International Journal of Mathematical Analysis
Vol. 18, 2024, no. 1, 11 - 19
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ijma.2024.912544

Explainable Artificial Intelligence and Mathematics: New Frontiers (and Challenges) of Research Not Only as "AppliedMath"

Massimiliano Ferrara

University Mediterranea of Reggio Calabria
&
Bocconi University, Italy

Abstract

The increasing complexity of artificial intelligence (AI) models has led to a rising demand for explainability in AI (XAI). Explainable Artificial Intelligence aims to make AI's decision-making processes transparent and understandable to humans. This paper examines the integral connection between XAI and mathematics, highlighting how mathematical principles can enhance the interpretability, transparency, and trustworthiness of AI models. We explore the mathematical foundations that underpin XAI techniques, examine case studies where mathematics has improved explainability, and propose future directions for integrating mathematics into XAI frameworks.

1 Introduction

AI systems, promoted by advanced algorithms and massive datasets, have demonstrated remarkable capabilities across various domains. However, the "black-box" nature of many AI models, especially deep learning, presents challenges in understanding their decision-making processes. Explainable AI (XAI) addresses these challenges by providing insights into how and why AI systems make certain decisions. Mathematics, being the language of precision and structure, plays a pivotal role in constructing and elucidating these explanations.

1.1 Aims and Scientific Motivation

The need for explainability in AI has grown alongside the adoption of AI in critical areas like healthcare, finance, and autonomous driving. Understanding AI decisions is essential for trust, regulatory compliance, and ethical AI deployment. Trust in AI involves knowing not only the outcomes but also the pathways and reasons leading to those outcomes.

This study aims to: (a) Identify the mathematical foundations critical to XAI; (b) Explore case studies demonstrating the application of mathematics in XAI; (c) Propose future research directions for enhancing XAI with mathematical principles.

2 Mathematical Foundations of XAI

Mathematics provides the bedrock upon which many XAI methods are built. From linear algebra and calculus to more complex fields like information theory and topology, mathematical concepts facilitate the extraction of meaningful information from AI models.

2.1 Linear Algebra and Matrix Decompositions

Linear algebra is fundamental in model interpretation, particularly in techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). These methods reduce data dimensionality while preserving variance, making it easier to visualize and interpret high-dimensional data.

2.1.1 Principal Component Analysis (PCA)

PCA transforms data by projecting it onto orthogonal vectors that maximize variance. The transformation of a dataset (X) using PCA involves computing its covariance matrix (Σ), and then deriving its eigenvalues and eigenvectors. The principal components are the eigenvectors corresponding to the largest eigenvalues.

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

$$\Sigma v = \lambda v$$

Here, v represents the eigenvectors (principal components), and λ the eigenvalues.

2.1.2 Singular Value Decomposition (SVD)

SVD generalizes PCA and decomposes a matrix into singular vectors and singular values. For a given matrix (A), SVD can be represented as:

$$A = U\Sigma V^T$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix of singular values.

2.2 Calculus and Optimization

Gradient-based optimization techniques, derived from calculus, are essential for training AI models. Understanding gradients and Hessian matrices helps in explaining how models learn from data, and in identifying critical features and decision boundaries.

2.2.1 Gradient Descent

Gradient descent minimizes a function $f(\theta)$ by iteratively moving in the direction of the steepest descent, defined by the negative gradient. The update rule is given by:

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$$

where η is the learning rate, and $\nabla f(\theta_t)$ is the gradient of the function at θ_t.

2.2.2 Hessian Matrices and Curvature

The Hessian matrix (H) of a function $f(\theta)$ at point θ is a square matrix of second-order partial derivatives, representing the local curvature:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_n \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_n^2} \end{bmatrix}$$

2.3 Information Theory

Information theory quantifies uncertainty and information gain, aiding in the development of metrics such as entropy and mutual information. These metrics are vital for feature selection and model interpretability.

2.3.1 Entropy and Information Gain

Entropy ($H(X)$) measures the uncertainty in a random variable (X):

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

Information gain measures the reduction in entropy when a dataset is split based on an attribute.

$$IG(Y|X) = H(Y) - H(Y|X)$$

2.3.2 Mutual Information

Mutual information ($I(X;Y)$) quantifies the amount of information obtained about one random variable through another:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

2.4 The Contribution of Game Theory from Our Own Perspective

Game theory is a mathematical field that studies strategic interactions between rational agents and its potential application across a wide range of disciplines, including artificial intelligence. In the context of explainable artificial intelligence, game theory can offer a fundamental approach to understanding and improving transparency in AI models.

One of the crucial aspects of game theory is the conceptualization of strategic interactions as "games," where participants make rational decisions to maximize their objectives. By applying these notions to AI explainability, we can consider the decision-making process of AI models as a game between the artificial system and human users who seek to understand its actions.

Game theory can provide a conceptual framework for analyzing the strategies used by AI models to communicate their decisions clearly and understandably. For example, through concepts like Nash equilibrium, it is possible to evaluate how AI models and human users can work together optimally to ensure effective explanations of the decisions made by the system.

Moreover, game theory can help model scenarios where the explainability of AI might conflict with other goals, such as computational efficiency or predictive performance. Through the analysis of multi-user games and strategic trade-offs, we can develop strategies to balance these different considerations and design explainable AI models that meet a range of competing requirements.

Lastly, incorporating game theory into the realm of explainable artificial intelligence can lead to new perspectives and approaches to tackling challenges related to the transparency and interpretability of artificial systems. By using fundamental concepts of game theory to analyze and optimize interactions between AI models and human users, we can promote the development of intelligent systems that are not only powerful and accurate but also understandable and acceptable to society.

2.4.1 Shapley Values

Shapley values, originating from cooperative game theory, ensure fair distribution of payoffs among players. In the context of XAI, Shapley values attribute the contribution of each feature to the overall prediction. The Shapley value for a feature (i) is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

where N is the set of all features, and $v(S)$ is the value function representing the prediction when the subset (S) of features is present.

2.4.2 Application in SHAP

SHapley Additive exPlanations (SHAP) apply Shapley values to provide consistent and verifiable feature attributions. Delve into the mathematical formulation of SHAP using the Shapley value equation above and demonstrate with an example.

3 Case Studies

To illustrate the synergy between mathematics and XAI, we consider several case studies where mathematical techniques have enhanced explainability.

3.1 LIME and SHAP

Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are popular XAI methods that rely on mathematical principles. LIME uses locally weighted linear regression to approximate a model's behavior around a specific prediction, while SHAP leverages cooperative game theory to distribute contributions of features fairly.

3.1.1 LIME

Detail the mathematical methodology behind LIME, including the optimization of local surrogates and interpretability of linear approximations. Provide a detailed example showcasing a step-by-step application of LIME to a specific prediction instance.

3.1.2 SHAP

Discuss SHAP's foundation in Shapley values from cooperative game theory. Highlight the mathematical derivation of Shapley values and their contribution to fair attribution of feature importance. Include a case study that rigorously applies SHAP to a real-world dataset, illustrating how feature contributions are computed and interpreted.

3.2 Decision Trees and Rule Extraction

Decision trees, inherently interpretable models, use recursive partitioning based on feature values to generate easily understandable rules. Techniques like Decision Tree Surrogate Models create interpretable approximations of complex models.

3.2.1 Recursive Partitioning

Explain the mathematical basis of recursive partitioning, including impurity measures like Gini impurity and entropy in the context of decision trees. Provide a case study that demonstrates the construction of a decision tree and the derivation of decision rules from the model.

$$Gini(S) = 1 - \sum_{i=1}^{n} (p_i)^2$$

3.2.2 Rule Extraction Methods

Detail methods for extracting rules from black-box models, such as model distillation and surrogate decision trees, with mathematical explanations of each approach. Include examples of rule extraction processes, illustrating the transformation of complex model outputs into human-understandable rules.

3.3 Bayesian Networks

Bayesian networks utilize probability theory to represent and reason about the dependencies among variables. These networks simplify the visualization and understanding of probabilistic relationships, aiding in the interpretability of predictions.

3.3.1 Probabilistic Graphical Models

Discuss the mathematical foundation of Bayesian networks, including concepts of conditional independence and factorization of joint distributions. Provide an example application of Bayesian networks in a specific domain, highlighting how probabilistic dependencies are modeled and interpreted.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))$$

4 Future Directions

The integration of advanced mathematical techniques into XAI is an ongoing field of research. Future work may involve:

4.1 Topological Data Analysis (TDA)

TDA applies concepts from algebraic topology to uncover the shape and structure of data. Persistent homology, a key tool in TDA, can reveal robust features that contribute to model explanations.

4.1.1 Persistent Homology

Explain persistent homology's mathematical foundation and its utility in identifying significant data features that persist across multiple scales. Include examples of how TDA has been applied to complex datasets and the insights it has provided.

4.2 Causal Inference

Mathematical techniques from causal inference can help distinguish causation from correlation in AI models, providing deeper insights into the underlying mechanisms driving predictions.

4.2.1 Causal Models

Introduce causal models and the mathematical formulation of causal relationships (e.g., do-calculus). Discuss applications in interpreting model decisions, providing examples of causal inference techniques applied to real-world AI predictions.

4.3 Information Geometry

Information geometry examines the differential-geometric structure of statistical models. This perspective can enhance our understanding of model parameter spaces and improve interpretability.

4.3.1 Geometric Understanding of Models

Explain the mathematical principles of information geometry, including divergence measures and their role in interpreting statistical models. Provide examples of how information geometry can be applied to examine and understand deep learning models.

5 Conclusions

Mathematics serves as a critical pivot for the development of explainable AI. By leveraging mathematical principles, we can create more transparent, interpretable, and trustworthy AI systems. As AI continues to evolve, the collaboration between XAI and mathematics will be essential in bridging the gap between complex models and human understanding. In the near future the horizontal combination between these two Knowledge driver will produce new platforms through which to promote latest generation Decision Support Systems and Expert Systems where the dual (in the philosophical interpretation of ancient Greek) Man-Machine they are two distinct sets that can develop a union of intersections for a new era of humanity.

References

[1] Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Leanpub, 2019.

[2] Lundberg, S. M., and Lee, S.-I., A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 2017.

[3] Ribeiro, M. T., Singh, S., and Guestrin, C., ”Why Should I Trust You?” Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, 2016. https://doi.org/10.1145/2939672.2939778

[4] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[5] Shapley, L. S., A Value for N-Person Games, *Annals of Mathematics Studies, Contributions to the Theory of Games (AM-28), Volume II*, 1953.

[6] Pearl, J., *Causality: Models, Reasoning, and Inference*, (2nd ed.), Cambridge University Press, 2009. https://doi.org/10.1017/cbo9780511803161

[7] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, 2016.

[8] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[9] Lee, J. A., and Verleysen, M., *Nonlinear Dimensionality Reduction*, Springer, 2007. https://doi.org/10.1007/978-0-387-39351-3

[10] Edelsbrunner, H., and Harer, J., *Computational Topology: An Introduction*, American Mathematical Society, 2010.
https://doi.org/10.1090/mbk/069

Received: June 15, 2024; Published: August 2, 2024