Minimax Group Fairness: Algorithms and Experiments

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth

ediana@wharton.upenn.edu

June 23, 2022

Introduction

- Machine learning researchers and practitioners have often focused on achieving group fairness with respect to protected attributes (race, gender, ethnicity, etc.)
- **Equality of error rates** is one of most intuitive and well-studied group fairness notions
- But in practice, equalizing error rates and similar notions may require **artificially inflating error** on easier-to-predict groups and **may be undesirable** for a variety of reasons

Motivation

- There are many social applications of machine learning in which most/all of the targeted population is disadvantaged
- Might be interested in ensuring predictions are roughly equally accurate across racial groups, income levels, geographic location, etc
  - But, if this can only be achieved by raising lower group error rates, then we have worsened overall social welfare
- Therefore, might be preferable to consider the alternative fairness criterion of **minimax group error**, recently proposed by [Martinez, 2020]
  - Seek not to equalize error rates, but to minimize largest group error rate, making sure that **the worst-off group is as well-off as possible**

Contributions

1. Propose two algorithms, both two player zero-sum games:
    1. MinimaxFair: Finds a minimax group fair model from a given statistical class
    2. MinimaxFairRelaxed: Finds a model that minimizes overall error subject to the constraint that all group errors must be below a predetermined threshold
        - **Navigates tradeoffs** between a relaxed notion of minimax fairness and overall accuracy

Contributions

3. Prove that both algorithms converge and are oracle efficient. We also study their generalization properties.
4. Show how our framework can be extended to handle different types of error rates, such as false positive (FP) and false negative (FN) rates, as well as overlapping groups
5. Provide a thorough experimental analysis of our two algorithms under different prediction regimes

MinimaxFair vs. Equal Errors for Regression

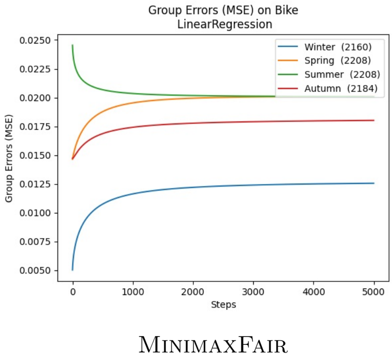
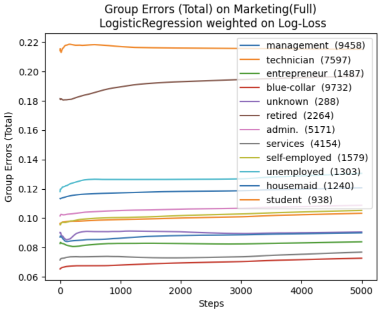


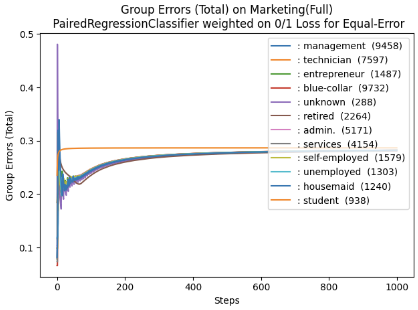
Figure: Comparison of Minimax and Equal Error Solutions on Seoul Bike Dataset

Public bikes rented at each hour in Seoul Bike sharing system
Label: Rented bikes (normalized), *Group*: Season

MinimaxFair vs. Equal Errors for Classification



MinimaxFair

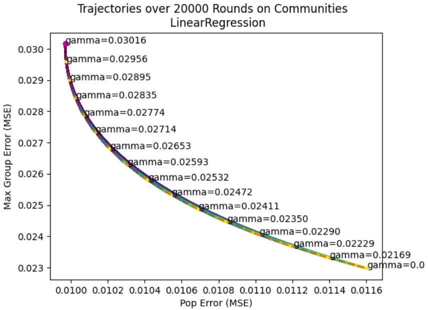


Equal Errors

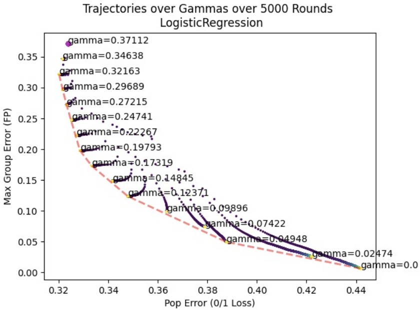
Figure: Comparison of Minimax and Equal Errors on Marketing Dataset

Direct marketing campaigns (phone calls) of a Portuguese bank
Label: client subscribes term deposit, *Group*: Job

Fairness Accuracy Tradeoff with MinimaxFairRelaxed



Linear Regression on Communities Dataset



Classification (FP) on COMPAS Dataset

Figure: Fairness Accuracy Tradeoff Curves

Communities and Crime: US Communities, 1990 - 1995
Label: Violent crimes per population, *Group*: Race
COMPAS: Arrest data from Broward County, Florida
Label: Two year recidivism, *Groups*: Race, sex

Generalization Results

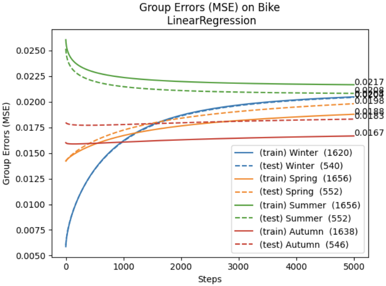
- With probability $1-\delta$, generalization gap per group bounded by

$$O\left(\sqrt{\frac{\log\frac{1}{\delta}+d\log n_i}{n_i}}\right)$$

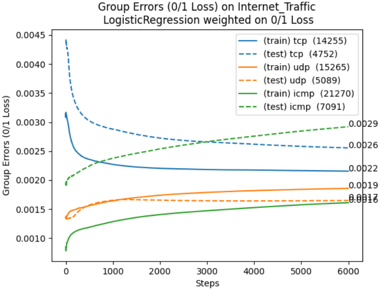
  where d is VC dimension of class H, and n_i is sample size of group i
- Generalization gap for *minimax group* is bounded by

$$O\left(\max_i\sqrt{\frac{\log\frac{K}{\delta}+d\log n_i}{n_i}}\right)$$

  i.e. dominated by sample size of the *smallest* group



Bike Dataset



Internet Traffic Dataset

Figure: Train vs. Test Performance of MinimaxFair

Network connection data used to distinguish between 'bad' connections, called intrusions or attacks, and 'good' normal connections.
Label: Connection Legitimacy, *Group*: Protocol Type

References

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu (2018)
Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness
Proceedings of the 35th International Conference on Machine Learning,Stockholm, Sweden, PMLR 80, 2018.

Natalie Martinez, Martin Bertran, and Guillermo Sapiro (2020)
Minimax Pareto Fairness: A Multi Objective Perspective
Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

Dheeru Dua and Casey Graff (2017)
UCI Machine Learning Repository
University of California, Irvine, School of Information and Computer Sciences

Yoav Freund and Robert E. Schapire (1996)
Game Theory, On-line Prediction and Boosting
Proceedings of the Ninth Annual Conference on Computational Learning Theory, 1996.

Alekh Agarwal and Alina Beygelzimer and Miroslav Dudík and John Langford and Hanna Wallach (2018)
A Reductions Approach to Fair Classification
Proceedings of the 35th International Conference on Machine Learning, 2018.

Alexandra Chouldechova and Diana Benavides-Prado and Oleksandr Fialko and Rhema Vaithianathan (2018)
A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions
Conference on Fairness, Accountability and Transparency, 2018.