Fairness Warnings provide interpretable boundary conditions for when a fairly trained model *may* behave unfairly.

Fairness Warnings

Dylan Slack, Sorelle Friedler, and Emile Givental

MOTIVATION

- When *shouldn't* you use your fair machine learning tool?
- Example: Suppose there's a fair recidivism tool trained in Chicago, will it behave fairly in Philadelphia?

METHODS

Idea: Practitioners often have access to *covariate information*

1. Perturb data set.
2. Label perturbation covariates by binary notion of model fairness (e.g. < 80% demographic parity).
3. Predict covariate shift fairness behavior with interpretable model.

COMPAS SLIM EXAMPLE

Predict UNFAIR DEMOGRAPHIC PARITY if SCORE < -1			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
priors_count	3.2 priors	20 points / prior	+.....
age	34.5 years	-2 points / year	+.....
ADD POINTS FROM ROWS 1 to 2 (Warning accuracy: 88%)		SCORE	=.....

Predict UNFAIR EQUAL OPPORTUNITY if SCORE < -19			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
priors_count	3.2 priors	24 points / prior	+.....
age	34.5 years	-2 points / year	+.....
ADD POINTS FROM ROWS 1 to 2 (Warning accuracy: 86%)		SCORE	=.....

LIMITATIONS

- Warnings only suggest that there *may* be unfairness in particular application; fairness warnings are based on summary statistics and may not capture true covariate shift behavior.

HAVERFORD COLLEGE

UCI University of California, Irvine

Check us out at FAT* 2020:

Fairness Warnings & Fair-MAML: Learning Fairly from Minimal Data

Scan code for arXiv.