Themenblock 10 Explainable/Hybrid/Robust AI 10_01Frage_Taxonomy

Taxonomie von Methoden zur Interpretierbarkeit

Erarbeitet von
Marc Feger M.Sc.

Lernziele

- Du verstehst den Unterschied zwischen intrinsischer und post-hoc Interpretierbarkeit in KI-Modellen, einschließlich der Anwendung und Bedeutung beider Ansätze
- Du kannst modellspezifische und modellagnostische Interpretationsmethoden unterscheiden und weißt, wie sie in verschiedenen KI-Kontexten eingesetzt werden
- Du erkennst, wie lokale und globale Interpretierbarkeit funktioniert, um einzelne KI-Entscheidungen zu erklären und das Gesamtverhalten von KI-Modellen zu verstehen

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Da wir bereits die wichtige Rolle der interpretierbaren KI in unseren vorherigen Diskussionen besprochen haben, werden wir heute in die faszinierenden Details ihrer Taxonomie eintauchen – das „Wie“ hinter den Mechanismen, die KI-Entscheidungen verständlich machen.

Intrinsische vs. Pos-Hoc Interpretierbarkeit

Quelle [1, 2, 3, 4, 5, 6, 7]

Beginnen wir mit „Intrinsischer vs. Post-hoc-Interpretierbarkeit“. Intrinsische Modelle sind von Natur aus transparent. Stellt euch diese wie Glasuhren vor, bei denen man alle Zahnräder arbeiten sehen kann. Beispiele hierfür sind lineare Regressionen und Entscheidungsbäume. Diese Modelle sind unkompliziert – sie ermöglichen es uns, leicht nachzuvollziehen, wie sich Eingabedaten auf das Ergebnis auswirken.

Im Gegensatz dazu sind Post-hoc-Modelle wie komplizierte Maschinen in einer Box. Dazu gehören komplexe Modelle wie neuronale Netzwerke oder Zufallswälder. Ihre inneren Abläufe sind nicht sofort ersichtlich. Um diese Modelle zu verstehen, wenden wir Techniken wie LIME oder SHAP-Werte an, nachdem das Modell seine Entscheidungen getroffen hat. Diese Methoden zerlegen den Output des Modells und geben uns eine Analyse im Nachhinein, warum eine bestimmte Entscheidung getroffen wurde.

Modellspezifische vs. Modellagnostische Methoden

Quelle [1, 2, 4, 5, 6, 7]

Damit einher geht auch die Frage nach „Modellspezifischen vs. Modellagnostischen“ Methoden. Modellspezifische Methoden sind maßgeschneiderte Werkzeuge, die für bestimmte KI-Modelle entwickelt wurden. Diese Methoden sind wie speziell angefertigte Schlüssel, die dazu dienen, Einblicke in einen spezifischen Modelltyp zu gewähren. Zum Beispiel fällt die Untersuchung der Koeffizienten in linearen Modellen unter die Kategorie der modellspezifischen Methoden.

Andererseits sind modellagnostische Methoden universell einsetzbar. Sie sind wie Generalschlüssel, die Einblicke in jedes KI-Modell ermöglichen, unabhängig von seinem Typ. Zu dieser Kategorie gehören leistungsstarke Techniken wie die Permutations-Feature-Importanz, LIME oder SHAP, die verwendet werden können, um den Einfluss verschiedener Features auf die Vorhersagen eines Modells zu bewerten, unabhängig von der Architektur des Modells.

Lokale vs. Globale Interpretierbarkeit

Quelle [1, 2, 4, 5, 6]

Zuletzt haben wir „Lokale vs. Globale Interpretierbarkeit."
Lokale Interpretierbarkeit bezieht sich auf das Verständnis spezifischer Entscheidungen, die vom Modell getroffen wurden. Es ist so, als würde man sich einen einzelnen Pinselstrich in einem Gemälde ansehen, um die Technik des Künstlers in diesem speziellen Teil zu verstehen. Techniken wie LIME helfen in diesem Zusammenhang, indem sie erklären, warum das Modell eine bestimmte Entscheidung für eine einzelne Instanz getroffen hat. Im Gegensatz dazu geht es bei der globalen Interpretierbarkeit darum, das gesamte Bild zu sehen. Es ist, als würde man zurücktreten, um das gesamte Gemälde zu betrachten, um den Gesamtstil und die Muster zu verstehen. Dies beinhaltet Techniken wie Feature-Importanz-Rankings, die uns helfen, das Verhalten des Modells über alle Instanzen hinweg zu erfassen.

Take-Home Message

Letztlich bietet die Taxonomie der interpretierbaren KI eine strukturierte Möglichkeit, KI-Entscheidungen zu verstehen. Ob es sich um die Klarheit intrinsischer Modelle, die Vielseitigkeit modellagnostischer Methoden oder die detaillierten Einblicke aus lokalen Interpretierbarkeitstechniken handelt, jeder Aspekt bringt uns die komplexe Welt der KI näher. Vielen Dank für eure Aufmerksamkeit, und ich hoffe, diese Sitzung hat euer Verständnis für die komplexe Welt der interpretierbaren KI vertieft!

Quellen

Quelle [1] Molnar, C. (2024). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Quelle [2] Molnar, C., Casalicchio, G., Bischl, B. (2021). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. https://doi.org/10.1007/978-3-030-65965-3_28

Quelle [3] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. https://arxiv.org/abs/2103.11251

Quelle [4] Allen, G. I., Gan, L., & Zheng, L. (2023). Interpretable Machine Learning for Discovery: Statistical Challenges & Opportunities. https://arxiv.org/abs/2308.01475

Quelle [5] Permutation Feature Importance: https://scikit-learn.org/stable/modules/permutation_importance.html

Quelle [6] Local Interpretable Model-Agnostic Explanations (LIME): https://lime-ml.readthedocs.io/en/latest/#

Quelle [7] SHapley Additive exPlanations (SHAP):
https://shap.readthedocs.io/en/latest/index.html

Weiterführendes Material

Molnar, C. (2021). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Disclaimer

Transkript zu dem Video „Themenblock 10 Explainable/Hybrid/Robust AI 10_01Frage_Taxonomy“, Marc Feger.