1. Fair Machine Learning

Fairness

When deploying machine learning systems in social-sensitive setting you may have to consider not only *performance/accuracy* but also *fairness*.

Take for instance a bank application for *loan selection*. Let your data matrix be:

$\mathbf{X}=$

	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

We want to model a decision Y that maximize the bank's profits as a function of the data $\mathbf{X}$:

$$Y = f(\mathbf{X})$$

Bias

What if $Y = f(\text{Ethnicity})$ varies strongly as a function of the ethnicity of the customer?

- The data set we learned from is *historically biased* and our system would then *reinforce* an existing social bias;
- The data set we learned from is *observationally biased* and our system would then *introduce* a new social bias.

The correlation between a sensitive variable (like ethnicity) and the output (like profit) is real in the data, and it helps maximize our objective. Yet, for ethical reasons, we do not want to exploit and worsen this bias.

Protected Attributes

Let us distinguish our features between *sensitive* or *protected* attributes $\mathcal{A}$ and standard features $\mathcal{X}$

	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

$$\mathcal{A} = \{\text{Ethnicity}\}$$
$$\mathcal{X} = \{\text{Postcode, Univ Degree, Monthly Income}\}$$

Fairness is defined with respect to these protected attributes. Definition is complex and subject to debate.

Case Study 1: Fairness through unawareness (is not fair!)

Let us discard *protected* attributes $\mathcal{A}$ and train the model only on the standard features $\mathcal{X}$.

Why is this not fair?

Case Study 1: Fairness through unawareness (is not fair!)

Let us discard *protected* attributes $\mathcal{A}$ and train the model only on the standard features $\mathcal{X}$.

ID	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

Even if we ignore *protected* attributes (like Ethnicity), some standard features (like Postcode) may be **highly correlated** with the protected attribute [3].

The same biases would then be re-inforced or introduced.

Case Study 2: COMPAS

Northpointe developed a model that given a set of attributes $\mathbf{X}$ of a defendant, would predict the degree of recidivism Y.

Image from propublica.org

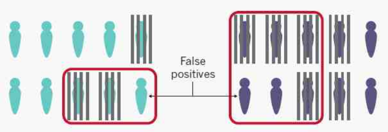


Image from [2]

ProPublica accused the tool of being *unfair*, with respect to *false positives*: more black defendants, later proved innocent, were classified as high risk.

Northpointe argued that their tool was *fair*, with respect to *prediction*: accuracy in classification among white/black defendants was the same.

Different measures of fairness may be **inconsistent** [1].

2. Casuality in Machine Learning

Correlation is not causation

It is well-known that machine learning systems learn *correlations*, not *causation*.

Take for instance an application to predict number of thefts. Let your data matrix be:

$\mathbf{X}=$

Ice-cream sold	Number of thefts
210	22
209	21
12	2
11	1
...	...

We want to model Theft $= f$ (Ice).

Prediction and Intervention

Is it correct to use the model in which the number of thefts is a *function* of the number of ice-cream sold?

Prediction and Intervention

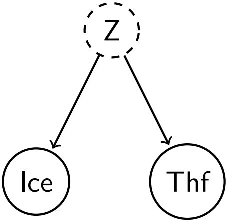
Is it correct to use the model in which the number of thefts is a *function* of the number of ice-cream sold?

We know that the number of ice-cream sold *does not cause* the number of thefts.
Yet:

- If you want only to **predict**, then the model is enough.
We captured a *predictive regularity*: from the cause we infer the effect, from the effect we infer the cause.
- If you want to **intervene**, then the model is not enough.
We need to know *relationship of cause and effect*: acting on the cause will change the effect, acting on the effect will leave the cause untouched.

Causal Models

Reasoning about causality is not trivial: it requires its own theory, its own statistical algorithms, its modelling practices [4].



Graphical models[1] are versatile tools to understand and reason about relationships of cause and effect.

[1]These DAGs are causal models and they are endowed with a semantics explained by the theory of causality.

Questions?

Feel free to ask questions at fabiomz@ifi.uio.no

References I

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[2] R Courtland. Bias detectives: the researchers striving to make algorithms fair., 2018.

[3] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[4] Judea Pearl. *Causality*. Cambridge university press, 2009.