CS 270: Combinatorial Algorithms and Data Structures Spring 2023

Lecture 24 — April 13, 2023

Guest Lecturer: Michael Kim *Scribe: Alex Fu*

1 Overview

This lecture will be on more recent work in algorithms for machine learning, in particular questions of fairness in the context of prediction as well as multicalibration.

2 Motivating example

Suppose we are building a medical risk predictor, which takes as input an individual's health record and outputs a *predicted probability* (of some health outcome) in $[0, 1]$. The following cartoon, from a 2019 paper by Obermeyer et al. [1], illustrates an example of unfairness and miscalibration — the predicted probailities failed to accurately capture the actual probabilities. In order to be considered for advanced medical care, black patients needed to be much sicker than white patients to qualify.

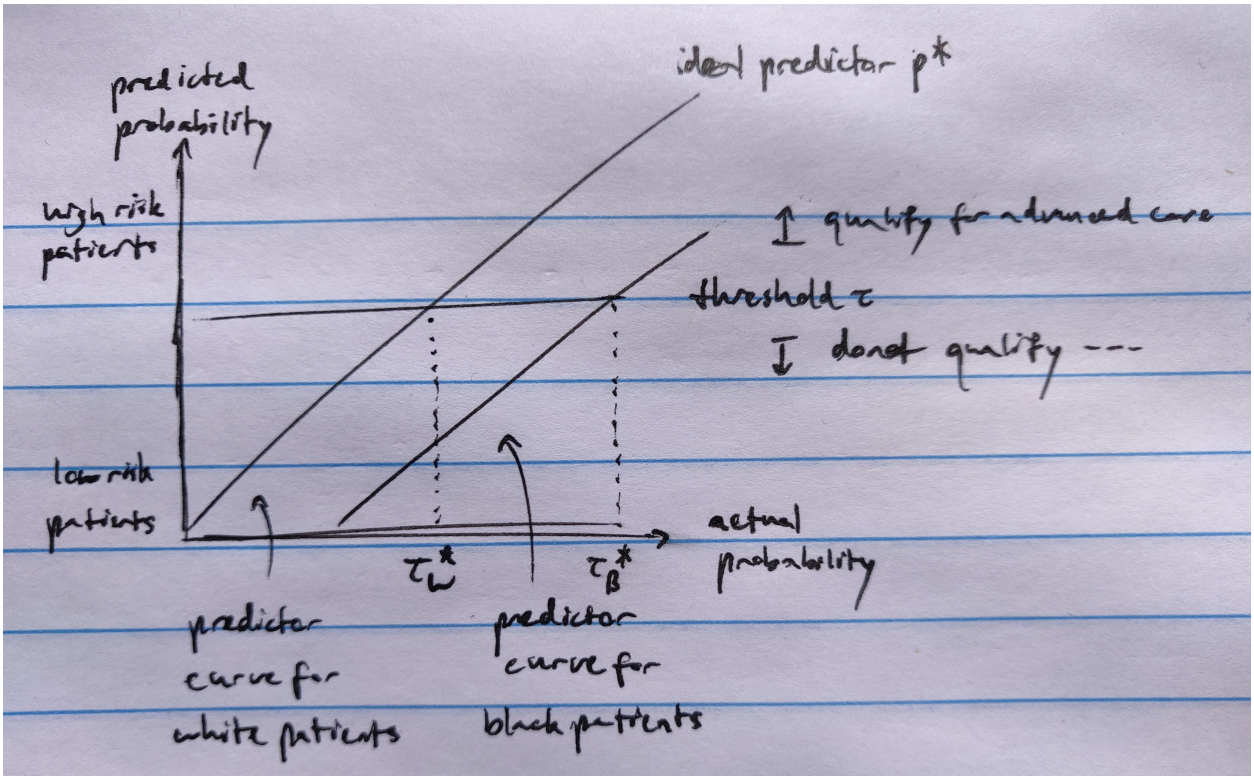

Figure 1: The predictor curves for different demographics of patients.

3 Formal setup

- A domain $X \subseteq \{0,1\}^d$, in which an individual $x \in X$ is given by d boolean features.
- An outcome space $Y = \{0,1\}$.
- A distribution $\mathcal{D}$ on $X \times Y$.
- A predictor $p\colon X \to [0,1]$.
- An optimal predictor $p^*(x) = \mathbb{P}_{\mathcal{D}}(y = 1 \mid X = x)$, the "actual probability" from the example.

We will assume that p^* has a distribution complicated enough that we cannot directly learn it to high precision. Typically, in supervised learning:

- Fix a hypothesis class of predictors $\mathcal{H} \subseteq \{h\colon X \to [0,1]\}$.
- *Goal*: find $\operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[(h(x) - y)^2]$.
- *Concern*: this framework only cares about minimizing a single objective, so it may have errors concentrated on important demographic groups and may bias towards larger populations.
- *New goal*: calibration.

Definition 3.1. A predictor $p\colon X \to [0,1]$ is **calibrated** if for all $v \in [0,1]$,

$$\mathop{\mathbb{P}}_{\mathcal{D}}(y = 1 \mid p(x) = v) \approx_{\varepsilon} v.$$

p is calibrated for $S \subseteq X$ if $\mathbb{P}_{\mathcal{D}}(y = 1 \mid p(x) = v, x \in S) \approx_{\varepsilon} v$. For a more precise definition, a predictor p is *α-calibrated* if for all $v \in [0,1]$,

$$\left| \mathop{\mathbb{E}}_{\mathcal{D}}[(y - v) \cdot \mathbb{1}\{p(x) = y\}] \right| \leq \alpha(\varepsilon, v).$$

Some motivation for Definition 3.1 comes from forecasting: for a calibrated meteorologist, 70% of the days where they predicted a 70% risk of rain should be rainy days.

We can phrase our objection to unfair predictors through calibration. For instance, the predictor from [1] is not well-calibrated for the subpopulation S of black patients: $\mathbb{E}[y \mid p(x) = v, x \in S] \gg v$.

Question 3.2. Is calibration enough?

Consider the predictor

$$p^\dagger(x) = \begin{cases} p^*(x) & x \notin S \\ \mu_S & x \in S, \end{cases}$$

where $\mu_S \coloneqq \mathbb{E}[y \mid x \in S]$ is the expected outcome of the group S. This is known as **algorithmic stereotyping**, where the algorithm overlooks the actual variation in the group and treats the group as a monolith.

Claim 3.3. $p^\dagger$ is 0-calibrated over S.

Proof. Note that the only supported value of $p^\dagger$ on S is μ_S.

$$\begin{aligned}\mathbb{E}[y \cdot \mathbb{1}\{p^\dagger(x) = \mu_S, x \in S\}] &= \mathbb{E}[y \cdot \mathbb{1}\{x \in S\}] \\ &= \mathbb{E}[y \mid x \in S] \cdot \mathbb{P}(x \in S) \\ &= \mu_S \cdot \mathbb{P}(x \in S) \\ &= \mathbb{E}[\mu_S \cdot \mathbb{1}\{p^\dagger(x) = \mu_S, x \in S\}].\end{aligned}$$

□

In calibration, we can get away with treating a group as a monolith. Thus, calibration may be a necessary condition for fairness in prediction, but it is not enough alone.

4 Multicalibration

- The individual-level p^* is out of reach.
- Group calibration is too weak.

We will propose something in the middle: **multicalibration**. We calibrate not just on the whole population or on protected groups, but on groups we could hope to identify in the hypothesis class.

Definition 4.1. Fix a collection of subpopulations $\mathcal{C} \subseteq 2^X$. A predictor $\tilde{p}$ is $(\mathcal{C}, \alpha)$**-calibrated** if for all $S \in \mathcal{C}$ and for all $v \in \text{support}(\tilde{p}) \subseteq [0, 1]$,

$$\left|\mathbb{E}[(y - \tilde{p}(x)) \cdot \mathbb{1}\{\tilde{p}(x) = v, x \in S\}]\right| \leq \alpha. \tag{1}$$

Question 4.2. For which $\mathcal{C}$ can we actually enforce calibration?

Answer: "**Computationally identifiable**" $\mathcal{C}$.

- Set membership should be efficient. That is, for all $S \in \mathcal{C}$, $\mathbb{1}\{x \in S\}$ should be able to be evaluated efficiently.
- Auditing should be possible. Given a predictor p, we should be able to answer the question

"Does there exist a subpopulation $S \in \mathcal{C}$ that violates Eq. (1)?"

5 MCBoost algorithm

The following *multicalibration boosting* or *HKRR* algorithm is an iterative learning algorithm originating from a paper by Hebert-Johnson, Kim, Reingold, and Rothblum [2].

Algorithm 5.1 (MCBoost)**.**

1. Initialize $p_0(x) = \frac{1}{2}$ for all $x \in X$.
2. Repeat for $t = 0, 1, 2, \ldots$:

3. If there exists $S \in \mathcal{C}$ and $v \in [0,1]$ such that $|\mathbb{E}[(y - p_t(x)) \cdot \mathbb{1}\{p_t(x) = v, x \in S\}]| > \alpha$:

4. Update $p_{t+1}(x) \leftarrow p_t(x) - \eta_t \cdot \mathbb{1}\{p_t(x) = v, x \in S\}$.

5. Else:

6. Return $\tilde{p} \coloneqq p_t$.

Claim 5.2. If MCBoost terminates, then $\tilde{p}$ is (C, α)-multicalibrated.

This is clear by Definition 4.1, so the real question is why MCBoost terminates:

- Number of iterations T.
- Complexity of auditing.

Claim 5.3. Suppose for all $S \in \mathcal{C}$, $\mathbb{1}\{x \in S\}$ is computable in time s. Then $\tilde{p}(x)$ is computable in time $O(T \cdot s)$.

Lemma 5.4. *MCBoost terminates in $T \leq O(\alpha^{-2})$ iterations.*

Proof. Define the potential function $\phi(p) \coloneqq \mathbb{E}[(p(x) - p^*(x))^2] \geq 0$. Because $\phi(p_0) = \frac{1}{4} \in O(1)$, it suffices to show that $\phi(p_t) - \phi(p_{t+1}) \geq \Omega(\alpha^2)$.

$$
\begin{aligned}
&\phi(p_t) - \phi(p_{t+1}) \\
&= \phi(p_t) - \mathbb{E}[((p_t(x) - y) - \eta_t \cdot \mathbb{1}\{p_t(x) = v, x \in S\})^2] \\
&= \cancel{\phi(p_t)} - \cancel{\phi(p_t)} - \eta_t^2 \cdot \mathbb{E}[\mathbb{1}\{p_t(x) = v, x \in S\}] + 2\eta_t \cdot \mathbb{E}[(p_t(x) - y) \cdot \mathbb{1}\{p_t(x) = v, x \in S\}] \\
&\geq -\alpha^2 + 2\alpha^2 \\
&= \alpha^2,
\end{aligned}
$$

where $|\mathbb{E}[(p_t(x) - y) \cdot \mathbb{1}\{p_t(x) = v, x \in S\}]| > \alpha$ by the auditing step, so we can choose $\eta_t \in \{\pm\alpha\}$ according to the sign of $\mathbb{E}[(p_t(x) - y) \cdot \mathbb{1}[p_t(x) = v, x \in S]]$. □

References

[1] Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[2] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. *Proceedings of the 35th International Conference on Machine Learning*, 80:1939–1948, 2018.