Interpretierbarkeit

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Die Interpretierbarkeit einer machine learning (ML)-Methode ist das Ausmaß, in dem ein menschlicher Benutzer den von ihr ausgeführten Berechnungsprozess verstehen kann. Interpretierbarkeit ist eng mit der Vorhersagbarkeit verbunden: Der Benutzer sollte in der Lage sein, die Vorhersagen der Methode auf einem Testdatensatz zu antizipieren. Interpretierbarkeit unterscheidet sich von der Erklärbarkeit, die das Verständnis spezifischer Vorhersagen mithilfe bereitgestellter Erklärungen betrifft.

Definition

Man betrachte einen Trainingsdatensatz monatlicher Verkaufszahlen und eine daraus gelernte lineare Hypothese. Die Berechnung der Vorhersage für den nächsten Monat aus dieser Hypothese erfordert eine Steigung, einen Achsenabschnitt und eine Multiplikation, und ein Benutzer kann die Berechnung auf Papier durchführen. Bei d Merkmale hat die angepasste lineare Abbildung ein Gewicht pro Merkmal, und die Vorhersage erfordert d Multiplikationen, immer noch eine Berechnung, die der Benutzer durchführen kann. Dieselbe Vorhersage, die von einem tiefes Netz geliefert wird, kommt ohne eine nachvollziehbare Erklärung für den Benutzer an. Interpretierbarkeit ist das, was die beiden trennt.

Eine machine learning (ML)-Methode ist für einen menschlichen Benutzer interpretierbar, wenn dieser den Berechnungsprozess der Methode verstehen kann. Der internationale Terminologiestandard für artificial intelligence (AI) definiert Interpretierbarkeit nicht (International Organization for Standardization and International Electrotechnical Commission, 2022). Ihr nächstverwandter Begriff ist die Vorhersagbarkeit, die Eigenschaft eines KI-Systems, die es Benutzern ermöglicht, zuverlässige Annahmen über die von einer ML-Methode gelieferten Vorhersagen zu treffen (International Organization for Standardization and International Electrotechnical Commission, 2022, Abschn. 3.5.8 und 5.15.7).

Das Verständnis einer Berechnung kann nicht direkt beobachtet werden, daher wird Interpretierbarkeit durch Vorhersagbarkeit beurteilt: Der Benutzer antizipiert die Vorhersagen, die die Methode liefert, und die in der Literatur vorgeschlagenen Evaluationen messen, wie oft diese Antizipationen richtig sind (Chen et al., 2018; Colin et al., 2022; Doshi-Velez and Kim, 2017; Hase and Bansal, 2020). Vorhersagbarkeit ist schwächer als Interpretierbarkeit: Eine Methode,

deren Berechnung der Benutzer nachvollziehen kann, ist für diesen Benutzer vorhersagbar, aber eine Methode, deren Vorhersagen der Benutzer antizipiert, muss nicht eine sein, deren Berechnung er durchführen könnte. Der Terminologiestandard macht denselben Punkt: Ein Benutzer kann sich auf die Vorhersagen eines KI-Systems verlassen, ohne verstehen zu können, wie die Methode diese erzeugt hat (International Organization for Standardization and International Electrotechnical Commission, 2022, Abschn. 5.15.7).

Abb. 1 zeigt einen Test der Vorhersagbarkeit. Ein Trainingsdatensatz wird von zwei verschiedenen Methoden verwendet, die Hypothesen $\widehat{h}$ und $\widehat{h}'$ lernen. Der Benutzer antizipiert die Vorhersagen auf einem Testdatensatz, indem er für jeden seiner Datenpunkte ein Pseudo-Label bereitstellt, durch visuelle Extrapolation des linearen Trends des Trainingsdatensatz. Die Vorhersagen von $\widehat{h}$ stimmen mit diesen Antizipationen überein: Die Methode, die $\widehat{h}$ erzeugt, ist für den Benutzer vorhersagbar. Die Vorhersagen von $\widehat{h}'$ weichen davon ab: Die Methode, die $\widehat{h}'$ erzeugt, besteht den Test der Vorhersagbarkeit nicht und kann daher keine sein, deren Berechnung der Benutzer versteht. Die Übereinstimmung allein begründet keine Interpretierbarkeit, da der Benutzer die Vorhersagen von $\widehat{h}$ antizipierte, ohne deren Berechnung durchzuführen. Die Methode, die $\widehat{h}$ erzeugt, ist dennoch auch für einen Benutzer interpretierbar, der mit dem Konzept einer lineare Abbildung vertraut ist: Ein solcher Benutzer könnte die Berechnung von $\widehat{h}$ verstehen.

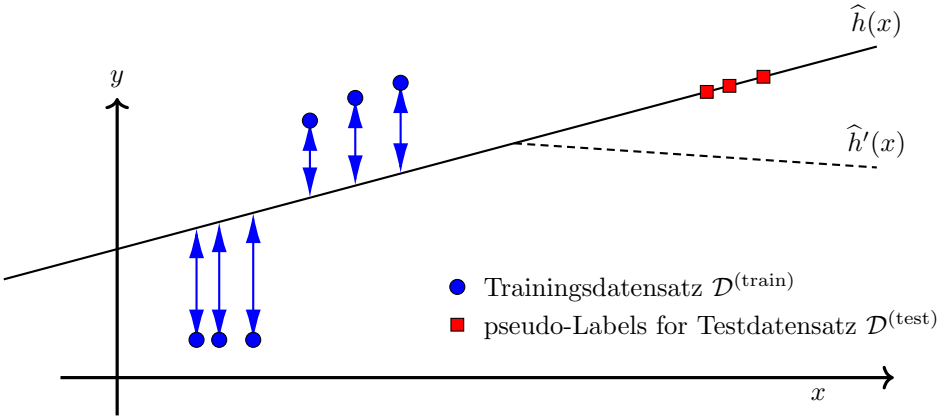


Abbildung 1: Ein menschlicher Benutzer stellt Pseudo-Labels (gefüllte rote Quadrate) für das Testdatensatz $\mathcal{D}^{(\text{test})}$ bereit, indem er den linearen Trend des Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ (gefüllte blaue Kreise) visuell extrapoliert; die vertikalen Doppelpfeile markieren die Vorhersage-Fehler von $\widehat{h}$ auf dem Trainingsdatensatz. Die Vorhersagen der Hypothese $\widehat{h}$ stimmen mit diesen Pseudo-Labels überein, während die Vorhersagen einer anderen Hypothese $\widehat{h}'$ davon abweichen. Der Benutzer antizipiert hier die Vorhersagen, ohne die Berechnung (eine Steigung, einen Achsenabschnitt und eine Multiplikation) durchzuführen, was zum Verständnis der Methode erforderlich wäre.

Neben dem schrittweisen Nachvollziehen der Berechnung, wie bei der lineare Abbildung des Verkaufsbeispiels, kann ein Benutzer eine gelernte Hypothese durch ihre Zerlegung in Komponenten verstehen, die selbst interpretierbar sind (Lipton, 2018). Die Hypothese $\widehat{h}$ in Abb. 1 zerlegt sich in eine Steigung und einen Achsenabschnitt: Die Steigung gibt an, wie sich die Vorhersage mit dem Merkmal ändert. Ein trainiertes tiefes Netz entbehrt einer solchen Zerlegung, da die Aktivierung eines einzelnen Neuron typischerweise keinem menschlich interpretierbaren Konzept entspricht. Mechanistic interpretability zielt darauf ab, interpretierbare Komponenten aus einem trainierten tiefes Netz wiederherzustellen (Elhage et al., 2021; Olah et al., 2020) durch Zerlegungsmethoden wie Sparse Coding von Aktivierungen (Cunningham et al., 2024).

Das AI Risk Management Framework des US National Institute of Standards and Technology unterscheidet Interpretierbarkeit von Erklärbarkeit (National Institute of Standards and Technology, 2023), der Eigenschaft, die von erklärbare künstliche Intelligenz (XAI) angestrebt wird (Gunning and Aha, 2019). Im Gegensatz zur Interpretierbarkeit erfordert die Erklärbarkeit, dass mit jeder Vorhersage eine Erklärung bereitgestellt wird. Diese Erklärungen können die Form von Saliency Maps oder Referenzbeispielen aus dem Trainingsdatensatz annehmen. Eine Herausforderung für XAI-Methoden besteht darin, dass eine berechnete Erklärung der gelernten Hypothese untreu sein und den Benutzer

dadurch irreführen kann. Eine interpretierbares maschinelles Lernen-Methode vermeidet dies: Die Beschreibung ihrer Berechnung ist die Berechnung selbst, sodass die Beschreibung ihr nicht untreu sein kann (Rudin, 2019).

Siehe auch: Erklärbarkeit, XAI, interpretierbares maschinelles Lernen, vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI), Regularisierung, local interpretable model-agnostic explanations (LIME), mechanistic interpretability, Transparenz.

Querverweise

Literatur

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 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.