Oracle Labs

Unlocking Fairness: a Trade-off Revisited

Michael Wick, Swetasudha Panda, Jean-Baptiste Tristan

Oracle Labs, MA

Introduction

- Much work studies the relationship between fairness and accuracy.
- A common conclusion is the relationship is a trade-off.
- But it is important to clarify that the underlying assumption is neither the data or labels are biased.
- However, fairness can arise because either the data or labels are biased.
- And if we evaluate accuracy against biased ground-truth, *then the accuracy is biased too.*

Contributions

- We study the relationship between fairness and accuracy, but accounting for bias in the data and labels.
- When accounted for, we find that fairness can often *improve* accuracy.
- Inspired by semi-supervised approaches like GE and posterior regularization, we propose a semi-supervised fairness method that harnesses fairness as training signal.
- We find that the method can impart beneficial qualities of unlabeled data to unfair training data and surpassing.

Fairness regimes

In machine learning theory we often assume

- a data distribution $\mathcal{D}$,
- and a labeling function f.

Either or both of which could be biased.

- Because of **selection bias**, the data distribution might be wrong ($\mathcal{D}'$)
- Because of **label bias**, the labeling function might be wrong (f').

For example, due to implicit bias, a manager might make hiring or promotional decisions that are unfair to individuals with a protected attribute such as gender or race or age. This means that any accuracy evaluated against such labels must also be biased.

Usually when concluding that accuracy and fairness is a tradeoff, there is an implicit assumption that the labels are *correct.*

In this work, we wonder whether the relationship between fairness and accuracy would actually change if we recognized that the labels are *incorrect.*

Fairness

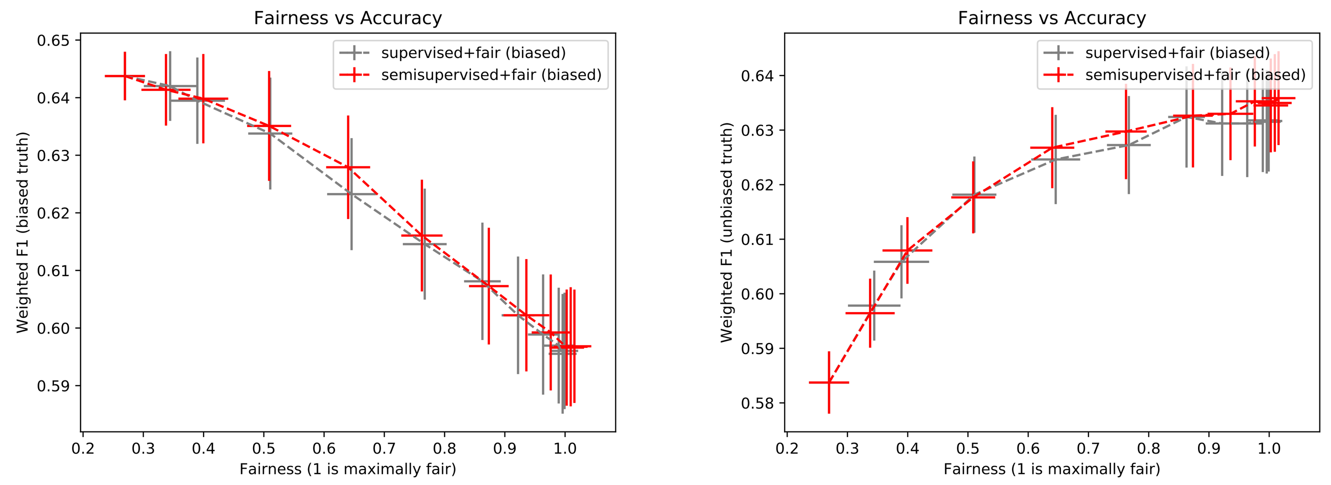
Fairness: we employ demographic parity, which measures the ratio between favorable outcomes between protected and unprotected classes.
Semi-supervised fairness: We employ a soft version of this as a training constraint: the ratio of the probability of the classifier assigning the favorable outcome on unlabeled data should be one.

Experimental Strategy

- Train on biased data (label bias, selection bias).
- Evaluate data on both biased and unbiased data.
- Key challenge: we don't have access to unbiased labels.

Illustrative experiment

Suppose we vary the extent to which a classifier enforces fairness and plot the accuracy for different amounts of fairness.



(a) Truth has label bias (b) Truth has no label bias

Systems

All classifiers are trained on biased labels except for the oracle baseline which is trained on the unbiased labels.

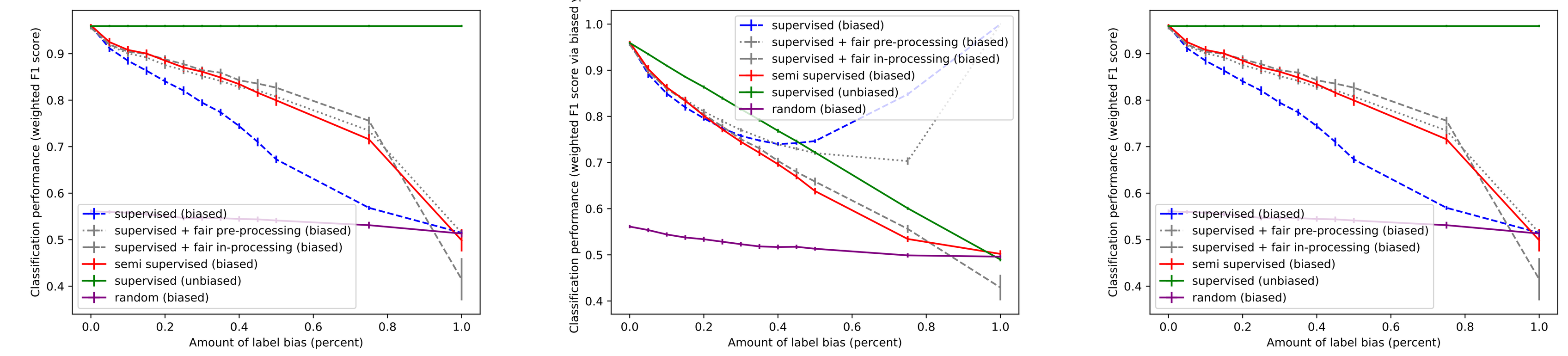
- Supervised logistic regression.
- Fair logistic regression (in-processing).
- Fair logistic regression ("reweighing" pre-processing; Kamiran & Calders 2012).
- Fair logistic regression (semi-supervised).
- Random (choose label proportionally at random).
- Oracle: supervised logistic regression trained on unbiased labels.

Data Generating Model

Want: Data of the form $\mathcal{D} = \{x, \rho, z, y\}$ in which x is the vector of unprotected attributes, ρ is the binary protected attribute, z is the (typically unobserved) label that has no label bias and y is the (typically observed) label that may have label bias.
Problem: we do not have access to unbiased labels z in the real-world.
Solution: assume a probabilistic model of label bias $y \sim g(y|z, \rho, x, \beta)$

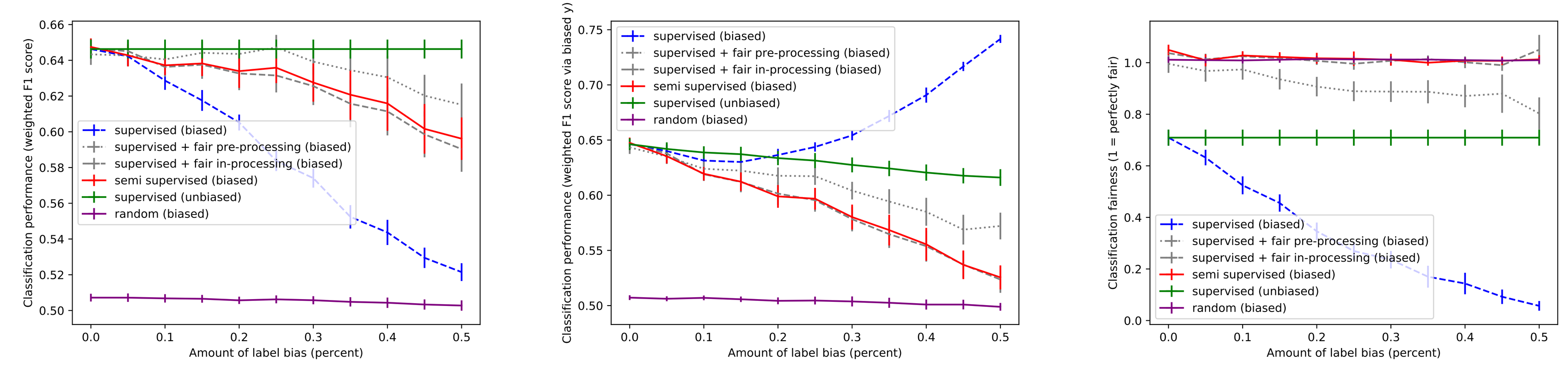
- Simulate data from scratch from $P(\mathcal{D}) = P(x, \rho, z, y) = g(y|z, \rho, x, \beta)P(z, \rho, x)P(\beta)$.
- Simulate labels for an existing data set (COMPAS) by assuming the labels are correct and simulating incorrect labels from g.

Experiment: Varying Label Bias



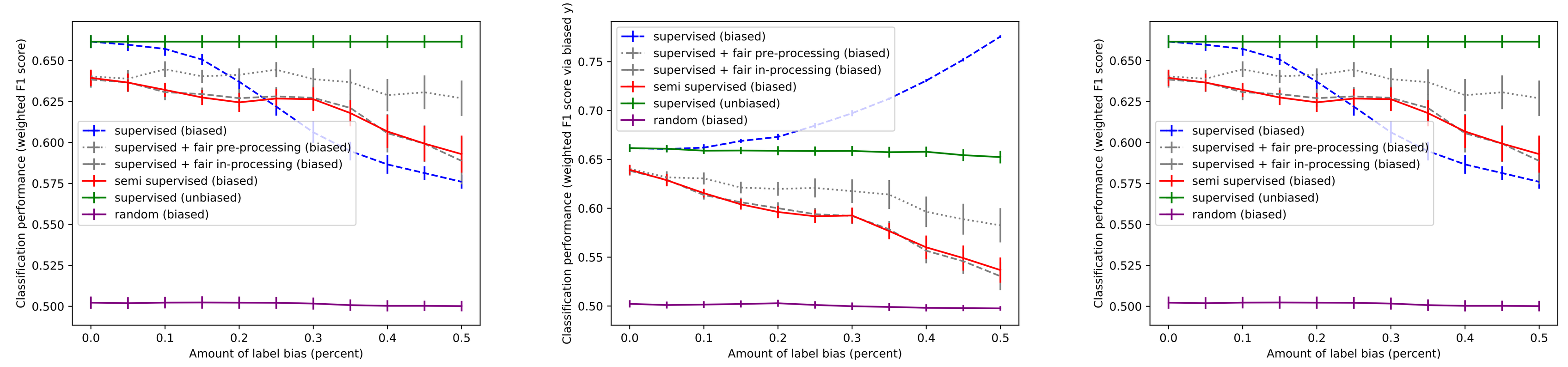
(a) F1 (unbiased truth) (b) F1 (biased truth) (c) Fairness

Figure: Classifier accuracy (F1) and fairness as a function of the amount of label bias.



(a) F1 (unbiased truth) (b) F1 (biased truth) (c) Fairness

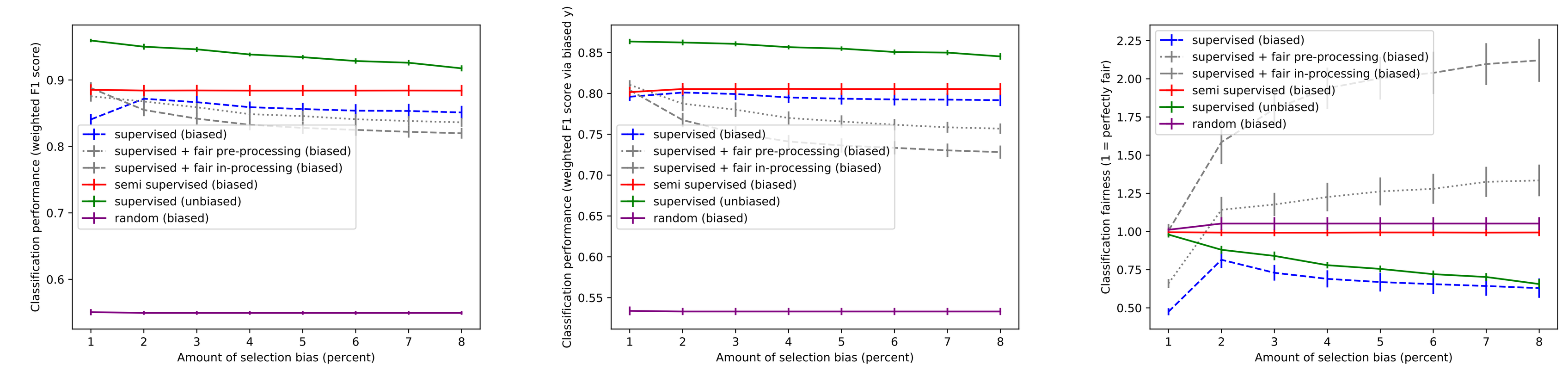
Figure: Varying label bias on COMPAS (assumption holds).



(a) F1 (unbiased truth) (b) F1 (biased truth) (c) F1 (unbiased truth (v))

Figure: Varying label bias on COMPAS (assumption violated).

Experiment:Varying selection bias



(a) F1 (unbiased truth) (b) F1 (biased truth) (c) Fairness

Figure: Classifier F1 and fairness as a function of the amount of selection bias.

URL: labs.oracle.com

Email: {michael.wick}@oracle.com