Fairness in Machine Learning

Michael Yang   Prof. Anand Sarwate

DIMACS REU 2018

June 4, 2018

The debate over recidivism scores

The debate over recidivism scores

COMPAS is an algorithmic tool that predicts how likely jailed individuals will commit another crime

The debate over recidivism scores

COMPAS is an algorithmic tool that predicts how likely jailed individuals will commit another crime

In 2016, ProPublica published an analysis asserting that COMPAS treated black and white individuals differently

The debate over recidivism scores

COMPAS is an algorithmic tool that predicts how likely jailed individuals will commit another crime

In 2016, ProPublica published an analysis asserting that COMPAS treated black and white individuals differently

Resulted in long exchange between ProPublica, authors of COMPAS, and computer science community

Some ML/stats notation

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

y: what the ML program or decision-maker *thinks* Y is

What COMPAS got right

- Scores were *well-calibrated* (also called *equal positive predictive values*):

  $$E[Y = 1 \mid y = 0, A = \text{black}] = E[Y = 1 \mid y = 0, A = \text{white}]$$

  Translation: Black people with a score of 7 were as likely to recidivate as white people with a score of 7

What COMPAS got wrong

- Unequal false negative rates:

  $$E[y = 0 \mid Y = 1, A = \text{black}] \neq E[y = 0 \mid Y = 1, A = \text{white}]$$

  Translation: White people who would actually recidivate almost twice as likely to be scored "low risk"
- Unequal false positive rates:

  $$E[y = 1 \mid Y = 0, A = \text{black}] \neq E[y = 1 \mid Y = 0, A = \text{white}]$$

  Translation: Black people who would not actually recidivate almost twice as likely to be scored "higher risk"

New data: Mortgages

New data: Mortgages

Opportunity to take lessons from COMPAS and apply them to a new, different dataset

New data: Mortgages

Opportunity to take lessons from COMPAS and apply them to a new, different dataset

- 11.9 million observations (compared to 18,000 in COMPAS data)

New data: Mortgages

Opportunity to take lessons from COMPAS and apply them to a new, different dataset

- 11.9 million observations (compared to 18,000 in COMPAS data)
- Missing Y, the ground truth

New data: Mortgages

Opportunity to take lessons from COMPAS and apply them to a new, different dataset

- 11.9 million observations (compared to 18,000 in COMPAS data)
- Missing Y, the ground truth
- Loan approvals are decided using a combination of human and computer decision-making

New data: Mortgages

Opportunity to take lessons from COMPAS and apply them to a new, different dataset

- 11.9 million observations (compared to 18,000 in COMPAS data)
- Missing Y, the ground truth
- Loan approvals are decided using a combination of human and computer decision-making

End goals:
Understand (different kinds of) fairness on a new set of data.
Make it easier for new researchers to get caught-up with the fair ML conversation.

Additional/future technical directions

Other kinds of technical fairness:

Additional/future technical directions

Other kinds of technical fairness:

- Parity between metrics (e.g. equivalent *predictive accuracy* between groups)

Additional/future technical directions

Other kinds of technical fairness:

- Parity between metrics (e.g. equivalent *predictive accuracy* between groups)
- Conditional independence (e.g. acceptance is independent of race conditional on SAT score or $y \perp A \mid X$)

Additional/future technical directions

Other kinds of technical fairness:

- Parity between metrics (e.g. equivalent *predictive accuracy* between groups)
- Conditional independence (e.g. acceptance is independent of race conditional on SAT score or $y \perp A \mid X$)
- Absence of causal chains (best visualized with probabilistic graphical models)

Additional/future technical directions

Other kinds of technical fairness:

- Parity between metrics (e.g. equivalent *predictive accuracy* between groups)
- Conditional independence (e.g. acceptance is independent of race conditional on SAT score or $y \perp A \mid X$)
- Absence of causal chains (best visualized with probabilistic graphical models)

Relationship between the above kinds of fairness

Additional/future technical directions

Other kinds of technical fairness:

- Parity between metrics (e.g. equivalent *predictive accuracy* between groups)
- Conditional independence (e.g. acceptance is independent of race conditional on SAT score or $y \perp A \mid X$)
- Absence of causal chains (best visualized with probabilistic graphical models)

Relationship between the above kinds of fairness

Learning fair classifiers/predictors in addition to accurate ones

Acknowledgments

Funding Received from:
National Science Foundation CCF-1559855