Faires Machine Learning

Handout

Autor:innen: Fiona Fischer, Linda Fernsel
im Projekt “Fair Enough?” (16DHB4002)
der Forschungsgruppe Informatik und Gesellschaft,
an der Hochschule für Technik und Wirtschaft (HTW) Berlin

Version: November 2023

Inhalt

Dieses Handout behandelt das Themengebiet “Faires Machine Learning” am Beispiel von Gesichtserkennungstechnologie. Es beginnt mit einer Einführung zu Gesichtserkennungstechnologie und zeigt den Nutzen der kritischen Betrachtung von Machine Learning an Hand der Studie “Gender Shades” auf. Dann wird erklärt, wie Machine Learning Modelle aus sozialer und aus technologischer Perspektive kritisch betrachtet werden können.

IdentiTOP Use Case & Jupyter Notebook

Die Inhalte des Handouts werden an Hand des fiktiven Use Case *IdentiTOP* in einem Python Jupyter Notebook demonstriert. Google Colab bietet die Möglichkeit, Jupyter Notebooks mit nur einem Browser nutzen zu können.

Das Notebook ist unter `bit.ly/identitop` verfügbar.

Lernziele

- Die Teilnehmenden kennen Anwendungsfälle für Gesichtserkennungstechnologie.
- Die Teilnehmenden können gängige Risiken bei der Anwendung von Machine-Learning-Modellen identifizieren (Soziale Perspektive).
- Die Teilnehmenden kennen Qualitätsmetriken für Machine Learning Modelle (Technologische Persoektive).
- Die Teilnehmenden können zu Risiken passende Qualitätsmetriken auswählen (Technologische Perspektive.
- Die Teilnehmenden können Qualitätsmetriken basierend auf der Wahrheitsmatrix berechnen (Technologische Perspektive).
- Die Teilnehmenden können die Bedeutung der Qualitätsmessergebnisse für die Risiken von Machine Learning Modellen interpretieren (Technologische Perspektive + Soziale Perspektive).

1. Gesichtserkennungstechnologie

Mit Hilfe von Machine Learning können wir Modelle entwickeln, die Daten anhand von Eigenschaften klassifizieren oder darüber Vorhersagen treffen. Solche Modelle werden zum Beispiel für Gesichtserkennungstechnologie verwendet. Gesichtserkennungstechnologie hat drei Stufen [3]:

Gesichtserkennung
Erkennen, ob sich ein Gesicht im Bild befindet. 1️⃣

Eigenschaftenerkennung
Erkennen um was für ein Gesicht es sich handelt. 2️⃣

Identitätserkennung
Erkennen, wessen Gesicht es ist. 3️⃣

Beispielhafte Anwendungsfälle für Gesichtserkennungstechnologie sind:

1️⃣ Gesichtserkennung	2️⃣ Eigenschaftenerkennung	3️⃣ Identitätserkennung
🤳 Foto-Filter bei Apps wie SnapChat [1]	💲 Geschlecht und Alter erkennen um zur Zielgruppe passende Werbung vorzuschlagen [2]	👮 Identifikation polizeilich gesuchter Personen bei Videoü berwachung [5]
👩‍💻 Überwachung ob bei einer Online-Prüfung eine Person alleine vor dem PC ist [8]	🧑‍🏫 Interesse im Unterricht messen als Feedback für Lehrkräfte [7]	🪪 Identifikation von Teilnehmenden bei einer Online-Prüfung [6]

Schau in’s Jupyter Notebook für die Beschreibung von *IdentiTOP*.

2. Gender Shades

“Gender Shades” ist eine von vielen Studien, die Machine Learning kritisch betrachten. Bei dieser Studie wurden automatisierte Gesichtserkennungsdienste hinsichtlich ihrer Fairness bezüglich Geschlecht und Hautfarbe untersucht [4]. Die Analyse der Gesichtsanalysealgorithmen ergab, dass Frauen im Vergleich zu Männern und dunkelhäutige Personen im Vergleich zu hellhäutigen Personen öfter falsch klassifiziert wurden [4]. Dunkelhäutige Frauen wurden dabei am häufigsten falsch klassifiziert [4]. Die Wissenschaftler:innen stellten fest, dass dies vor allem an den Trainingsdaten der Algorithmen lag: Bestehende Datensätze umfassten überwiegend hellhäutige Männer [4].

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Abbildung 1: Ergebnisse der “Gender Shades”-Studie. Quelle: gendershades.org

3. Machine Learning kritisch betrachten

Die kritische Betrachtung kann auf drei aufeinander aufbauenden Stufen erfolgen, wovon hier hauptsächlich Stufe 1 behandelt wird:

Wie fair ist das Machine-Learning-System? 1

Was sind Ursachen für Unfairness? 2

Wie kann Unfairness behoben werden? 3

Die kritische Betrachtung auf Stufe 1 folgt vier Schritten:

1 🔔 Modellrisiken identifizieren → 2 🔎 Qualitätsmetriken auswählen → 3 📏 Qualitätsmetriken berechnen → 4 🧠 Qualitätsmetriken interpretieren

Zum Schluss kann man noch einen weiteren Schritt anschließen, der jedoch bereits Stufe 2️⃣ und 3️⃣ anschneidet:

5️⃣

🛠️
Verbesserungsvorschläge erarbeiten

Alle Schritte werden im Folgenden genauer erklärt

Schritt 1️⃣ Modellrisiken identifizieren 🔔

Folgende Fragen können helfen, um die Risiken eines Modells zu identfiizieren:

- Was sind mögliche **Nachteile** und **negative Folgen** eines Systems?
- Welche **Fehler** können auftreten?
- Was sind **Auswirkungen** dieser Probleme auf Personen?
- Können diese Probleme **strukturell** auftreten und **Diskriminierung** verursachen?

Im folgenden werden die zuvor vorgestellten Anwendungsfällen für Gesichtserkennungstechnologie in Hinblick auf ihre Risiken diskutiert:

Foto-Filter bei Apps wie SnapChat [1]
Die Verwendung von Gesichtserkennungstechnologie in Apps wie SnapChat ermöglicht es Benutzer:innen, interaktive, unterhaltsame oder verschönernde Filter auf ihre Gesichter anzuwenden. Es können jedoch Fehler bei der Erkennung von Gesichtsmerkmalen auftreten, was zu unerwünschten Filtereffekten führt. Außerdem könnte der Gebrauch verschönernder Filter das eigene Selbstbild und das der anderen Nutzer:innen stören.

Geschlecht und Alter erkennen um zur Zielgruppe passende Werbung vorzuschlagen [2]
Das Erkennen von Geschlecht und Alter ermöglicht eine gezielte Werbung, die potentiell relevanter für die angesprochene Person ist. Falsche Erkennung von Geschlecht oder Alter könnte zu irrelevanten Werbeanzeigen führen. Dies könnte außerdem psychische Folgen haben, z.B. bei Trans-Personen. Auch bei korrekter Erkennung könnten Personen sich missverstanden oder stereotypisiert fühlen. Es ist wichtig, sicherzustellen, dass keine Stereotypen verstärkt werden und dass Personen die Möglichkeit haben, ihre Präferenzen manuell anzugeben.

Identifikation polizeilich gesuchter Personen bei Videoüberwachung [5]
Die Identifikation polizeilich gesuchter Personen könnte die öffentliche Sicherheit erhöhen und die Polizeiarbeit effizienter machen. Falsche Identifikationen könnten unschuldige Bürger:innen ins Visier der Strafverfolgung bringen. Unschuldige könnten festgenommen oder belästigt werden, was zu psychischen und körperlichen Problemen führen kann [10], sowie zu einem Vertrauensverlust in die Strafverfolgungsbehörden. Die Technologie sollte also mit großer Vorsicht und in Kombination mit menschlicher Überprüfung eingesetzt werden.

Überwachung der alleinigen Anwesenheit einer Person bei einer Online-Prüfung am PC [8]
Die Überwachung, ob bei einer Online-Prüfung eine Person alleine vor dem PC ist, soll sicherstellen, dass Betrug bei Prüfungen verhindert wird. Die Technologie könnte jedoch andere Objekte fälschlicherweise als zweite Person interpretieren, oder wiederum andere Personen nicht als solche erkennen. Prüflinge könnten ungerechterweise des Betrugs beschuldigt werden, was schwerwiegende akademische Konsequenzen haben könnte. Außerdem könnte das Gefühl, überwacht zu werden, Prüflinge unter Druck setzen, wodurch sie sich anders verhalten als gewöhnlich.

Interesse im Unterricht messen als Feedback für Lehrkräfte [7]
Wenn das Interessiertheit der Lernenden im Unterricht gemessen wird, könnten Lehrpersonen wertvolles Feedback erhalten, um ihren Unterricht anzupassen und zu verbessern. Nicht alle Gesichtsausdrücke spiegeln jedoch das tatsächliche Interesse oder das Verständnis wider. Dies betrifft z.B. Personen aus unterschiedlichen Kulturkreisen, die unterschiedliche Mimik haben, oder Personen, die ihre Gefühle auf Grund von anderen inhärenten Eigenschaften anders zum Ausdruck bringen als andere. Lernende könnten ungerechterweise als desinteressiert oder abgelenkt eingestuft werden, was zu ungerechtfertigten pädagogischen Maßnahmen führen könnte. Diese Technologie sollte also nicht als alleiniges Mittel zur Bewertung des Interesses verwendet werden.

Identifikation von Teilnehmenden bei einer Online-Prüfung [6]
Die Identifikation von Teilnehmenden bei einer Online-Prüfung stellt sicher, dass die richtige Person die Prüfung ablegt. Eine falsche Erkennung könnte dazu führen, dass eine Person ungerechterweise von einer Prüfung ausgeschlossen würde, was zu akademischen und emotionalen Belastungen führen könnte.

Schau in's Jupyter Notebook für die Identifizierung möglicher Risiken von *IdentiTOP*.

Schritt2️⃣ Qualitätsmetriken auswählen 🔎

Wir haben im vorigen Schritt festgestellt, dass jedes Modell individuelle Risiken mit sich bringt. Es ist daher notwendig, solche Metriken auszuwählen, die helfen, die spezifischen Risiken einzuschätzen. Dafür muss zu jedem Risiko die passende Metrik gefunden werden. Viele Metriken basieren auf der Wahrheitsmatrix (en: confusion matrix).

Die Wahrheitsmatrix

Eine Wahrheitsmatrix gibt einen Überblick über die Richtigkeit der Ergebnisse eines Modells, indem sie die Modellvorhersagen mit der Wirklichkeit (also dem richtigen Ergebnis) vergleicht. Die Ergebnisse eines Machine Learning Modells können dabei in “positiv” und “negativ” eingeteilt werden.

Die Wahrheitsmatrix enthält die Anzahl richtiger Vorhersagen: die Anzahl richtig positiver Fälle (en: true positive; kurz: **TP**) und die Anzahl richtig negative Fälle (true negative; kurz: **TN**). Außerdem enthält sie die Anzahl falscher Vorhersagen: die Anzahl falsch positiver Fälle (en: false positive; kurz: **FP**) und die Anzahl falsch negativer Fälle (true negative; kurz: **FN**).

Konkret ist die Wahrheitsmatrix nach folgendem Schema aufgebaut [11]:

Vorhersage → Wirklichkeit ↓	Als positiv vorhergesagt	Als negativ vorhergesagt
Wirklich positiv	TP	FN
Wirklich negativ	FP	TN

Eine hohe Modellqualität ist gegeben, wenn TP und TN hoch und FP und FN niedrig sind.

Schau in’s Jupyter Notebook für die Wahrheitsmatrix von *IdentiTOP*.

Welche Qualitätsmetriken gibt es?

Gängige Metriken zur Prüfung der Modellqualität sind die Richtigkeit (en: accuracy), Genauigkeit (en: precision) oder Trefferquote (en: recall) [9]. Die Richtigkeit bezieht sich auf alle Vorhersagen (TP + TN + FP + FN), die Genauigkeit nur auf die Fälle mit positiver Vorhersage (TP + FP) und die Trefferquote bezieht sich auf die wirklich positiven Fälle (TP + FN). Diese Metriken werden wie folgt anhand der Wahrheitsmatrix berechnet [11]:

```
Richtigkeit = (TP + TN) / (TP + TN + FP + FN)
```

```
Genauigkeit = TP / (TP + FP)
```

```
Trefferquote = TP / (TP + FN)
```

Zusätzlich gibt es weniger gängige Metriken, z.B. die **Spezifität** (en: specificity), die Falsch-negativ-Rate (en: false negative rate; kurz: **FNR**), die Falsch-positiv-Rate (en: false positive rate; kurz: **FPR**), die Falscherkennungsrate (en: false discovery rate; kurz: **FDR**) und die Falschauslassungsrate (en: false omission rate; kurz: **FOR**), die auf der Wahrheitsmatrix basieren [11]. Diese Metriken werden wie folgt an Hand der Wahrheitsmatrix berechnet [11]:

```
Spezifität = TN / (TN + FP)
```

```
FNR = FN / (FN + TP)
```

```
FPR = FP / (FP + TN)
```

```
FDR = FP / (FP + TN)
```

```
FOR = FN / (FN + TN)
```

Fairness

Fairness kann man mit Hilfe mathematischer Fairness-Definitionen messen. Fairness-Definitionen sagen aus, dass ein Modell für verschiedene Gruppen gleich gut funktioniert.

Was bedeutet “für verschiedene Gruppen gleich gut”?

Gruppen	**gleich**	**gut**
Die Unterteilung von Personen nach verschiedenen **Attributen**, z.B. “Geschlecht”. Eine Gruppe wäre in diesem Fall “Frauen”. Gruppen können auch “intersektional” sein. Das bedeutet, dass Personen nach mehreren Attributen aufgeteilt wurden, z.B. “Geschlecht” und “Hautfarbe”. In diesem Fall wäre z.B. “dunkelhäutige Frauen” eine Gruppe.	Die Qualität des Modells soll für alle Gruppen ähnlich sein. Maße für Ähnlichkeit zweier Zahlen sind die **Differenz** (z.B. FPR von Gruppe 1 minus FPR von Gruppe 2) oder das **Verhältnis** (z.B. FPR von Gruppe 1 geteilt durch FPR von Gruppe 2).	Die messbare **Qualität** des Modells, z.B. die FPR. Die Qualitätsmetrik wird je nach Risiko ausgewählt.

Einige bekannte, konkrete Fairness-Definitionen sind z.B. [13]:

- **Vorhersageparität** (en: Predictive Parity), das entspricht der Genauigkeitsgleichheit. Das heißt, die Genauigkeit für jede Gruppe muss gleich hoch sein.
- **Chancengleichheit** (en: Equal Opportunity), das entspricht der Falsch-negativ-Raten-Gleichheit (kurz: FNR-Gleichheit). Das heißt, die FNR für jede Gruppe muss gleich niedrig sein.

- **Vorhersagegleichheit** (en: Predictive Equality), das entspricht der Falsch-positiv-Raten-Gleichheit (kurz: FPR-Gleichheit). Das heißt, die FPR für jede Gruppe muss gleich niedrig sein.

Schau in's Jupyter Notebook für die Festlegung von Qualitätsmetriken für *IdentiTOP*.

Schritt 3 Qualitätsmetriken berechnen 📏

An Hand der Modellergebnisse (Vorhersagen und Wahrheit) werden nun die ausgewählten Metriken berechnet.

Für die Fairness-Risiken werden die passenden Qualitätsmetriken für jede (intersektionale) Gruppe separat berechnet. Diese separate Qualitätsberechnung nennt man auch "Schnittanalyse" (en: slicing analysis).

Für den Vergleich der Qualitätsmetriken zweier Gruppen wird die Differenz oder das Verhältnis der Ergebnisse zueinander berechnet. Zum Vergleich der Qualität bei mehr als zwei Gruppen kann man zum Beispiel den Wert jeder Gruppe mit der am besten abschneidenden Gruppe vergleichen. Auch ein Vergleich mit dem durchschnittlichen Qualitätswert aller Gruppen wäre eine sinnvolle Variante.

⚠ Achtung: Nicht jede Metrik lässt sich direkt an Hand der Modellergebnisse berechnen. Es müssen möglicherweise weitere Daten erhoben bzw. abgefragt werden. Um Fairness-Definitionen prüfen zu können, müssen z.B. demographische Daten vorhanden sein.

Schau in's Jupyter Notebook für die Berechnung der Qualitätsmetriken von *IdentiTOP*.

Schritt 4 Qualitätsmetriken interpretieren 🧠

Als letzter Schritt werden die Berechnungen interpretiert. Lassen die Werte auf eine Erfüllung der Qualitätskriterien schließen? Welche Risikohypothesen haben sich bewahrheitet?

Ab welchem Schwellenwert man schließt, dass die Qualität ähnlich genug ist, um fair zu sein, muss je nach Kontext bestimmt werden. Häufig orientiert man sich an der "Disparate Impact"-Regel [12]. Danach sollte das Verhältnis zweier Qualitätsergebnisse zwischen 0,8 und 1,2 liegen, bzw. nach anderer Interpretation die Differenz zwischen -0,2 und 0,2 liegen. Wenn die Qualitätswerte sehr klein sind oder durch 0 geteilt werden müsste, bietet es sich an, die Differenz dem Verhältnis als Maß für Gleichheit vorzuziehen.

Häufig fällt die Aussage unterschiedlich aus, je nachdem welche Vergleichsart und welche Toleranzgrenzen gewählt wurden. Daher sollten stets alle Berechnungen offengelegt und ganzheitlich interpretiert werden.

Schau in's Jupyter Notebook für die Interpretation der Ergebnisse für *IdentiTOP*.

Schritt 5: Verbesserungsvorschläge erarbeiten 🛠️

Basierend auf den Ergebnissen der Qualitätsmessung kann man überlegen, was Gründe für die Ergebnisse sind (siehe auch [14]) und Empfehlungen abgeben, wie versucht werden könnte, das System zu verbessern.

Schau in's Jupyter Notebook für die Empfehlungen für die Verbesserung von *IdentiTOP*.

Quellen

1) BesteTipps (2023). Gesichtserkennung auf Snapchat: wie geht das? Lösung. Verfügbar unter: https://www.bestetipps.de/computer/handy/gesichtserkennung-auf-snapchat-wie-geht-das-loesung/ (Zugriff zuletzt: 02.08.2023)
2) Vossen, R. (2013). Zukunft der Werbung: „Cara“ erkennt Alter und Geschlecht – und zeigt das passende Plakatmotiv. Verfügbar unter: https://www.basicthinking.de/blog/2013/05/22/zukunft-der-werbung-cara-erkennt-alter-und-geschlecht-und-zeigt-das-passende-plakatmotiv/ (Zugriff zuletzt: 02.08.2023)
3) adaLearning (2019). Joy Boulamwini on Face Recognition Technology. Verfügbar unter: https://www.youtube.com/watch?v=rWMLcNaWfe0 (Zugriff zuletzt: 07.08.2023)
4) Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research (Vol. 81, pp. 1-15). Conference on Fairness, Accountability, and Transparency. (Zugriff zuletzt: 07.08.2023)
5) Leisegang, D. (2020). Biometrische Videoüberwachung: Die neue Hochrisikotechnologie. Blätter für deutsche und internationale Politik. Verfügbar unter: https://www.blaetter.de/ausgabe/2020/maerz/biometrische-videoueberwachung-die-neue-hochrisikotechnologie (Zugriff zuletzt: 07.08.2023)
6) Stiebert, J. (2022). Klage gegen Gesichtserkennung bei Uni-Prüfungen. Posteo News. Verfügbar unter: https://posteo.de/news/klage-gegen-gesichtserkennung-bei-uni-pr%C3%BCfungen (Zugriff zuletzt: 07.08.2023)
7) Wang, N. (2019). Totale Überwachung in Chinas Schulen: Wenn Kameras jede Gesichtsregung auswerten. Der Tagesspiegel. Verfügbar unter: https://www.tagesspiegel.de/politik/wenn-kameras-jede-gesichtsregung-auswerten-4657504.html (Zugriff zuletzt: 07.08.2023)
8) Schiller, A. (2021). Universitäten spähen Studenten mit Software aus. Frankfurter Allgemeine Zeitung. Verfügbar unter: https://www.faz.net/aktuell/karriere-hochschule/proctorio-und-wiseflow-hochschulen-spaehen-studenten-aus-17455837.html (Zugriff zuletzt: 07.08.2023)
9) Oppermann, A. (2021): Accuracy, Precision, Recall, F1-Score und Specificity. Verfügbar unter: https://artemoppermann.com/de/accuracy-precision-recall-f1-score-und-specificity/ (Zugriff zuletzt: 08.08.2023)
10) Hill, K., (2023): Eight Months Pregnant and Arrested After False Facial Recognition Match. Verfügbar unter: https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html (Zugriff zuletzt: 09.08.2023)
11) Wikipedia (2023): Confusion matrix. Verfügbar unter: https://en.wikipedia.org/wiki/Confusion_matrix (Zugriff zuletzt: 19.10.2023)
12) Wikipedia (2023): Disparate impact. Verfügbar unter: https://en.wikipedia.org/wiki/Disparate_impact (Zugriff zuletzt: 19.10.2023)
13) Verma, S. und Rubin, J. (2018). Fairness Definitions Explained. FairWare. Verfügbar unter: https://doi.org/10.1145/3194770.3194776
14) Suresh, H. und Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. EEAMO’21: Equity and Access in Algorithms, Mechanisms, and Optimization. Verfügbar unter: https://doi.org/10.1145/3465416.3483305