Faculty of Mathematics and Computer Science

Machine learning course (ML)

Explainability in Machine Learning

Murariu Tudor Cristian

Department of Computer Science, Babeş-Bolyai University
1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania
E-mail: tudor.murariu@stud.ubbcluj.ro

Machine learning and Deep Learning models, particularly complex ones such as neural networks, are often perceived as ”black boxes”, making it difficult for humans to understand their decision-making processes.

Explainability in ML refers to the ability to describe, in understandable terms, how a model makes decisions. Explainable Artificial Intelligence aims to make machine learning models more interpretable without sacrificing their performance.

There are two primary types of ML explainability:

1. Global Explainability: Understanding how a model behaves across all inputs. This gives an overall understanding of the model's decision logic.
2. Local Explainability: Understanding how the model arrived at a specific decision for a given input. Techniques like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (Shapley Additive Explanations) are commonly used for this.

Explainability is critical for several reasons:

1. Trust and Transparency: Users and stakeholders need to trust that a model's decisions are reasonable and unbiased.
2. Regulatory Compliance: In fields like finance and healthcare, regulations may require explanations for automated decisions that affect individuals.
3. Debugging and Model Improvement: Understanding how the model makes predictions can help developers find issues, mitigate biases, and improve model performance.
4. Learning and Comprehension: Understanding how a model makes predictions can significantly enhance the learning process for students and researchers.

[1] [2] [5] [4] [3]

References

[1] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103, doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[2] Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317. doi:https://doi.org/10.1613/jair.1.12228.

[3] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1312. doi:https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1312. [Correction added on 11 June 2019: "explainabilty" corrected to "explainability" in the title.].

[4] Jansen, F.J., Monteiro, M.d.S., 2020. Do ml experts discuss explainability for ai systems? a discussion case in the industry for a domain-specific solution. arXiv preprint arXiv:2002.12450 , 7URL: https://doi.org/10.48550/arXiv.2002.12450. presented at IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20).

[5] Sarhan, M., Layeghy, S., Portmann, M., 2022. Evaluating standard feature sets towards increased generalisability and explainability of ml-based network intrusion detection. Big Data Research 30, 100359. URL: https://www.sciencedirect.com/science/article/pii/S2214579622000533, doi:https://doi.org/10.1016/j.bdr.2022.100359.