interpretability

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

The interpretability of a machine learning (ML) method is the extent to which a human user can comprehend its behavior. Interpretability is closely related to predictability: the user should be able to anticipate the predictions of the method on a test set. Interpretability differs from explainability, which concerns understanding with the help of provided explanations.

Definition

Consider a training set of monthly sales and a straight line fitted through it. Reading next month's prediction off that line takes a slope, an intercept and one multiplication, and a user can carry the computation out on paper. The same prediction delivered by a deep net arrives with no account a user could follow. Interpretability is what separates the two.

Interpretability is often defined, informally, as the ability to explain, or to present in understandable terms, to a human (Doshi-Velez and Kim, 2017). The international terminology standard for artificial intelligence (AI) does not define interpretability (International Organization for Standardization and International Electrotechnical Commission, 2022). Its closest notion is predictability, the property of an artificial intelligence system (AI system) that enables reliable assumptions by users about the predictions delivered by a machine learning (ML) method (International Organization for Standardization and International Electrotechnical Commission, 2022, Sect. 3.5.8 and 5.15.7).

An ML method is interpretable for a human user if they can comprehend the computational process of the method. Comprehension of a computation cannot be observed directly, so interpretability is judged through predictability: the user anticipates the predictions the method delivers, and the evaluations proposed in the literature measure how often those anticipations are right (Chen et al., 2018; Colin et al., 2022; Doshi-Velez and Kim, 2017; Hase and Bansal, 2020). Predictability is the weaker of the two: the terminology standard notes that a user may rely on the predictions of an AI system without being able to explain the factors behind them (International Organization for Standardization and International Electrotechnical Commission, 2022, Sect. 5.15.7). A method whose computation the user can follow is predictable to that user; a method whose predictions the user anticipates need not be one whose computation they could carry out.

Fig. 1 depicts a training set and a test set. The training set is used by two different methods that learn hypotheses $\hat{h}$ and $\hat{h}'$. The ML method producing the hypothesis $\hat{h}$ is interpretable to a human user familiar with the concept of a linear map. Since $\hat{h}$ is a linear map, the user can anticipate the predictions of $\hat{h}$ on the test set. In contrast, the ML method producing $\hat{h}'$ is less interpretable: the behavior of $\hat{h}'$ deviates from the user's expectations.

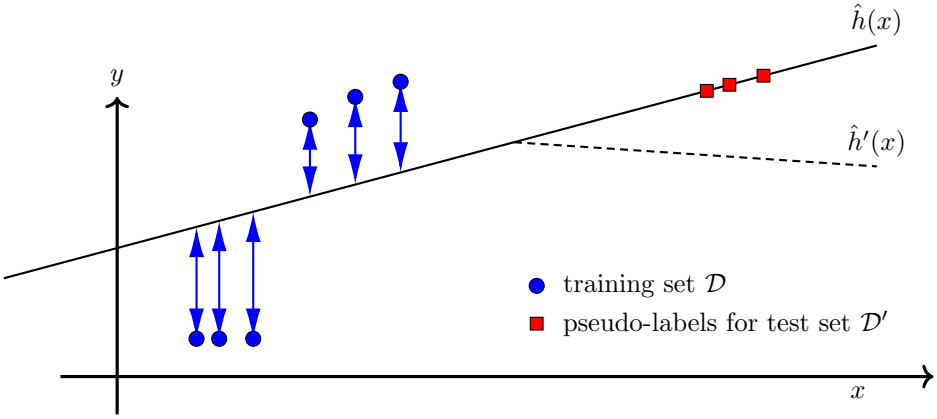

Figure 1: Predictability, the weaker half: a human user provides pseudo-labels (squares) for the test set $\mathcal{D}'$ by extrapolating the linear trend of the training set $\mathcal{D}$ (circles). The predictions of the hypothesis $\hat{h}$ agree with these pseudo-labels, while the predictions of another hypothesis $\hat{h}'$ deviate from them. The user anticipates outputs here without carrying out the computation, which comprehending the method would additionally require

The AI Risk Management Framework of the US National Institute of Standards and Technology distinguishes interpretability from explainability (National Institute of Standards and Technology, 2023), the property pursued by explainable artificial intelligence (XAI) (Gunning and Aha, 2019). In contrast to interpretability, explainability requires that an explanation is provided along with each prediction. These explanations may take the form of saliency maps or reference examples from the training set. A challenge for XAI methods is that a computed explanation can be unfaithful to the learned hypothesis and thereby mislead the user (Rudin, 2019).

Interpretability can also be obtained by decomposing a learned hypothesis into components that are themselves interpretable (Lipton, 2018). The hypothesis $\hat{h}$ in Fig. 1 decomposes into a slope and an intercept: the slope states how the prediction changes with the feature. A trained deep net lacks such a decomposition, since the activation of an individual neuron typically does not correspond to a human-interpretable concept. Mechanistic interpretability aims to recover interpretable components from a trained deep net (Olah et al., 2020; Elhage

et al., 2021) by decomposition methods such as sparse coding of activations (Cunningham et al., 2024).

Cross References

explainability, XAI, interpretable ML, trustworthy AI, regularization, LIME, mechanistic interpretability, transparency

References

1. Chen et al. (2018). Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In: Proc. 35th Int. Conf. Mach. Learn., pp. 883–892.

2. Colin et al. (2022). What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. In: Adv. Neural Inf. Process. Syst., pp. 2832–2845.

3. Cunningham et al. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. In: 12th Int. Conf. Learn. Representations.

4. Elhage et al. (2021). A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread.

5. Gunning and Aha (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), pp. 44–58.

6. International Organization for Standardization and International Electrotechnical Commission (2022). ISO/IEC 22989:2022 — Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. International Standard, 1st edition. URL: https://www.iso.org/standard/74296.html.

7. Lipton (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), pp. 31–57.

8. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. URL: https://doi.org/10.6028/NIST.AI.100-1.

9. Olah et al. (2020). Zoom In: An Introduction to Circuits. Distill, 5(3).

10. Doshi-Velez and Kim (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. URL: https://arxiv.org/abs/1702.08608.

11. Hase and Bansal (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?. In: Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, pp. 5540–5552.

12. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.