Fairness by Adjusting Thresholds?

Janek Willeke

Master's thesis

Introduction

Fair Machine Learning

- Decisions that algorithms make increasingly impact people's wellbeing
- Fair Machine learning aims to make these decisions fair
- Different notions of fairness exist along with models optimizing for them

Disparate impact

One such notion of discrimination is that a classifier's prediction $\hat{y} \in \{-1, 1\}$ should not depend on the sensitive variable s. **Disparate impact** quantifies the degree of dependence.

$$di(\hat{y}) = \frac{min(P(\hat{y} = 1, s = 0), P(\hat{y} = 1, s = 1))}{max(P(\hat{y} = 1, s = 0), P(\hat{y} = 1, s = 1))}$$

Fair algorithms

- A range of classifiers have been proposed that jointly optimize for accuracy and fairness
- But: any margin-based classifier can be made fair on training data by **thresholding**.
- **Thresholding**: adjusting group thresholds such that the predicted positive rate for both sensitive groups is equal
- provably optimal under strong assumptions [KLMR18]

Research Objective

Research objective

Compare performance of fair in-processing classifiers with thresholded classifiers

- Relate performance differences to classifier characteristic
- attempt to explain how classifier characteristics cause performance differences

Classifiers

Classifiers and Constraints

- **Adversarially constrained models**[ES15][ZLM18] predict labels and prevent adversary from predicting sensitive group from logits or penultimate layer activations
- **Variational Fair Autoencoder**[LSL+15] is a gaussian latent variable model. In the generative model, latent variables are independent of s. Additional MMD penalty on penultimate layer
- Adversarially constrained models and variational fair autoencoder enforce **independence between margins and** s in the input space (logistic regression) or penultimate layer(neural networks)

Classifiers and Constraints

- **Penalized Logistic Regression** constrains sums of logits [KAAS12] or sums of distances to the decision boundary [ZVRG17] to be equal across groups
- **Thresholding** shifts the decision boundary for each sensitive group s such that accuracy is maximized given that $di \leq p$, where p is the desired fairness constraint
- Thresholding only **constrains the cdfs of the logits at the tresholds** to be equal across sensitive groups

Evaluating performance

Evaluating Performance

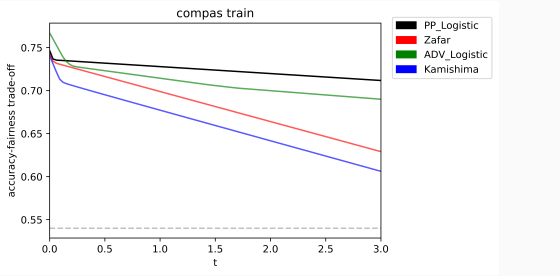
- Models selected on validation data based on their performance for varying values of the fairness-accuracy trade-off t
- 5 data sets of varying size: n = 999 to n= 45222

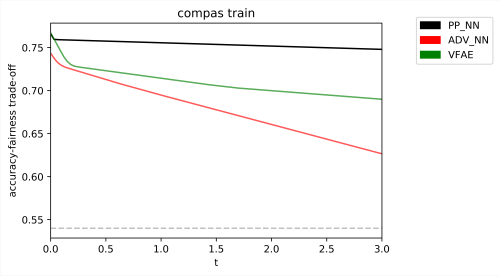
Performance of a fair classifier

$$perf = \frac{1}{N} \sum^{N} 1_{(y_i = \hat{y}_i)} - t * (1 - di)$$

Results

Thresholded models always achieves higher performance on training data

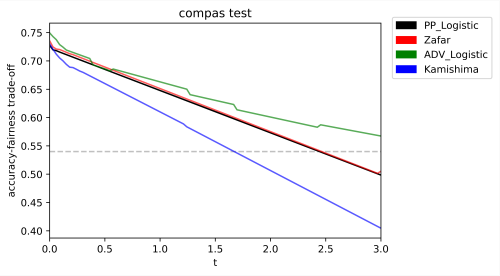


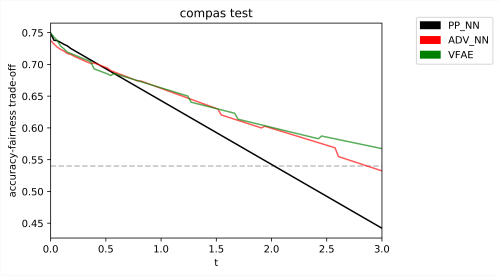


X-axis are fairness-accuracy trade-offs, y-axis performance.**Thresholded model in black**.

Results

Yet on test data performance is inferior to the models enforcing indepedence between the margins and s

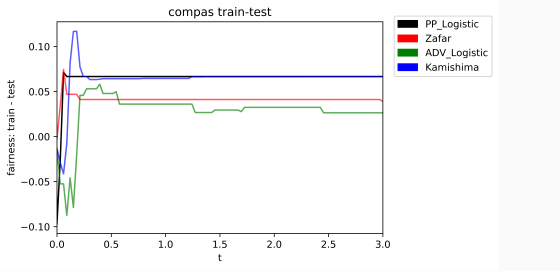


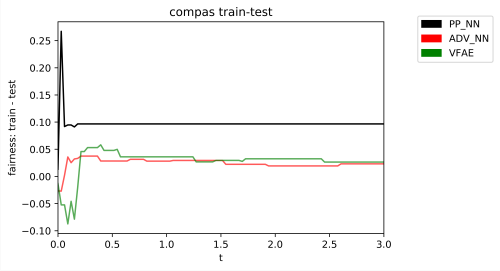


X-axis are fairness-accuracy trade-offs, y-axis performance.**Thresholded model in black**.

Results

This can be explained by a **drop in fairness** going from train to test set. Models reducing dependencies between predictions and sensitive variable do generalize better

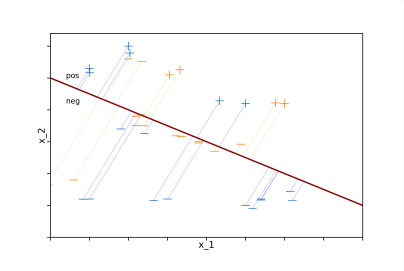


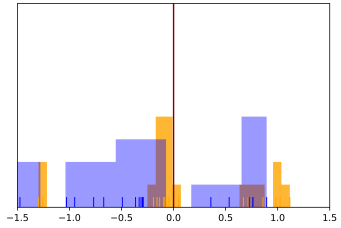


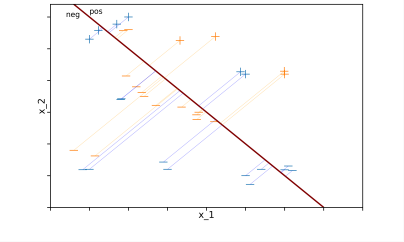
X-axis are fairness-accuracy trade-offs, y-axis is generalization error. **Thresholded model in black**

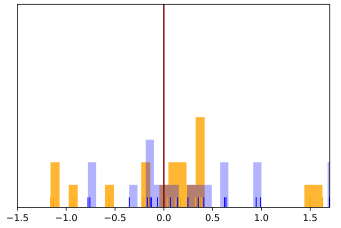
Explaining the Generalization Gap?

Margin Distribution and Generalization







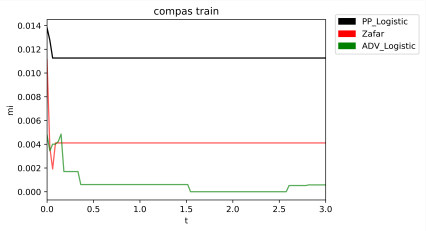


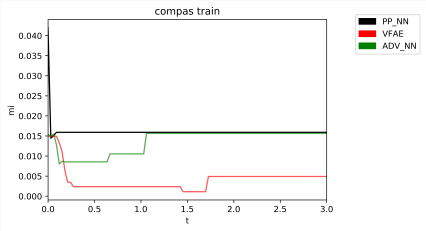
Orange and blue are sensitive groups, (+,-) labels

Margin Distribution and Generalization

Hypothesis: Mutual information between m(x,s) and s low $\Longrightarrow$ Good generaliazion

- If hypothesis were true, then variance in mutual information of trained models should explain variance in generalization





Y-axis is mutual information, x-axis is s, **Thresholded model in black**

Conclusions

Conclusions

- Thresholding performs well on training data but does not generalize
- There exists a **fairness generalization gap** and there is no research that can explain it
- Yet fair algorithms need to generalize well in order to work in the real world

- Harrison Edwards and Amos Storkey, *Censoring representations with an adversary*, arXiv preprint arXiv:1511.05897 (2015).
- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma, *Fairness-aware classifier with prejudice remover regularizer*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2012, pp. 35–50.
- Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan, *Algorithmic fairness*, AEA Papers and Proceedings **108** (2018), 22–27.
- Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, *The variational fair autoencoder*, arXiv preprint arXiv:1511.00830 (2015).

- Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, *Mitigating unwanted biases with adversarial learning*, Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2018, pp. 335–340.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi, *Fairness constraints: Mechanisms for fair classification*, Artificial Intelligence and Statistics, 2017, pp. 962–970.