Multicalibration

Abdellah Aznag

Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

Today, we present...

Multicalibration: Calibration for the (Computationally-Identifiable) Masses

Úrsula Hébert-Johnson [1] **Michael P. Kim** [1] **Omer Reingold** [1] **Guy N. Rothblum** [2]

Motivation

- **Task:** Learning a predictor f on a population $\mathcal{X}$, based on samples i from a ground truth D.
- **Classification:** Each i holds a (random) boolean value with expectation p_i^*

- **Desirable properties:**
  - Given a particular subpopulation (in danger of being discriminated) $S \subset \mathcal{X}$, we want $E_{i\sim S} f = E_{i\sim S} p_i^*$ (No bias)
  - ...One problem: Might discriminate the groups with high variance.
  - One alternative: Consider the levels $L_v = \{i \mid f_i = v\}$: We want for all levels $v \in [0,1]$, $E_{i\sim L_v \cap S} f = E_{i\sim L_v \cap S} p_i^*$.

  This is the idea behind **calibration**.

- **Main takeaway:** Zero bias (or even low bias), is not enough to ensure "fairness" for a subpopulation. Enforcing zero bias across all the levels $L_v \cap S$ is better but (much) stronger.

Requirements

- Unbiased is unrealistic: instead, we want $(1) |E_{i \sim L_v \cap S}(f - p_i^*)| \leq \alpha$.
- We want **calibration** across multiple sets: We want it to be satisfied for a collection $\mathcal{C}$ of subpopulations.
- We want a **tractability**: Testing for all values $v \in [0,1]$ is unrealistic. We only test the levels v for a **discretized** version of $[0,1]$, of precision λ, denoted $\Lambda[0,1]$.
- We want **feasibility**: We can only hope for (1) to be true on a $(1-\alpha)$ −fraction of S.

Strongest definition of multicalibration

Definition (($\mathcal{C}, \alpha, \lambda$)-multicalibration). *Let* $\mathcal{C} \subseteq 2^{\mathcal{X}}$ *be a collection of subsets of* $\mathcal{X}$. *For any* $\alpha, \lambda > 0$, *a predictor* f *is* $(\mathcal{C}, \alpha, \lambda)$-multicalibrated *if for all* $S \in \mathcal{C}$, $v \in \Lambda[0,1]$, *and all categories* $S_v(f)$ *such that* $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S_v(f)] \geq \alpha\lambda \cdot \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S]$, *we have*

$$\mathop{\mathbf{E}}_{i \in S_v(f)}[f_i - p_i^*] \leq \alpha.$$

« Normal » definition of multicalibration

Definition (Calibration). *For any* $v \in [0,1]$, $S \subseteq \mathcal{X}$, *and predictor* f, *let* $S_v = \{i : f_i = v\}$. *For* $\alpha \in [0,1]$, f *is* α-calibrated *with respect to* S *if there exists some* $S' \subseteq S$ *with* $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S'] \geq (1-\alpha) \cdot \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S]$ *such that for all* $v \in [0,1]$,

$$\left| \mathop{\mathbf{E}}_{i \sim S_v \cap S'} [f_i - p_i^*] \right| \leq \alpha. \qquad (2)$$

Note: Normal multicalibration is just normal calibration on all the subpopulations of the collection

Overview of results:

- Is it possible? Answer: Yes! as long as we have a certificate of membership for each $S \in \mathcal{C}$
- Can it be done efficiently? Answer: Yes! as long as:
  - We have enough samples
  - The certificate of membership can be done efficiently
  - We have an implicit representation of $\mathcal{C}$
- Bonus: (γ: Ratio of samples from S)
  - Time complexity $\sim t|\mathcal{C}|\,\text{poly}\left(\frac{1}{\alpha}, \frac{1}{\gamma}\right)$
  - Sample complexity $\sim \log|\mathcal{C}|\,\text{poly}\left(\frac{1}{\alpha}, \frac{1}{\gamma}\right)$
- How much do we lose in accuracy? Very little!
- On a high level, multicalibration is as hard as Weak Agnostic Learning

How do we interact with data?

Definition (Guess-and-check oracle). *Let* $\tilde{q} : 2^{\mathcal{X}} \times [0,1] \times [0,1] \to [0,1] \cup \{\checkmark\}$. $\tilde{q}$ *is a* **guess-and-check oracle** *if for* $S \subseteq \mathcal{X}$ *with* $p_S = \mathbf{E}_{i \sim S}[p_i^*]$, $v \in [0,1]$, *and any* $\alpha > 0$, *the response to* $\tilde{q}(S, v, \omega)$ *satisfies the following conditions:*

- *if* $|p_S - v| < 2\omega$, *then* $\tilde{q}(S, v, \omega) = \checkmark$
- *if* $|p_S - v| > 4\omega$, *then* $\tilde{q}(S, v, \omega) \in [0,1]$
- *if* $\tilde{q}(S, v, \omega) \neq \checkmark$, *then*

$$p_S - \omega \leq \tilde{q}(S, v, \omega) \leq p_S + \omega.$$

Note: typo, replace α with ω

Intuition: We have an inner mechanism that

- Checks if a subpopulation is close to a given level
- Returns a more accurate level for that subpopulation

Bridge:

- Between Differential-Privacy and Adaptive Data Analysis

The Algorithm

Algorithm 1 – Learning a $(\mathcal{C}, \alpha)$-multicalibrated predictor

Let $\alpha, \lambda > 0$ and let $\mathcal{C} \subseteq 2^{\mathcal{X}}$.
Let $\tilde{q}(\cdot,\cdot,\cdot)$ be a guess-and-check oracle.

- Initialize: $f = (1/2, \ldots, 1/2) \in [0,1]^{\mathcal{X}}$
- Repeat:
  - ○ For each $S \in \mathcal{C}$ and $v \in \Lambda[0,1]$:
    - Let $S_v = S \cap \{i : f_i \in \lambda(v)\}$
    - if $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S_v] < \alpha\lambda \cdot \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S]$: **continue**
    - Let $\bar{v} = \mathbf{E}_{i \sim S_v}[f_i]$
    - Let $r = \tilde{q}(S_v, \bar{v}, \alpha/4)$
    - If $r \neq \checkmark$: **update** $f_i \leftarrow f_i + (r - \bar{v})$ for all $i \in S_v$ (project onto $[0,1]$ if necessary)
  - ○ If no S_v updated: **exit**
- For $v \in \Lambda[0,1]$:
  - ○ Let $\bar{v} = \mathbf{E}_{i \sim \lambda(v)}[f_i]$
  - ○ For $i \in \lambda(v)$: $f_i \leftarrow \bar{v}$
- Output f

Intuition: As long as we can find a potential uncalibrated, we

- Call the oracle. Effect: Either we check if it's calibrated, or:
- We gain a better understanding of the true levels on this part of the population
- Improve our predictor based on this new knowledge

Performance

Theorem 1. *Suppose* $\mathcal{C} \subseteq 2^{\mathcal{X}}$ *is collection of sets where for* $S \in \mathcal{C}$*, there is a circuit of size* s *that computes membership in* S *and* $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S] \geq \gamma$*. For any* $p^* : \mathcal{X} \to [0, 1]$*, there is a predictor that is* $(\mathcal{C}, \alpha)$*-multicalibrated implemented by a circuit of size* $O(s/\alpha^4\gamma)$.

Theorem 2. *Suppose* $\mathcal{C} \subseteq 2^{\mathcal{X}}$ *is collection of sets such that for all* $S \in \mathcal{C}$*,* $\mathbf{Pr}_{i \sim \mathcal{D}}[i \in S] \geq \gamma$*, and suppose set membership can be evaluated in time* t*. Then Algorithm 1 run with* $\lambda = \alpha$ *learns a predictor of* $f : \mathcal{X} \to [0, 1]$ *that is* $(\mathcal{C}, 2\alpha)$*-multicalibrated for* p^* *from* $O(\log(|\mathcal{C}|)/\alpha^{11/2}\gamma^{3/2})$ *samples in time* $O(|\mathcal{C}| \cdot t \cdot \mathrm{poly}(1/\alpha, 1/\gamma))$.

Best-in-class prediction

Theorem 5. *Suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a collection of subsets of $\mathcal{X}$ and $\mathcal{H}$ is a set of predictors. There is a predictor f that is α-multicalibrated on $\mathcal{C}$ such that*

$$\mathop{\mathbf{E}}_{i\sim\mathcal{X}}[(f_i - p_i^*)^2] - \mathop{\mathbf{E}}_{i\sim\mathcal{X}}[(h_i^* - p_i^*)^2] < 6\alpha,$$

where $h^* = \mathrm{argmin}_{h\in\mathcal{H}} \mathbf{E}_{i\sim\mathcal{X}}[(h-p^*)^2]$. Further, suppose that for all $S \in \mathcal{C}$, $\mathbf{Pr}_{i\sim\mathcal{D}}[i \in S] \geq \gamma$, and suppose that set membership for $S \in \mathcal{C}$ and $h \in \mathcal{H}$ are computable by circuits of size at most s; then f is computable by a circuit of size at most $O(s/\alpha^4\gamma)$.

Intuition:

- We calibrate $\mathcal{C}$ and the levels of all the elements in the set of predictors
- The prediction error increases by a small additive term

Multicalibration and WAL

- Informal definition of Weak Agnostic Learning:

  Given a concept class and a hypothesis class, an algorithm is a weak agnostic learner if, whenever there is a non-trivial concept that correlates with the data, the algorithm can produce a non-trivial hypothesis that also correlates with the data

- Informal result:

  Weak Agnostic Learning is as hard as Multicalibration. Meaning we can reduce one to the other.

  - If our collection admits a WAL, then we can construct a multicalibrated algorithm
  - If our collection has a multicalibrated predictor, then we can construct a WAL

Definition (Weak agnostic learner). *Let $\rho \geq \tau > 0$, $\mathcal{C} \subseteq 2^{\mathcal{X}}$, and $\mathcal{H} \subseteq [-1,1]^{\mathcal{X}}$. A (ρ, τ)-*weak agnostic learner *$\mathcal{L}$ for a concept class $\mathcal{C}$ with hypothesis class $\mathcal{H}$ solves the following promise problem: given a collection of labeled samples $\{(i, y_i)\}$ where $i \sim \mathcal{D}$ and $y_i \in [-1,1]$, if there is some $c \in \mathcal{C}$ such that $\mathbf{E}_{i \sim \mathcal{D}}[c_i \cdot y_i] > \rho$, then $\mathcal{L}$ returns some $h \in \mathcal{H}$ such that $\mathbf{E}_{i \sim \mathcal{D}}[h_i \cdot y_i] > \tau$.*

Theorem 3. *Let $\rho, \tau > 0$ and $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be some concept class. If $\mathcal{C}$ admits a (ρ, τ)-weak agnostic learner that runs in time $T(|\mathcal{C}|, \rho, \tau)$, then there is an algorithm that learns a predictor that is $(\mathcal{C}, \alpha)$-multicalibrated on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S] \geq \gamma\}$ in time $O(T(|\mathcal{C}|, \rho, \tau) \cdot \mathrm{poly}(1/\alpha, 1/\lambda, 1/\gamma))$ as long as $\rho \leq \alpha^2 \lambda \gamma / 2$ and $\tau = \mathrm{poly}(\alpha, \lambda, \gamma)$.*

Theorem 4. *Let $\alpha, \gamma > 0$ and suppose $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a concept class. If there is an algorithm for learning a $(\mathcal{C}', \alpha)$-multicalibrated predictor on $\mathcal{C}' = \{S \in \mathcal{C} : \mathbf{Pr}_{i \sim \mathcal{D}}[i \in S] \geq \gamma\}$ in time $T(|C|, \alpha, \gamma)$ then we can implement a (ρ, τ)-weak agnostic learner for $\mathcal{C}$ in time $O(T(|C|, \alpha, \gamma) \cdot \mathrm{poly}(1/\tau))$ for any $\rho, \tau > 0$ such that $\tau \leq \min\{\rho - 2\gamma, \rho/4 - 4\alpha\}$.*

Intuition: We have an inner mechanism that

- Checks if a subpopulation is close to a given level
- Returns a more accurate level for that subpopulation

Bridge:

- Between WAL and Fairness

In practice?

- Holistic problem:
  - Multicalibration vs Equalizing error rates
    - Tay-Sach disease in Askhenazi population
  - Fairness in data vs Fairness in outcome

Discussion