The Science of Fair ML
Things That Can Go Wrong

- Distribution of $x$ differs between groups
- Some group underrepresented in data $S$
- Features in $x$ are less predictive for some group
- Some group is just harder to predict
- The $y$'s are biased in the first place
Fairness in ML

- Typically a property of a model (ML algo output)
- Exceptions: online decision-making, RL, bandit settings
- Multiple types of fairness definitions
Types of Model Fairness

- Group fairness (most common)
- Individual fairness
- Interpolations between the two
- Others (causal, fair representations, ...
Group Fairness Notions

Start by identifying:

• groups or attributes we wish to "protect" (e.g. race, gender)

• what constitutes harm (e.g. error, false pos/neg)

Choices are subjective & domain-specific
Then seek to equalize rates of harm across groups.

Example:
- domain: consumer lending
- groups: male & female
- harm: false rejection (negs)

Want to find model $h(x)$ s.t.

$$FN(h, \text{male}) \approx FN(h, \text{female})$$

\[ \text{allows for optimization of overall error} \]
Confusion Matrices

\[ F_{N}(h, \text{male}) = \Pr_{(x,y) \sim P} [h(x) = -1/ \text{male} \land y = +1] \]

\[
\begin{array}{c|c|c}
\hline
h(x) & +1 & -1 \\
\hline
+1 & a & b \\
-1 & c & d \\
\hline
\end{array}
\]

Group 1
\[
\frac{c}{a+c}
\]

Group 2
\[
\frac{c'}{a'+c'}
\]
\textbf{Note:} We can achieve FN rates by randomization.

\textit{If individual \textit{x}, predict} \( \hat{y} = + \) \textit{with prob.} \( p \)

\textit{If} \( y = - \), \textit{can't be a FN}
\textit{If} \( y = + \), \( \hat{y} = - \) \textit{w. p.} \( 1-p \)

\[ \therefore \text{FN}(p,*) = 1-p \]
If we are given a model $h(x)$ and have access to group membership, easy to audit $h(x)$ for fairness.

How can we learn a fair model $h(x)$?

Why won't standard ML algorithms work?
A Post-Processing Approach ("bolt on")

- Start with non-fair $h(x)$, want to use M/F error rates
- Build a probabilistic classifier on top of $h(x)$:

\[
\begin{array}{c|cc}
 & M & F \\
\hline
P & p & q \\
p & r & s
\end{array}
\]

$h(x)$: 

\[
\begin{aligned}
\tilde{h}(x) &= + & \text{prob.} \\
&= + & \text{(closed under mixtures)}
\end{aligned}
\]
\( p = q = r = s = 0 : \)
\( \langle \bar{h} \rangle = h , \langle \bar{h} \rangle = 3 \)

\( p = q = r = s = \frac{1}{2} : \)
\( \langle \bar{h} \rangle = \frac{1}{2} \)

\( p = r = \frac{1}{2} , q = s = 1 : \)
error on men = \( \frac{1}{2} \)
error on women = same as \( h \)

etc.
Equalizing Error Rates

• Let group error rates be \( \varepsilon_M, \varepsilon_F \)

• Suppose \( \frac{1}{2} \geq \varepsilon_F \geq \varepsilon_M \)

• Pick \( g = 1, s = 0 \) \( \Rightarrow \hat{\varepsilon}_F = \varepsilon_F \)

• Pick \( r = 1-p \), agree with \( h \) with prob \( p \) pointwise

\[ \hat{\varepsilon}_M = p \varepsilon_M + (1-p)(1-\varepsilon_M) \]

\[ = p \varepsilon_M + 1 - \varepsilon_M - p + p \varepsilon_M \]

\[ = 2p \varepsilon_M - p + 1 - \varepsilon_M \]

\[ = p(2 \varepsilon_M - 1) + 1 - \varepsilon_M \]
Now solve for $p$:

\[
p(2\varepsilon m - 1) + 1 - \varepsilon m = \varepsilon_F - \varepsilon_F
\]
\[
p(2\varepsilon m - 1) = \varepsilon_F + \varepsilon m - 1
\]
\[
p = \frac{\varepsilon_F + \varepsilon m - 1}{2\varepsilon m - 1}
\]

**Sanity Checks:**

- $\varepsilon m, \varepsilon F \leq \frac{1}{2} \Rightarrow \text{num and denom} \leq 0
  \Rightarrow p \geq 0$
- $\varepsilon m = \varepsilon F \Rightarrow p = \frac{0}{0} = 1, r = 0$
- $\varepsilon F = \frac{1}{2} \Rightarrow p = \frac{1}{2} + \varepsilon m - 1 = \frac{\varepsilon m - \frac{1}{2}}{2\varepsilon m - 1} = \frac{\varepsilon m - \frac{1}{2}}{2(\varepsilon m - \frac{1}{2})} = \frac{1}{2}, r = \frac{1}{2}$
Feasible Region & Pareto Frontier

\[ \alpha \varepsilon_F + (1 - \alpha) \varepsilon_M < \varepsilon_F \]
Finding the Frontier

Feasible Region

allowed unfairness

smallest possible $\varepsilon$ given $\gamma$

EF

3F

EE

EF - $\varepsilon$N

SE

$\varepsilon$M

$\varepsilon$M
Algorithm

- Problem of finding \( \hat{h} \) than minimizes \( \text{EL}( \hat{h} ) \)

Subject to

- \( y\text{-axis} \leq z \)

is a linear program in \( p, q, r, s \).

(Framework of result due to Hardt, Price, Srebro.)
Recap

- Introduce randomized $\bar{h} = \langle p, q, r, s \rangle$ "on top" of $h$
- Allows different tradeoffs between $\bar{\epsilon}$ (error) and $\bar{\epsilon}_F - \bar{\epsilon}_M$ (unfairness)
- Optimal tradeoffs given by Pareto Frontier
- Fast algorithm
- Generalizes to FP/FN rates, more groups
Any drawbacks?
What More Could We Want?

- Suppose original $h$ came from $H$
- $h \in H$ chosen to have small error
- But $H$ might contain better error-unfairness tradeoffs than Pareto Frontier of $h$
The H Frontier

Feasible Region

H frontier
Challenges

- \( H \) may not be closed under mixtures
- Model space for \( h \) was just \( \langle p, q, r, s \rangle \), had fast Pareto algo
- In general, even finding low-error \( H \) is "hard" even absent fairness
- But ML does have good non-fair heuristics...
**Oracle/Reduction Approach**

- Assume we have a subroutine ("oracle") \( L \) for (approx.) minimizing \( \varepsilon(h) \) in \( H \) (non-fair)
- Assume \( L \) handles reweightings of data
- Use repeated calls to \( L \) to find \( H \)-frontier
The Set-Up

- L outputs $h \in H$, but we will output mixtures on distributions $\Delta(H)$ (needs to be sparse)

- Want to solve:

  minimize $E(h)$ for $h \in A(H)$ subject to:

  $|E_F(h) - E_M(h)| \leq \varepsilon$

large $\tau$

small $\tau$
Game Theory Framework

• Algo structured as a two-player, zero-sum, repeated game

• Learner plays some $\alpha \Delta(H)$ at each round

• Regulator plays a distribution on data:
  - Weight $\ell$ on $F$
  - $1-\ell$ on $M$
Payoffs

- Payoff to Learner:
  \[-[\mathcal{E}(h) + (\text{overall error}) \max(0, |2E_F(h) - (1-2)E_M(h)| - \gamma)]\]
  \[(\text{fairness violation})\]

- Payoff to Regulator:
  \[+ [\text{same}]\]

- Important/deep fact:
  Nash equilibrium of game
  = solution (\Delta(h)) to
  optimization problem
Algorithm (Sketch)

- Reg. starts by letting $\mathcal{T} = \omega t.$ of $F$ in $S$
- Lrn. starts by calling $L$ on $S \rightarrow h_1$
- For $t = 1, 2, \ldots, T - 1$:
  - Reg. updates $\omega t s.$ to improve payoff on
    \[
    \frac{1}{t} (h_1 + h_2 + \ldots + h_t)
    \]
  - Lrn. calls $L$ on $S$ & current $\omega t s. \rightarrow h_{t+1}$

Final output: $h = \frac{1}{T} (h_1 + h_2 + \ldots + h_T)$
"Theorem": If \( L \) outputs (near) optimal \( h_t \) at each round, then after \( T \) rounds, \( h \) is within \( -\frac{1}{\sqrt{T}} \) of opt. soln to:

\[
\begin{align*}
\text{minimize } & \quad \mathbb{E}(h) \\
\text{subject to } & \quad |E_f(h) - \mathbb{E}_m(h)| \leq \epsilon
\end{align*}
\]
• So we have reasonably general/practical algos for training group fair models

• What about individual fairness guarantees?

• Q: Why not let each \( x \) be its own “group”?

• A: E.g. to equalize errors, only rates are 0 (perfect), 1 (opposite) or \( \frac{1}{2} \) (random)
**Better: Metric Fairness**

- Posit a distance metric $d(x, x')$ between individuals
- Learn a real-valued model $h(x)$
- Fairness constraint: for every $x, x'$:
  $$|h(x) - h(x')| \leq \alpha d(x, x')$$ where $\alpha$ is a constant
- "Similar individuals treated similarly"
Challenges

• Where/from whom do we get \( d(x, x') \) ?

• Pushes the problem elsewhere?

• \( d(x, x') \) closed form?

• Thresholding \( h(x) \) may lose fairness

• No practical algos*
Interpolating between group & individual fairness
• Suppose we enforce group fairness by race, gender, age, disability, income...
• Might still discriminate disabled Hispanic women over 55 making ≤ $40K
• "Fairness gerrymandering"
• Let's enforce group fairness for all combos* of protected features
Game Theory II

• Learner still chooses (mixture) model $h$

• Regulator now finds combos on which $h$ is unfair—itslef a learning problem!

• Both players use heuristic/oracle $L$

• Same kind of theorem—get $1/\sqrt{T}$ of optimal, on *significant subgroups
Average Individual Fairness

- Suppose we make many decisions about $x$ over time
- E.g. product recs, ads, ...
- Let $E_x(h)$ be error rate across decisions for individual $x$
- Ask that all $E_x(h)$ approx. equal
- Game Theory III: Regulator now picks $x$ s.t. $E_x(h)$ is largest
Figure 2: (a) Error-unfairness trajectory plots illustrating the convergence of algorithm **AIF-Learn**. (b) Error-unfairness tradeoffs and individual errors for **AIF-Learn** vs. simple mixtures of the error-optimal model and random classification. Gray dots are shifted upwards slightly to avoid occlusions.
(Subjective) Fairness Elicitation

• What if fairness is complex and subjective?

• Elicitation:
  - "x & x' should receive same outcome"
  - "x should receive at least as good an outcome as x'"

• Game Theory IV: Regulator enforces elicited constraints
### 5.2 Subjective Fairness Elicitation

In your view, as a matter of fairness, should the following two individuals receive the same recidivism prediction, or is it okay to give them different predictions?

<table>
<thead>
<tr>
<th>sex</th>
<th>age</th>
<th>race</th>
<th>juven. felony count</th>
<th>juven. misdemeanor count</th>
<th>juven. other count</th>
<th>priors count</th>
<th>severity of charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>25</td>
<td>Caucasian</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>Felony</td>
</tr>
</tbody>
</table>

vs.

<table>
<thead>
<tr>
<th>sex</th>
<th>age</th>
<th>race</th>
<th>juven. felony count</th>
<th>juven. misdemeanor count</th>
<th>juven. other count</th>
<th>priors count</th>
<th>severity of charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>29</td>
<td>African-American</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>Felony</td>
</tr>
</tbody>
</table>

Should be treated equally | Ok to treat differently, or no opinion

**Figure 1:** Screenshot of sample subjective fairness elicitation question posed to human subjects.
Other Strengthenings

- Minimax group fairness
  - Make worst group error as small as possible
  - If $\varepsilon_1 \geq \varepsilon_2 \geq \ldots \geq \varepsilon_k$ minimax
    $\varepsilon_1' = \varepsilon_2' = \ldots = \varepsilon_k'$ equalized
    then $\varepsilon_1 \leq \varepsilon_1', \ldots, \varepsilon_k \leq \varepsilon_k'$

- Lexicographic group fairness
Fair (Supervised) ML: Recap

- Group fairness notions
- Bolt-on/post-processing
- Pareto frontiers/tradeoffs
- In-processing: reduction to non-fair; game theory
- Individual fairness
- Group $\rightarrow$ Individual
Other Settings

• Fair RL/Control
• Fair representations
• Causal approaches
• Fair clustering
• Fair rankings
• Fair labelings
The Real World

- Open-ended AI services
- Unknown downstream uses
- Data curation/collection
- External/"activist" audits