

Fair Machine Learning

Juliet Fleischer

11. Januar 2025



Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate reduzieren

Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate reduzieren
- Trainingsdaten: vergangene Polizeistops

Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate reduzieren
- Trainingsdaten: vergangene Polizeistops
- Zielvariable: Straftat begangen (0 = Nein, 1 = Ja)

Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate reduzieren
- Trainingsdaten: vergangene Polizeistops
- Zielvariable: Straftat begangen (0 = Nein, 1 = Ja)
- **Protected Attribute:** Ethnie

Gleiche Vorhersageraten zwischen Gruppen sind fair

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

Gleiche Vorhersageraten zwischen Gruppen sind fair

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Verständnis von Fairness: Personen erfahren aufgrund ihrer Gruppenzugehörigkeit Diskriminierung

Gleiche Vorhersageraten zwischen Gruppen sind fair

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Verständnis von Fairness: Personen erfahren aufgrund ihrer Gruppenzugehörigkeit Diskriminierung
- Vorhersageraten zwischen Gruppen sollen gleich sein

Gleiche Vorhersageraten zwischen Gruppen sind fair

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Verständnis von Fairness: Personen erfahren aufgrund ihrer Gruppenzugehörigkeit Diskriminierung
- Vorhersageraten zwischen Gruppen sollen gleich sein
- z.B. Statistical Parity

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b)$$

Fairness anhand der Fehlermatrix konstruieren

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

- **Sufficiency:** Zuverlässigkeit der Vorhersage soll gleich sein

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

- **Sufficiency:** Zuverlässigkeit der Vorhersage soll gleich sein
- z.B. Predictive Parity

$$P(Y = 1|A = a, \hat{Y} = 1) = P(Y = 1|A = b, \hat{Y} = 1)$$

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Zahlreiche Variationen von Gruppen-Metriken

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Sowohl Separation als auch Sufficiency können statt mit \hat{Y} auch mit S definiert werden, z.B. Calibration

$$P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$$

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Sowohl Separation als auch Sufficiency können statt mit \hat{Y} auch mit S definiert werden, z.B. Calibration

$$P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$$

- oder Well-calibration

$$P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s) = s$$

- Sufficiency nimmt Perspektive des Entscheidenden an
- Separation gut, wenn Y durch einen objektiv wahren Prozess entstanden ist
- Independence gut, wenn Form der Gleichheit erzwungen werden soll

- Sufficiency nimmt Perspektive des Entscheidenden an
 - Separation gut, wenn Y durch einen objektiv wahren Prozess entstanden ist
 - Independence gut, wenn Form der Gleicheit erzwungen werden soll
- ⇒ normalerweise nicht miteinander vereinbar

Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium

$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium
 - $$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
 - ▶ Definition des Distanzmaßes d_X im Feature Space ist eine Herausforderung

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium
- $d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$
- ▶ Definition des Distanzmaßes d_X im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
 - ▶ Definition des Distanzmaßes d_X im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding
 - ▶ Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
 - ▶ Definition des Distanzmaßes d_X im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding
 - ▶ Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden
 - ▶ Keine eindeutige mathematische Definition, sondern verschiedene Ansätze zum Testen von FTU

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness (FTA)
 - ▶ Lipschitz-Kriterium
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
 - ▶ Definition des Distanzmaßes d_X im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding
 - ▶ Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden
 - ▶ Keine eindeutige mathematische Definition, sondern verschiedene Ansätze zum Testen von FTU
 - ▶ Problem der **Proxis** (Variablen, die mit PA hoch korreliert sind)

Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

Wie sorgen wir für algorithmische Fairness?

- Preprocessing : Daten vor dem Training bearbeiten
z.B. (Re-)Sampling, Transformation

Wie sorgen wir für algorithmische Fairness?

- Preprocessing : Daten vor dem Training bearbeiten
z.B. (Re-)Sampling, Transformation
- Inprocessing: Trainingsprozess anpassen, Optimierungsproblem modifizieren
z.B. Regulaisierung

Wie sorgen wir für algorithmische Fairness?

- Preprocessing : Daten vor dem Training bearbeiten
z.B. (Re-)Sampling, Transformation
- Inprocessing: Trainingsprozess anpassen, Optimierungsproblem modifizieren
z.B. Regulaisierung
- Postprocessing:Vorhersagen nach dem Training bearbeiten
z.B. Thresholding

Wie sorgen wir für algorithmische Fairness?

- Preprocessing : Daten vor dem Training bearbeiten
z.B. (Re-)Sampling, Transformation
- Inprocessing: Trainingsprozess anpassen, Optimierungsproblem modifizieren
z.B. Regulaisierung
- Postprocessing:Vorhersagen nach dem Training bearbeiten
z.B. Thresholding
- Interpretable ML Methoden können hier auch sehr helfen!

Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

Woher kommt Bias?

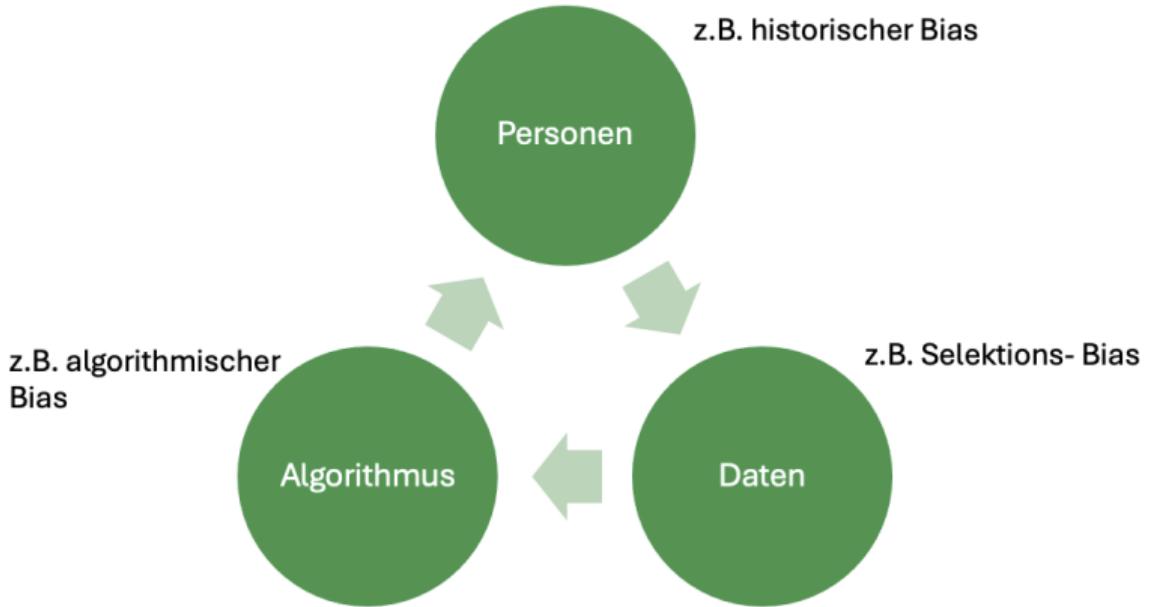


Abbildung: Quellen von Bias in der Daten, Nutzer, Algorithmus Feedback Loop

Wie geht es weiter?

- binäre Klassifikation, ein PA ist simpelster Fall

Wie geht es weiter?

- binäre Klassifikation, ein PA ist simpelster Fall
- in Praxis eher mehrere PAs und vielfältige Aufgaben
→ regression, unsupervised learning, ...

Wie geht es weiter?

- binäre Klassifikation, ein PA ist simpelster Fall
- in Praxis eher mehrere PAs und vielfältige Aufgaben
→ regression, unsupervised learning, ...
- Es gibt (noch) nicht, die **eine** Definition von Fairness

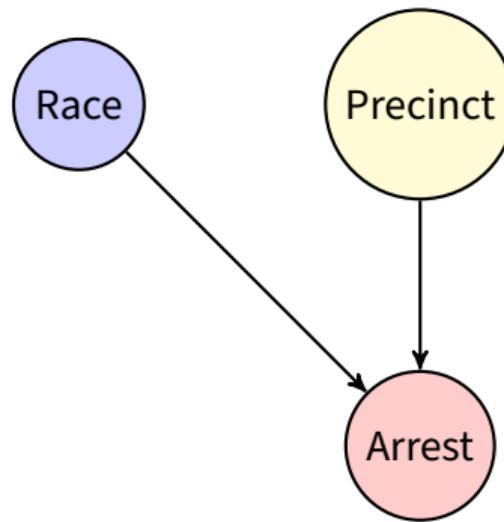
Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

- Gruppen Metriken spiegeln einfach Verteilung von Y, A, \hat{Y}, X in den Daten wider
- Accuracy und Fairness Trade-Off
- Folgen von Fairness Interventionen nicht sicher - profitiert die geschützte Gruppe?

Ist die Gruppenzugehörigkeit der Grund für die Festnahme?

- kausale Definitionen



- Gruppen Fairness: FACE, FACT (on average or on conditional average level)
- Individuelle Fairness: counterfactual fairness, path-based fairness

- Interaktives Tool:
<https://research.google.com/bigpicture/attacking-discrimination-in-ml/>
- Einführung in Fairness mit mlr3.fairness:
<https://journal.r-project.org/articles/RJ-2023-034/>
- Fairness and Machine Learning Buch: <https://fairmlbook.org/>

POC werden überproportional häufig gestoppt

Verteilung der Ethnie in NYC (2023)

<https://www.census.gov/quickfacts/newyorkcitynewyork>

Tabelle: Verteilung der Ethnie in SQF Daten

race	prop
BLACK	58.61%
WHITE HISPANIC	20.32%
BLACK HISPANIC	10.13%
WHITE	5.48%
OTHER	2.67%
NA	2.79%