FAIR FORESTS:
REGULARIZED TREE INDUCTION TO MINIMIZE MODEL BIAS

Edward Raff, Jared Sylvester & Steven Mills

Booz Allen Hamilton, Strategic Innovation Group

ACM/AAAI Conference on AI, Ethics & Society, 2 February 2018, New Orleans

arxiv.org/abs/1712.08197

Abstract

The potential lack of fairness in the outputs of machine learning algorithms has recently gained attention both within the research community as well as in society more broadly. Surprisingly, there is no prior work developing tree-induction algorithms for building fair decision trees or fair random forests. These methods have widespread popularity as they are one of the few to be simultaneously interpretable, non-linear, and easy-to-use. In this paper we develop, to our knowledge, the first technique for the induction of fair decision trees. We show that our "Fair Forest" retains the benefits of the tree-based approach, while providing both greater accuracy and fairness than other alternatives, for both "group fairness" and "individual fairness." We also introduce new measures for fairness which are able to handle multinomial and continues attributes as well as regression problems, as opposed to binary attributes and labels only. Finally, we demonstrate a new, more robust evaluation procedure for algorithms that considers the dataset in its entirety rather than only a specific protected attribute.

Contributions

- Introduce a new learning algorithm to produce fair Decision Trees and Random Forests
- Define new measures of discrimination for:
  - Multinomial features
  - Continuous features
  - Regression problems
- Present a new evaluation procedure to assess fairness with respect to all features in a dataset

The fair Decision Trees (DT^F) & fair Random Forests (RF^F) produced with our technique are:

- Non-linear
- Interpretable
- Easy to to use (no parameter tuning required)
- Applicable to numeric features & classes
- Achieve high accuracy
- Achieve high group fairness
- Achieve high individual fairness

Tree Induction Algorithm

- *Goal:* make decisions that are accurate but not based on protected attributes such as race, gender or age
- "Fairness through unawareness" is insufficient
  - Even if the protected attribute is completely removed from the dataset, other features may be highly correlated with it and function as proxies
- Standard decision trees pick features based on *Information Gain* (IG), i.e. "how easy does this feature make it to predict the target?"
- We introduce a new criteria to choose tree nodes:

$$IG_{\text{fair}}(T,b) = IG^{y}(T,b) - IG^{a}(T,b)$$

IG with respect to the target
minus
IG with respect to the protected attribute

- This chooses features which make it *easy to predict the target, but hard to predict the protected attribute*

$$I^{a}_{\text{Gini}}(T) = \frac{1 - \sum_{\forall T_i \in \text{splits}(a)} \left(\frac{|T_i|}{|T|}\right)^2}{1 - |\text{splits}(a)|^{-1}}$$

IG is defined via Gini impurity, which can be calculated based on splitting the protecting attribute a, not the target value y, as is standard.

- We can use the same IG$_{\text{fair}}$ splitting criteria when constructing Random Forests
- They maintain the benefits of standard DTs/RFs:
  - Easy to use: no hyperparamaters to tune
  - Not a black box: easy to interpret & explain
  - Powerful: can learn non-linear decision boundaries
- Also define IG with respect to *protected numeric features*
  - No need to select threshold to discretize features
  - Can provide fairness w.r.t. any potential threshold

$$G_r(T,b) = 1 - \frac{1}{3} \sum_{T_i \in \text{split}} \frac{|T_i|}{|T|} \min\left(\frac{|\mu_{b,T} - \mu_{b,T_i}|}{\sigma_{b,T}}, 3\right)$$

Metrics

- *Discrimination* (a.k.a. "group fairness")
  - Difference between average predicted scores for each protected attribute value
  - Previous definitions are limited to binary features
  - We extend to attributes with >2 values
  - Also extend to continuous target variables for regression problems ("MaxD")
    - No need to discretize outputs
  - These two definitions can be used to evaluate any fair learning algorithms in the future

$$\text{Discrimination} = \frac{2}{k} \sum_{i=1}^{k} \left| \frac{\sum_{x_j \in T} \hat{y}_j}{|T|} - \frac{\sum_{x_j \in T_i} \hat{y}_j}{|T_i|} \right|$$

$$\text{MaxD} = \arg\max_{t} \left| \frac{\sum_{x_i(a_p) < t} \hat{y}_i}{|x_i(a_p) < t|} - \frac{\sum_{x_i(a_p) \geq t} \hat{y}_i}{|x_i(a_p) \geq t|} \right|$$

- *Inconsistency* (a.k.a. "individual fairness")
  - Similar samples should receive similar outputs
- *Accuracy*
- *Delta* = Accuracy – Discrimination

Results

- Using these canonical datasets (Zemel, et al. 2013):
  - German Credit (protect age ≥ 25)
  - Adult Income (protect gender)
  - Heritage Health (protect age ≥ 65)
- Comparisons with:
  - Baselines: Decision Trees (DT), Random Forests (RF), Logistic Regression (LR)
  - Learned Fair Representations (LFR; Zemel et al., 2013)
  - Fair Logistic Regression (LR^F; Kamishima et al., 2011)
  - Fair Naïve Bayes (NB^F; Kamiran & Calders, 2009).

(Our techniques)

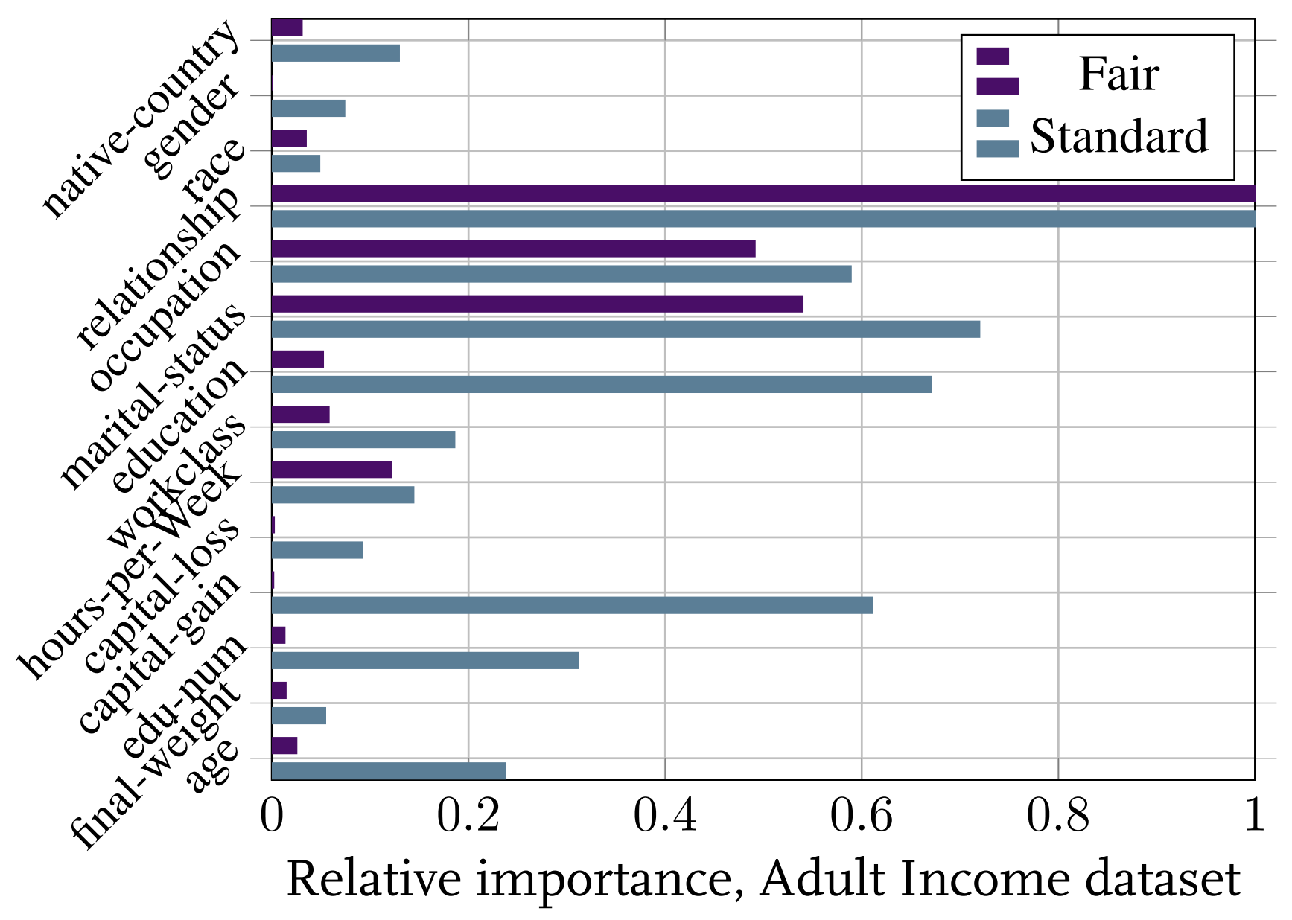
	German Credit				Adult Income				Heritage Health			
	Acc	Delta	Discrim	Incon	Acc	Delta	Discrim	Incon	Acc	Delta	Discrim	Incon
DT	0.6890	0.6509	0.0381	0.2140	0.8364	0.4801	0.3563	0.4417	0.8404	0.8196	0.0207	0.2062
DT^F	*0.6990*	0.6908	0.0082	*0.0070*	0.7511	0.7444	0.0067	*0.0033*	**0.8474**	*0.8473*	*0.0001*	*0.0001*
RF	0.6970	*0.6911*	0.0059	0.0020	**0.8501**	0.5463	0.3038	0.3944	0.8472	0.8464	0.0007	0.0005
RF^F	**0.7000**	**0.7000**	**0.0**	**0.0**	0.7530	*0.7530*	**0.0**	**0.0**	**0.8474**	**0.8474**	**0.0**	**0.0**
NB^F	0.6888	0.6314	0.0574	0.3132	*0.7847*	**0.7711**	0.0136	0.4366	0.6878	0.5678	0.1200	0.4107
LR	0.6790	0.5517	0.1273	0.3050	0.6787	0.4895	0.1892	0.2703	0.7547	0.6482	0.1064	0.2767
LR^F	0.5953	0.5842	0.0111	0.1284	0.6758	0.6494	0.0264	0.2234	0.7212	0.7038	0.0174	0.3777
LFR	0.5909	0.5867	0.0042	0.0592	0.7023	0.7018	*0.0006*	0.1892	0.7365	0.7365	**0.0000**	**0.0000**

Binary protected features & binary target

(See table above)

- *Fair Forests (RF^F) reduce discrimination & inconsistency to zero*
  - Fair Forests achieve the lowest discrimination & inconsistency on all three datasets
- *Accuracy improves* compared to standard RF on two datasets
  - For the German Credit & Heritage Health datasets, there is no cost-of-fairness when using Fair Forests
  - This is not the case when using LR & LR^F
  - Caution: we do not expect there to be no cost-of-fairness in general, for this or any other technique
- Switching from standard DT to standard RF reduces discrimination and inconsistency by itself

Interpretability & Feature Importance



- Examine which features were selected by RF vs. RF^F
- Gender importance goes to zero *(yay!)*
- Capital-gain and capital-loss importance also go to zero
  - Too highly correlated w/ gender? Problem w/ data collection? Deeper societal issue?
  - RF^F can tells us if something interesting is going on even if we don't want to use it to make final decisions

Continuous protected feature & binary target

(Results in paper)

- Fair Forests can protect a numeric feature directly without discretization
  - Protected attribute is binary in all prior work we found
- Test on German & Health datasets, where protected attribute (age) was originally numeric
- Continuous version of Fair Forests was also able to reduce discrimination & inconsistency to zero
- No loss of accuracy compared to standard RF

Continuous Target

(Results in paper)

- Can also use Fair Forests for regression problems
- DT^F & RF^F perform just as well with continuous targets

Full Evaluation Process

- Evaluate Fair Forests' ability to *protect any attribute*, not just the ones pre-identified as sensitive
  - First time this has been done
  - Strong demonstration of robustness
  - Future-proofing: if we decide to eschew discrimination w.r.t. new & different features, we'll be ready
  - This evaluation technique is applicable to other fairness algorithms, not just ours
- *We can reduce discrimination to zero for any attribute* in the German Credit & Heritage Health datasets
  - … and the accuracy always comes out the same.
  - For the Adult Income dataset, zero discrimination is only possible for gender, but discrimination is halved on average when considering all features.

Booz | Allen | Hamilton