CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 19: Predictions and Calibration

Nov 4th

Lecturer: Nika Haghtalab *Readings: Foster and Vohra [1998], Hart [2025]*
Scribe: Suho Kang, Tonghun Lee, Jennifer Zhao

1 Calibrated Forecasting

We've spent the last modules in the adversarial world – online learning with regret guarantees, bandits under partial feedback, and zero-/general-sum games (including Stackelberg games and swap-regret dynamics). The common lens was to pick actions so that, against any sequence of outcomes our cumulative loss is close to the best comparator in hindsight or to an equilibrium value. Today we pivot from choosing actions to judging predictions.

1.1 Motivation

The motivating question is "How should we evaluate a probabilistic weather forecaster"? Suppose the outcome of the weather is either sunny or rainy and the forecaster wants to predict the chance of rain for tomorrow. If the forecaster predicts a 100% chance of rain, then assigning an error of 0 when it actually rains and 1 when it is sunny seems reasonable.

However, if the forecaster instead predicts a 30% chance of rain, and it does indeed rain, what error should we assign to the forecaster? More generally, how should we evaluate forecasters that communicate their uncertainty about the world events and how should we account for the underlying uncertainty in the world.

One might consider evaluating a forecaster by averaging his predictions over the course of a year. One might at least hope for internal consistency in the prediction of the forecaster. For example, if, when averaging on all days where the predictor predicts a 0.3 chance of rain, significantly more or significantly less than 30% of those days are rainy, then the forecaster's predictions are not reliable. This notion of reliability of forecast, and whether it's possible to produce reliable forecasts, is the subject of our study.

1.2 Evaluating Predictions

We have motivated the need for a more principled measure for evaluating forecasts. Before stating the definition of calibration error, we provide the following minimum requirement.

Definition 1.1 (Calibrated Forecasts [Foster & Vohra]). Let $c_1, c_2, \ldots, c_T \in [0, 1]$ be sequence of forecasts and let $a_1, a_2, \ldots, a_T \in \{0, 1\}$ be actual realizations. Then we say $c_{1:T}$ is ϵ-calibrated

with respect to $a_{1:T}$ if

$$\lim_{T\to\infty} \underbrace{\sum_{p\in[0,1]} \frac{n_T(p)}{T} |\bar{a}_T(p) - p|}_{\text{(Expected) Calibration Error}} \leq \epsilon,$$

where

$$n_T(p) = \sum_{t=1}^{T} \mathbf{1}\{c_t = p\}, \quad \text{and} \quad \bar{a}_T(p) = \frac{1}{n_T(p)} \sum_{t=1}^{T} \mathbf{1}\{c_t = p\} \cdot a_t.$$

In other words, $\bar{a}_T(p)$ is the empirical realized distribution.

1.3 Example: Weather Forecasting

	t					
	1	2	3	4	5	6
a_t	rain	sunny	rain	rain	sunny	rain
c_t	0.75	0.50	0.75	0.50	0.75	0.75

Table 1: Forecast with zero calibration error

The calibration error is as follows:

$$\begin{aligned}
\text{Cal Error} &= \frac{n_T(0.5)}{T}|\bar{a}(0.5) - 0.5| + \frac{n_T(0.75)}{T}|\bar{a}(0.75) - 0.75| \\
&= \frac{2}{6}|0.5 - 0.5| + \frac{4}{6}|0.75 - 0.75| \\
&= 0
\end{aligned}$$

	t					
	1	2	3	4	5	6
a_t	rain	sunny	rain	rain	sunny	rain
c_t	0.8	0.4	0.8	0.4	0.8	0.8

Table 2: Forecast with nonzero calibration error

The calibration error is as follows:

$$\begin{aligned}\text{Cal Error} &= \frac{n_T(0.4)}{T}|\bar{a}(0.4) - 0.4| + \frac{n_T(0.8)}{T}|\bar{a}(0.8) - 0.8| \\ &= \frac{2}{6}|0.5 - 0.4| + \frac{4}{6}|0.75 - 0.8| \\ &\approx 0.66\end{aligned}$$

2 Sequential Calibration even against Adversaries

The setting is as follows. It involves T sequential rounds between a forecasting algorithm (denoted as ALG) and an adversary (denoted as ADV). In this process, we have a sequence of forecasts $c_1, \ldots, c_T$, where each $c_t \in [0, 1]$, and corresponding realization $a_1, \ldots, a_T$ with each $a_t \in \{0, 1\}$.

For each round $t \in \{1, \ldots, T\}$, based on the complete history of the forecasts and realizations, denoted as $h_{t-1} = (c_1, a_1; \ldots, c_{t-1}, a_{t-1};)$, the forecasting algorithm picks (possibly non-deterministically) its prediction c_t. Similarly, the adversary picks the realization a_t, knowing the history h_{t-1} and ALG. The forecasting algorithm's goal is to select its forecasts/predictions to minimize the overall expected calibration error across all T rounds, while the adversary's goal is to choose the realizations to maximize this error, making the forecasting algorithm perform as poorly as possible. Our goal is to prove the following theorem:

Theorem 2.1. *Even when a_t's are generated by an adaptive adversary who knows the forecaster's algorithm and history, for every $\epsilon > 0$, there exists a forecasting algorithm that is $\epsilon-$calibrated.*

For a fixed $\epsilon > 0$, let $N = \mathcal{O}(\frac{1}{\epsilon})$, and we restrict the algorithm (ALG) to making forecasts only from the discretized set of numbers, $D = \{\frac{1}{2N}, \ldots, \frac{2N-1}{2N}\}$. By imposing this restriction, we can view the interaction between the algorithm and the adversary (ADV) as a T-stage game with finite action sets. This framing allows us to directly apply the Minimax Theorem. Now, even under this restriction on ALG, to prove Theorem 2.1, it is sufficient to prove the following:

Theorem 2.2. *Let $T \geq N^3$. There exists a randomized algorithm for the forecaster that guarantees that $\mathbb{E}[K_T] \leq \frac{1}{N}$ against any mixed strategy employed by the adversary.*

Theorem 2.2 can be viewed as the following minimax problem:

$$\min_{\text{ALG}} \max_{\substack{\text{random} \\ \text{sequence}}} K_T \leq v$$

By the Minimax Theorem, it is sufficient to the following "dual" version of the above theorem:

Theorem 2.3. *Let $T \geq N^3$. For every mixed strategy of the adversary, there exists a strategy of the forecaster that guarantees $\mathbb{E}[K_T] \leq \frac{1}{N}$. For $N = \frac{1}{\epsilon}$, the calibration error is of order $T^{\frac{2}{3}}$.*

This is because Theorem 2.3 can be viewed as the following maximin problem:

$$\max_{\substack{\text{random} \\ \text{sequence}}} \min_{\text{ALG}} K_T \leq v$$

The intuition is that we can view this as a game over a finite T rounds between the forecaster and the adversary. At each round t, the forecaster chooses an action based on the full history and the adversary chooses an action based on the same history. Both have perfect recall of past, so strategies are history dependent. Thus, a mixed strategy of the forecaster is a randomized algorithm (or a distribution over history-dependent mappings to the next prediction) and a mixed strategy of the adversary is a distribution over all deterministic history-dependent adversarial choices (distribution over sequences).

In the minimax setting (Theorem 2.2), where the sequence is randomized, we seek a randomized algorithm that minimizes K_T. On the other hand, in the maximin setting (Theorem 2.3), it suffices to consider a deterministic algorithm against a distribution over sequences. This is because the minimizer (a randomized algorithm) can be regarded as a convex combination of deterministic algorithms, and therefore, without loss of generality, we may assume the minimizer itself is deterministic.

2.1 Proof of Theorem 2.3

Let's recall that for each $d \in D$, we define

$$n_T(d) := \sum_{t=1}^{T} \mathbf{1}\left(c_t = d\right),$$

and

$$\bar{a}_T(d) := \frac{1}{n_T(d)} \sum_{t=1}^{T} \mathbf{1}\left(c_t = d\right) a_t.$$

Given a mixed strategy of the adversary, the probability of the realization being 1 in round t, conditioned on the past history h_{t-1}, can be computed as $p_t = \Pr\left[a_t = 1 | h_{t-1}\right]$. The forecasting algorithm (ALG) then sets its prediction c_t by taking this probability p_t and rounding it to the nearest value within the discretized set D. By design, this rounding ensures that two conditions are met: $c_t \in D$, and $|c_t - p_t| \leq \frac{1}{2N}$.

The main challenges of bounding K_T is from (i) discretization error, and (ii) error by randomness (variance) in the sequence of outcomes. For each $d \in D$, we denote the impact of error by randomness as

$$G(d) := n(d)(\bar{a}_T(d) - d) = \sum_{t=1}^{T} \mathbf{1}(c_t = d)(a_t - c_t).$$

Note that each $G(d)$ is an unsigned term, and we can rewrite $K_T = \frac{1}{T} \sum_{d \in D} |G(d)|$. To isolate the impact of discretization, for each $d \in D$, we define

$$\tilde{G}(d) := \sum_{t=1}^{T} \mathbf{1}(c_t = d)(a_t - p_t),$$

and

$$\tilde{K}_T = \frac{1}{T} \sum_{d \in D} |\tilde{G}(d)|.$$

First, we bound the discretization error.

Claim 2.4. $|K_T - \tilde{K}_T| \leq \frac{1}{2N}$.

Proof. $K_T - \tilde{K}_T \leq \sum_{t=1}^{T} \frac{1}{T}|p_t - c_t| \leq \sum_{t=1}^{T} \frac{1}{T}\frac{1}{2N} = \frac{1}{2N}$. □

Now, all we have to bound is the impact of randomness. We bound it by bounding the variance of $\tilde{G}(d)$.

Claim 2.5. *For each* $d \in D$, $\mathbb{E}[\tilde{G}(d)^2] \leq \frac{1}{4}\mathbb{E}[n(d)]$.

Proof. Let $Z_t = a_t - p_t$, which is a random variable with $\mathbb{E}[Z_t|h_{t-1}] = 0$. Since a_t is a Bernoulli variable, we have $\mathbb{E}[Z_t^2|h_{t-1}] \leq \frac{1}{4}$. For any $s < t$, since $\mathbb{E}[Z_s Z_t|h_{t-1}] = Z_s\mathbb{E}[Z_t|h_{t-1}] = 0$, we have

$$\mathbb{E}[\tilde{G}(d)^2] = \mathbb{E}[\sum_{t=1}^{T} \mathbf{1}(c_t = d)Z_t^2] \leq \frac{1}{4}\mathbb{E}[n_T(d)].$$

□

Claim 2.6. $\mathbb{E}[\tilde{K}_T] \leq \frac{1}{2}\sqrt{\frac{N}{T}}$.

Proof. By Jensen's inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}[\tilde{K}_T] &= \frac{1}{T}\sum_{d \in D} \mathbb{E}[|\tilde{G}(d)|] \\
&\leq \frac{1}{T}\sum_{d \in D} \sqrt{\mathbb{E}[\tilde{G}(d)^2]} \quad \text{(Jensen's Inequality)} \\
&\leq \frac{1}{T} \cdot \frac{1}{2} \cdot \sum_{d \in D} \sqrt{\mathbb{E}[n_T(d)]} \quad \text{(Claim 3.2)} \\
&\leq \frac{1}{2T}\sqrt{T}\sqrt{\left(\sum_{d \in D} \mathbb{E}[n_T(d)]\right)} \quad \text{(Cauchy-Schwarz Inequality)} \\
&= \frac{1}{2}\sqrt{\frac{N}{T}}.
\end{aligned}$$

□

Now, by Claim 3.1 and Claim 3.3, we have $|K_T| \leq |K_T - \tilde{K}_T| + |\tilde{K}_T| \leq \frac{1}{2N} + \frac{1}{2}\sqrt{\frac{N}{T}} \leq \frac{1}{N} < \epsilon$, proving Theorem 2.3. □

References

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Sergiu Hart. Calibrated forecasts: The minimax proof. In *Matching, Dynamics and Games for the Allocation of Resources: Essays in Celebration of David Gale's 100th Birthday*, pages 153–159. Springer, 2025.