Your Title

Your Name

January 2, 2025

Outline

1. Daten und Kontext
2. Fairness Definitionen
3. Quellen von Bias
4. Methoden für Fairness

Outline

1 Daten und Kontext

2 Fairness Definitionen

3 Quellen von Bias

4 Methoden für Fairness

Outline

1. Daten und Kontext
2. Fairness Definitionen
3. Quellen von Bias
4. Methoden für Fairness

Outline

1 Daten und Kontext

2 Fairness Definitionen

3 Quellen von Bias

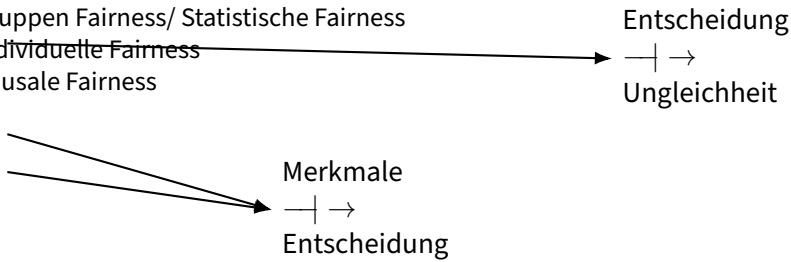
4 Methoden für Fairness

Daten und Kontext

- "begründeter Verdacht" ("reasonable suspicion") als Grund für Polizeikontrolle
- "wahrscheinliche Ursache" ("probable cause") als Grund für Festnahme
- Mögliche Resultate:
  - Keine weiteren Maßnahmen
  - Schnelle Abtastung (frisk)
  - Durchsuchung (search)
  - Festnahme (arrest)
- Kritik, dass Strategie diskriminierend gegenüber Schwarzen und Latinos ist
- Modellieren des Entscheidungsprozesses für Festnahme mit RF
- `race` als protected attribute (PA)

Fairness Definitionen

- Fairness Definitionen
  - Gruppen Fairness/ Statistische Fairness
  - ~~Individuelle Fairness~~
  - Kausale Fairness



Gruppen Fairness

Idee: Gruppen von Personen erfahren Diskriminierung aufgrund dieser Gruppenzugehörigkeit
Unterkategorien:

- Unabhängigkeit: $\hat{Y} \perp A$
- Separierbarkeit: $\hat{Y} \perp A|Y$
- Suffizienz: $Y \perp A|\hat{Y}$

Unabhängigkeit

Statistical Parity

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

Separierbarkeit

Fokus liegt auf gleichen Fehlerraten Herleitung anhand der Fehlermatrix

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	True Positive (TP)	False Negative (FN)
$Y = 0$	False Positive (FP)	True Negative (TN)

Table: Error Matrix

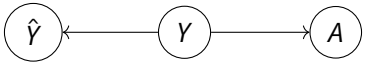


Figure: Simple Graphical Model

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

Suffizienz

Positive/negative Vorhersage soll die selbe Bedeutung für alle Gruppen haben

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

Erweiterung davon bedingt auf den Score, nicht das Label: Calibration

Advantages and Disadvantages

+	−
Leicht verständlich Leicht einzubauen in existierende Optimierungskriterien	Infragmarginality Observational

Table: Advantages and Disadvantages

–> Scheint erstmals sinvoller und simpler Ansatz zu sein. Schaut man etwas tiefer stellt sich die Fragem ob die Metriken wirklich geeignet sind um Fairness zu erhöhen Aktuelle Literatur weist darauf hin, dass sie eher Ausdruck der Zufallsverteilung von $\hat{Y}, Y, A, X$ sind

Individuelle Fairness

1. Fairness through Unawareness (FTU) = Blinding
   - Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden
   - Keine eindeutige mathematische Definition, sondern verschiedene Ansätze zum Testen von FTU
   - Erweiterung: PA und alle Proxies dürfen nicht im Entscheidungsprozess verwendet werden
2. Fairness through Awareness (FTA)
   - Ähnlichkeit im Feature Space soll zu Ähnlichkeit im Prediction Space führen
   - Lipschitz-Kriterium:
   $$d(\hat{y}_i, \hat{y}_j) \leq \lambda d(x_i, x_j)$$
   - Lässt sich in ein lineares Optimierungsproblem umformulieren
   - Definition des Distanzmaßes d im Feature Space ist eine Herausforderung

Kausale Fairness

Ziel ist hier die kausalen Struktur der Daten zu verstehen Fairness als pipeline Problem (Fairness + ML)

- Counterfactual Fairness
- Definition über DAGs

Fairness Methoden

- Preprocessing Daten vor dem Training bearbeiten
- Inprocessing Trainingsprozess anpassen, Optimierungsproblem modifizieren
- Postprocessing Vorhersagen nach dem Training bearbeiten