(un)Fairness in Machine Learning

Aaron Roth

@aaroth

Penn
University of Pennsylvania
PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME
Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 21, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, “That’s my kid’s stuff.” Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the theft.
What Does “Unfairness in Machine Learning” Mean?

• We’ll be more specific soon, but...

• Machine learning algorithms will inevitably make mistakes

• “Fairness” speaks to how those mistakes are distributed across individuals.
  • E.g. do certain sub-populations suffer a disproportionately high rate of errors?
Why might machine learning be “unfair”? 

• Many reasons:
  • Data might encode existing biases.
    • E.g. labels are not “Committed a crime?” but “Was arrested.”
  • Data collection feedback loops.
    • E.g. only observe “Paid back loan” if loan was granted.
  • Different populations with different properties.
    • E.g. “Graduated High School” might correlate with label differently in wealthy populations.
  • Less data (by definition) about minority populations.
Upshot: To be “fair”, the algorithm may need to explicitly take into account group membership. Else, optimizing accuracy fits the majority population.
But what is “fairness”? 
Individual Fairness Notions

For classification problems (e.g. lending)

• “Metric Fairness” [Dwork et al.]:
  • “Similar Individuals should be treated similarly”
    \[ \forall x, x': \left| \Pr[h(x) = 1] - \Pr[h(x') = 1] \right| \leq d(x, x') \]

• “Weakly Meritocratic Fairness” [Joseph et al.]:
  • “Less creditworthy individuals should not be favored over more creditworthy individuals.”
    \[ \forall (x, y), (x', y'): \Pr[h(x) = 1] > \Pr[h(x') = 1] \Rightarrow y > y' \]
Statistical Fairness Notions

Partition the world into a small number of protected subgroups (often 2): $G_1, \ldots, G_k$, and ask for equality of some statistic across groups:

- **Statistical parity:** Positive Classification Rates [Calders et al, Dwork et al.]

$$\Pr_{(x,y) \sim D} [h(x) = 1 \mid x \in G_i] = \Pr_{(x,y) \sim D} [h(x) = 1 \mid x \in G_j] \forall i,j$$

- **Equalized Odds:** False Positive/Negative Rates [Kleinberg et al, Chouldechova, Hardt et al, Zafar et al.]

$$\Pr_{(x,y) \sim D} [h(x) = 1 \mid y = 0, x \in G_i] = \Pr_{(x,y) \sim D} [h(x) = 1 \mid y = 0, x \in G_j] \forall i,j$$

- **Equal Calibration:** Positive Predictive Value [Kleinberg et al, Chouldechova]:

$$\Pr_{(x,y) \sim D} [y = 1 \mid h(x) = 1, x \in G_i] = \Pr_{(x,y) \sim D} [y = 1 \mid h(x) = 1, x \in G_j] \forall i,j$$
A Case Study: the COMPAS recidivism prediction tool.

- A cartoon of the COMPAS tool:
  - People have features $x$ and true label $y \in \{\text{R}(eoffend), \text{D}(id not r}\}$
  - The tool makes a prediction $f(x) \in \{R, Y\}$

The COMPAS risk tool is unfair. It has a higher false positive rate on the black population compared to the white population.

The COMPAS risk tool is fair. It has equal positive predictive value on both the black and white population.
A Case Study: the COMPAS recidivism prediction tool.

• Two proposed definitions of fairness, in the same style. Fix a collection of “protected” groups $G_1, \ldots, G_k$

• Propublica: A classifier is fair if for every pair of groups $G_i, G_j$:
  \[
  \Pr_{(x,y)} [f(x) = R | y = D, x \in G_i] = \Pr_{(x,y)} [f(x) = R | y = D, x \in G_j]
  \]

• Northpointe: A classifier is fair if for every pair of groups $G_i, G_j$:
  \[
  \Pr_{(x,y)} [y = R | f(x) = R, x \in G_i] = \Pr_{(x,y)} [y = R | f(x) = R, x \in G_j]
  \]

• Both reasonable. But.. [Chouldechova 16], [Kleinberg, Mullainathan, Raghavan 16]
  • No classifier can simultaneously satisfy both conditions if the base rates in the two populations differ, and the classifier is not perfect.

1. And equalize false negative rates
Why does equalizing FP rates (sometimes) correspond to “fairness”? 

- Being incorrectly labelled as “High Risk” constitutes a harm. FP rate is your probability of being harmed if you are born as a uniformly random “Low Risk” member of a population.
Why does equalizing FP rates (sometimes) correspond to “fairness”?

• Being incorrectly labelled as “High Risk” constitutes a harm. FP rate is your probability of being harmed if you are born as a uniformly random “Low Risk” member of a population.

Equal FP rates => Indifference between groups when behind veil of ignorance.
(Statistically) easy to check and enforce without making any assumptions, but semantics have limited meaning when applied to individuals…

- Toy example: Protected subgroups are “Men”, “Women”, “Blue”, “Green”. Labels are independent of protected attributes.
- The following allocation achieves statistical parity and equalized odds:
Statistical Fairness Notions

• The problem: Statistical constraints averaged over coarse subgroups can be “gerrymandered” to pack unfairness into structured subgroups.
  • No reason to expect it won’t happen with standard optimization techniques: we will see it does.

• Just add “green men”, “blue women”, etc. as protected subgroups?
  • What about other groups?
  • Can’t ask for this constraint on every definable subgroup without overfitting.
Statistical Fairness Notions

• A Middle Ground: Ask for equal FP rates across all possible divisions of the data that can “reasonably be identified” with a simple decision rule.
  • E.g. all ways of combining protected features with an “AND”
    • Protect not just black people, women, old people, disabled people, but...
    • old black women, young white disabled people, old women, etc.
    • Lots of such groups: If $d$ protected features, $2^d$ combinations...

• Mitigates the “gerrymandering problem”
  • Now asking for statistical fairness over many finely defined subgroups.
Learning for Subgroup Fairness

• Goal: find the optimal fair distribution over classifiers in $H$ subject to equalizing FP rates across subgroups defined by a class of functions $G$.

• Main result: reduction to a short sequence of learning problems over $G$ and $H$.

Key idea: simulate a zero-sum game between an Auditor and a Learner.

• Auditor’s objective: Find a subgroup on which Learner’s rule is maximally discriminatory.

• Learner’s objective: Minimize weighted combination of error and FP rates on subgroups specified by Auditor.

• Optimal solution corresponds to equilibrium in this game.
Main Theoretical Result

Theorem (Informal): There is an oracle efficient algorithm that makes $\text{poly}(d, \frac{1}{\epsilon})$ oracle calls, and outputs a distribution $D$ over $H$ that achieves $\epsilon$-optimal error (with respect to the best $\gamma$-fair distribution over classifiers), and is $(\gamma + \epsilon)$-fair with respect to $G$.

So does it work?
Experiments

Data set: Communities and Crime

- Census and other data on 2000 US communities.
- Target prediction: High vs. low crime rate.
- 122 real valued features; 18 sensitive (racial demographics of community, police department, etc)
- Protected groups: Any group that can be defined as a linear threshold function of the 18 protected attributes.

- Heuristic learning oracle: regression.
trajectory: gamma = 0.006
Pareto Curves: (error, γ)

Across a range of datasets unfairness starts around .02-.03
We are able to drive near 0 with only a 3-6% increase in error
Pareto curve for \((\text{error, } \gamma)\): varying the features

Pareto frontier for varying number of protected features, communities and crime

- 18 protected features
- 10 protected features
- 4 protected features
Flattening the Discrimination Heatmap

![Diagram showing the flattening of a discrimination heatmap.](image)
Subgroup Fairness!
Does fairness to marginal subgroups achieve finer subgroup fairness? No!
Summary

• “Unfairness” can be a real problem that arises naturally in ML applications, without any malicious design intent.

• Due to impossibility results, must pick and choose even amongst fairness desiderata.

• Asking for “fairness” is not without cost.
  • Quantitative definitions of fairness induce fairness/error “Pareto Curves” for classification tasks.
  • Does not answer the question: Which point on the Pareto frontier should we pick? This requires domain expertise and stakeholder input.
For more:

• Me:
  • [http://www.cis.upenn.edu/~aaroth](http://www.cis.upenn.edu/~aaroth)
  • @aaroth

• Subgroup Fairness:
  • “Preventing Fairness Gerrymandering”, Kearns, Neel, Roth, Wu. ICML 2018.

• *The Ethical Algorithm*, Kearns and Roth.
  • Forthcoming in 2019 from Oxford University Press