Minimax AUC Fairness: Efficient Algorithm with Provable Convergence

Machine Learning in Practice Reading Group

Duke B&B

January 29, 2024

Presented by Luke Wang

Section 1: Introduction

Purpose

- Current AUC based algorithms aiming to improve fairness of the prediction model fail to account for all possible disparate effects.
- Find a scoring function that maximize the minimum of inter-group and intra-group AUC
- The proposed algorithm addresses both inter-group and intra-group disparities, with proven convergence. It demonstrates an improvement in prediction fairness while maintaining high prediction accuracy across various datasets.

Intuition: The probability of ranking a positive individual before negative individual should have minimal dependence on whether these individuals belong to the same demographic group or not.

Potential Applications

- Recidivism, Loan Approval......
- Identify high-risk group and guide healthcare resource allocation
- Application of xCI

Group Conditioned AUC

- $AUC_{z,z'}(f_\theta) = \mathbb{E}\left[\mathbb{I}\left(f_\theta(X) > f_\theta(X')\right) | Y = 1, Y' = -1, Z = z, Z' = z'\right]$, where Z is the group sensitive attribute eg. sex, race, f_θ is the prediction model, Y indicates the true label, and X dentes covariates
- $AUC_{a,a}(f_\theta) = AUC_{a,b}(f_\theta) = AUC_{b,a}(f_\theta) = AUC_{b,b}(f_\theta)$ indicates perfect fairness
- The chance of a qualified candidate from any gender ranking higher than an unqualified candidate from any gender is the same (Rawlsian Principle of Maximin Welfare for Distributive Justice)
- The interpretation depends on the purpose of the prediction model
- "The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric" by Nathan Kallus and Angela Zhou

Section 3: Methods - Problem Setup and Notation

- They are attempting to optimize:

$$\max_{\theta \in \Theta} \min_{z, z' \in \mathcal{Z}} AUC_{z,z'}(f_\theta)$$

- Since we only have access to empirical estimates, denoted by $A\hat{U}C$:

$$A\hat{U}C_{z,z'}(f) = AUC(f; S^{z+}, S^{z'-}) = \frac{1}{n^{z+} + n^{z'-}} \sum_{i=1}^{n^{z+}} \sum_{j=1}^{n^{z'-}} \left[\mathbb{I}(f(\mathbf{x}_i^{z+}) > f(\mathbf{x}_j^{z'-})) \right]$$

- As the indicator function $\mathbb{I}$ is not differentiable, it is substituted with a surrogate loss function for optimization.

$$\hat{R}_\ell(\cdot; S) = \left(\hat{R}_\ell(\cdot; S^{z+}, S^{z'-}) \right)_{z,z' \in \{a,b\}}$$

- The problem then becomes:

$$\min_{\theta \in \Theta} \max_{z,z' \in \mathcal{Z}} \hat{R}^{\ell}_{z,z'}(\theta)$$

- Again this is not differentiable, we ease this into a zero-sum game:

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} F'(\theta, \lambda) = \lambda^T \hat{R}'_{\ell}(\theta) = \sum_{z,z' \in \mathcal{Z}} \lambda_{z,z'} \hat{R}^{\ell}_{z,z'}(\theta)$$

$$\text{where } \Lambda = \left\{ \lambda \in \mathbb{R}^4 \,\middle|\, \sum_{z,z' \in \mathcal{Z}} \lambda_{z,z'} = 1, \lambda_{z,z'} \geq 0 \right\} \text{ is a } 2 \times 2\text{-dimensional simplex}$$

Section 3: Methods - MiniMax Algorithm

Algorithm 1 MinimaxFairAUC

Inputs: Training set S with label Y and protected attribute Z, model f_θ, number of iterations T, batch size m, learning rates $\{\eta_\theta, \eta_\lambda\}$
Initialize $\theta_0 \in \Theta$ and $\lambda_0 \in \Lambda$ with $\lambda_{z,z'} = \frac{n^{z+} + n^{z'-}}{n^+ + n^-}$ for all $z, z' \in \mathcal{Z}$
for $t = 1$ to $T - 1$ **do**
    $B_t = \text{StratifiedSampler}_m(S; Y, Z)$
    $\theta_t = \theta_{t-1} - \eta_\theta \nabla_\theta \hat{R}'_\ell(\theta_{t-1}; B_t)$
    $\gamma_t = \lambda_{t-1} \exp(\eta_\lambda \nabla_\lambda \hat{R}'_\ell(\theta_{t-1}; B_t))$
    $\lambda_t = \gamma_t / \|\gamma_t\|_1$
end for
Outputs: $\theta_T \sim \text{Unif}(\{\theta_t\}_{t=1}^T)$

Section 4: Theoretical Results - Assumption for Guaranteed Convergence

- **Assumption 1**
  For any $\theta \in \Theta$ and $\lambda \in \Lambda$, the gradients of F are bounded by G_θ and G_λ respectively, i.e.,

$$\|\lambda^T \nabla_\theta R'(\theta; S)\|_2 \leq G_\theta, \quad \text{and} \quad \|R'(\theta; S)\|_\infty \leq G_\lambda.$$

- **Assumption 2**
  The objective F is L_θ and L_λ smooth respectively, i.e.,

$$\|\lambda^T \nabla_\theta R'(\theta; S) - \lambda^T \nabla_\theta R'(\theta'; S)\|_2 \leq L_\theta \|\theta - \theta'\|_2,$$

$$\text{and} \quad \|R'(\theta; S) - R'(\theta'; S)\|_\infty \leq L_\lambda \|\lambda - \lambda'\|_1$$

  for any $\theta, \theta' \in \Theta$ and $\lambda, \lambda' \in \Lambda$.

- **Assumption 3**
  For any fixed $\theta \in \Theta$, $\lambda \in \Lambda$ and randomly sampled pair ξ, the variances of the stochastic gradients are bounded by σ_θ^2 and σ_λ^2 respectively.

Theorem 2 (Informal). Suppose Assumption 1, 2 and 3 hold true. Then the output θ_T of Algorithm 1 satisfies

$$\mathbb{E}\left[\|\nabla P_{1/2L}(\theta_T)\|_2\right] \leq \epsilon(T, \eta_\theta, \eta_\lambda),$$

where $\epsilon(T, \eta_\theta, \eta_\lambda)$ is an absolute constant. In particular, to achieve some small $\epsilon = \epsilon(T, \eta_\theta, \eta_\lambda)$, one chooses $\eta_\theta = \Theta(\epsilon^4)$, $\eta_\lambda = \Theta(\epsilon^2)$ and $T = \Theta(\epsilon^{-8})$. Furthermore, there exists $\hat{\theta} \in \Theta$ such that $\mathbb{E}\left[\|\hat{\theta} - \theta_T\|_2\right] \leq \epsilon/2L$ and it satisfies

$$\mathbb{E}\left[\min_{\xi \in \partial P(\hat{\theta})} \|\xi\|_2\right] \leq \epsilon.$$

Section 5: Experiment Results - Overview

- The author implemented algorithms on four datasets – Bank, Adult, Compas, and Default

Name	# instances	# attributes	Group ratio	Class ratio
Adult	48,842	15	0.48:1	3.03:1
Bank	41,188	21	0.05:1	7.55:1
Compas	11,757	53	1.86:1	1.94:1
Default	30,000	24	1.52:1	3.52:1

- They used fully connected NN of 2 hidden layer with ReLU activation and normalization
- They compared their metric with four algorithms – AUC Max, MiniMaxFair, InterFairAUC, EqualAUC

Section 5: Experimental Results

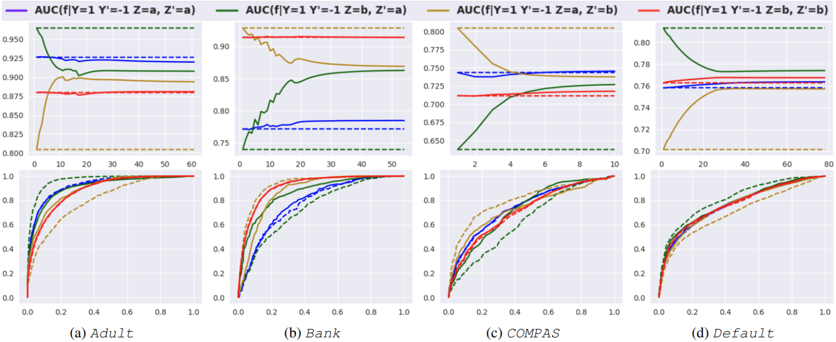

Figure: Convergence plots on training set (u pper half) and ROC plots on test set (lower half) of Algorithm 1 (solid curves) versus AUCMax (dashed curves). For convergence plots, the x-axis indicates the number of epochs, and the y-axis indicates the AUC score. For ROC plots, the x-axis indicates the FPR and the y-axis indicates the TPR.

Section 5: Experimental Results

Algorithm \ Metric	Adult		Bank		Compas		Default	
	Overall	Min/Max	Overall	Min/Max	Overall	Min/Max	Overall	Min/Max
AUCMax	**.902 ± .002**	.823 ± .005	.910 ± .002	.780 ± .018	.732 ± .004	.779 ± .041	.763 ± .005	.871 ± .017
MinimaxFair	.894 ± .007	.905 ± .010	.885 ± .003	.827 ± .004	.730 ± .001	.913 ± .029	.753 ± .002	.909 ± .021
InterFairAUC	.894 ± .004	.950 ± .003	**.912 ± .001**	.836 ± .018	.738 ± .003	.939 ± .014	.763 ± .003	.952 ± .024
EqualAUC	.886 ± .003	.953 ± .004	.866 ± .005	**.891 ± .025**	.731 ± .003	.956 ± .012	.761 ± .002	**.972 ± .020**
Algorithm 1	.901 ± .004	**.953 ± .002**	.907 ± .004	.858 ± .014	**.741 ± .004**	**.961 ± .012**	**.767 ± .002**	.968 ± .013

Figure: Comparison of Algorithm 1 versus baselines. 'Overall' is the AUC score on the full dataset, measuring the utility. 'Min/Max' is the minimum group-level AUC score over the maximum one, measuring the fairness. The numbers are reported as 'Mean ± Standard Deviation'. Best results at each column are highlighted in bold. Second best are highlighted in underline