sensitives Merkmal

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

Ein sensitives Merkmal eines Datenpunkts ist ein Attribut, wie etwa das Geschlecht oder die ethnische Herkunft, das die Vorhersage, die ein ML-System (ML-System) für diesen Datenpunkt liefert, nicht beeinflussen darf, unabhängig davon, ob es zum Merkmalsvektor gehört oder nicht. Der Begriff hat zwei Bedeutungen, die auseinandergehalten werden: eine regulatorische, bei der das Datenschutzrecht festlegt, welche Kategorien personenbezogene Datener Daten besonders sind, und eine im Sinne des maschinelles Lernen (ML)s, bei der die Anwendung das Attribut s benennt, über das hinweg ein Fairness (maschinelles Lernen)-Kriterium Vorhersagen vergleicht. Das Weglassen von s aus dem Merkmalsvektor entfernt es nicht aus den Daten, da die verbleibenden Merkmale davon abhängig sein können. Gelernte Merkmale, wie sie etwa die Hauptkomponentenanalyse (PCA) liefert, sind keine Ausnahme: Das Kriterium, das sie lernt, erwähnt s nie, und dennoch können sie vorhersagekräftig dafür sein. Wie viel von s sie in sich tragen, wird durch die Rekonstruktionsgenauigkeit gemessen, die Genauigkeit, mit der ein linearer Klassifikator s aus ihnen für Datenpunkte vorhersagt, aus denen er nicht gelernt wurde. Für das Wetter in Krems an der Donau (Österreich) rekonstruieren zwei gelernte Merkmale die Jahreshälfte, zu der ein Tag gehört, mit einer Rekonstruktionsgenauigkeit von 0.94, gegenüber einem Mehrheitsklassenanteil von 0.51. Die Projektion der gelernten Merkmale auf das orthogonale Komplement der Richtung, die der linearer Klassifikator verwendet, senkt diese Genauigkeit, ohne sie auf den Mehrheitsklassenanteil zu drücken, sodass ein sensitives Merkmal leichter zu messen als zu entfernen ist.

Definition

Betrachtet sei eine Bank, die über Kreditanträge anhand des Einkommens, des Berufs und der Adresse eines Antragstellers entscheidet. Eine Aufsichtsbehörde fragt, ob die Entscheidungen vom Geschlecht des Antragstellers abhängen. Die Bank antwortet, dass das Geschlecht nie als Merkmal erfasst wurde: Der Merkmalsvektor $\mathbf{x}$, den die gelernte Hypothese $\widehat{h}$ einliest, enthält diese drei Zahlen und nichts weiter. Diese Antwort klärt nichts. Der Beruf und die Adresse eines Antragstellers können vorhersagekräftig für das Geschlecht sein, und eine aus $\mathbf{x}$ berechnete Vorhersage hängt dann dennoch vom Geschlecht ab.

Ein sensitives Merkmal eines Datenpunkts ist ein Attribut, das die für diesen Datenpunkt gelieferte Vorhersage nicht beeinflussen darf, unabhängig davon, ob es zum Merkmalsvektor gehört oder nicht. Ein Datenpunkt umfasst dann drei Teile: den Merkmalsvektor $\mathbf{x} \in \mathcal{X}$, das Label y und das sensitive Merkmal s, das typischerweise nur wenige Werte annimmt. Das Geschlecht eines Kreditantragstellers wird im Folgenden als $s \in \{0, 1\}$ geschrieben.

Welches Attribut als sensitiv gilt, wird auf zwei unterschiedliche Weisen festgelegt, und die beiden Bedeutungen werden auseinandergehalten. Im regulatorischen Sinne wird die Antwort durch das Gesetz gegeben. Die Datenschutz-Grundverordnung (DSGVO) untersagt, mit Ausnahmen, die Verarbeitung personenbezogene Datener Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie von genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer Person, Gesundheitsdaten und Daten zum Sexualleben oder der sexuellen Orientierung (European Parliament and Council of the European Union, 2016, Art. 9(1)). Der KI-Verordnung übernimmt diese Liste unter der Bezeichnung besondere Kategorien personenbezogene Datener Daten (European Parliament and Council of the European Union, 2024, Art. 3(37)). Im Sinne des maschinelles Lernen (ML)s entscheidet die Anwendung, und ihre Wahl muss nicht zu dieser Liste gehören: Das Geschlecht des Kreditantragstellers ist im obigen Beispiel ein sensitives Merkmal und stellt nach der DSGVO keine besondere Kategorie dar. Ein Fairness (maschinelles Lernen)-Kriterium vergleicht die Vorhersagen von $\widehat{h}$ über die Werte von s hinweg (Barocas et al., 2023), während der Datenschutz stattdessen verlangt, dass der Ausgabe des ML-System (ML-System)s keinen Rückschluss auf s zulässt, wobei ein solcher Rückschluss durch die Datenschutzleckage gemessen wird.

Die naheliegende Abhilfemaßnahme ist diejenige, die die Bank angewandt hat: s aus dem Merkmalsvektor wegzulassen. Dadurch wird s zwar aus dem Merkmalsvektor, nicht aber aus den Daten entfernt. Der Datenpunkt wird als Realisierung von Zufallsvariablen (ZV) modelliert, wobei der Merkmalsvektor und das sensitive Merkmal gemeinsam gezogen werden. Das Attribut ist nach wie vor vorhanden, wann immer die beiden voneinander abhängig sind,

$$\mathbb{P}\left(s = 1 \mid \mathbf{x} \in \mathcal{A}\right) \neq \mathbb{P}\left(s = 1\right) \text{ für ein } \mathcal{A} \subseteq \mathcal{X}. \tag{1}$$

Ein Merkmal, für das die Abhängigkeit (1) gilt, ist eine Ersatzvariable (Proxy) für s. Solche redundanten Kodierungen sind es, die eine Entscheidungsregel, welche s ignoriert, dennoch diskriminieren lassen (Dwork et al., 2012).

Das Repräsentationslernen hebt die Abhängigkeit (1) nicht auf. Es ersetzt die ursprünglichen Merkmale durch gelernte, $\mathbf{z} = \mathbf{\Phi}(\mathbf{x}) \in \mathbb{R}^{d'}$, die von einer selbst aus einem Datensatz gelernten Merkmalstransformation $\mathbf{\Phi}$ geliefert werden. Die Hauptkomponentenanalyse (PCA) lernt eine lineare Abbildung $\mathbf{\Phi}$, deren Zeilen die d' Eigenvektoren der Stichproben-Kovarianzmatrix zu den größten Eigenwerten sind, ein Kriterium, in dem s niemals vorkommt. Die gelernten Merkmale können dennoch vorhersagekräftig für s sein, und wie viel sie davon in

sich tragen, ist eine messbare Größe. Ein linearer Klassifikator $\widehat{h}^{(s)}$, der s aus $\mathbf{z}$ vorhersagt, wird mittels der Methode der kleinste Quadrate auf einem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ gelernt, und seine Rekonstruktionsgenauigkeit ist seine Genauigkeit auf einem Testdatensatz $\mathcal{D}^{(\text{test})}$ von Datenpunkten außerhalb von $\mathcal{D}^{(\text{train})}$,

$$\operatorname{acc}\left(\widehat{h}^{(s)} \mid \mathcal{D}^{(\text{test})}\right) := 1 - \frac{1}{\left|\mathcal{D}^{(\text{test})}\right|} \sum_{(\mathbf{z},s)\in\mathcal{D}^{(\text{test})}} L^{(0/1)}\left((\mathbf{z},s),\widehat{h}^{(s)}\right), \qquad (2)$$

mit dem 0/1-Verlust, der bei einer falschen Vorhersage den Wert 1 und andernfalls 0 annimmt. Die Rekonstruktionsgenauigkeit (2) wird dem Mehrheitsklassenanteil gegenübergestellt, also der Genauigkeit, die sich ergibt, wenn für jeden Datenpunkt der häufigere Wert von s vorhergesagt wird.

Die Merkmale, welche die PCA aus den Wetterdaten in Krems an der Donau (Österreich) lernt, veranschaulichen, was diese Messung aussagt. Jeder Tag vom 6. Januar bis zum 31. Dezember 2024 ist ein Datenpunkt, und seine ursprünglichen Merkmale sind die jeweils acht täglichen Messwerte der fünf vorangegangenen Tage, also vierzig Zahlen. Sein Attribut s gibt an, ob der Tag in der kalten Jahreshälfte von Oktober bis März liegt; dies trifft auf 178 der 361 Tage zu, sodass der Mehrheitsklassenanteil 0.51 beträgt. Das Halbjahr befindet sich nicht unter den vierzig ursprünglichen Merkmalen. Jedes ursprüngliche Merkmal wird zunächst auf einen Stichprobenmittel von null und eine Stichprobenvarianz von eins skaliert, da Temperaturen, Niederschlag und Luftdruck unterschiedliche Einheiten aufweisen. Zwei gelernte Merkmale rekonstruieren das Halbjahr mit einer Rekonstruktionsgenauigkeit (2) von 0.94, gemittelt über zwanzig zufällige Aufteilungen der Tage in zwei Hälften, $\mathcal{D}^{(\text{train})}$ und $\mathcal{D}^{(\text{test})}$, und ein einziges gelerntes Merkmal liefert bereits 0.94. Die Entscheidungsgrenze eines solchen linearer Klassifikators ist eine Gerade in der Ebene der beiden gelernten Merkmale (Abb. 1(a)). Lässt man sämtliche Temperaturmessungen, die offensichtliche Ersatzvariable für die Jahreszeit, aus den ursprünglichen Merkmalen weg, verbleiben fünfundzwanzig Zahlen pro Tag. Die Rekonstruktionsgenauigkeit beträgt dann 0.82 mit zwei gelernten Merkmalen und 0.89 mit fünf. Die beiden Kurven in Abb. 1(b) vergleichen die beiden Sätze ursprünglicher Merkmale für $d' = 1, \ldots, 10$. Das Halbjahr eines Tages ist kein sensitives Merkmal. Es dient hier als Stellvertreter für ein solches, und die Berechnung hängt nicht davon ab, welche Bedeutung das Attribut hat.

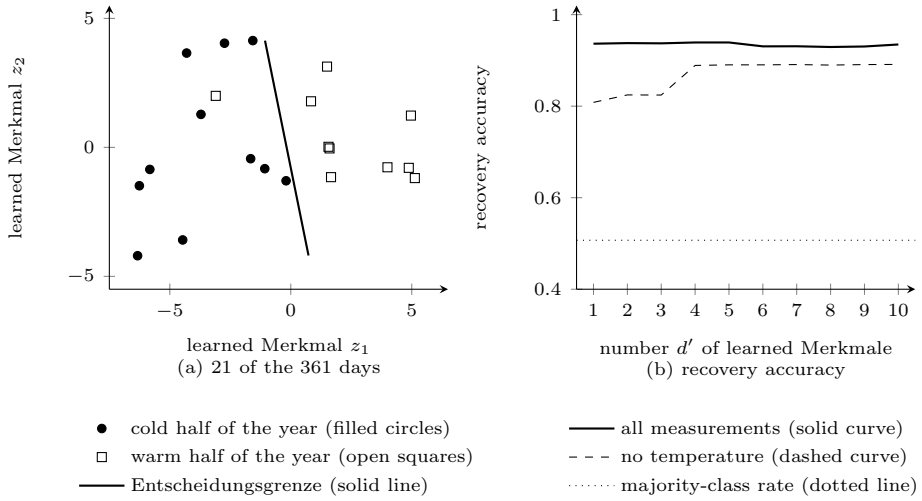


Abbildung 1: Ein Attribut, das kein Merkmal ist, rekonstruiert aus gelernten Merkmalen des Wetters in Krems an der Donau. (a) Jeder 18. der 361 Tage des Jahres 2024 in kalendarischer Reihenfolge, insgesamt 21 Tage, angeordnet anhand zweier gelernter Merkmale und markiert nach der Jahreshälfte, zu der sie gehören, der kalten (ausgefüllte Kreise) oder der warmen (offene Quadrate), zusammen mit der Entscheidungsgrenze (durchgezogene Linie) des aus 180 der 361 Tage gelernten linearer Klassifikators. (b) Die Rekonstruktionsgenauigkeit (2) in Abhängigkeit von der Anzahl d' gelernter Merkmale, für alle acht täglichen Messwerte (durchgezogene Kurve) sowie unter Weglassung sämtlicher Temperaturmessungen (gestrichelte Kurve), im Vergleich zum Mehrheitsklassenanteil (gepunktete Linie). Daten erzeugt mit `pythondemos/sensattr.py`

Ein sensitives Merkmal ist leichter zu messen als zu entfernen. Der linearer Klassifikator $\widehat{h}^{(s)}$ nutzt eine einzige Richtung des gelernten Merkmalsraums, nämlich diejenige, die durch seine gelernten Modellparameter $\widehat{\mathbf{w}}$ gegeben ist. Die Projektion der gelernten Merkmale auf das orthogonale Komplement dieser Richtung, den Nullraum der lineare Abbildung $\mathbf{z} \mapsto \widehat{\mathbf{w}}^\top \mathbf{z}$, entfernt genau diese Richtung und nicht mehr. Bei zehn gelernten Merkmalen der Tage in Krems erreicht ein auf den projizierten Merkmalen neu gelernter linearer Klassifikator bei einer der zwanzig Aufteilungen immer noch 0.84, gegenüber 0.93 vor der Projektion, sodass auch die anderen Richtungen das Attribut in sich tragen. Das Lernen einer Merkmalstransformation unter einem Strafterm für die Rekonstruktion von s setzt am selben Problem direkt beim Lernen der Merkmale an (Zemel et al., 2013). Jeder dieser Schritte setzt voraus, dass s erfasst wird, weshalb der KI-Verordnung einem Anbieter eines Hochrisiko-KI-Systems die Verarbeitung besonderer Kategorien personenbezogene Datener Daten gestattet, wenn dies zur Erkennung und Korrektur von Verzerrungen (Bias) unter Angabe von Garantien unbedingt erforderlich ist (European Parliament and Council of the European Union, 2024, Art. 10(5)).

Synonyme: geschütztes Merkmal, geschütztes Attribut.

Siehe auch: Fairness (maschinelles Lernen), vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI), Datenschutz, Datenschutzleckage, Repräsentationslernen, PCA, linearer Klassifikator, DSGVO, KI-Verordnung, personenbezogene Daten, Nullraum, Genauigkeit.

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

5. Zemel et al. (2013). Learning fair representations. In: International Conference on Machine Learning, pp. 325–333.