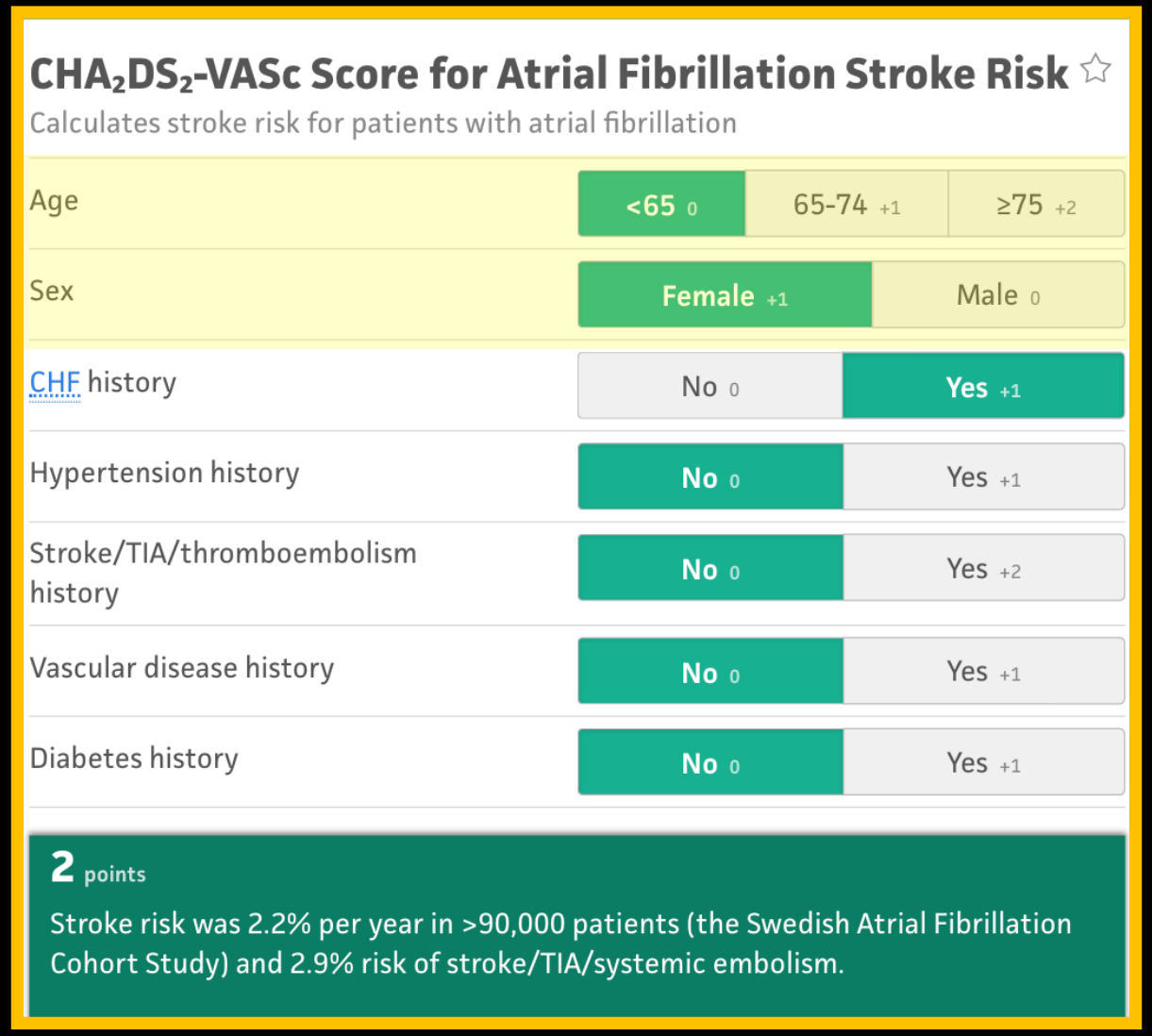
Fairness without Harm

Decoupled Classifiers with Preference Guarantees

Berk Ustun
Harvard University

Joint work with Yang Liu and David Parkes

Medical Diagnostics



CHA_2DS_2-VASc Score for Atrial Fibrillation Stroke Risk ☆

Calculates stroke risk for patients with atrial fibrillation

Age	**<65** 0	65-74 +1	≥75 +2
Sex	**Female** +1	Male 0	
CHF history	No 0	**Yes** +1	
Hypertension history	**No** 0	Yes +1	
Stroke/TIA/thromboembolism history	**No** 0	Yes +2	
Vascular disease history	**No** 0	Yes +1	
Diabetes history	**No** 0	Yes +1	

2 points

Stroke risk was 2.2% per year in >90,000 patients (the Swedish Atrial Fibrillation Cohort Study) and 2.9% risk of stroke/TIA/systemic embolism.

- Data includes group attributes like `age` & `gender`
- Models make use of groups attributes for prediction
- Model performance can vary between groups

http://www.mdcalc.com/cha2ds2-vasc-score-atrial-fibrillation-stroke-risk

Relevant Ethical Principles

Beneficence
do the best in one's ability

Non-Maleficence
do no harm

Goals for Fair ML

train most accurate model for each group without harming any group

Hard to Capture Group Heterogeneity

Group A			
x	n^+	n^-	h^*_A
0	50	0	+
1	0	50	−

Group B			
x	n^+	n^-	h^*_B
0	0	50	−
1	50	0	+

Pooled with z			
(x, z)	n^+	n^-	h^*_0
(0,0)	0	50	+
(1,0)	50	0	−
(0,1)	50	0	−
(1,1)	0	50	+

best models for each group makes 0 mistakes

no linear classifier can predict XOR
☞ any linear classifier makes 50 mistakes

Standard Techniques Can Harm Groups

	Training Error for everyone
LR with No Attributes	27.9%
LR with 1-Hot Encoding	27.0%
Change in Error	-0.9%

`adult` dataset with 12 groups based on
`gender` × `marital_status` × `immigration_status`

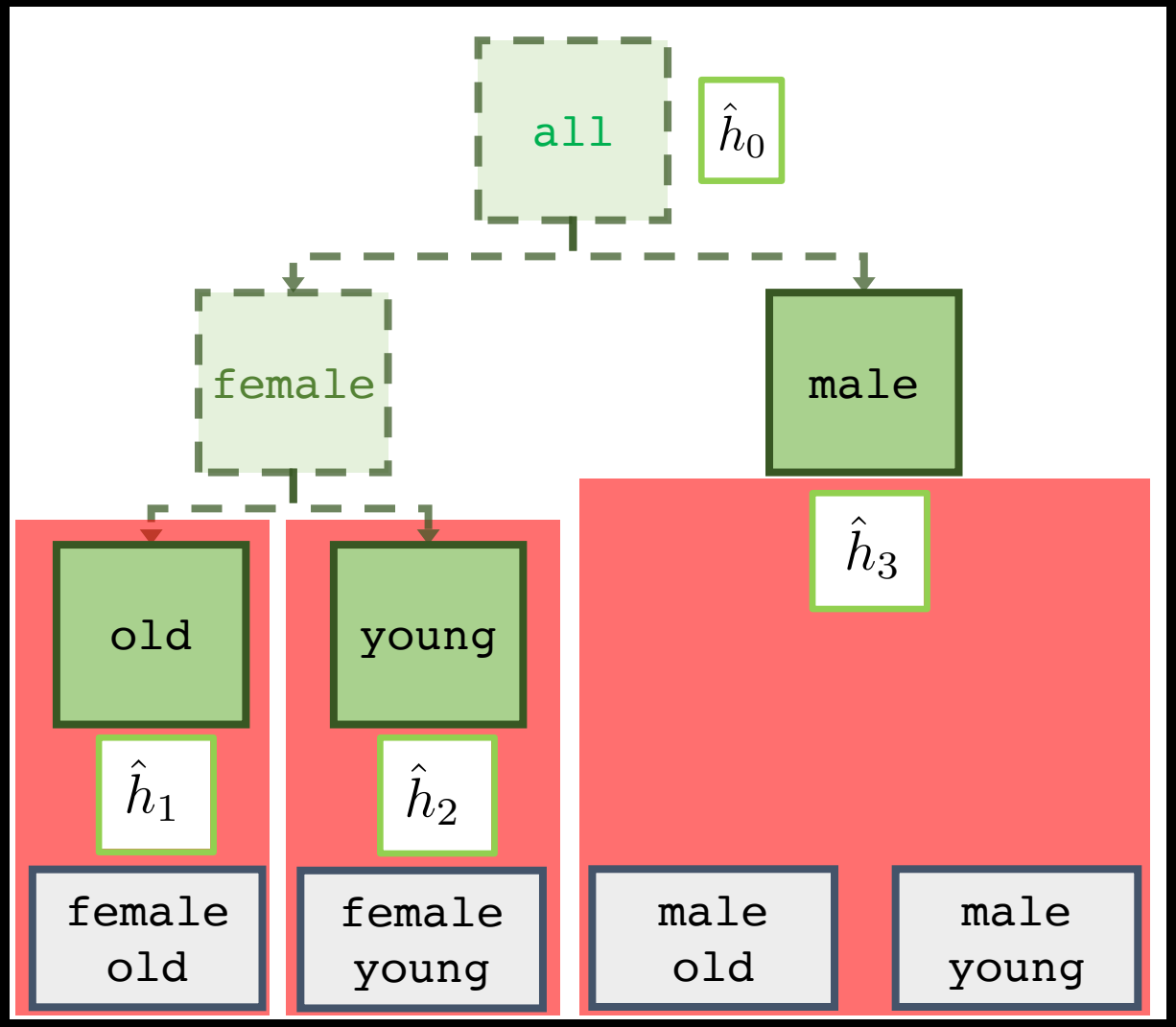
Standard Techniques Can Harm Groups

	Training Error for everyone	Training Error (female, married, resident)
LR with No Attributes	27.9%	33.5%
LR with 1-Hot Encoding	27.0%	35.3%
Change in Error	-0.9%	+1.8%

groups should **not** be worse off when we use their sensitive attributes

adult dataset with 12 groups based on
gender × marital_status × immigration_status

Decoupled Classifiers with Preference Guarantees



Beneficence

train classifiers for intersectional subgroups by "recursively decoupling"

Non-Maleficence

check classifiers satisfy preferences guarantees to ensure the fair use of group attributes

Preference Guarantees

Rationality

each group has better test accuracy with own model vs. blind model

Rationality Violation

majority of members would rather ***not report*** *group membership*

Envy-Freeness

each group has better test accuracy with own model vs. model of another group

Envy-freeness Violation

majority of members would rather ***misreport*** *group membership*

Fairness without Harm: Decoupled Classifiers with Preference Guarantees

Berk Ustun [1] **Yang Liu** [2] **David C. Parkes** [1]

Abstract

In domains such as medicine, it can be acceptable for machine learning models to include *sensitive attributes* such as gender and ethnicity. In this work, we argue that when there is this kind of *treatment disparity* then it should be in the best interest of each group. Drawing on ethical principles such as beneficence ("do the best") and non-maleficence ("do no harm"), we show how to use sensitive attributes to train decoupled classifiers that satisfy *preference guarantees*. These guarantees ensure the majority of individuals in each group prefer their assigned classifier to (i) a pooled model that ignores group membership (*rationality*), and (ii) the model assigned to any other group (*envy-freeness*). We introduce a recursive procedure that adaptively selects group attributes for decoupling, and present formal conditions to ensure preference guarantees in terms of generalization error. We validate the effectiveness of the procedure on real-world datasets, showing that it improves accuracy without violating preference guarantees on test data.

1. Introduction

When machine learning systems are deployed in human-facing applications (e.g., lending, hiring, medical decision support), their performance may vary over groups defined by *sensitive attributes* such as gender and ethnicity. Such performance disparities are now regularly reported (Angwin et al., 2016; Dastin, 2018), eliciting calls for fairness in machine learning (Crawford, 2013), and prompting the development of technical solutions (Zliobaite, 2015; Barocas et al., 2018; Corbett-Davies & Goel, 2018).

Many of the proposed methods for fair machine learning have aimed to build models that predict or perform in the same way across groups (e.g., Hardt et al., 2016; Zafar et al., 2017a; Feldman et al., 2015; Zafar et al., 2017c; Agarwal et al., 2018; Narasimhan, 2018). Such methods can be broadly viewed as methods to achieve fairness by *parity* (see Zafar et al., 2017b, for a discussion). Parity is an appropriate notion of fairness for applications such as hiring or sentencing, where a model that exhibits disparate treatment or disparate impact may be viewed as a system to perpetrate wrongful discrimination (see Arneson, 2006; Hellman, 2008; Barocas & Selbst, 2016).

In comparison, less work has sought to articulate suitable notions of fairness for domains with different ethical principles (with some exceptions, see e.g., Chen et al., 2018). In medical applications, for example, the relevant ethical principles are *beneficence* (do the best in one's ability) and *non-maleficence* (do no harm) (see e.g., Beauchamp et al., 2001). Accordingly, methods for fair machine learning should be designed to produce the most accurate model for each group (beneficence) without harming any group (non-maleficence).

These goals represent new challenges for the fair use of sensitive attributes in machine learning. Consider, for example, training a medical diagnostic using a dataset with sensitive attributes such as age, gender and ethnicity. In this case, a model that ignores group membership may not be beneficent as it may impose inevitable performance trade-offs between heterogeneous groups (see Figure 1). In practice, heterogeneity may arise due to intrinsic differences between groups, or discrepancies in the quality or amount of data.

While these issues motivate the need to build models that explicitly consider group membership (see Corbett-Davies et al., 2017; Lipton et al., 2018), it is not clear *how to do this in a way that is fair to each group*. As shown in Figure 2, simple approaches such as a "one-hot encoding" may not recover the most accurate model for each group. Conversely, one could harm groups by fitting a model from a hypothesis class that is overly complex (e.g., by overfitting), or that one that cannot adequately capture the heterogeneity (e.g., by "gerrymandering" along intersectional subgroups as discussed in Kearns et al., 2018; Hébert-Johnson et al., 2018).

[1]Harvard University, Cambridge, MA, USA [2]UC Santa Cruz, Santa Cruz, CA, USA. Correspondence to: Berk Ustun <berk@seas.harvard.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Poster

#133

Code

https://github.com/ustunb/dcptree

Contact

www.berkustun.com

@berkustun