**Intro**

Dadurch dass ML Algorithmen immer mehr bei institutionellen Entscheidungen eingesetzt werden und in unsere Leben treten, gilt es sich zu Fragen inwieweit Vorhersagen von 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 dem Bereich FairML beschäftigt (wie viele andere auch). In diesem Vortrag trage ich zusammen, was ich über FairML gelernt habe um euch einen Einstieg in das Feld zu geben. Wir widmen uns dafür erstmal der Frage 1) Woher Bias kommen kann? Dann bringe ich euch die 2) zwei grundlegenderen Ideen über Fairness in ML zu denken nahe (Fokus der Präsentation) . Und zuletzt gebe ich euch einen kurzen 3) Überblick über Methoden um algorithmische Fairness herzustellen?

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

Running Example:

*ADM um Bewerberinnen für Universtität zu sortieren. Wollen besten Bewerber, höchste chance der von guten noten, guter arbeit etc.*

**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

In Software packeten umgesetzt

Als idfferenz oder Quotient

**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)