Fair Classification

Outline

1. Definitions of fairness
   - Statistical parity
   - Predictive rate parity
   - Equalized odds

   (with a guest appearance of calibration later in the lecture)

2. Impossibility results
   - Not even any 2 of these 3 notions can be satisfied simultaneously, except in degenerate cases

3. Fairness-accuracy tradeoff

4. Using social choice to generalize ML fairness

Model

- Running examples: Think of…
  - A bank deciding which loan applications to approve
  - A judge deciding which alleged offenders to grant bail

- Model:
  - $X \in \mathbb{R}^d$ = non-sensitive attributes (e.g., income, education, …)
  - $A \in \{0,1\}$ = sensitive attribute (e.g., race or gender)
  - $Y \in \{0,1\}$ = target variable (e.g., would they truly repay the loan? would the alleged offender commit a crime before their trial?)
  - $C \in \{0,1\}$ = binary classification

Model

- Classifier: function f which takes X as input and returns C

- Goal: match $C = Y$ without discriminating based on A
  - At deployment time, we do not know Y. We are only given X.
  - But in the training data, we may have (partial) access to Y.

- Notation:
  - Evaluate f on a distribution over (X, Y)
  - $\Pr[C = c \mid A = a]$ = probability that a random individual with sensitive attribute value a receives classification outcome c
  - $\Pr[C = c \mid Y = y, A = a], \Pr[Y = y \mid C = c, A = a]$ etc. defined similarly

Confusion Matrix

		Predicted Value C	
		Positive	Negative
Target Value Y	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- False Positive Rate $FPR = \frac{FP}{FP+TN} = \Pr[C = 0 | Y = 0]$
- True Positive Rate $TPR = \frac{TP}{TP+FN} = \Pr[C = 1 | Y = 1]$
- Positive Predictive Value $PPV = \frac{TP}{TP+FP} = \Pr[Y = 1 | C = 1]$
- Intuitively, we want all three metrics to match across the two groups, but that is impossible! (except in special cases)
    - This isn't the impossibility result we'll see a proof of (but has a similar proof)

Kleinberg, Mullainathan, Raghavan.
"Inherent Trade-Offs in the Fair Determination of Risk Scores"

Chouldechova.
"Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments"

Recidivism

- Northpointe designed COMPAS, a tool to assess recidivism risk
- Judges in New York, Wisconsin, California, and Florida started using it when making bail decisions
- ProPublica published an article showing that COMPAS had wildly different FPR and TPR across white and black populations
- Northpointe argued back, suggesting that COMPAS was fair because PPV was nearly equal in both populations

Northpointe. [“Response to ProPublica: Demonstrating accuracy equity and predictive parity”]

ProPublica. [“Machine Bias”]

Three Fairness Notions

1. Statistical Parity
   - $\Pr[C = 1 \mid \mathrm{A} = 0] = \Pr[C = 1 \mid A = 1]$
     - Shorthand: $\Pr[C \mid A] = \Pr[C]$ or $C \perp A$
   - Equal rate of positive (or negative) prediction across the groups

2. Predictive Rate Parity
   - $\Pr[Y = y \mid C = c, A = 0] = \Pr[Y = y \mid C = c, A = 1], \forall y, c \in \{0,1\}$
     - Shorthand: $\Pr[Y \mid C, A] = \Pr[Y \mid C]$ or $Y \perp A \mid C$
   - Among those predicted positively (or negatively), equal truly positive (or negative) across the groups

3. Equalized Odds
   - $\Pr[C = c \mid Y = y, A = 0] = \Pr[C = c \mid Y = y, A = 1], \forall y, c \in \{0,1\}$
     - Shorthand: $\Pr[C \mid Y, A] = \Pr[C \mid Y]$ or $C \perp A \mid Y$
   - Among those truly positive (or negative), equal fractions predicted positively (or negatively) between the groups

Impossibility Theorems

- Theorem: Statistical parity + predictive rate parity are mutually incompatible unless $A \perp Y$

- Intuition: $A \perp C$ and $A \perp Y \mid C \Rightarrow A \perp Y$

- Proof:
  - $C \perp A$: $\Pr[C \mid A] = \Pr[C]$
  - $Y \perp A \mid C$: $\Pr[Y \mid A, C] = \Pr[Y \mid C]$
  - Combining:

$$\begin{aligned}
&\Pr[Y = y \mid A = a] \\
&= \Sigma_{c \in \{0,1\}} \Pr[C = c \mid A = a] \cdot \Pr[Y = y \mid A = a, C = c] \\
&= \Sigma_{c \in \{0,1\}} \Pr[C = c\,] \cdot \Pr[Y = y \mid C = c] \\
&= \Pr[Y = y]
\end{aligned}$$

Impossibility Theorems

- Theorem: Statistical parity + equalized odds are mutually incompatible unless $A \perp Y$ or $C \perp Y$ (lol)

- Intuition: $C \perp A$ and $C \perp A \mid Y \Rightarrow A \perp Y$ or $C \perp Y$

- Proof:

$$\begin{aligned}&\Pr[C=c]=\Pr[C=c|A=a]\\&=\Sigma_{y\in\{0,1\}}\Pr[Y=y|A=a]\cdot\Pr[C=c|Y=y,A=a]\\&=\Sigma_{y\in\{0,1\}}\Pr[Y=y|A=a]\cdot\Pr[C=c|Y=y]\end{aligned}$$

$$\Pr[C=c]=\Sigma_{y\in\{0,1\}}\Pr[Y=y]\cdot\Pr[C=c|Y=y]$$

- ➢ $\Sigma_{y\in\{0,1\}}\Pr[C=c|Y=y]\cdot(\Pr[Y=y|A=a]-\Pr[Y=y])=0$
- ➢ Second terms for $y=0$ and $y=1$ are negatives of each other!

Impossibility Theorems

- Theorem: Statistical parity + equalized odds are mutually incompatible unless $A \perp Y$ or $C \perp Y$ (lol)
- Intuition: $C \perp A$ and $C \perp A \mid Y \Rightarrow A \perp Y$ or $C \perp Y$
- Proof:
  - $\Sigma_{y \in \{0,1\}} \Pr[C = c | Y = y] \cdot (\Pr[Y = y | A = a] - \Pr[Y = y]) = 0$
  - Second terms for $y = 0$ and $y = 1$ are negatives of each other!
  - $(\Pr[C = c | Y = 0] - \Pr[C = c | Y = 1]) \cdot (\Pr[Y = 0 | A = a] - \Pr[Y = 0]) = 0$
  - $\Pr[C = c | Y = 0] = \Pr[C = c | Y = 1]$ $(C \perp Y)$
    or $\Pr[Y = 0 | A = a] = \Pr[Y = 0]$ $(Y \perp A)$

Impossibility Theorems

- Theorem: Predictive rate parity + equalized odds are mutually incompatible unless $A \perp Y$
- Intuition: $A \perp C \mid Y$ and $A \perp Y \mid C \Rightarrow A \perp (Y, C) \Rightarrow A \perp Y$
- Similar proof as before

Calibration

- Ideal: $X \to C \in \{0,1\}$
- In practice: $X \to R \in [0,1] \to C \in \{0,1\}$

- Calibration:
  - $\Pr[Y = 1 \mid R = r] = r, \forall r \in [0,1]$

- Calibration by group:
  - $\Pr[Y = 1 \mid R = r, A = a] = r, \forall r \in [0,1], a \in \{0,1\}$
  - Then, $\Pr[Y = 1 \mid R = r, A = a] = \Pr[Y = 1 \mid R = r]$
    - $Y \perp A \mid R$, which is predictive rate parity for score functions

Fairness-Accuracy Tradeoff(?)

- Example: Suppose $Y = 1 \Leftrightarrow A = 1$
  - Two groups are highly "unequal"
  - Statistical parity would require a classifier to deliberately return incorrect classifications

- Easy lower bound
  - $error \geq d_{TV}\big(D_0(Y), D_1(Y)\big)$, where $D_a(Y)$ is the distribution of Y conditioned on $A = a$

- The same does not hold for equalized odds
  - "The perfectly accurate classifier is fair"

Issues with such notions

1. Entitlement: What if the groups of interest are not equally entitled?
   - Equalized odds, can work with differing base rates, but what if the differences in entitlements go beyond just the difference in base rates?
   - E.g., what if you want to be “fair” to job applicants with different commuting distances?

2. Non-binary outcomes: What if the outcomes are not binary?
   - E.g., “no bail”, $500 bail”, “$1000 bail”, “$1500 bail”, …
   - Converting to binary might allow a classifier to “gerrymander” fairness by being fair in binary decisions while discriminating in the bail amount*

3. Preferences: Who are we being fair to? What do they want?
   - What if the end users (stakeholders) have heterogeneous preferences over non-binary outcomes?

*Arnold. “Racial Bias in Bail Decisions”

Mitigating the issues

- (Individual) Metric Fairness
  - Take as input a distance metric d over individuals
  - Classifier f maps an individual to a multi-dimensional vector (non-binary soft classification)
  - Example: (probabilistically) assigning students to job interviews
- "Treat equals equally"
  - Classification outcomes shouldn't differ much for individuals close to each other
  - $|f(x) - f(y)|_{TV} \leq L \cdot d(x, y), \forall x, y$
- Attempts solving two of the three problems
  - Different entitlements and non-binary outcomes
  - But this still ignores end user preferences!
    - Why must we assign two students with similar profiles the same job interviews if they prefer different jobs?
  - This is easy to address*, but still a less popular approach due to the difficulty of obtaining a reasonable metric in many practical applications

Dwork, Hardt, Pitassi, Reingold, Zemel. "Fairness Through Awareness"

*Kim, Korolova, Rothblum, Yonal. "Preference-Informed Fairness"

How do we take stakeholder preferences into account?

Envy-Freeness

- Envy-free classification
  - $u_i(f(x_i)) \geq u_i(f(x_k)) - \epsilon, \forall i, k$
  - "My value for my classification should be (almost) at least my value for your classification outcome"

- *Example application: Advertisement
  - It may not be unfair to show different ads to Bob than to Alice if each of them genuinely prefer seeing the ads they're shown to the ads the other person is shown.
  - Assumes equal entitlements

*Balcan, Dick, Noothigattu, Procaccia. "Envy-free Classification"

Envy-Freeness

- Envy-free classification
  - $u_i\big(f(x_i)\big) \geq u_i\big(f(x_k)\big) - \epsilon, \forall i, k$
  - "My value for my classification should be (almost) at least my value for your classification outcome"

- [+]Example application: Assigning students to job interviews
  - It may not be unfair to assign two students with similar profiles to different job interviews if they prefer the interviews that they're assigned to those that the other person is assigned
  - $u_i\big(f(x_i)\big) \geq u_i\big(f(x_k)\big) - L \cdot d(x_i, x_k)$
  - Extends envy-free classification to incorporate differing entitlements

[+]Kim, Korolova, Rothblum, Yona. "Preference-Informed Fairness"

Average Group Envy-Freeness

- "**On average,** individuals from one group shouldn't envy individuals from another group."
  - Weaker, not stronger, than individual envy-freeness

- $\mathbb{E}_{i \in G_1, k \in G_2}\left[u_i\big(f(x_i)\big) - u_i\big(f(x_k)\big)\right] \geq -\epsilon$

- Can be imposed over exponentially many given pairs of groups with only polynomial training sample complexity

- Example application: loan & bail decisions, where envy is unavoidable at the individual level

Hossain, Mladenovic, Shah. "Designing Fairly Fair Classifiers Via Economic Fairness Notions"