EMBA Program
MB-511

Introduction to Explainability and Fairness in Machine Learning

Instructor
Anant Prakash Awasthi

XAI (Adding explainability to Machine Learning)

EMBA Program
MB-511

Explainable AI (XAI) refers to the development of artificial intelligence systems whose decisions and behaviors can be easily understood and interpreted by humans. While AI models like deep neural networks have demonstrated remarkable performance in various tasks, they often function as "black boxes," meaning it's difficult for users to understand why they make specific decisions. This lack of transparency can be a significant barrier in critical applications, especially those where trust, accountability, and regulatory compliance are essential.

By making AI systems more transparent and interpretable, Explainable AI aims to enhance trust, facilitate collaboration between humans and machines, improve decision-making processes, and ensure accountability in critical domains like healthcare, finance, criminal justice, and autonomous vehicles.

Why XAI?

EMBA Program
MB-511

Traditional AI models, particularly those built on complex machine learning algorithms, often operate as black boxes. These methods ingest data, process it through intricate layers, and produce an output, a prediction, classification, or recommendation, without revealing its reasoning.

Explainable AI (XAI) comes in, a field dedicated to demystifying the inner workings of intelligent systems and building trust in their applications.

Why XAI (Issues with Traditional Algorithms) ?

EMBA Program
MB-511

- Lack of Trust: If users cannot understand how an AI system makes a decision, they may hesitate to trust its recommendations. This is particularly critical in high-stakes domains like healthcare or finance.
- Debugging and Improvement: When an AI model produces an erroneous outcome, a lack of explainability hinders efforts to identify the root cause and rectify the issue.
- Bias and Fairness: AI systems are susceptible to inheriting biases in the data they are trained on. XAI techniques can help uncover these biases and mitigate their impact.
- Regulatory Concerns: As AI becomes more pervasive, regulations demanding transparency and accountability will likely emerge. XAI paves the way for responsible AI development that adheres to ethical and legal frameworks.

XAI (Adding explainability to Machine Learning)

EMBA Program
MB-511

Explainable AI seeks to address this challenge by designing AI systems that provide explanations for their decisions in a human-understandable manner. These explanations can take various forms, such as:

- Feature importance: Highlighting which features or variables had the most significant influence on the AI's decision.
- Model visualization: Representing the AI model's internal structure and decision-making process in a comprehensible format, such as decision trees or flowcharts.
- Local explanations: Providing explanations for specific predictions or decisions made by the AI, rather than offering a global overview of the entire model.
- Natural language explanations: Expressing the rationale behind AI decisions in plain language, making it easier for non-experts to understand.
- Counterfactual explanations: Describing how changes in input data could lead to different outcomes, helping users grasp the sensitivity of the AI's decisions to different factors.

XAI Applications (Domain Specific)

EMBA Program
MB-511

XAI holds immense potential across various sectors where AI is making significant inroads:

- Healthcare: XAI can explain why a diagnostic tool flagged a particular patient for further examination, fostering better communication between doctors and patients.
- Finance: Explainable loan approval models can improve transparency for loan applicants and ensure fair lending practices.
- Criminal Justice: XAI can ensure that AI models are not biased against certain demographics when applied to risk assessment or recidivism prediction.
- Autonomous Vehicles: Understanding how self-driving cars make decisions in critical situations is paramount for building trust and ensuring safety.

These are just a few examples, and as AI continues to permeate various aspects of our lives, the demand for XAI solutions will only grow.

XAI Techniques - Model-Agnostic Techniques

EMBA Program
MB-511

These methods work with any model, regardless of its internal structure. Techniques include:

- Feature Importance: This approach highlights the features in the data that contribute most significantly to the model's output.
- Local Interpretable Model-Agnostic Explanations (LIME): LIME creates a simplified explanation for an individual prediction by approximating the original model locally around that specific instance.
- SHapley Additive exPlanations (SHAP): SHAP assigns a contribution score to each feature based on its impact on the model's prediction.

XAI Techniques - Model-Specific Techniques

EMBA Program
MB-511

These techniques leverage the specific architecture of the model to provide explanations:

- Decision Trees: Decision trees' strength is their inherent interpretability, as the decision-making process is explicitly encoded in the tree structure.
- Rule-Based Models: Like decision trees, rule-based models represent knowledge in human-readable rules, making them inherently explainable.

XAI – Tools and Platforms

EMBA Program
MB-511

- LIME (Local Interpretable Model-agnostic Explanations): A popular model-agnostic technique that creates simplified, interpretable models to explain individual predictions.
- SHAP (SHapley Additive exPlanations): Another powerful model-agnostic approach based on game theory. SHAP provides feature importance scores, indicating how each feature influences the model's output.
- AIX360 (AI Explainability 360): A comprehensive open-source toolkit from IBM Research that offers diverse explainability algorithms and fairness metrics for evaluating and mitigating bias.
- InterpretML: A Microsoft-developed toolkit that offers both model-agnostic and model-specific explainability techniques, with a focus on interpretable glass-box models.
- Skater: A Python library that provides a unified interface for various XAI techniques, facilitating comparison and selection of the most appropriate tools.

XAI – References

EMBA Program
MB-511

- Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective (https://arxiv.org/abs/2401.04374)
- Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (https://arxiv.org/abs/1910.10045)
- The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations (https://arxiv.org/abs/2307.00364)
- Axe the X in XAI: A Plea for Understandable AI (https://arxiv.org/abs/2403.00315)
- Explainable Artificial Intelligence: Precepts, Methods, and Opportunities for Research in Construction (https://arxiv.org/abs/2211.06579)
- Explainable AI (XAI) — A guide to 7 Packages in Python to Explain Your Models (Link)

Fairness in Machine Learning

EMBA Program
MB-511

Fairness in machine learning refers to the principle of ensuring that the decisions and predictions made by machine learning models do not systematically favor or discriminate against particular individuals or groups based on sensitive attributes such as race, gender, age, or socioeconomic status. Fairness aims to prevent biases and discrimination from being perpetuated or amplified through automated decision-making systems.

Fairness in machine learning can be understood through various perspectives and metrics, including:

- Individual Fairness: Ensuring that similar individuals receive similar predictions or treatment from the model regardless of their sensitive attributes.
- Group Fairness: Ensuring that the model's predictions or decisions are equitable across different groups defined by sensitive attributes. Common group fairness criteria include demographic parity, equal opportunity, and disparate impact.
- Intersectional Fairness: Considering the intersections of multiple sensitive attributes (e.g., race and gender) to avoid discrimination against individuals belonging to multiple marginalized groups.
- Procedural Fairness: Ensuring that the processes involved in developing, deploying, and evaluating machine learning models are transparent, accountable, and inclusive.

RUAI – References

EMBA Program
MB-511

- Responsible Artificial Intelligence -- from Principles to Practice (https://arxiv.org/abs/2205.10785)
- Responsible Artificial Intelligence: A Structured Literature Review (https://arxiv.org/abs/2403.06910)
- Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence (https://arxiv.org/abs/2312.01555)
- Towards organizational guidelines for the responsible use of AI (https://arxiv.org/abs/2001.09758)
- Responsible AI by Design in Practice (https://arxiv.org/abs/1909.12838
- Example Notebooks - https://fairlearn.org/v0.10/auto_examples/index.html

EMBA Program
MB-511

Have a question?

Feel Free to Reach out at

- +91-88846-52929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)