interpretierbares maschinelles Lernen

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

Interpretierbares maschinelles Lernen (interpretierbares ML) bezieht sich auf machine learning (ML)-Methoden, deren Hypothesenraum nur Hypothesen enthält, die ein menschlicher Benutzer direkt verstehen kann, wie z.B. dünnbesetzte lineare Modelle oder flache Entscheidungsbäume. Solche Methoden lassen einen Benutzer erkennen, wie sich eine Vorhersage ändert, wenn sich die Merkmale eines Datenpunkt ändern, und wie sich die Vorhersagen ändern, wenn ein anderes Trainingsdatensatz verwendet wird. Eine interpretierbare Hypothese ist ihre eigene Erklärung; nachträgliche Erklärungen der Vorhersagen einer undurchsichtigen gelernten Hypothese können stattdessen unzuverlässig sein. Die Beschränkung des Hypothesenraum auf interpretierbare Hypothesen wirkt auch als eine Form der Regularisierung. Ein Strafterm kann eine verständliche Hypothese aus einem Hypothesenraum auswählen, der hauptsächlich andere enthält: least absolute shrinkage and selection operator (Lasso) setzt einige Gewichte eines lineares Modell auf exakt null, wie viele, hängt von seinem Regularisierungsparameter ab, wodurch eine Hypothese mit wenigen Nicht-Null- Gewichte übrig bleibt, und explainable empirical risk minimization (EERM) bestraft Abweichungen von den Vorhersagen, die ein Benutzer während des Training liefert.

Definition

Ein Arzt liest eine Regel, die das Risiko eines Patienten aus einer Handvoll Messungen vorhersagt, und kann jeden Schritt davon manuell überprüfen. Interpretierbares ML bezieht sich auf machine learning (ML)-Methoden, deren Hypothesenraum nur solche Hypothesen enthält, die ein menschlicher Benutzer direkt verstehen kann (siehe Interpretierbarkeit). Solche Methoden lassen einen Benutzer sowohl nachvollziehen, was eine gelernte Hypothese mit ihren Eingaben macht, als auch wie das Training sie erzeugt hat. Eine interpretierbare ML-Methode ermöglicht es dem Benutzer zu verstehen, wie sich eine Vorhersage ändert, wenn sich die Merkmale eines Datenpunkt ändern, und wie sich die Vorhersagen ändern, wenn ein anderes Trainingsdatensatz verwendet wird. Daher erfordern interpretierbare ML- Methoden geeignete Entscheidungen für den Hypothesenraum und den Training-Algorithmus.

Beispiele für interpretierbares ML umfassen empirische Risikominimierung (ERM) mit lineare Modelle, die eine kleine Anzahl bedeutungsvoller Merkmale kombinieren, flache Entscheidungsbäume, deren Vorhersagen aus wenigen expliziten Tests folgen, verallgemeinerte additive Modelle (GAM) und Listen von Entscheidungsregeln (Molnar, 2025). Abb. 1 zeigt einen Entscheidungsbaum für die medizinische Triage in einem Krankenhaus. Dieses System umfasst zwei Tests von Vitalparametern, wodurch das Personal jede Vorhersage durch Lesen des Baumes nachvollziehen kann.

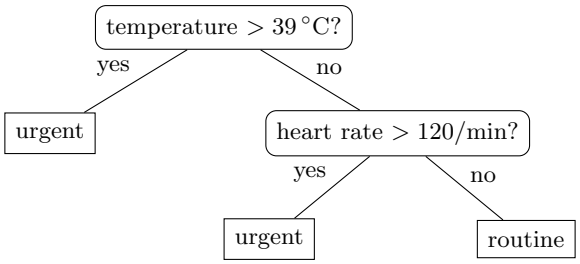


Abbildung 1: Eine interpretierbare Hypothese: ein flacher Entscheidungsbaum für die medizinische Triage. Jede Vorhersage folgt aus höchstens zwei expliziten Tests der Merkmale, sodass ein menschlicher Benutzer die Vorhersagen nachvollziehen und antizipieren kann.

Eine interpretierbare Hypothese macht die Auswirkung von Merkmal-Änderungen explizit. Für ein lineares Modell $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ ändert das Ändern des Merkmal x_j um einen Betrag Δ die Vorhersage um genau $w_j \Delta$. Jedes Gewicht gibt an, wie stark ein Merkmal die Vorhersage beeinflusst. Für den Entscheidungsbaum in Abb. 1 ändert sich eine Vorhersage nur, wenn ein Merkmal eine der expliziten Testschwellen überschreitet. Ein Patient mit 38.5 °C und einer Herzfrequenz von 110/min ist Routine; ein Anstieg auf 38.9 °C lässt diese Vorhersage unverändert, während ein Anstieg auf 39.5 °C sie dringend macht.

Eine interpretierbare ML-Methode macht auch die Auswirkung von Trainingsdatensatz-Änderungen nachvollziehbar. Diese Abhängigkeit wird unter stability untersucht, die eine ML-Methode als eine Abbildung von einem Trainingsdatensatz zu einer gelernten Hypothese formalisiert. Für lineare Regression ist diese Abbildung in geschlossener Form über die Normalgleichungen verfügbar, wann immer die Merkmalsmatrix vollen Spaltenrang hat (Hastie et al., 2009, Abschn. 3.2, Gl. (3.6)), sodass der Benutzer berechnen kann, wie ein gestörtes Label im Trainingsdatensatz die Vorhersagen verschiebt (siehe lineare Regression).

Eine interpretierbare Hypothese ist ihre eigene Erklärung: Es wird keine separate Erklärung für sie konstruiert. Die Erklärung, die mit einer Vorhersage geliefert wird, kann dann eine Beschreibung der gelernten Hypothese selbst sein: die Gewichte eines lineares Modell, das wenige Merkmale kombiniert,

oder das Flussdiagramm eines flachen Entscheidungsbaum wie in Abb. 1. Ein Hypothesenraum, der nur solche Hypothesen enthält, wird intrinsisch interpretierbar genannt (Molnar, 2025). Dies steht im Gegensatz zu erklärbare künstliche Intelligenz (XAI), das nachträgliche Erklärungen für die Vorhersagen einer möglicherweise undurchsichtigen gelernten Hypothese konstruiert. Für Anwendungen mit hohen Einsätzen, wie die medizinische Triage in Abb. 1, kann die Verwendung eines interpretierbaren Modell der Erklärung eines undurchsichtigen Modell vorzuziehen sein, da nachträgliche Erklärungen der Hypothese, die sie erklären, unzuverlässig sein können (Rudin, 2019).

Die Beschränkung von $\mathcal{H}$ auf interpretierbare Hypothesen ist Modellbeschneidung (model pruning), eine der drei Wege, die die Regularisierung unterscheidet. Der Weg über den Strafterm liefert eine interpretierbare gelernte Hypothese, ohne den Hypothesenraum vorher einzuschränken. Ein Strafterm, der zur Zielfunktion von ERM hinzugefügt wird, bestraft die Hypothesen, denen ein Benutzer nicht folgen kann, sodass das Training eine verständliche aus einem Hypothesenraum auswählt, der hauptsächlich die anderen enthält. Least absolute shrinkage and selection operator (Lasso) tut dies für lineare Modelle, die nicht verständlicher sind als ein tiefes Netz, sobald sie Hunderte von Merkmale kombinieren. Lasso verwendet die ℓ_1-Norm $\|\mathbf{w}\|_1$ der Gewichte $\mathbf{w}$ als den Strafterm, der einige Gewichte auf exakt null setzt, wie viele, hängt vom Regularisierungsparameter α ab (Hastie et al., 2009, Abschn. 3.4.2). Die gelernte Hypothese ist dann eine gewichtete Summe der verbleibenden Merkmale, sodass ein Benutzer die Merkmale benennen kann, die die Vorhersage verwendet, und wie stark jedes einzelne sie ändert.

Eine Methode, die diesem Weg nahekommt, aber kein interpretierbares ML ist, ist explainable empirical risk minimization (EERM), die die Idee auf das Training eines beliebigen Modell ausdehnt. Sein Strafterm wird aus den Vorhersagen gebildet, die ein spezifischer Benutzer für die Datenpunkte des Trainingsdatensatz liefert (siehe EERM). Es liefert daher Erklärbarkeit an diesen Benutzer anstatt Interpretierbarkeit, und es belässt den Hypothesenraum wie er ist. Die Benutzer-Vorhersagen müssen nicht einzeln für jeden Datenpunkt erhoben werden. Sie könnten als die Vorhersagen eines einfacheren Proxy-Modell erhalten werden, das der Benutzer als interpretierbar betrachtet, sodass der Strafterm die Abweichung von diesem Proxy misst. In Abb. 2 werden die Benutzer- Vorhersagen aus einem einfachen lineares Modell erhalten, das an das Trainingsdatensatz angepasst wurde.

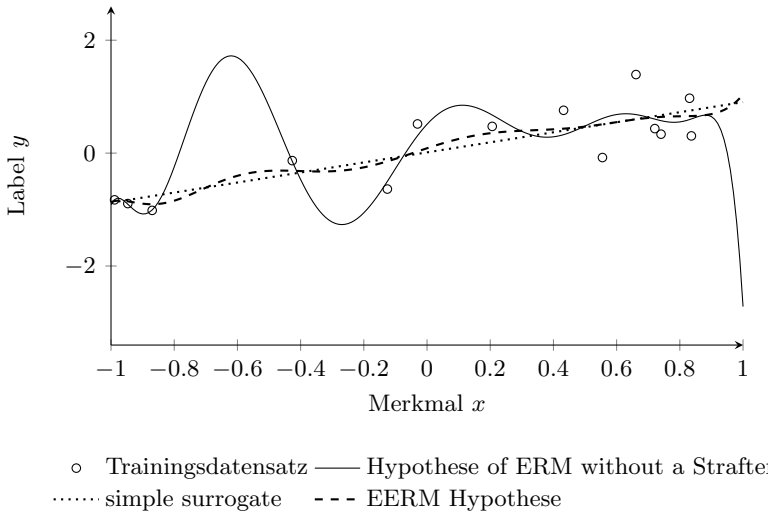


Abbildung 2: Die Idee von EERM (Zhang et al., 2024): Training eines hochdimensionalen Modell (Polynome vom Grad 10) mit einem Strafterm, der die Abweichung der Vorhersagen der gelernten Hypothese von den Benutzer- Vorhersagen misst. Diese Benutzer-Vorhersagen werden aus einem einfachen Ersatz- lineares Modell (gepunktete Linie) gewonnen. ERM ohne einen Strafterm (dünne durchgezogene Linie) überanpasst das Trainingsdatensatz, während EERM eine Hypothese (gestrichelte Linie) liefert, die nahe an den Benutzer-Vorhersagen bleibt. Daten generiert von `pythondemos/interpretableml.py`

Siehe auch: Interpretierbarkeit, Erklärbarkeit, XAI, Entscheidungsbaum, Entscheidungsregel, GAM, Lasso, stability, EERM.

Querverweise

Literatur

1. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.

2. Zhang et al. (2024). Explainable empirical risk minimization. Neural Comput. Appl., 36(8), pp. 3983–3996.

3. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

4. 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.