

Calibration and its many applications



Recall calibration

Mean Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

multicalibration of a predictor \bar{y} , $\forall v \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$

$$\mathbb{E}[\bar{y} - y \mid (x, y) \in S, \bar{y} = v] \approx 0$$

“A sequence of predictions \bar{y}_t is multicalibrated on a sequence of examples (x_t, y_t) with respect to a set of demographic groups G if for every $S \in G$ the predictions are calibrated on the subsequence:

$$\{(x_t, y_t) : x_t \in S\}”$$

Mean Consistency

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

- Mean Consistency

- Given a distribution P over $X \times [0,1]$ and a mean predictor $\bar{\mu}: X \rightarrow [0,1]$

- Given a set $S \subseteq X$ write:

$$\mu(S) = \mathbb{E}[y \mid x \in S] \quad \bar{\mu}(S) = \mathbb{E}[\bar{\mu}(x) \mid x \in S]$$

- A predictor μ is ϵ -mean consistent on a set S if:

$$\left| \mu(S) - \bar{\mu}(S) \right| \leq \frac{\epsilon}{\Pr[x \in S]}$$

Multicalibration (Rephrased)

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

- Calibration

- Given a discretization parameter m and a grid $\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1\}$.

- Write $\bar{\mu}(x) \in B(i)$ if $\left| \bar{\mu}(x) - \frac{i}{m} \right| \leq \frac{1}{2m}$

- Given $S \subseteq X$: $S(\bar{\mu}, i) = \{x \in S : \bar{\mu}(x) \in B(i)\}$

$\bar{\mu}: X \rightarrow [0,1]$ is ϵ -calibrated if for every $i \in [m]$:

$\bar{\mu}$ is ϵ -mean consistent on $X(\bar{\mu}, i)$

- Multi-calibration

- Given: An arbitrary collection of overlapping sets $G \subseteq 2^X$

$\bar{\mu}: X \rightarrow [0,1]$ is ϵ -multi-calibrated w.r.t. G if for every $i \in [m]$, $S \in G$:

$\bar{\mu}$ is ϵ -mean consistent on $S(\bar{\mu}, i)$

Many variants... and many uses!

To endow sequential predictions with uncertainty estimates without making assumptions about the data.

Both for mean and variance

To guarantee some level of robustness against covariate shift

A simple goal

To endow sequential predictions with uncertainty estimates without making assumptions about the data.

What do/should reported uncertainty estimates mean?

Given your features x , our model predicts your expected disease severity in two days time is $f(x)$.

How sure are you?

I have a 95% prediction interval that your severity will be in $[\ell(x), u(x)]$.

Hmmm...



What do/should reported probability estimates mean?

Ideally

$$f(x) = \mathbb{E}[y \mid x]$$

$$\Pr_y \left[y \in [\ell(x), u(x)] \mid x \right] = 0.95$$

Randomness is entirely over the unrealized/unmeasured randomness of the world, conditional on all of your observable attributes.

More likely

$$\text{Calibration: } f(x) = \mathbb{E}_{(x,y)} [y \mid f(x)]$$

$$\text{Marginal Coverage: } \Pr_{(x,y)} \left[y \in [\ell(x), u(x)] \right] = 0.95$$

Randomness is *averaging over people*.

What do/should reported probability estimates mean?

- True conditional expectations and prediction intervals are too strong in rich feature spaces.
 - If we have never seen x before we have no information at all about $y | x$.
- Standard Solutions:
 - Parametric assumptions. E.g. assume $\mathbb{E}[y | x] = \langle \theta, x \rangle$, form confidence regions around θ which translate into prediction intervals
 - If we believe the model..
 - From conditional to marginal guarantees.
 - Calibration, conformal prediction.

Marginal Guarantees.



$[\ell(x), u(x)]$ is a 95% marginal prediction interval.

But I'm part of a demographic group representing less than 5% of the population...

Marginal Guarantees.



What about for people like me?

For African Americans under the age of 50
the 95% prediction interval is [a,b]

For women with a family history of diabetes
the 95% prediction interval is [c,d]

What does this mean
for me?

For people with egg allergies and no history of
smoking, the 95% prediction interval is [e, f].

Distributional Assumptions

- Standard Assumption for Conformal Prediction:
 - Exchangeability (e.g. iid data): The future must look like the past.
- But this is often violated:
 - Covariate shift: e.g. as a disease moves through a population, the demographics of patients can change suddenly in unanticipated ways.
 - Label shift: e.g. as treatments improve, the distribution on y *conditional* on x can change.
 - Strategic effects: In e.g. lending, hiring, college admissions, people may manipulate their features to optimize for a deployed classifier, which may frequently be retrained.
 - Time series data: Predictions about correlated data --- e.g. disease severity by time.

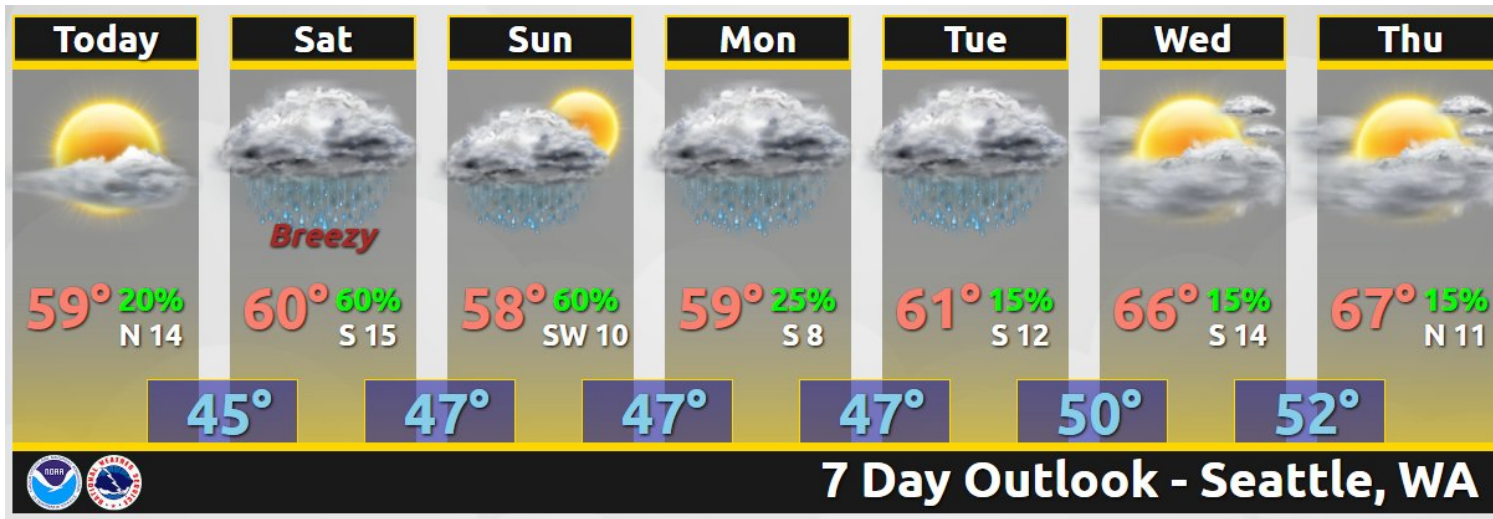
Our Goal

To mitigate both of these problems for sequential prediction:

1. By making *stronger than marginal* guarantees, and
2. Assuming *nothing* about the data

For prediction intervals... *and similar results for predicting means, variances, and higher moments of the label distribution*

Calibration



Mean Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

“A sequence of predictions \bar{y}_t is multicalibrated on a sequence of examples (x_t, y_t) with respect to a set of demographic groups G if for every $S \in G$ the predictions are calibrated on the subsequence:

$$\{(x_t, y_t) : x_t \in S\}”$$

Mean Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

- Mean Consistency

- Given: a sequence of examples $\left((x_1, y_1), \dots, (x_T, y_T) \right) \in \left(X \times [0,1] \right)^T$
and a sequence of predictions $(\bar{y}_1, \dots, \bar{y}_T) \in [0,1]^T$
- Given a set $S \subseteq X$ write:

$$\mu(S) = \sum_{t: x_t \in S} y_t \quad \bar{\mu}(S) = \sum_{t: x_t \in S} \bar{y}_t$$

A predictor μ is α mean-consistent on a set S if:

$$\left| \mu(S) - \bar{\mu}(S) \right| \leq \alpha T$$

Mean Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum '18]

- Multi-Calibration

- Given a discretization parameter m and a grid $\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1\}$.
- Write $\bar{y} \in B(i)$ if $\left| \bar{y} - \frac{i}{m} \right| \leq \frac{1}{2m}$
- Given $S \subseteq X$: $S(i) = \{x_t \in S : \bar{y}_t \in B(i)\}$

Given: An arbitrary collection of overlapping sets $G \subseteq 2^X$

$(\bar{y}_1, \dots, \bar{y}_T)$ is (α, m) -multicalibrated w.r.t. G if for every $i \in [m]$, $S \in G$:
it is α -mean consistent on $S(i)$

The Online Problem

- For $t = 1, \dots, T$
 - An adversary selects $x_t \in X$ and $y_t \in [0,1]$ and shows x_t to the learner.
 - The learner makes a (possibly randomized) prediction \bar{p}_t
(Of the mean, variance, a prediction interval, etc.)
 - The learner observes y_t .
- goal: prediction player guarantees multicalibration/
multi-validity in the worst-case over adversaries.

Mean Multicalibration

- For each time step c , group $S \in G$ and prediction bucket $i \in [m]$:

$$V_C^{S,i} = \sum_{t=1}^C 1[x_t \in S, \bar{y}_t \in B(i)] \cdot (y_t - \bar{y}_t)$$

(If $\left| \frac{1}{T} V_T^{S,i} \right| \leq \alpha$ for all $S \in G, i \in [m]$, we are (α, m) -multicalibrated.)

The Basic Idea

Taking inspiration from [Fudenberg and Levine 95]

- We want to bound $L_T = \max_{i,S} \left| V_T^{S,i} \right|$.
- This quantity depends on the whole history --- hard to grapple with.

Instead, try to greedily bound its increase after we see x_t :

$$\Delta_t(\bar{y}_t) = \mathbb{E} \left[L_t - L_{t-1} \mid x_t, \bar{y}_t \right]$$

(If we can bound this for every history, we can bound $E[L_T]$ by telescoping...)

An Interlude: Zero Sum Games

- A Zero Sum Game is defined by:

1. A *minimization player* (the learner) with finite strategy space A_1
2. A *maximization player* (the adversary) with finite strategy space A_2
3. A utility function $u: A_1 \times A_2 \rightarrow \mathbb{R}$.

Extended to distributions in the natural way. For $Q_1 \in \Delta A_1, Q_2 \in \Delta A_2$:

$$u(Q_1, Q_2) = \mathbb{E}[u(a_1, a_2)]$$

- Von Neumann's Minimax Theorem:

$$\min_{Q_1 \in \Delta A_1} \max_{a_2 \in A_2} u(Q_1, a_2) = \max_{Q_2 \in \Delta A_2} \min_{a_1 \in A_1} u(a_1, Q_2)$$

"Order of play doesn't matter"



The Basic Idea

Taking inspiration from [Fudenberg and Levine 95]

- Define a game between the learner and adversary with:

$$u(\bar{y}_t, y_t) = \Delta_t(\bar{y}_t)$$

- Bound the value of the game by imagining the adversary goes first.
 - (Clear the learner can do well: if the adversary shows you his distribution, predict its mean)
- Apply the minimax theorem to conclude the Learner can do just as well against a worst-case label.

(Non-constructive argument)

Upshot

Theorem: There exists an algorithm that for any set of groups G , and against any adversary, guarantees (α, m) -multicalibration for:

$$\alpha \leq (4 + \epsilon) \sqrt{\frac{2 \ln \left(\frac{2|G|m}{\delta} \right)}{T}}$$

With probability $1 - \delta$

And the algorithm?

Just compute the Minimax Equilibrium.

• For $t = 1$ to T :

• Compute $C_{t-1}^i(x_t) = \sum_{S \in G(x_t)} \exp(\eta V_{t-1}^{S,i}) - \exp(-\eta V_{t-1}^{S,i})$ for $i \in [m]$

• If $C_{t-1}^i(x_t) > 0$ for all i then predict $\bar{y}_t = 1$

• If $C_{t-1}^i(x_t) < 0$ for all i then predict $\bar{y}_t = 0$

• Otherwise:

• find i^* s.t. $C_{t-1}^{i^*}(x_t) \cdot C_{t-1}^{i^*+1}(x_t) \leq 0$

• Let $p_t \in [0,1]$ be s.t. $p_t \cdot C_{t-1}^{i^*}(x_t) + (1 - p_t) \cdot C_{t-1}^{i^*+1}(x_t) = 0$

• Predict $\bar{y}_t = \frac{i^*}{m} - \frac{1}{rm}$ with probability p_t , otherwise predict $\bar{y}_t = \frac{i^*}{m}$

Prediction Interval Multivalidity

“A sequence of 95%-prediction intervals $[\ell_t, u_t]$ is multivalid on a sequence of examples (x_t, y_t) with respect to a set of demographic groups G if for every $S \in G$ and for every interval $[\ell, u]$, the prediction intervals cover 95% of the labels in the set:

$$\{(x_t, y_t) : x_t \in S, [\ell_t, u_t] \approx [\ell, u]\}”$$

Similarly for Prediction Intervals

- We can invoke essentially the same arguments.
- If the adversary is forced to announce their label distribution, we can read off a prediction interval from its CDF.
 - *Assuming continuity
- A minimax argument gives the existence of an algorithm that does well.

The Algorithm

For $t = 1$ to T :

$$\text{Compute } C_{t-1}^{i,j}(x_t) = \sum_{S \in G(x_t)} \exp(\eta V_{t-1}^{S,(i,j)}) - \exp(-\eta V_{t-1}^{S,(i,j)}) \text{ for } i, j \in [m]$$

Using these, solve $LP(t)$ w Ellipsoid, separation oracle to obtain $Q_L^t \in \Delta A_1^{rm}$

Sample $(\ell, u) \sim Q_L^t$ and predict $(\ell_t, u_t) = (\ell, u)$

$LP(t)$:

Find $Q_L \in \Delta A_1^{rm}$ to minimize γ

Such that for every (ρ, rm) -smooth

distribution $Q_A \in A_2^{\rho, rm}$:

$$\mathbb{E}_{(\ell, u) \sim Q_L} \left[\mathbb{E}_{y \sim Q_A} \left[(1[y \in [\ell, u]] - (1 - \delta)) C_{t-1}^{i,j}(x_t) \right] \right] \leq \gamma$$

Moment Multicalibration

- Mean Calibration:

“The average label amongst all points for which we predicted mean $\frac{i}{m}$ should be $\frac{i}{m}$.”

- Moment Calibration:

“The variance on the set of points for which we predicted variance $\frac{i}{m}$ should be $\frac{i}{m}$ ”?

- No! Problem: this isn't feasible/desirable.

- We know mean multicalibration is attainable because $\bar{\mu}(x) = \mathbb{E}[y \mid x]$ obtains it for every G .

- But $\bar{m}_k(x) = \mathbb{E}\left[\left(y - \mathbb{E}[y \mid x]\right)^k \mid x\right]$ does not satisfy this moment condition.

- Consider $P = \{(x_1, 0), (x_2, 1)\}$...

Moment Multicalibration

The problem: higher moments don't combine linearly under mixtures

Observation *Let \mathcal{P} be a mixture distribution over m component distributions \mathcal{P}_ℓ with mixture weights $w_\ell \geq 0$, $\sum_{\ell=1}^m w_\ell = 1$. Let $\mu_\ell, m_{k\ell}$ be the mean and k^{th} moment associated with \mathcal{P}_ℓ . Then:*

$$m_k = \sum_{\ell=1}^m w_\ell \left(\sum_{a=0}^k \binom{k}{a} (\mu_\ell - \mu)^{k-a} m_{a\ell} \right).$$

If the mixture variables have the same mean, i.e. $\mu_\ell = \mu$ for all ℓ , then, the above expression reduces to:

$$m_k = \sum_{\ell=1}^m w_\ell m_{k\ell}.$$

Mean Conditioned Moment Multicalibration

- Fix a moment predictor $\bar{m}_k: X \rightarrow [0,1]$.

$$m_k(\mathcal{S}) = \mathbb{E}\left[(y - \mu(\mathcal{S}))^k \mid x \in \mathcal{S}\right] \quad \bar{m}_k(\mathcal{S}) = \mathbb{E}[\bar{m}_k(x) \mid x \in \mathcal{S}]$$

- A predictor is ϵ -moment consistent on a set \mathcal{S} if:

$$\left| m_k(\mathcal{S}) - \bar{m}_k(\mathcal{S}) \right| \leq \frac{\epsilon}{\Pr[x \in \mathcal{S}]}$$

Mean Conditioned Moment Multicalibration

- Let $S(\bar{\mu}, i, \bar{m}_k, j) = \{x \in S : \bar{\mu}(x) \in B(i), \bar{m}_k(x) \in B(j)\}$

Definition: A pair of predictors $(\bar{\mu}, \bar{m}_k)$ are ϵ -mean-conditioned moment-multicalibrated on a collection of sets G if for every $S \in G$ and for every $i, j \in [m]$:

1. $\bar{\mu}$ is ϵ -mean consistent on $S(\bar{\mu}, i, \bar{m}_k, j)$
2. \bar{m}_k is ϵ -moment consistent on $S(\bar{\mu}, i, \bar{m}_k, j)$.


Observe this is feasible: satisfied by the true distributional quantities.

Mean Conditioned Moment Multicalibration

- What does it mean?

Amongst all people who received the same prediction, their true (average) dosage was indeed \hat{v} , and the variance was $\hat{\sigma}$

And I can interpret this at my option as an average over any demographic group in G of which I am a member.



Your mean ideal dosage is \hat{v} and the variance is $\hat{\sigma}$.

Mean Conditioned Moment Multicalibration

- If you had real distributional moments, you'd get conditional prediction intervals for all x :

$$\Pr_y \left[y \in \left[\mu(x) - \left(\frac{m_k(x)}{\delta} \right)^{\frac{1}{k}}, \mu(x) + \left(\frac{m_k(x)}{\delta} \right)^{\frac{1}{k}} \right] \mid x \right] \geq 1 - \delta$$

- Mean conditioned moment multicalibrated estimates give you marginal prediction intervals simultaneously over all $S \in G, i, j \in [m]$:

$$\Pr_{(x,y)} \left[y \in \left[\bar{\mu}(x) - \left(\frac{\bar{m}_k(x)}{\delta} \right)^{\frac{1}{k}}, \bar{\mu}(x) + \left(\frac{\bar{m}_k(x)}{\delta} \right)^{\frac{1}{k}} \right] \mid x \in S(\bar{\mu}, i, \bar{m}_k, j) \right] \geq 1 - \delta$$

What about covariate shift?

Given a bunch of training data from one distribution D_s ,
how to design something which one can easily
transform to work well on a different distribution D_t
over X ?

Remember propensity weights?

$\Pr[X = x \mid D_s]$ versus $\Pr[X = x \mid D_t]$?

What about covariate shift?

“standard” fixes: assume a parametric model of e , or that it belongs to some class Σ

Let data be triples (x, y, z)

Suppose we are interested in the average value of y for our target, and we have

$$e_{st}(x) := \mathbb{P}[Z = s \mid x]$$

$$\mu_t^* = \mathbb{E}_{x, y \sim D_s} \left[\frac{1 - e_{st}(x)}{e_{st}(x)} y \right]$$

Error will come from how inaccurate our propensity weights are (both representation + generalization error)

The above will work well when given unlabeled data from both source and target, if there's enough data to estimate e .

But, this is assuming we have enough data to compute relative weight for both source and target.

Propensity scores + calibration

Suppose we had a predictor p of y which was multiaccurate w.r.t.

$$C(\Sigma) = \left\{ \frac{1 - \sigma(x)}{\sigma(x)} \mid \sigma \in \Sigma \right\}$$

for some distribution D :

$$\mathbb{E}[c(x)[p(x) - y] \mid (x, y) \sim D] \leq \alpha.$$

Then,

$\mathbb{E}_{(x,y) \sim D_t}[p(x)]$ is a good estimate of y on D_t

Some references

Omniprediction

Universal adaptability

Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold

Online Multivald Learning: Means, Moments, and Prediction Intervals

Varun Gupta, Chris Jung, George Noarov, Mallesh Pai, Aaron Roth

Moment Multicalibration for Uncertainty Estimation.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra

Taking a step back



Standard ML perspective

Fixed Dataset
(or distribution)

Fixed
Objective

Primary question:
how to optimize
objective subject
to some
constraints?

Standard ML perspective

Equitable

questions and checklist

Set of measurements X

Fixed Dataset
(or distribution)

- What set of measurements are you gathering?
- Are there differences in error for measuring/operationalizing the constructs these measurements capture?
- Does the distribution match your test distribution in demographics?

Labels Y

- What is it you are trying to predict?
- Do you actually observe Y all of the time?
- Is your target a proxy for what you actually observe?
- Does your dataset have the “right” amount of data from each population for each label?
- Is there strong enough correlation between X and Y for every population you care about?

Issues of sampling/discretization/preprocessing

- Optimal sampling for minimizing global loss s.t. some statement about population loss isn't iid.
- data processing will have different impact on different populations

Standard ML perspective

Equitable

questions and checklist

Fixed Dataset
(or distribution)

What happens if this distribution shifts?

- As a result of our predictions?
- Critical to take into account both
 - whether our predictions affect what we observe
 - and whether they actually change the ground truth
 - If so, how?

Standard ML perspective

Equitable

questions and checklist

- How did we choose this objective?
- Is it giving similar importance to performance on different demographic groups?
- This choice may have come from computational necessity, possibly at the expense of an objective which more accurately aligns with our real goals and values
- How are we handling uncertainty? Do we allow our models to present their level of uncertainty when they provide predictions?

Fixed
Objective

Standard ML perspective

Equitable

questions and checklist

- How do we audit/test for discriminatory models?
 - Do we have observational or interventional power?
 - Perhaps the correct intervention is not on demographics, but on discrimination due to demographics.
- Who was involved in making all of these choices?
- What are intended and possible unintended impacts of this model and its predictions?
- How should we intervene when we see models behaving in a possibly discriminatory manner?
- What variables do we consider resolving?

Primary question:
how do all of these choices impact the distribution of our predictions and errors across demographics?

Standard ML perspective

Equitable

questions and checklist

The broader ecosystem

How will this model be used?

Decision/action

