**Intro**

Dadurch dass ML Algorithmen immer mehr bei institutionellen Entscheidungen eingesetzt werden und in unsere Leben treten, gilt es sich zu Fragen inwieweit Vorhersagenvvon einem Algortihmus richtig sind. Richtig, im normativen Sinne. Also, sind die Voerhersagen fair? Frage nach Fairness hoch relevant, was zu dem eigenen Forschungsbereich FairML geführt hat.

Ich habe mich über die letzte Zeit mit damit beschäftigt, wie Fairness in Machine Learning definiert ist. Dabei auf zahlreiche Definitionen gestoßen. Nach diesem Vortrag habt ihr einen Einblick darüber, mit welchen Fragen sich das Feld FairML auseinandersetzt und vor allem wie wir über Fairness in Machine Learning denken.

*(Um alles etwas anschaulicher versetzen wir uns in den Kontext eines automatisiertes Entscheidungssystems. Wir wollen einen Algorithmus trainieren, der die Polizei dabei unterstützen soll zu entscheiden, ob eine Person festgenommen werden soll. Trainingsdaten kriegen wir von New Yorker Polizei. In New York gibt es die stop-and-frisk Polizeistrategie, d.h. Beamte dürfen Personen auf der Straße stoppen. Stop kann keine weiteren Maßnahmen mit sich ziehen oder die person kann durchsucht, oder verhaftet werden. Der Datensatz bietet uns Infos die Umstände des Stopps charakterisieren, und auch demographische Daten, wie Ethnie. Kann kritisch sein, aufgrund solcher Merkmale sollte nicht diskriminiert werden. Diese Art von Merkmalen nennen wir protected attribute.)*

*🡪 Ich halte es für sinnvoller tatsächlich für den Zweck der Klarheit doch den leichteren Datensatz zur Illustration zu nehmen. Ich würde das dementsprechend anpassen.*

Mit diesem Beispiel im Hinterkopf widmen wir uns erstmal der Frage 1) Woher Bias kommen kann? Dann bringe ich euch die 2) zwei grundlegenderen Ideen über Fairness in ML zu denken nahezubringen. Und zuletzt gebe ich euch einen kurzen 3) Überblick über Methoden um algorithmische Fairness herzustellen?

**1 Quellen von Bias**

* Nutzer/Personen 🡪 Daten 🡪 Algorithmus Kreislauf
* Vor allem wichtige Unterscheidung: Bias der in unseren gesellschaftlichen Strukturen eingebettet ist (historischer Bias) und Bias der durch unpassende statistische Methoden oder verzerrte Schätzer in Kreislauf kommt
* Mischung von mehreren auch möglich
* In unserem Kontext vermutlich historischer bias und selection bias
* Kreislauf da alle drei Stationen interagieren und bias sich verstärken kann, Feedback-Loop

Wichtig Gedanken darüber zu machen, welche Art von Bias für Situation relevant ist.

Kann beeinflussen, wie wir Fairness definieren

In ML zwei grundlegende Ansätze über Fairness zu denken

1. Gruppen Fairness 🡪 Gleichheit zwischen Gruppen
2. Individuelle Fairness 🡪 Gleichheit innerhalb der Gruppe

Auf Basis dieser Ideen haben sich zahlreiche Formalisierungen von Fairness ergeben, können hier nicht auf alle eingehen, sondern wichtigste Untergruppen + Idee dahinter

Wenn ich konkrete Definition nenne, dann mit englischem Namen, um Verwirrung zu vermeiden

Wir nehmen Standard Rahmen an, in dem auch in der Literatur Fairness erstmals formalisiert wird: binäre Klassifizierung, ein PA

**2.1 Gruppen Fairness - Independence**

Verständnis von Fairness, das zugrunde liegt: Es ist unfair, wenn Personne aufrgrund ihrer Gruppenzugehörigkeit (definiert durch PA) Diskriminierunge erfahren

Gruppenzugehörigkeit soll unabhängig von Vorhersage sein

z.B. Demographic parity fordert gleiche positive Vorhersageraten zwischen Gruppen d.h. in Gruppe a wird der gleiche Anteil als als positiv vorhergesagt wie in Gruppe b

**2.2 Gruppen Fairness – Separation + Sufficiency**

Sowohl Separation als auch Sufficiency können anhand Fehlermatrix hergeleitet werden

Bei Separation bedingen wir auf wahres Label Y 🡪 Fokus auf Fehlerraten

Für Sufficiency bedingen wir auf Vorhersage 🡪 Fokus auf Zuverlässigkeit der Vorhersage

**3. Individuelle Fairness**

FTA

FTU

Es kann komplexe Zusammenhänge zwischen Daten geben (Problem der Proxis, das simples FTU nicht adressiert). Das wird in gewisser Weise von kausalen Definitionen aufgenommen. Dritte große Gruppe, gehören zu den individuellen Fairness Metriken. Bedürfen eigener, ausführlicher Theorie, deshalb hier nicht näher eingehen (bei interesse gerne nach paper fragen)