Erklärbarkeit

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

Erklärbarkeit ist die Eigenschaft einer maschinelles Lernen (ML)-Methode, deren Erklärungen es einem menschlichen Nutzer ermöglichen, die gelieferten Vorhersagen vorherzusehen. Die Methode liefert zu jeder Vorhersage eine Erklärung, und diese Erklärung ist wirksam, wenn sie dem Nutzer erlaubt, die Vorhersage für einen Datenpunkt vorherzusehen; das Verstehen der Erklärung ist für diesen Test nicht erforderlich. Erklärbarkeit ist immer relativ zu einem bestimmten Nutzer oder einer Nutzergruppe: Eine Erklärung kann für den einen wirksam und für den anderen nutzlos sein. Erklärbarkeit kann gemessen werden, indem die Vorhersagen einer gelernten Hypothese mit den Erwartungen eines Nutzers vor und nach Erhalt einer Erklärung verglichen werden.

Definition

Eine Radiologin, der gezeigt wird, welche Pixel eines Scans einen Bild-Klassifikator zu seiner Vorhersage geführt haben, kann beim nächsten Scan vorhersagen, was der Klassifikator sagen wird. Eine maschinelle Lernen-Methode ist für einen menschlichen Nutzer erklärbar, wenn die von ihr gelieferten Erklärungen es dem Nutzer ermöglichen, die gelieferten Vorhersagen vorherzusehen (Colin et al., 2022; Jung and Nardelli, 2020). Erklärbarkeit umfasst damit den Begriff der Erklärung: Zu jeder Vorhersage wird eine Erklärung für diese spezifische Vorhersage geliefert (siehe Abb. 1), etwa die Merkmal-Werte, die am meisten dazu beigetragen haben, oder bei Bild-Daten die relevanten Pixel. In der Abbildung ist die Erklärung für die Vorhersage an einem Datenpunkt mit Merkmalsvektor $\mathbf{x}$ eine lineare Abbildung, die die gelernte Hypothese in der Nähe von $\mathbf{x}$ approximiert. Die Definition verlangt das Vorliegen einer Erklärung, nicht einer bestimmten: Erklärbarkeit ist nachgewiesen, wenn eine Erklärung existiert, die funktioniert, und eine Erklärung, der der Nutzer nicht folgen kann, zeigt nicht, dass die Methode unerklärbar ist. Erklärbarkeit ist relativ zu einem bestimmten Nutzer oder einer Nutzergruppe: Dieselbe gelernte Hypothese kann für einen Nutzer erklärbar und für einen anderen undurchschaubar sein.

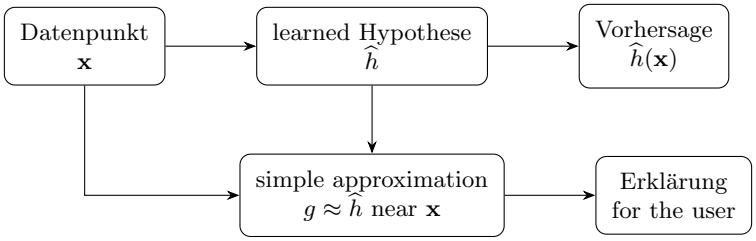


Abbildung 1: Erklärbarkeit bezieht sich auf Erklärungen. Die gelernte Hypothese $\widehat{h}$ liefert die Vorhersage $\widehat{h}(\mathbf{x})$ für einen Datenpunkt mit Merkmalsvektor $\mathbf{x}$. Eine Erklärung wird aus einer einfachen Hypothese g konstruiert, z. B. einer lineare Abbildung, die $\widehat{h}$ in der Nähe von $\mathbf{x}$ approximiert. Die Erklärung wird dem Nutzer zusammen mit der Vorhersage geliefert.

Die Nützlichkeit einer Erklärung kann daran gemessen werden, wie sehr sie den Nutzer in die Lage versetzt, die Vorhersagen auf einem kuratierten Testdatensatz vorherzusehen: Der Nutzer gibt für jeden Datenpunkt im Testdatensatz eine Erwartung an. Wenn der Nutzer die Erklärungen versteht, sollten diese Erwartungen den Vorhersagen der gelernten Hypothese nahekommen (Colin et al., 2022). Eine solche Maßzahl wird als Vorhersagbarkeit bezeichnet (International Organization for Standardization and International Electrotechnical Commission, 2022, Abschn. 5.15.7). Wie bei der Interpretierbarkeit ist die Vorhersagbarkeit schwächer als das Verstehen der Erklärungen: Das Vorhersehen der Vorhersagen erfordert nicht das Verstehen der gegebenen Erklärungen.

Eine zweite Maßzahl dieser Vorhersagbarkeit verwendet die bedingte Differentialentropie, die für Zufallsvariablen (ZV) definiert ist. Wird der Datenpunkt als Realisierung einer ZV modelliert, so sind auch die Vorhersage und die Erwartung des Nutzers Zufallsvariablen, und ihre gemeinsame Wahrscheinlichkeitsverteilung ergibt die bedingte Differentialentropie der Vorhersagen gegeben die Erwartungen (Jung and Nardelli, 2020, Abschn. II). Die bedingte Differentialentropie quantifiziert die Unsicherheit, die über die Vorhersagen verbleibt, wenn die Erwartungen bekannt sind; ein kleinerer Wert bedeutet also, dass die Erwartungen die Vorhersagen stärker bestimmen (Cover and Thomas, 2006, Abschn. 8.4). In der Praxis ist diese bedingte Differentialentropie unbekannt und muss durch einen Schätzer ersetzt werden. Ein solcher Schätzer unterteilt die Vorhersagen und die Erwartungen auf einem Testdatensatz in Bins und nimmt den Anteil der Test-Datenpunkte, die in jedes Bin-Paar fallen, als gemeinsame Wahrscheinlichkeitsverteilung. Eine verwandte Maßzahl, die wechselseitige Information zwischen einer ausgewählten Teilmenge der Merkmale und der Vorhersage, wird durch den Erklärer von Chen et al. (2018), Abschn. 2.1 maximiert.

Eine Erklärung erhöht die Erklärbarkeit, wenn der Nutzer damit argumentieren kann. Abb. 2 illustriert dies für einen Nutzer, der in Begriffen von lineare Abbildungen denkt und die Vorhersagen einer undurchsichtigen Hypothese (Gaussian Kernel-Ridge-Regression, eine Kernel-Methode) vorhersieht. Ohne Erklärungen antizipiert der Nutzer mit einer einzigen lineare Abbildung, die an

einige Vorhersagen angepasst ist, und die Erwartungen weichen stark von den Vorhersagen ab. Die Erklärung für einen Test- Datenpunkt mit Merkmal-Wert x_0 ist eine lokale lineare Approximation der Hypothese um diesen Punkt (vgl. lokale interpretierbare modellunabhängige Erklärungen (LIME)), eine Steigung und ein Achsenabschnitt, die der Nutzer bei x_0 auswertet. Mit dieser Erklärung stimmen die Erwartungen mit den Vorhersagen fast exakt überein. Das Beispiel setzt voraus, dass der Nutzer weiß, wie eine lineare Approximation anzuwenden ist: Die Erwartungen verbessern sich nur, weil der Nutzer die Erklärung für einen Datenpunkt auswerten kann. Ein Nutzer, der das nicht kann, gewinnt nichts daraus.

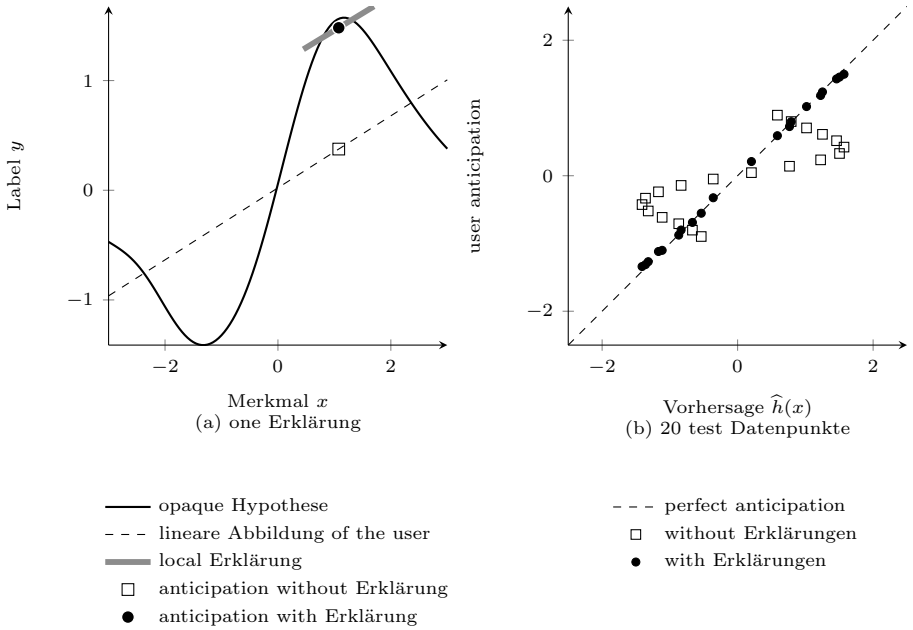


Abbildung 2: Eine undurchsichtige Hypothese $\widehat{h}$ (Gaussian Kernel-Ridge-Regression, durchgezogene Kurve) und ein Nutzer, der in Begriffen von lineare Abbildungen denkt. (a) Ohne Erklärungen antizipiert der Nutzer mit einer einzigen lineare Abbildung, die an einige Vorhersagen angepasst ist (gestrichelte Linie), was beim Test- Datenpunkt x_0 weit danebenliegt (offenes Quadrat). Die Erklärung für x_0 ist eine lokale lineare Approximation von $\widehat{h}$ (dicker grauer Abschnitt); der Nutzer wertet sie bei x_0 aus und antizipiert fast genau (gefüllter Kreis). (b) Dieselben beiden Erwartungen gegenüber den Vorhersagen für 20 der 40 Datenpunkte des Testdatensatz (jeweils jeden zweiten). Ohne Erklärungen (offene Quadrate) weichen sie stark ab; mit der lokalen linearen Approximation als Erklärung (gefüllte Kreise) folgen sie der Diagonalen der idealen Antizipation. Die geschätzte bedingte Differentialentropie der Vorhersagen gegeben die Erwartungen, über alle 40, sinkt von 1.22 auf 0.09 Bit. Daten erzeugt von `pythondemos/explainability.py`

Eine Erklärung kann auch in die Irre führen. Eine berechnete Erklärung kann der gelernten Hypothese untreu sein, sodass Erwartungen, die auf ihr beruhen, mit hoher Sicherheit falsch sind (Rudin, 2019). Saliency Maps, Intensitätskarten über die Pixel eines Eingabebildes (siehe erklärbare künstliche Intelligenz (XAI)), ändern sich bei manchen Methoden kaum, wenn das trainierte tiefes Netz durch eines mit zufälligen Gewichte ersetzt wird (Adebayo et al., 2018, Abschn. 3). Der Vorhersagbarkeitstest erkennt eine solche Erklärung: Erwartungen, die daraus abgeleitet werden, weichen von den Vorhersagen ab.

Die internationale Terminologienorm für Künstliche Intelligenz (KI) definiert Erklärbarkeit als die Eigenschaft eines KI-System, wichtige Faktoren, die seine Ergebnisse beeinflussen, so auszudrücken, dass Menschen sie verstehen können. Eine Erklärung soll die Frage beantworten, warum das KI-System ein Ergebnis geliefert hat, ohne zu behaupten, dass das Ergebnis optimal war (International Organization for Standardization and International Electrotechnical Commission, 2022, Abschn. 3.5.7). Das AI Risk Management Framework des US National Institute of Standards and Technology (NIST) verknüpft Erklärbarkeit mit einer Darstellung der Mechanismen, die dem Betrieb eines KI-System zugrunde liegen. NIST reserviert Interpretierbarkeit für die Bedeutung der Ausgabe im Kontext des vorgesehenen Zwecks (National Institute of Standards and Technology, 2023). Die beiden Dokumente teilen die Begriffe also unterschiedlich auf. Die ausgedrückten Faktoren, die ein Ergebnis beeinflussen und die die Terminologienorm unter Erklärbarkeit fasst, sind dem NIST-Begriff der Interpretierbarkeit nahe; der NIST-Begriff der Erklärbarkeit entspricht stattdessen eher der Interpretierbarkeit als Verstehen der Berechnung der Methode.

Regulatorisch wird Erklärbarkeit als Bestandteil der Transparenz behandelt. Die KI-Verordnung verlangt, dass Hochrisiko-KI-Systeme hinreichend transparent sind, damit Betreiber die Ausgabe eines Systems interpretieren und angemessen nutzen können (European Parliament and Council of the European Union, 2024, Art. 13(1)). Für einzelne automatisierte Entscheidungen berechtigt das Recht auf Erklärung eine betroffene Person zu einer klaren und verständlichen Darstellung der Rolle, die ein KI-System bei der Entscheidung gespielt hat (European Parliament and Council of the European Union, 2024, Art. 86). Der Colorado Automated Decision-Making Technology Act von 2026 verlangt, dass der Betreiber dem betroffenen Verbraucher eine verständliche Beschreibung der Entscheidung und der Rolle der Technologie darin gibt, zusammen mit der Möglichkeit, weitere Informationen anzufordern (Colorado General Assembly, 2026, Abschn. 6-1-1704(3)). Was als automatisierte Entscheidung gilt, ist weit gefasst: Der Europäische Gerichtshof entschied, dass eine Auskunftei, die die Fähigkeit einer Person zur Begleichung künftiger Zahlungen berechnet, bereits eine solche trifft, wenn ein Dritter sich maßgeblich auf diesen Wert stützt (Court of Justice of the European Union, 2023).

Siehe auch: Erklärung, Interpretierbarkeit, XAI, erklärbare empirische Risikominimierung (EERM), LIME, Transparenz, Recht auf Erklärung, vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI), Regularisierung.

Querverweise

Literatur

1. Adebayo et al. (2018). Sanity Checks for Saliency Maps. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 9505–9515.

2. Court of Justice of the European Union (2023). Judgment of 7 December 2023, SCHUFA Holding (Scoring), C-634/21, ECLI:EU:C:2023:957. Court of Justice of the European Union, First Chamber. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0634.

3. Colorado General Assembly (2026). Senate Bill 26-189 — Automated Decision-Making Technology. Colorado General Assembly, 2026 Regular Session, signed May 14, 2026. URL: https://leg.colorado.gov/bills/sb26-189.

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

5. 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.

6. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

7. 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.

8. Jung and Nardelli (2020). An Information-Theoretic Approach to Personalized Explainable Machine Learning. IEEE Signal Process. Lett., 27, pp. 825–829.

9. 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.

10. Cover and Thomas (2006). Elements of Information Theory. 2nd ed., Wiley, Hoboken, NJ, USA.

11. 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.