Fairness in Machine Learning
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 positive)

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.

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

allows for optimization of overall error
Note: We can achieve FN rates by randomization.

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

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

\[ \Rightarrow FN(p,*;*) = p. \]
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?
Ways Things Go Wrong

• Have much less data on some group (fine if groups all “same”)

• Different groups have different distributions

• Our features are less predictive on some group

• Some group inherently less predictable

• Our data is biased in the first place
Algós for Fair ML:
Bias Mitigation
A Post-Processing Approach ("bolt on")

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

\[
\begin{array}{c c c}
M & F \\
\hline
+ & p & q \\
- & r & s
\end{array}
\]

\[ \hat{h}(x) = + \]

(\text{closed under mixtures})
\[ p = q = r = s = 0 : \]
\[ \langle h \rangle = h, \langle 3h \rangle = 3h \]
\[ p = q = r = s = \frac{1}{2} : \]
\[ \langle h \rangle = \frac{1}{2}, \text{ perfectly fair} \]
\[ p = r = \frac{1}{2}, q = s = 1 : \]
\[ \text{error on women} = \frac{1}{2} \]
\[ \text{error on women} = \text{same as h} \]

etc.
Set of all \( \langle \text{pig, girl, cow} \rangle \) gives Pareto frontier of \( h \).

\[ \text{error on men} \]

\[ \text{error on women} \]

\[ \text{overall error} \]

\[ \varepsilon(h) \]
Algorithm

- Problem of finding $h$ that minimizes $\ell(h)

subject to

$y-axis \leq T$

is a linear program

in $p_i,q_i,s_i$.

(Framework & result due to Hardt, Price, Srebro.)
What more could we want?

- Imagine $h \in H$ (NNs, DTs, ...) by some learning algo

Can we find $H$-frontier?
Well...

...even finding $h^*$ is intractable in worst case.

...but we do have effective non-fair heuristics.
The Reductions/Oracle Approach

- Assume we have a black-box subroutine $L$ for learning $h \in H$ w.r.t. $E\left(h\right)$ only (non-fair).

- But $L$ is "pretty good" in general (can solve weighted class problems in $H$).

Show we can use $L$ for fair learning.
Constrained optimization

\[
\min_{h \in \Delta(H)} \{ \epsilon(h) \} \quad \text{s.t.}
\]

Fairness constraints:

1. \(|\epsilon(h, \text{white}) - \epsilon(h, \text{black})| \leq \alpha\)
2. \(|\epsilon(h, \text{white}) - \epsilon(h, \text{hispanic})| \leq \alpha\)
3. \(|\epsilon(h, \text{black}) - \epsilon(h, \text{hispanic})| \leq \alpha\)
   \vdots
K. (usually small, but...)

Introduce variables for weights in \(\Delta(\text{H})\) & constraints \(\Rightarrow\)

Huge LP.
Game Theory Formulation

- Learner plays mixed strategy \( p \in \Delta(H) \)
- Regulator plays mixed strategy \( q \) over fairness constraints
- Zero-sum game on:
  \[ E(p) + \text{constraint violations}(p,q) \]
  payoff to Regulator
  = \(-\) payoff to Learner
  Nash equl = constrained opt solution
A Classic Theorem (Freund & Schapire)

If $L$ & $R$ play iteratively:

1) $L$ best responds to $g_t$
2) $R$ updates $g_{t+1}$ using no-regret algo

Then converge to $\frac{1}{\sqrt{t}}$-optimal solution.
(2) Usually easy
(1) Often reduces to weighted classification with wts. given by \( \frac{e^t}{\text{"oracle" } L} \)

\( \Rightarrow \) "principled heuristics" that are implementable.
Towards Individual Fairness
Q: Why not treat each individual \( x \) as their own "group"?

A: Error (or FP, FN, ...)

"rate" on \( x \) is either 0 or 1.

But there are other approaches...
Metric Fairness

- Posit a distance metric $d(x, x')$ between pairs of individuals.
- $h(x)$ our real-valued prediction.
- Then constrain $h(x)$ to obey $\forall x, x'$:

$$|h(x) - h(x')| \leq \alpha \cdot d(x, x')$$
Difficulties

• Where do we get \( d(x, x') \)?
• Closed form?
• Usually want to threshold \( h(x) \), lose fairness
• Practical challenges
Subgroup

Fairness
• Suppose we ask for group fairness by all of race, gender, disability, age, income,...

• Might still discriminate against disabled Hispanic women over age 55 making ≤ 20K/yr.
Framework

- Model class $H$
- Group membership class $G$
- For $g \in G$, $g(x) \in \{0,1\}$ indicates if $x$ is in $g$ (e.g. disabled Hispanic...)
- Now allowing $G$ to be large or infinite
Game Theory II

- Learner plays $h \in H$
- Regulator plays $g \in G$, finds most violated $g$ (e.g. $h$ has high error on $g$)

Reduce to non-fair case; $L$ no-regret, $R$ best response
Another Approach: Average Individual Fairness
Suppose we will make many decisions about $x$ over time.
- E.g. product rec's
- Then any $h$ has error rate $E_x(h)$ across problems.
- Ask that all $E_x(h)$ be equal across individuals $x$.
- Game Theory III.
Fairness

Elicitation
• What if fairness isn’t “simple”...
• ...but we can elicit empirical fairness judgements.
• E.g., "Alice & Bob should receive same treatment" "Alice should be treated at least as well as Bob"
**Framework**

- **Outcome data** $S = \{x_i, y_i\}_{i \geq 1}$
- **Fairness data** $F$ of form $x_i = x_j$, $x_i \equiv x_j$
- Find $h \in H$ that min's error on $S$ subject to $F$
- Generalize to dist's of $S$ & $F$
- **Game Theory** IV
Beyond Equalization

- Problem: may achieve by heedlessly inflating harm to advantaged

- Alternative: minimax group fairness:

\[\min \max_{g \in G} \min_{h \in H} \mathbb{E}_g(h)\]

- Game Theory
Other Learning Settings
Fairness in Bandits

- Ground truth data \(<x, y>\)
- Loan \(\sim \mathcal{L}_R\), prob. of repayment
- Unknown linear map
  \[ y = \theta \cdot x + \text{noise} \] (linear regress)
- Meritocratic fairness:
  If \( y_1 \geq y_2 \), must have
  prob. of loan to \( x_1 \) \( \geq \) prob. of loan to \( x_2 \)
• Bandit setting: each day $x_1, ..., x_k$ arrive, must choose loans fairly

• Standard algo: LIN-UCB

\[
\begin{align*}
\tilde{y} = \frac{\tilde{y}}{C_E} & \quad \begin{cases}
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
\text{Give loan(s) to highest UCBS} \Rightarrow \text{fast convergence to opt}
\end{cases}
\end{align*}
\]

\underline{Not fair}
Fair Modification

- Interval chaining
- May even choose non-overlapping intervals

\[ \{ \text{I I I I I I} \} \text{ treat equally} \]

\[ x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ldots \]

- Choose interval
  \( \Rightarrow \) more data
  \( \Rightarrow \) chains fragment
  \( \Rightarrow \) fast convergence to opt
Other Topics

- Fair RL (e.g. meritocratic wrt Q-values)
- Fair Representations
- Causal Approaches
- Fair Clustering
- Fair Rankings
Some Resources

- "Frontiers of Fairness in Machine Learning" Chouldechova & Roth
- "Fairness and ML" Barocas, Hardt, Narayanan, fairmlbook.org
- "The Ethical Algorithm" Kearns & Roth
Privacy in ML
What Do We Want?

- Not addressing preventing data breaches, unwanted access, etc. - domains of cryptography and security
- Rather, prevent inferences and exfiltration from trained model
(Bad) Examples

- K-NN models
- SVMs
- Neural Networks
- Any model with confidence ratings

- Even black-box access problematic
- “Anonymizing” data doesn’t work
High-Level Idea

Shouldn't reveal "anything" about your data - even with additional computation & data

Training data $D$
Differential Privacy

Say algo $A$ is $\varepsilon$-DP if for all neighboring $D, D'$ and set $S \subseteq \text{range}(A)$:

$$\Pr[A(D') \in S] \leq e^\varepsilon \Pr[A(D) \in S]$$

wrt randomization of $A$ only.