50 Years of Test (Un)fairness: Lessons for Machine Learning

Ben Hutchinson and Margaret Mitchell

Presented by:
Aditya Jain and Manish Ravula

History doesn’t repeat itself, but it often rhymes

United States Civil Rights Act of 1964: Title VI and Title VII prevented discrimination.

Education and Employment Testing Research	**ML Fairness Research**
1966-1976	2011 -
Questions	Features
Responses	Feature Values
Linear Model of scoring	All models *

The spurt of research on fairness issues that began in the late 1960s had results that were ultimately disappointing.No generally agreed upon method to determine whether or not a test is fair was developed. No statistic that could unambiguously indicate whether or not an item is fair was identified. There were no broad technical solutions to the issues involved in fairness

-Nancy Cole ‘73

Notations:

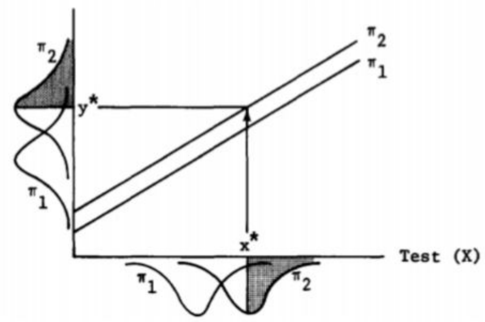
- **A** - Protected Variable (Eg: Race, Gender, etc.,)
- **R** - Score in a test that’s used as a proxy for true ability. (Eg: Credit Score, SAT Score, Interview Score, etc.,)
- **Y** - Number representing true ability. (Eg: Credit worthiness, Educational Merit etc.,)

Cleary’s view

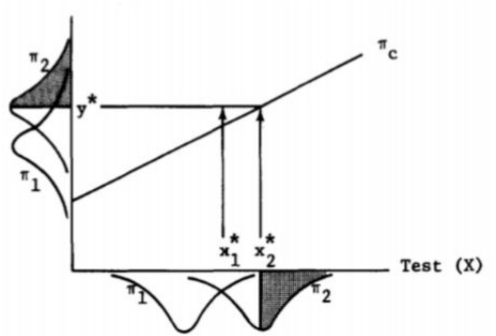
> ***“The test is biased if the criterion score predicted from the common regression line is consistently too high or tool ow for members of the subgroup.”***
>
> ***-Cleary[1960]***

- A definition of ‘unfairness’
- Very rudimentary.

Cleary’s view cont.



(a) Labels on regression lines indicate which subgroup they fit.



(b) The regression line labeled π_c fits both subgroups separately (and hence also their union).

Examples of regression lines for an unfair test

Examples of regression line for a fair test

Paradigm Shifts

- Fairness in Model → Fairness in Model usage
  - Cleary failed to take into account the differing false positive and false negative rates that occur when subgroups have different base rates.

- Focusing on Unfairness → Focusing on Fairness
  - Attributed to Darlington’s formulation.

Thorndike’s view

> *“A judgment on test-fairness must rest on the inferences that are made from the test rather than on a comparison of mean scores in the two populations. One must then focus attention on fair use of the test scores, rather thanon the scores themselves”* **-** *Thorndike [1960]*

- Sharing a common regression line is not important, as one can achieve fair selection goals by using different regression lines and different selection thresholds for the two groups.

- As long as the ratio of Predicted Positives to Ground Truth Positives remains same for each sub-group, the test can be considered fair. - The following quantity should remain same for each sub-group.

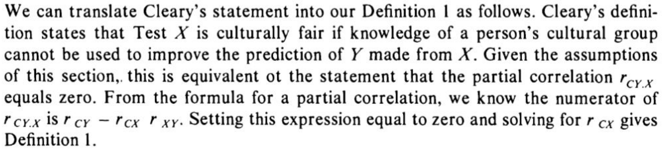
$$(TP + FP)/(TN + FN)$$

Darlington's attempt at unification

Unification in terms of correlations and partial correlations

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}.$$

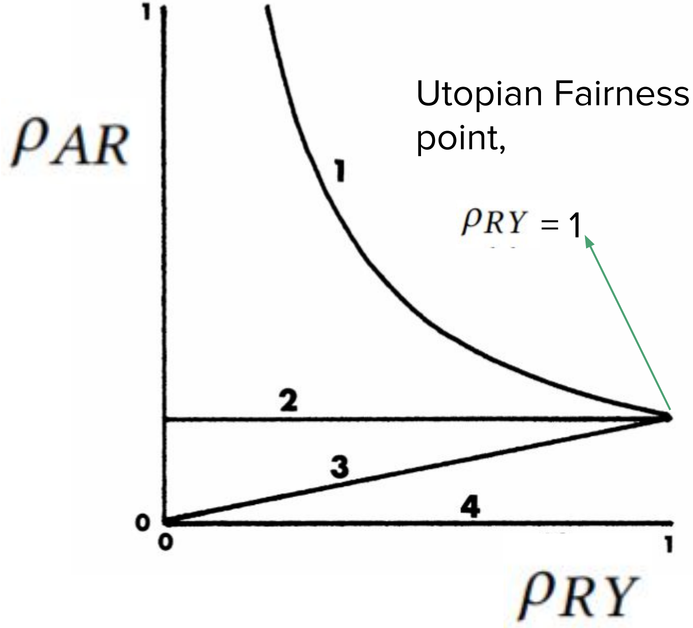
- Cleary’s: $\rho_{AR} = \rho_{AY}/\rho_{RY}$
- Throndike’s: $\rho_{AR} = \rho_{AY}$
- New criterion 1: $\rho_{AR} = \rho_{AY} \times \rho_{RY}$ and $\rho_{AR.Y} = 0$
- New criterion 2: $\rho_{AR} = 0$



We can translate Cleary’s statement into our Definition 1 as follows. Cleary’s definition states that Test X is culturally fair if knowledge of a person’s cultural group cannot be used to improve the prediction of Y made from X. Given the assumptions of this section,. this is equivalent ot the statement that the partial correlation $r_{CY.X}$ equals zero. From the formula for a partial correlation, we know the numerator of $r_{CY.X}$ is $r_{CY} - r_{CX}\ r_{XY}$. Setting this expression equal to zero and solving for r_{CX} gives Definition 1.

Darlington's attempt at unification, cont.

- Cleary's: $\rho_{AR} = \rho_{AY} / \rho_{RY}$
- Throndike's: $\rho_{AR} = \rho_{AY}$
- New criterion 1: $\rho_{AR} = \rho_{AY} \times \rho_{RY}$ and $\rho_{AR.Y} = 0$
- New criterion 2: $\rho_{AR} = 0$

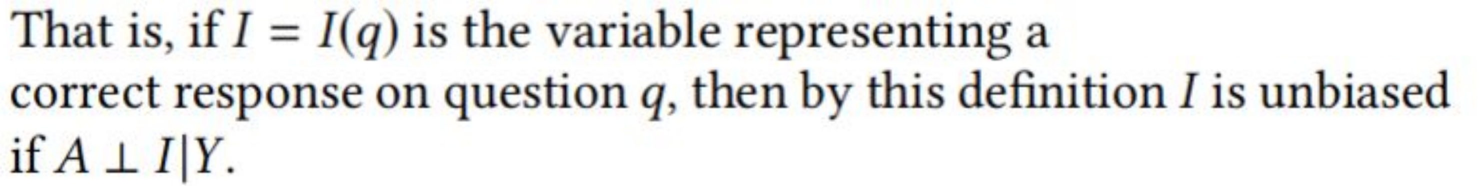


Utopian Fairness point,

$\rho_{RY} = 1$

A new approach: Differential Item Functioning

> *“A major change from focusing primarily on fairness in a domain, where so many factors could spoil the validity effort, to a domain where analyses could be conducted in a relatively simple, less confounded way. ... In a DIF analysis, the item is evaluated against something designed to measure a particular construct and something that thetest producer controls, namely a test score.”*

- Check fairness of individual items in a test than the whole test.
- 
  That is, if $I = I(q)$ is the variable representing a correct response on question q, then by this definition I is unbiased if $A \perp I | Y$.

Fairness notions in ML

- Sufficiency: $Y \perp A \mid R$
  - R satisfies sufficiency when the sensitive attribute A and target variable Y are clear from the context.

- Seperation: $R \perp A \mid Y$
  - Criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable.

- Independence: $R \perp A$
  - Criterion simply requires the sensitive characteristic to be statistically independent of the score.

Regression and Correlation

- ❑ **Regression models** (Anne Cleary ‘68) have played a more influential role in *testing research* than in the current ML Fairness research

Scope: Use Regression models to establish Fairness previously used for Unfairness. Further using models in general like Threshold Detection*

- ❑ **Correlation Vs Independence in Fairness**
  - ○ Correlation is a descriptive statistic as compared to Independence which is an inferential statistic
  - ○ Being a descriptive statistic, estimating correlation is a computationally easier and requires fewer assumptions as compared to estimating Independence which is usually non trivial.

Scope: Include Correlation as a Regularizing term while training.

Fairness Vs Unfairness

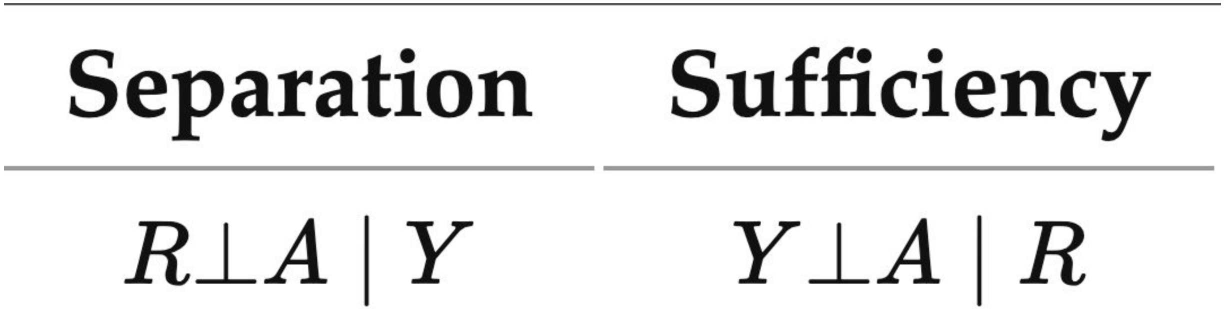
- ***What is fair*** is really hard to determine universally ?
- Obtaining more than **one of separation, sufficiency and independence is impossible** except in special circumstances *(Barocas '18 and others)*
- *Friedler et al. in 2016 and Thorndike* pointed out the **tension between individual and group notions of fairness:** "the two definitions of fairness ….. **will always be in conflict."**

- Individuals seeking justice do so when they believe that something has been unfair. Psychological studies have shown *How Our Brains Are Wired to Resist Unfair Treatment*. ***High quality labeled unfairness data*** *?*
- Earlier work on test fairness included measurements for unfair discrimination and unfair bias while **no measures exist today**.
- A shift in focus from outputs and outcomes to inputs and processes? **What causes unfairness?**

SCOPE: Defines equivalent fairness measures to encapsulate unfairness

Switching between Target (Y) and Model Score (R)

Different fairness criterion can be transformed into another by swapping the symbols R (Score) and Y (Target) and this correspondence is called ***Converse***.



Separation	**Sufficiency**
$R \perp A \mid Y$	$Y \perp A \mid R$

Separation is the converse of sufficiency.

Converse Cleary

Cleary's Regression model

Model Y (Target) ∼ R (Model Score) and look for consistent negative errors

Converse

Model R (Target) ∼ Y (Model Score) and conclude unfairness for a subgroup if the regression lines have positive errors

Converse Calibration

Calibration is defined as:

$$P(Y = 1 | R = r, A = a) = r$$

⇩

$$E(Y - r | R = r, A = a) = 0$$

⇩

$$E(R - y | Y = y, A = a) = 0$$

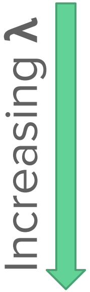
Converse Calibration: For each subgroup and level of ground truth performance Y = y,the expected error in R's prediction of the value y is zero.

SCOPE: Using the Converse framework on other fairness measures.

Parametrizing Fairness Measures

Darlington’s work to define sufficiency and separation in terms of correlations can be mapped to a parameterized space of of metrics

$$\rho_{AR} = \rho_{AY} \cdot \rho_{RY}^{\lambda}$$



Increasing **λ**

$$\rho_{AR} = \rho_{AY} / \rho_{RY}$$

Sufficiency (**λ** = -1)

$$\rho_{AR} = \rho_{AY}$$

$$\rho_{AR} = \rho_{AY} \times \rho_{RY}$$

Separation (**λ** = 1)

The framework elegantly maps contrasting metrics of separation and sufficiency through a unified parameterized model **representing compromises** between competing notions which can be further **used to interpolate the space**

Parameterizing Thorndike’s Fairness

A classifier is considered fair if it satisfies both

a) $\frac{TP + \lambda_1 FP}{TP + \lambda_2 FN}$ is constant for all values of A b) $\frac{TN + \lambda_1 FN}{TN + \lambda_2 FP}$ is constant for all values of A.

- Different fairness metrics can be derived for different values **λ_1** and **λ_2** and by choosing which or both conditions to satisfy.
- Like (1, 0)Thorndikian fairness is equivalent to sufficiency while (0, 1)Thorndikian fairness is equivalent to separation.

SCOPE: Fairness is a multifaceted measure. Its parameterization gives researchers the ability to encode competing interests of different parties, which in some scenarios might relate to the public’s notion of fairness

Summary

- ❑ Testing Criterions and their relation to fairness. Also different notions are often competing
- ❑ Model Vs Model Use Vs Fairness of Item (DIF) : Where does Fairness lie?
- ❑ Fairness Vs Unfairness: Asking a different question
- ❑ Correlation: Defining fairness in terms of correlation
- ❑ Frameworks of *Converse and Compromises (Parameterization)*
- ❑ *Values - Values encoded by technical definitions should be made explicit. By concretely relating fairness debates to ethical theories and value systems, we can make discussions more accessible to the general public and to researchers of other disciplines

Helpful Resources

- Richard B Darlington. 1971. Another Look at Cultural Fairness
- Nancy S Petersen and Melvin R Novick. 1976. An evaluation of some models for culture-fair selection
- Solon Barocas, Moritz Hardt, and Arvind Naranayan. 2018. Fairness in Machine Learning. http://fairmlbook.org