Fairness (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

Fairness (maschinelles Lernen) ist die Eigenschaft einer gelernten Hypothese oder der darauf aufbauenden Entscheidungsregel, Datenpunkte mit unterschiedlichen Werten eines sensiblen Attributs, wie etwa Geschlecht oder Alter, auf vergleichbare Weise zu behandeln. Ein kontrafaktisches Kriterium verlangt, dass die Vorhersage unverändert bleibt, falls das sensitive Merkmal einen anderen Wert angenommen hätte. Ein individuelles Kriterium verlangt ähnliche Vorhersagen für Datenpunkte, die gemäß einer aufgabenspezifischen Metrik ähnlich sind. Gruppenkriterien setzen ein Maß über die Gruppen hinweg gleich: die Bewilligungsraten (demografische Parität), die Raten der richtig-positiven und falsch-positiven Ergebnisse (chancengleiche Quoten; engl. *equalized odds*) oder die Wahrscheinlichkeit des Labels gegeben den Score (Kalibrierung). Die Kriterien sind im Allgemeinen miteinander unvereinbar: Kalibrierung und chancengleiche Quoten können nicht gleichzeitig gelten, wenn sich die Gruppen in ihrem Anteil an positiven Labels unterscheiden. Ein gewähltes Kriterium wird durch eine Entwurfsentscheidung für eine Komponente der empirischen Risikominimierung erfüllt: das Ausschließen des sensiblen Attributs und der damit korrelierten Merkmale aus dem Modell, das Hinzufügen eines Strafterms oder einer Nebenbedingung zur Verlustfunktion oder das Neuabtasten der Trainingsdatensatz.

Definition

Betrachtet sei eine Bank, die einen Kredit bewilligt, wenn die von einer gelernten Hypothese $\widehat{h}$ vorhergesagte Kreditwürdigkeit einen Schwellenwert τ überschreitet. Der Merkmalsvektor $\mathbf{x}$ eines Antragstellers erfasst das Einkommen, das Geschlecht und weitere Attribute, und $\widehat{h}$ wurde aus der Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ von fünf abgeschlossenen Krediten gelernt. Ein Antragsteller mit Merkmalsvektor $\mathbf{x}'$ erhält $\widehat{h}(\mathbf{x}') = 2.94$, unterhalb von $\tau = 3.8$, und wird abgelehnt. Derselbe Antragsteller, bei dem lediglich das Geschlecht vertauscht wurde, $\mathbf{x}^{\star}$, erhält $\widehat{h}(\mathbf{x}^{\star}) = 3.93$ und wird bewilligt. Auf die Frage, ob dieses System fair ist, gibt es mehrere präzise Antworten, die nicht alle gleichzeitig gelten können. Fairness beim maschinellen Lernen ist die Eigenschaft einer gelernten Hypothese oder der darauf aufbauenden Entscheidungsregel, neben dem Liefern genauer Vorhersagen auch Datenpunkte mit unterschiedlichen Werten

eines sensiblen Attributs s, wie etwa Geschlecht, ethnische Zugehörigkeit oder Alter, auf vergleichbare Weise zu behandeln (Barocas et al., 2023). Sie ist eine der zentralen Anforderungen an vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI).

Es werden drei Familien von Fairnesskriterien unterschieden (Barocas et al., 2023). Ein *kontrafaktisches* Kriterium verlangt, dass sich die Vorhersage für einen Datenpunkt nicht ändern würde, falls sein sensitives Merkmal einen anderen Wert angenommen hätte, und zwar unter einem kausalen Modell dafür, wie s die anderen Merkmale beeinflusst (siehe Kausalität) (Kusner et al., 2017). Ein *individuelles* Kriterium verlangt, dass Datenpunkte, die gemäß einer aufgabenspezifischen Metrik – etwa einer, die s ignoriert – ähnlich sind, ähnliche Vorhersagen erhalten (Dwork et al., 2012). Ein *Gruppenkriterium* verlangt die Gleichheit eines bestimmten Maßes über die durch s definierten Gruppen hinweg. Demografische Parität setzt die Bewilligungsraten gleich,

$$\mathbb{P}\left(\widehat{h}(\mathbf{x}) \geq \tau \mid s = 0\right) = \mathbb{P}\left(\widehat{h}(\mathbf{x}) \geq \tau \mid s = 1\right), \tag{1}$$

wobei der Datenpunkt als Realisierung einer Zufallsvariable (ZV) modelliert wird. Für ein binäres Label $y \in \{0, 1\}$, etwa ob ein Kredit zurückgezahlt wurde, setzen chancengleiche Quoten (engl. *equalized odds*) die Raten der richtig-positiven und der falsch-positiven Ergebnisse über die Gruppen hinweg gleich (Hardt et al., 2016), und Kalibrierung setzt $\mathbb{P}\left(y = 1 \mid \widehat{h}(\mathbf{x}) = c, s\right)$ über die Gruppen hinweg für jeden Score c gleich (Chouldechova, 2017).

Abb. 1 zeigt das Beispiel der Kreditwürdigkeitsprüfung. Die horizontale Achse ist eine eindimensionale Skizze des Merkmalsvektors; der Wechsel des Geschlechts verschiebt den Antragsteller von $\mathbf{x}'$ (offenes Quadrat) zu $\mathbf{x}^\star$ (offenes Dreieck). Die Differenz $\widehat{h}(\mathbf{x}^\star) - \widehat{h}(\mathbf{x}') = 0.99$ (Doppelpfeil) wird allein durch das Geschlecht verursacht, sodass $\widehat{h}$ das kontrafaktische Kriterium verletzt. Unter einer Metrik, die s ignoriert, verletzt sie ebenso das individuelle Kriterium: $\mathbf{x}'$ und $\mathbf{x}^\star$ sind dann derselbe Datenpunkt, erhalten jedoch Vorhersagen auf entgegengesetzten Seiten von τ. Die Gruppenkriterien lassen sich anhand von fünf Krediten nicht überprüfen. Sie vergleichen die Vorhersagen von $\widehat{h}$ über die Grundgesamtheit der Antragsteller, welche durch die Wahrscheinlichkeitsverteilung der Bedingung für demografische Parität (1) beschrieben wird.

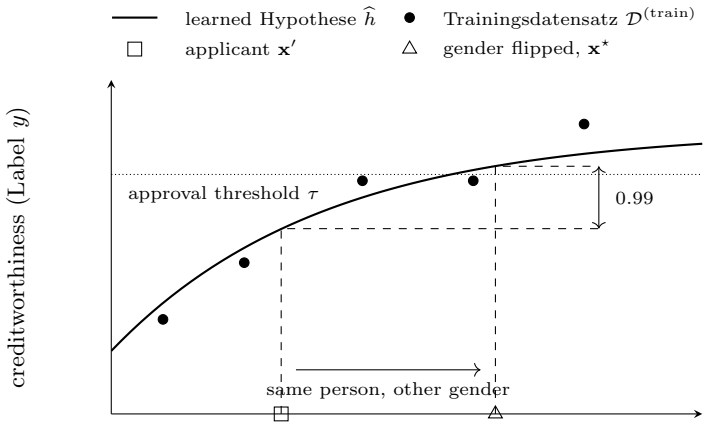


Abbildung 1: Das kontrafaktische Fairnesskriterium im Beispiel der Kreditwürdigkeitsprüfung. Die horizontale Achse ist eine eindimensionale Skizze des Merkmalsvektors. Die durchgezogene Kurve ist die aus der Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ von fünf abgeschlossenen Krediten (ausgefüllte Kreise) gelernte Hypothese $\widehat{h}$; ein Kredit wird bewilligt, wenn die vorhergesagte Kreditwürdigkeit den Schwellenwert τ (gepunktete Linie) überschreitet. Der Antragsteller mit Merkmalsvektor $\mathbf{x}'$ (offenes Quadrat) wird abgelehnt, während derselbe Antragsteller mit vertauschtem Geschlecht, $\mathbf{x}^\star$ (offenes Dreieck), bewilligt wird. Der vertikale Doppelpfeil überspannt die Differenz zwischen den beiden Vorhersagen, die allein durch das Geschlecht verursacht wird

Die Kriterien sind im Allgemeinen miteinander unvereinbar. Kleinberg et al. (2017), Thm. 1 und Chouldechova (2017), Gl. (2.6) zeigen, dass Kalibrierung und chancengleiche Quoten nicht gleichzeitig gelten können, wenn sich der Anteil an Datenpunkte mit $y = 1$ zwischen den Gruppen unterscheidet und $\widehat{h}$ nicht perfekt ist. Fairness beim maschinellen Lernen ist daher keine einzelne zu optimierende Eigenschaft, sondern eine Auswahl unter Kriterien. Die Auswahl hängt von der Anwendung ab: vom Schaden, den eine falsche Vorhersage verursacht, und vom geltenden Recht. Zwei Beispiele sind Art. 22 der Datenschutz-Grundverordnung (DSGVO) zu automatisierten Entscheidungen (European Parliament and Council of the European Union, 2016, Art. 22) und die Pflicht gemäß dem KI-Verordnung, die Trainingsdatensatz eines Hochrisiko-KI-Systems auf Verzerrungen zu untersuchen, die zu Diskriminierung führen (European Parliament and Council of the European Union, 2024, Art. 10) (siehe Daten-Governance).

Sobald ein Kriterium gewählt ist, wird es durch eine Entwurfsentscheidung für eine der Komponenten der empirischen Risikominimierung erfüllt (Pfau and Jung, 2026, Abschn. VIII). Das Modell kann das sensitive Merkmal ausschließen; $\mathbf{x}'$ und $\mathbf{x}^\star$ fallen dann zusammen und die Differenz in Abb. 1 verschwindet.

Das alleinige Ausschließen von s genügt nicht, da mit ihm korrelierte Merkmale, wie etwa die Postleitzahl, dieselbe Information tragen. Die Verlustfunktion kann einen Strafterm für die Abweichung enthalten, oder die Abweichung kann eine Nebenbedingung der restringierten Optimierungsaufgabe sein, welche die empirische Risikominimierung ersetzt. Die Trainingsdatensatz kann neu abgetastet werden, sodass die Gruppen gleichermaßen vertreten sind (siehe unausgewogene Daten). Jede dieser Möglichkeiten verringert die Abweichung bezüglich des gewählten Kriteriums, jedoch zu einem Preis. Die mit einem Strafterm oder einer Nebenbedingung gelernte Hypothese weist auf $\mathcal{D}^{(\text{train})}$ einen durchschnittlichen Verlust auf, der mindestens so groß ist wie jener der Lösung der empirischen Risikominimierung, da die empirische Risikominimierung diesen Verlust über ganz $\mathcal{H}$ minimiert. Das Verringern einer Abweichung kann zudem die Abweichung bezüglich eines anderen Kriteriums erhöhen: Ein Score, der innerhalb jeder Gruppe kalibriert ist und der positiven Klasse – sowie gleichermaßen der negativen Klasse – in beiden Gruppen denselben durchschnittlichen Score zuweist, existiert nur dann, wenn die Gruppen denselben Anteil an Datenpunkte mit $y = 1$ aufweisen oder der Score perfekte Vorhersagen liefert (Kleinberg et al., 2017, Thm. 1).

Siehe auch: sensitives Merkmal, vertrauenswürdige KI, Hypothese, Entscheidungsregel, Vorhersage, Kalibrierung, Metrik, Kausalität, Kontrafaktum, ERM, Strafterm, eingeschränktes Optimierungsproblem, Daten-Governance, KI-Verordnung, DSGVO, Hochrisiko-KI-System, unausgewogene Daten, erklärbare künstliche Intelligenz (XAI).

Querverweise

Literatur

1. 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.

2. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

3. Barocas et al. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

4. Chouldechova (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), pp. 153–163.

5. Dwork et al. (2012). Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226.

6. Hardt et al. (2016). Equality of opportunity in supervised learning. In: Adv. Neur. Inf. Proc. Syst., pp. 3315–3323.

7. Kleinberg et al. (2017). Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the 8th Innovations in Theoretical Computer Science Conference, pp. 43:1–43:23.

8. Kusner et al. (2017). Counterfactual Fairness. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 4066–4076.

9. Pfau and Jung (2026). Engineering Trustworthy AI: A Developer Guide for Empirical Risk Minimization. IEEE Trans. Artif. Intell., 7(5), pp. 2560–2576.