Lecture 17: Fairness

Machine Learning, Summer Term 2019

July 18, 2019

Michael Tangermann Frank Hutter Marius Lindauer

University of Freiburg

Acknowledgement of the sources for these slides

- Draft text book: Fairness and machine learning: Limitations and Opportunities by Solon Barocas, Moritz Hardt and Arvind Narayanan
- URL of the book: https://fairmlbook.org/
- NIPS 2017 Tutorial by Solon Barocas and Moritz Hardt

Lecture Overview

1. Motivation and Background
2. A Concrete Example
3. Two Definitions of Fairness
4. Wrapup

Connection of Machine Learning and Fairness

- Machine Learning is being used for making very important decisions

$\rightsquigarrow$ We need to make sure that these decisions are fair

Connection of Machine Learning and Fairness

- Machine Learning is being used for making very important decisions

$\rightsquigarrow$ We need to make sure that these decisions are fair

- What does it mean to be fair? This is a question for ethics and law:
  - Credit (Equal Credit Opprtunity Act)
  - Education (Civil Rights Act of 1964; Education Amendments of 1972)
  - Employment (Civil Rights Act of 1964)
  - Housing (Fair Housing Act)
  - Public Accomodation (Civil Rights Act of 1964)
- These laws extend to marketing and advertising in these domains!

Connection of Machine Learning and Fairness

- Machine Learning is being used for making very important decisions

⇝ We need to make sure that these decisions are fair

- What does it mean to be fair? This is a question for ethics and law:
  - Credit (Equal Credit Opprtunity Act)
  - Education (Civil Rights Act of 1964; Education Amendments of 1972)
  - Employment (Civil Rights Act of 1964)
  - Housing (Fair Housing Act)
  - Public Accomodation (Civil Rights Act of 1964)
- These laws extend to marketing and advertising in these domains!

- Machine learning allows for predictions and thus differentiation

⇝ Take objection to unjustified basis for differentiation
  - Practical irrelevance (e.g., bias in the training data)
  - Moral irrelevance

Legally recognized protected classes

- Race (Civil Rights Act of 1964)
- Color (Civil Rights Act of 1964)
- Gender (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familias status (Civil Rights Act of 1964)
- Disability status (Rehabilitation Act of 1973)
- Genetic information (Genetic Information Nondiscrimination Act)

Legally recognized protected classes

- Race (Civil Rights Act of 1964)
- Color (Civil Rights Act of 1964)
- Gender (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familias status (Civil Rights Act of 1964)
- Disability status (Rehabilitation Act of 1973)
- Genetic information (Genetic Information Nondiscrimination Act)

Note: society is far from being fair; with ML we aim to do better!

- Opportunity to objectively assess and remove bias
- Opportunity to get rid of implicit (even unconscious) biases

Issues to watch out for in machine learning

- Skewed sample
  - E.g., predictive policing
    - Future observations of crime confirm predictions
    - Riks of a feedback loop
      (less opportunities to observe crime that contradicts predictions)

Issues to watch out for in machine learning

- Skewed sample
  - E.g., predictive policing
    - Future observations of crime confirm predictions
    - Riks of a feedback loop
      (less opportunities to observe crime that contradicts predictions)
- Tainted examples
  - Learn to predict hiring decisions
  - Learn to predict job success (e.g., review scores)
  - Learn to predict objective score (e.g., sales)

Issues to watch out for in machine learning

- Skewed sample
  - E.g., predictive policing
    - Future observations of crime confirm predictions
    - Riks of a feedback loop
      (less opportunities to observe crime that contradicts predictions)
- Tainted examples
  - Learn to predict hiring decisions
  - Learn to predict job success (e.g., review scores)
  - Learn to predict objective score (e.g., sales)
- Limited features
  - Features may be informative for the majority group, but not for a minority group
  - Minority may even just be a minority in the training/validation data (example: skin cancer)
  - Different models with same validation accuracy can differ substantially concerning fairness

Issues to watch out for in machine learning

- Proxies
  - Many features are correlated with sensitive features
  - With rich data, sensitive features are largely encoded across other features

Issues to watch out for in machine learning

- Proxies
  - Many features are correlated with sensitive features
  - With rich data, sensitive features are largely encoded across other features
- Sample size disparity
  - By definition, there is more data for majority groups
  - Routinely, more data leads to smaller errors

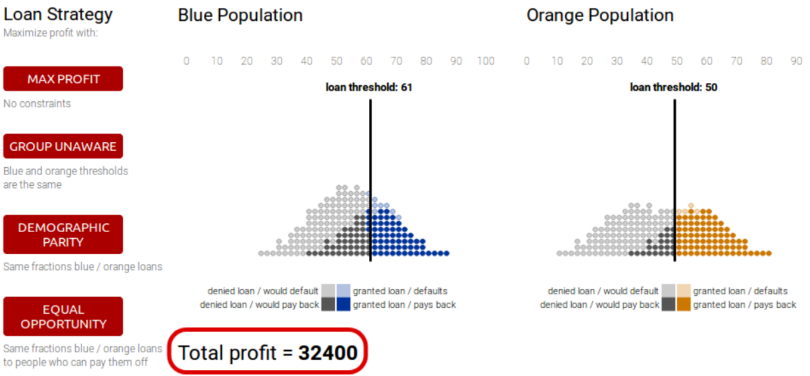
Three different problems

- Discovering unobserved differences in performance
  - Skewed sample
  - Tained examples
- Coping with observed differences in performance
  - Limited features
  - Sample size disparity
- Understanding the causes of disparities in predicted outcome
  - Proxies

Lecture Overview

1. Motivation and Background
2. A Concrete Example
3. Two Definitions of Fairness
4. Wrapup

Example: decisions about granting a loan or not



URL:

`https://research.google.com/bigpicture/attacking-discrimination-in-ml`

Lecture Overview

1. Motivation and Background
2. A Concrete Example
3. Two Definitions of Fairness
4. Wrapup

Example: job advertisement for software engineers

- X features of an individual (here: browsing history)
- A sensitive attribute (here: gender)
- $C = c(X, A)$ predictor (here: show ad or not)
- Y target variable (here: software engineer?)
- Notation: $P_a(E) = P(E \mid A = a)$

Two definitions of fairness

- Independence: C independent of A
- Separation: C independent of A conditional on Y

Recall notation:

- A sensitive attribute (example: gender)
- $C = c(X, A)$ predictor (example: show ad or not)
- Y target variable (example: software engineer?)

Independence

- Definition: C is independent of A
  - For all groups a and b and all outcomes c of C,

$$P_a(C = c) = P_b(C = c)$$

Independence

- Definition: C is independent of A
  - For all groups a and b and all outcomes c of C,

$$P_a(C = c) = P_b(C = c)$$

  - Other names: demographic parity, statistical parity
  - Approximate versions

$$\frac{P_a(C=1)}{P_b(C=1)} \geq 1 - \epsilon \qquad\qquad |P_a(C=1) - P_b(C=1)| \leq \epsilon$$

Independence

- Definition: C is independent of A
  - For all groups a and b and all outcomes c of C,

$$P_a(C = c) = P_b(C = c)$$

  - Other names: demographic parity, statistical parity
  - Approximate versions

$$\frac{P_a(C = 1)}{P_b(C = 1)} \geq 1 - \epsilon \qquad\qquad |P_a(C = 1) - P_b(C = 1)| \leq \epsilon$$

- Achieving independence
  - Post-processing
  - Training time constraint
  - Pre-processing, e.g., via representation learning

Independence

- Definition: C is independent of A
  - For all groups a and b and all outcomes c of C,

$$P_a(C = c) = P_b(C = c)$$

  - Other names: demographic parity, statistical parity
  - Approximate versions

$$\frac{P_a(C = 1)}{P_b(C = 1)} \geq 1 - \epsilon \qquad\qquad |P_a(C = 1) - P_b(C = 1)| \leq \epsilon$$

- Achieving independence
  - Post-processing
  - Training time constraint
  - Pre-processing, e.g., via representation learning
    - Find a representation $Z = f(X, A)$, aiming for $maxI(X; Z)$ and $minI(A; Z)$
    - Then we base our classifier C only on Z

Separation

- Define a score $r = R(A, X)$
- R independent of A conditional on Y
  - For all groups a and b, all outcomes r of R, and all outcomes y of Y:

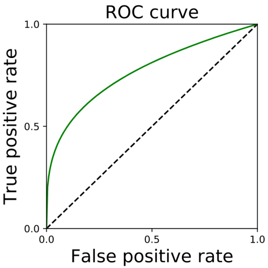
$$P_a(R = r \mid Y = y) = P_b(R = r \mid Y = y)$$

Separation

- Define a score $r = R(A, X)$
- R independent of A conditional on Y
  - For all groups a and b, all outcomes r of R, and all outcomes y of Y:

$$P_a(R = r \mid Y = y) = P_b(R = r \mid Y = y)$$

  - Again, you can define approximate versions like for independence
  - Achieving independence: training constraint or post-processing by looking at (TPR, FPR) for all possible thresholds on score r

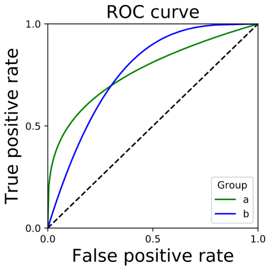


Separation

- Define a score $r = R(A, X)$
- R independent of A conditional on Y
  - For all groups a and b, all outcomes r of R, and all outcomes y of Y:

$$P_a(R = r \mid Y = y) = P_b(R = r \mid Y = y)$$

  - Again, you can define approximate versions like for independence
  - Achieving independence: training constraint or post-processing by looking at (TPR, FPR) for all possible thresholds on score r

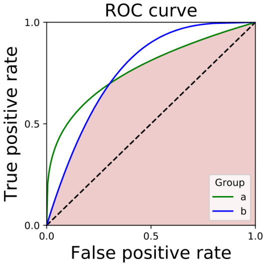


Separation

- Define a score $r = R(A, X)$
- R independent of A conditional on Y
  - For all groups a and b, all outcomes r of R, and all outcomes y of Y:

$$P_a(R = r \mid Y = y) = P_b(R = r \mid Y = y)$$

  - Again, you can define approximate versions like for independence
  - Achieving independence: training constraint or post-processing by looking at (TPR, FPR) for all possible thresholds on score r



Tradeoffs are Necessary

Theorem: impossibility of both independence and separation

If neither A nor R are independent of Y, then you cannot get both separation and independence at once.

- Someone (probably not a computer scientist) has to decide what fairness criteria we want to fulfill in a particular application
- Impossibility only holds for exact independence/separation
  - We can still aim for good tradeoffs of their approximate versions

Lecture Overview

1. Motivation and Background
2. A Concrete Example
3. Two Definitions of Fairness
4. Wrapup

Summary by learning goals

Having heard this lecture, you can now ...

- List some features that are illegal to use in some cases
- Motivate the need for research on fairness in machine learning
- Discuss two notions of algorithmic fairness and their relation