Fair Adversarial Gradient Tree Boosting

読み会@2020/11/10

楊明哲

論文情報

- 著者
  - Vincent Grari (Sorbonne University)
  - Boris Ruf (AXA)
  - Sylvain Lamprier (Sorbonne University)
  - Marcin Detyniecki (AXA)
- 出典: ICDM 2019

概要

どんな論文？

- 決定木のモデルで公平なモデルを獲得できるようにした.
- SOTAと同じくらいの性能を達成できた.
- 選んだ理由:
  - 公平なモデル獲得を目的として，決定木ベースの手法を提案していたため.
  - 勾配ブースティング決定木がテーブルデータに対して強力であるから.

概要

研究背景

- バイアスの軽減に対するタスクではよくニューラルネットワークが用いられているし，性能としてもいい.
- テーブルデータに対して，XGBoostやLightGBMなど勾配ブースティング木が非常に強力でよく用いられている.
- そこで，勾配ブースティング木でバイアスを軽減する(公平性を達成する)ような手法を提案する.

概要

既存研究について

- 既存研究ではほとんどがニューラルネットワークを用いて，バイアスを軽減してる.
- ニューラルネットワークでは性能は出せているが，なんでこうなったかが分かりにくい.
- テーブルデータではニューラルネットワークよりも勾配ブースティング木が好まれて使われている.

概要

貢献

- 一般的な分類モデルに対して，決定木を含む敵対的学習で公平を獲得する方法を提案する.
- 異なる公平性の規準でSOTAなモデルと同等の性能を達成した.

公平性配慮型機械学習

表記の定義

- $\mathbf{x} \in \mathbb{R}^d$： d次元の特徴量（非センシティブな特徴）
- s：センシティブな属性．今回は二値で考えている．
- y: 目的変数．今回は分類問題を扱うので二値で考える．
  - 特に観測値はy, 予測値は$\hat{y}$とする
- nこのサンプルのデータセット $(\mathbf{x_i}, s_i, y_i)_{i=1}^n$を用いて一般的な教師あり学習を行う．

公平性配慮型機械学習

公平性の定義

- Demographic Parity: センシティブ属性にかかわらず予測値の比率は一致する
- DP: $P(\hat{Y}=1\,|\,S=0)=P(\hat{Y}=1\,|\,S=1)$
- DPを使った評価P-ruleを次のように定義する

$$\text{P-rule}=min\left(\frac{P(\hat{Y}=1\,|\,S=0)}{P(\hat{Y}=1\,|\,S=1)},\frac{P(\hat{Y}=1\,|\,S=1)}{P(\hat{Y}=1\,|\,S=0)}\right)$$

公平性配慮型機械学習

公平性の定義

- もう一つの基準としてdispare impact（DI）を定義する.

$$\mathrm{DI} : |P(\hat{Y} = 1 \,|\, S = 1) - P(\hat{Y} = 1 \,|\, S = 0)|$$

- DIは0になるほど公平である.

公平性配慮型機械学習

公平性の規準 その2

- 観測値$\mathbf{Y}$も条件に入れる規準 Equalized Oddsがある.
- EO : $P(\hat{Y}=1 \mid S=0, Y=y) = P(\hat{Y}=1 \mid S=1, Y=y), \forall y \in 0,1$
- このEOによって求められるFPRとFNRの値が0に近いほど公平性が獲得できる.

提案手法

- 本研究では，一般的な勾配ブースティング木の枠組みに，敵対的学習を用いて，公平性を獲得していく．

提案手法

勾配ブースティング木(GTB)について

- 弱学習木をブースティングして予測性能の向上を目指している.
- 学習時

1. 観測値yと予測値$\hat{y}$　との誤差を減らすように決定木で学習を行う.

2. 1. を繰り返して，誤差を減らす.

- 予測時
  - 学習によって求めた残渣と使って予測を行う.

提案手法

Fair Adversarial Gradient Tree Boosting(FAGTB)

- FAGTBで公平性を獲得する方針として，目的変数とセンシティブ属性の予測精度を敵対的に学習を行い，目指していく．

- 目的変数の予測器と敵対的予測器(センシティブ属性の予測)をミニマックス法として最適化していく．

$$\arg\min_{F}\max_{\theta_A}\sum_{i=1}^{n}\mathbf{L}_{F_i}\left(F\left(x_i\right)\right)-\lambda\sum_{i=1}^{n}\mathbf{L}_{A_i}\left(F\left(x_i\right);\theta_A\right)$$

- λが敵対学習の割合を調整するハイパーパラメータ

- Lはそれぞれ損失関数である．今回は分類問題だから，負の対数尤度を採用．

提案手法

FAGTBのアルゴリズム

Algorithm 2 Fair Adversarial Gradient Tree Boosting

Input: training set $(x_i, s_i, y_i)_{i=1}^n$, a number of iterations M, an adversarial learning rate α, a differentiable loss function $\mathcal{L}_F$ for the output classifier and $\mathcal{L}_A$ for the adversarial classifier.

Initialize: Calculate the constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}_{F_i}(\gamma)$$

Initialize parameters θ_A of the neural network $A(x)$

for $m = 1$ **to** $M - 1$ **do**

(a) Calculate the pseudo residuals:

$$r_{im} = -\left[\frac{\partial \mathcal{L}_{F_i}(F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

(b) Calculate the pseudo residuals of the adversarial from the input $F_{m-1}(x_i)$:

$$t_{im} = -\left[\frac{\partial \mathcal{L}_{A_i}(F(x_i; \theta_A))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

(c) Calculate the training loss derivative:

$$u_{im} = r_{im} - \lambda * t_{im}$$

(d) Fit a classifier $h_m(x)$ to pseudo residuals using the training set $\{(x_i, u_{im})\}_{i=1}^n$

(e) Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}_{F_i}\left(F_{m-1}(x_i) + \gamma * h_m(x_i)\right) - \lambda * \mathcal{L}_{A_i}(F_{m-1}(x_i) + \gamma * h_m(x_i); \theta_A).$$

(f) Update the learning model:

$$F_m(x_i) = F_{m-1}(x_i) + \gamma_m * h_m(x_i)$$

(g) Fit the adversarial A to the using the new outputs (i.e., using the training set $\{(F_m(x_i), s_i)\}_{i=1}^n$)

$$\theta_A := \theta_A - \alpha * \frac{\partial \mathcal{L}_{A_i}(F_m(x_i); \theta_A)}{\partial \theta_A}$$

end do

提案手法

FAGTBの全体図

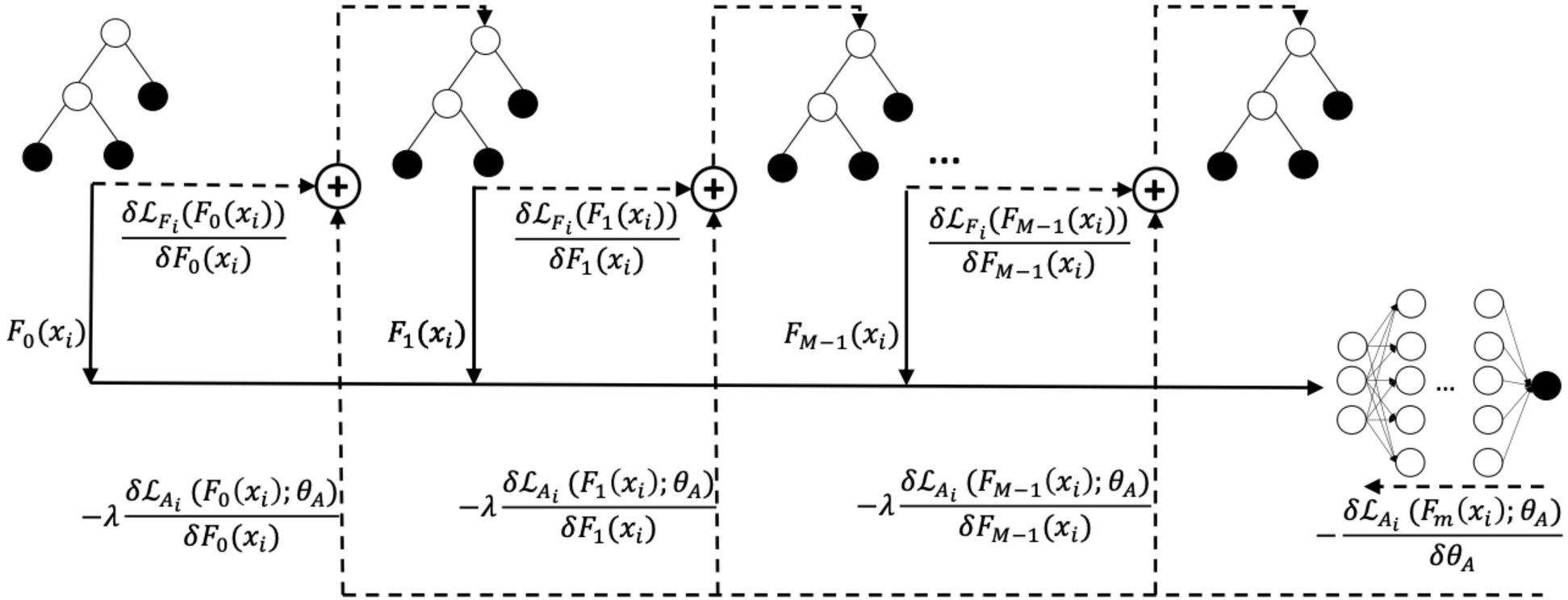


Fig. 1: The architecture of the Fair Adversarial Gradient Tree Boosting (FAGTB). 4 steps are depicted, each one corresponding to a tree h that is added to the global classifier F. The neural network on the right is the adversary that tries to predict the sensitive attributes from the outputs of the classifier. Solid lines represents forward operations, while dashed ones represent gradient propagation. At each step m, gradients from the prediction loss and the adversary loss are summed to form the target for the next decision tree h_{m+1}.

実験

実験概要

- 実験では，人工データと実データの両方を使って検証する.
- またSOTAな手法と比べて，有用性を示していく.
- 敵対的学習のハイパーパラメータλがfairnessとaccuracyでどういう影響を与えるか調べる

実験

人工データを用いる

- 状況設定として，自動車保険でクレームを受けるかどうかを考える.
- 特徴量として，車の色，年齢，性別，凶暴性，不注意さを持つ.
- 目的変数として，クレームをしたかどうかを二値で表す.
- センシティブ属性として，性別を用いる.
- それぞれの生成方法は右の通り

$$\begin{pmatrix} I_i \\ a_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 40 \end{pmatrix}, \quad \begin{pmatrix} 1 & 4 \\ 4 & 20 \end{pmatrix}\right]$$

$$A_i \sim \mathcal{N}(0,1)$$
$$c_i = (1.5 * s_i + A_i) > 1$$
$$y_i = \sigma(A_i + I_i + \epsilon_i) > 0.5$$
$$\epsilon_i \sim \mathcal{N}(0, 0.1)$$

The features are: age (a), agressivity (A), color (c), gender (s), inattention (I)

実験結果

人工データ

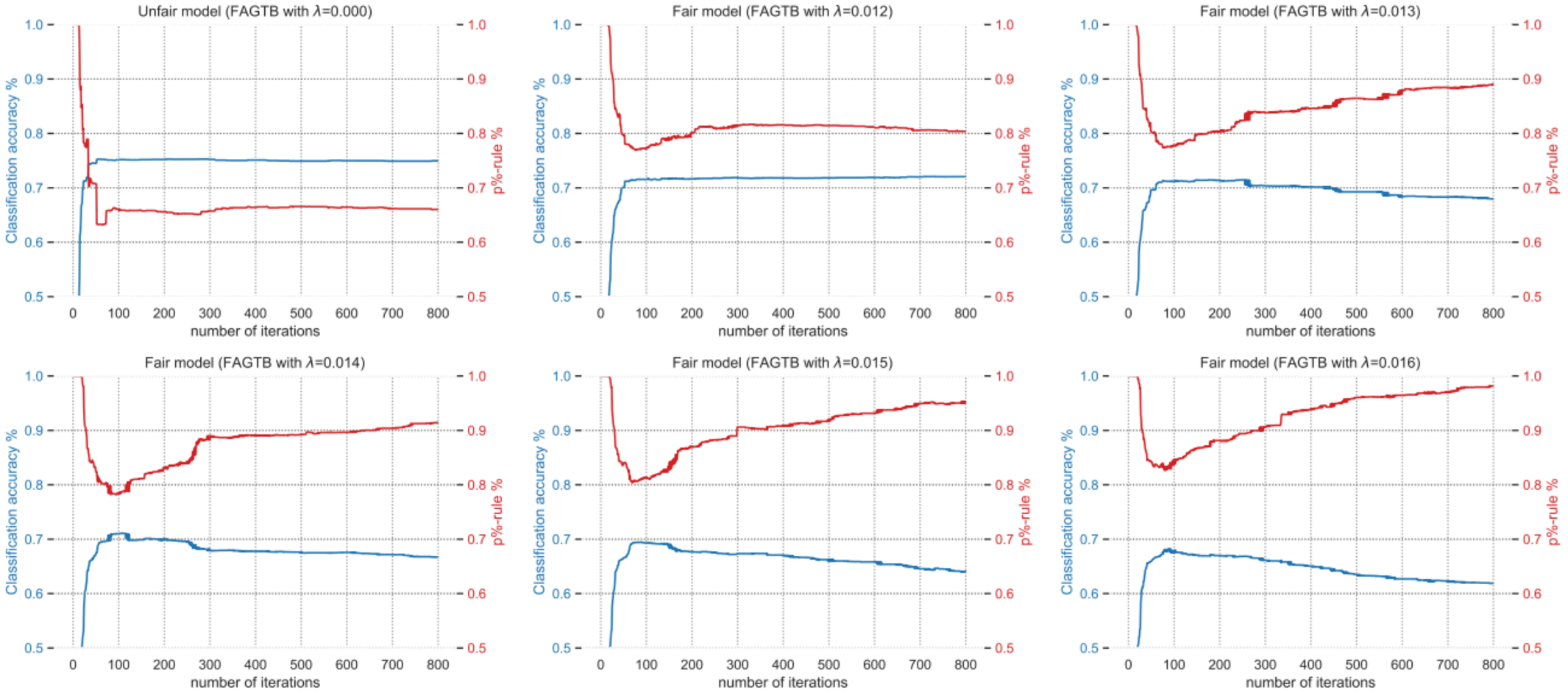


Fig. 2: Synthetic scenario: Accuracy and p-rule metric for a biased model ($\lambda = 0$) and for several fair models with varying values of λ optimized for demographic parity.

実験

実データ

- 既存のSOTAな手法と比べるため，実データ4つを使い比較を行う．
- 使用するデータセットはAdult, COMPAS, Default, Bankの４種類

TABLE II: Data sets used for the experiments

Data set	# Observations	# Features	Target	%Target	Sensitive	%Sensitive
Adult UCI	45,000	14	Income >= $50k	30.0%	Gender	58.0%
COMPAS	6,967	13	2-year recidivism	45.5%	Race	34.0%
Default	30,000	23	Defaulting on payments	22.1%	Gender	60.4%
Bank	45,211	16	Subscription to a term deposit	11.7%	Age	32.9%

Description of the data sets: number of observations, number of features, target, total share of the target, sensitive attribute, and total share of the sensitive attribute.

実験

問題設定

- DPを考えているときは，P-rule $= 90\,\%$ を満たしている状況のもとで，accuracyを用いて比較を行う.
- またEOを考えているときは，{FPR,FNR}が0.03以下を満たしているもとでのaccuarcyを比較する.

実験結果

Demographic Parity

TABLE III: Results for Demographic Parity

	Adult		COMPAS		Default		Bank	
	Accuracy	P-rule	Accuracy	P-rule	Accuracy	P-rule	Accuracy	P-rule
Standard GTB	86.8%	32.6%	69.1%	61.2%	82.9%	77.2%	90.8%	48.1%
Standard NN	85.3%	31.4%	67.5%	71.1%	82.1%	63.3%	90.3%	58.6%
FAGTB-1-Unit	84.4%	90.4%	61.8%	90.1%	81.5%	90.1%	90.1%	90.0%
FAGTB-NN	**84.9%**	90.3%	**64.5%**	90.0%	**82.2%**	90.2%	**90.2%**	90.0%
Wadsworth2018 [18]	83.1%	89.7%	63.9%	90.1%	81.8%	90.0%	**90.2%**	90.1%
Zhang2018 [17]	83.3%	90.0%	64.1%	89.8%	81.4%	90.0%	90.0%	90.0%
Zafar-DI [14]	82.2%	89.8%	63.9%	89.7%	80.7%	89.8%	89.2%	90.1%
Kamishima [28]	82.3%	89.9%	63.8%	90.0%	81.1%	90.0%	89.6%	89.9%
Feldman [8]	-	-	61.4%	90.1%	72.2%	90.2%	-	-

Comparing our approach with different common fair algorithms by accuracy and fairness (p-rule metric) for the Adult UCI, the COMPAS, the Default and the Bank data set.

実験結果

Equalized Odds

TABLE IV: Results for Equalized Odds

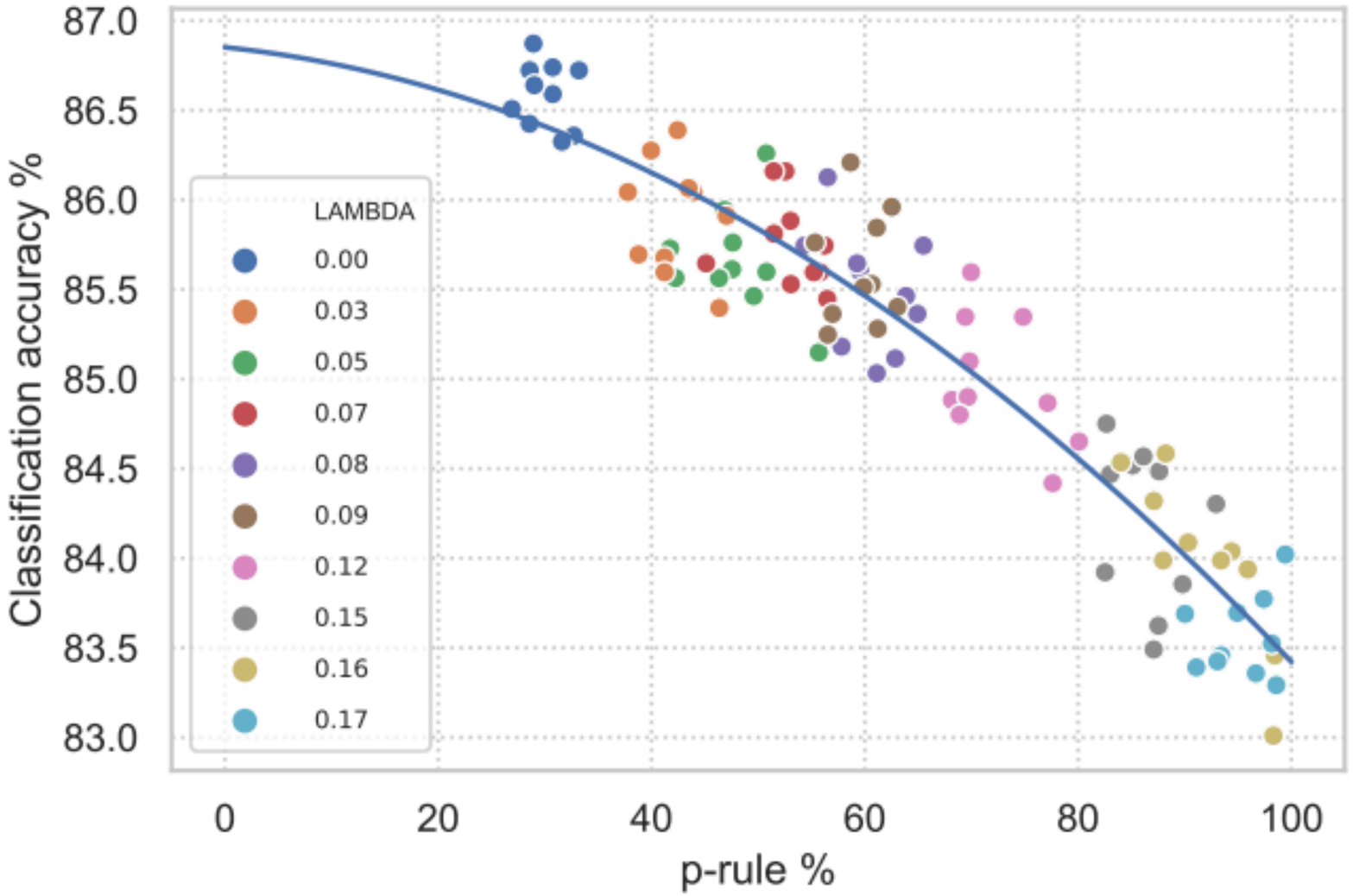
	Adult			COMPAS		
	Accuracy	DispFPR	DispFNR	Accuracy	DispFPR	DispFNR
Standard GTB	86.8%	0.06	0.07	69.1%	0.12	0.20
Standard NN	85.3%	0.07	0.10	67.5%	0.09	0.15
FAGTB-1-Unit	86.3%	0.02	0.02	65.1%	0.03	0.12
FAGTB-NN	**86.4%**	0.02	0.02	**66.2%**	0.01	0.02
Wadsworth2018 [18]	84.9%	0.02	0.03	65.4%	0.02	0.03
Zhang2018 [17]	84.8%	0.03	0.03	64.9%	0.03	0.02
Zafar-DM [10]	83.9%	0.03	0.09	64.3%	0.09	0.17
Kamishima [28]	82.6%	0.06	0.24	63.6%	0.08	0.11
Feldman [8]	80.6%	0.07	0.05	61.1%	0.03	0.03

	Default			Bank		
	Accuracy	DispFPR	DispFNR	Accuracy	DispFPR	DispFNR
Standard GTB	82.9%	0.02	0.04	90.8%	0.04	0.06
Standard NN	82.1%	0.02	0.05	90.3%	0.05	0.08
FAGTB-1-Unit	82.1%	0.00	0.01	89.7%	0.02	0.07
FAGTB-NN	**82.5%**	0.00	0.01	**90.3%**	0.01	0.07
Wadsworth2018 [18]	81.2%	0.01	0.02	89.4%	0.01	0.07
Zhang2018 [17]	81.9%	0.00	0.01	89.8%	0.00	0.07
Zafar-DM [10]	81.0%	0.00	0.03	89.5%	0.01	0.08
Kamishima [28]	80.5%	0.00	0.04	89.3%	0.00	0.08
Feldman [8]	71.8%	0.02	0.02	87.1%	0.05	0.06

omparing our approach with different common fair algorithms by accuracy and fairness (DispFPR, DispFNR) for the Adult UCI, the COMPAS, the Default and the Bank data set.

実験結果

ハイパーパラメータλ による影響



まとめ

- 決定木を用いて，公平性配慮型機械学習を行う手法を提案した.
- 本手法では既存のニューラルネットワークのモデルと同等の性能があることを示した.
- 決定木を用いることでNNモデルよりも説明性が増している.

感想

- 勾配ブースティング木を勉強できてよかった.
- ブースティングだと説明性があるのか疑問に思った.
- 人工データの作り方はいいなって思った.（AXAらしい)
- 深層ニューラルネット決定木 + 敵対学習 で公平性を獲得できたら自分のやりたいこと(公平性と説明性)を満たせる?
  - Deep Neural Decision Trees (https://arxiv.org/abs/1806.06988)

おまけ

勉強になった動画

- 勾配決定ブースティング木でとても分かりやすかったYoutuber(?)
  - StatQuest with Josh Starmer さん
  - https://www.youtube.com/watch?v=3CC4N4z3GJc