FAIRNESS WARNINGS & FAIR-MAML: LEARNING FAIRLY FROM MINIMAL DATA

Dylan Slack (UC Irvine)

Joint work with **Sorelle Friedler** and **Emile Givental**
(Haverford College)

Suppose you wish to use a group fair classification tool but have access to little or no training data.

Suppose you wish to use a group fair classification tool but have access to little or no training data.

➢ *Should you use a pre-trained fair model?*

Suppose you wish to use a group fair classification tool but have access to little or no training data.

- *Should you use a pre-trained fair model?*
- *Could you train a fair model with only a few labeled instances?*

RISK PREDICTION MOTIVATING EXAMPLE

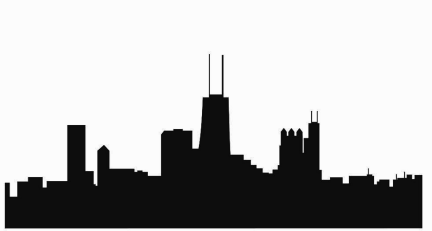
City of Philadelphia

$$\hat{f}(X)$$

RISK PREDICTION MOTIVATING EXAMPLE

City of Philadelphia

$\hat{f}(X)$



City of Chicago

$\hat{f}(?)$

Demographic differences
(Covariate shift)

OUR CONTRIBUTIONS

1. **Fairness Warnings:** provide ***interpretable boundary conditions*** for when a fair-ml model ***may*** behave unfairly.

OUR CONTRIBUTIONS

1. **Fairness Warnings:** provide ***interpretable boundary conditions*** for when a fair-ml model ***may*** behave unfairly.
2. **Fair-MAML:** train fair models from a ***minimal*** amount of data using ***meta-learning***.

OUR CONTRIBUTIONS

1. **Fairness Warnings:** provide ***interpretable boundary conditions*** for when a fair-ml model ***may*** behave unfairly.
2. **Fair-MAML:** train fair models from a ***minimal*** amount of data using ***meta-learning***.
3. ***Connect*** both methods by applying Fairness Warnings on Fair-MAML.

FAIRNESS WARNINGS

- Verifying fairness under distribution shift in lay user presentable way is difficult.

FAIRNESS WARNINGS

- Verifying fairness under distribution shift in lay user presentable way is difficult.
- Instead, provide ***warnings*** for when changes ***may results in unfairness*** according to ***summary statistics***.

EXAMPLE: COMPAS RECIDIVISM RISK PREDICTION USING SLIM [1,2]

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%, true positive rate: 88%, true negative rate: 89%)		SCORE	=.....

[1] Angwin et. at., 2016. Machine bias. ProPublica.
[2] Zeng et. al., 2017. Interpretable Classification Models for Recidivism Prediction. Journal of the Royal Statistical Society Series A.

EXAMPLE: COMPAS RECIDIVISM RISK PREDICTION USING SLIM [1,2]

Predict UNFAIR DEMOGRAPHIC PARITY if SCORE < -1 Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
priors_count	3.2 priors	20 points / prior	+-20
age	34.5 years	-2 points / year	+4
ADD POINTS FROM ROWS 1 to 2 (Warning accuracy: 88%, true positive rate: 88%, true negative rate: 89%)		SCORE	=-16

- Ex. Mean priors count decreases 1 prior, age decreases 2 years.

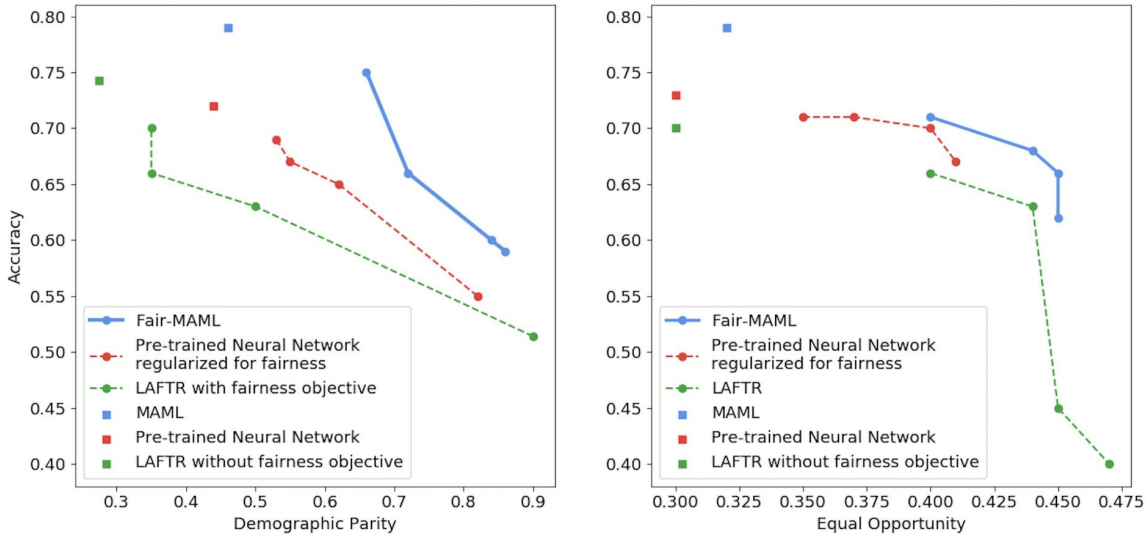
[1] Angwin et. at., 2016. Machine bias. ProPublica.
[2] Zeng et. al., 2017. Interpretable Classification Models for Recidivism Prediction. Journal of the Royal Statistical Society Series A.

FAIR-MAML

- There are ***minimal*** training examples from fairness task.
- Idea: train ***meta-model*** that can be fine-tuned to particular task using ***minimal data*** that achieves both ***fairness and accuracy*** using modified ***model agnostic meta-learning (MAML)*** [1]***.***

[1] Chelsea Finn et. al., 2017.

EXAMPLE: COMMUNITIES AND CRIME STATE BY STATE RISK PREDICTION



- Using small amount of fine-tuning data, works better than baselines.

THANKS!

- Check out (QR linked) paper and code for:
  - Joint Fairness Warnings, Fair-MAML experiments.
  - Extended experimenta evaluation.