learning with human knowledge. *Iscience*, 23(11), 2020.

[8] Royston Goodacre, Seetharaman Vaidyanathan, Warwick B Dunn, George G Harrigan, and Douglas B Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *TRENDS in Biotechnology*, 22(5):245–252, 2004.

[9] Lukas-Valentin Herm, Theresa Steinbach, Jonas Wanner, and Christian Janiesch. A nascent design theory for explainable intelligent systems. *Electronic Markets*, 32(4):2185–2205, 2022.

[10] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

[11] Atoosa Kasirzadeh. *Applying Mathematics to the Natural and Social World*. PhD thesis, University of Toronto (Canada), 2021.

[12] Harmanpreet Kaur. *Where are the Humans in Human-AI Interaction: The Missing Human-Centered Perspective on Interpretability Tools for Machine Learning*. PhD thesis, 2023.

[13] Eric Kuo, Michael M Hull, Ayush Gupta, and Andrew Elby. How students blend conceptual and formal mathematical reasoning in solving physics problems. *Science Education*, 97(1):32–57, 2013.

[14] Colton Ladbury, Reza Zarinshenas, Hemal Semwal, Andrew Tam, Nagarajan Vaidehi, Andrei S Rodin, An Liu, Scott Glaser, Ravi Salgia, and Arya Amini. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Translational Cancer Research*, 11(10):3853, 2022.

[15] Joong Gwang Lee, Ariamalar Selvakumar, Khalid Alvi, John Riverson, Jenny X Zhen, Leslie Shoemaker, and Fu-hsiung Lai. A watershed-scale design optimization model for stormwater best management practices. *Environmental Modelling & Software*, 37:6–18, 2012.

[16] David Leslie. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*, 2019.

[17] Hazel R Parry and Mike Bithell. Large scale agent-based modelling: A review and guidelines for model scaling. *Agent-based models of geographical systems*, pages 271–308, 2011.

[18] Hao Peng, Ruitong Zhang, Yingtong Dou, Renyu Yang, Jingyi Zhang, and Philip S Yu. Reinforced neighborhood selection guided multi-relational graph neural networks. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–46, 2021.

[19] Fredrik Rusek, Daniel Persson, Buon Kiong Lau, Erik G Larsson, Thomas L Marzetta, Ove Edfors, and Fredrik Tufvesson. Scaling up mimo: Opportunities and challenges with very large arrays. *IEEE signal processing magazine*, 30(1):40–60, 2012.

[20] Aaron Springer. *Accurate, Fair, and Explainable: Building Human-Centered AI*. University of California, Santa Cruz, 2019.

[21] Ehsan Toreini, Mhairi Aitken, Kovila PL Coopamootoo, Karen Elliott, Vladimiro Gonzalez Zelaya, Paolo Missier, Magdalene Ng, and Aad van Moorsel. Technologies for trustworthy machine learning: A survey in a socio-technical context. *arXiv preprint arXiv:2007.08911*, 2020.

[22] Thomas P Trappenberg. *Fundamentals of machine learning*. Oxford University Press, 2019.

[23] Warren J von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4):1607–1622, 2021.

[24] Thomas Wischmeyer. Artificial intelligence and transparency: opening the black box. *Regulating artificial intelligence*, pages 75–101, 2020.

[25] Xiaoge Zhang, Felix TS Chan, Chao Yan, and Indranil Bose. Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems*, 159:113800, 2022.

Email address: yliu1mat@gmail.com