Intro:

Wir haben jetzt viel darüber gehört wie ML Algorithmen interpretierbar gemacht werden können. Dadurch verstehen wir besser, wie die Entscheidung zustande kommt.

Was jetzt noch fehlt ist zu fragen, ob die Vorhersage des Algorithmus richtig im normativen Sinn ist? Ist die Vorhersage fair?

Dadurch dass ML Algorithmen immer mehr bei institutionellen Entscheidungen helfen sollen sehr relevant. Was zu dem eigenen Forschungsbereich fairML geführt hat.

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.

Mit diesem Beispiel im Hinterkopf widmen wir uns erstmal der Frage Woher Bias kommen kann? Der Fokus meiner Präsentation liegt dann darin euch die zwei grundlegenderen Ideen über Fairness in ML zu denken nahezubringen. Und zuletzt gebe ich euch einen kurzen Überblick über Methoden um algorithmische Fairness herzustellen?

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

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 🡪 Gleicheit 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 (Festnahme) , ein PA (Ethnie) das wir dichotomisieren

Gruppen Fairness

Verständnis von Fairness, das zugrundeliegt: Es ist unfair wenn Personne aufrgrund ihrer Gruppenzugehärigkeit (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 festgenommen vorhergesagt wie in Gruppe b