Notes on Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control

Catherine Chen cyc2152

December 2023

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control." arXiv, Sep. 29, 2022. Accessed: Dec. 18, 2023. [Online]. Available: http://arxiv.org/abs/2110.01052

Abstract: Learn Then Test (LTT) reframes risk-control as multiple hypothesis testing, to produce finite-sample guarantess on any predictive model, without assumptions on the model or true distribution of the underlying dataset.

1 Introduction

In LTT, begin with a learned model $\hat{f}$, then post-process the model using calibration data to make the final predictions. The post-processing is controlled by a low-dimensional parameter λ. Multiple values of the parameter are tested using the calibration data in order to find settings that control a user-chosen statistical error rate.

Conformal prediction, and risk-controlling prediction sets requires that λ is one-dimensional, an that the risk function is monotonic in λ. LTT does not require such assumptions, thus can control possibly non-monotonic risks.

1.1 Setting and Notation

Let $(X_i, Y_i)_{i=1,\ldots,n}$ be the calibration set, an i.i.d. set of variables, s.t. feature vectors $X_i \in \mathcal{X}$ and responses $Y_i \in \mathcal{Y}$, with pretrained machine learning model $\hat{f} : \mathcal{X} \mapsto \mathcal{Z}$. The raw model outputs in $\mathcal{Z}$ are post-processed to generate predictions $\mathcal{T}_\lambda(x)$ indexed by a low-dimensional parameter λ. Finally, $\hat{\lambda}$is determined by controlling a user-chosen error rate, independent of the quality of $\hat{f}$ or the data distribution.

In the general framework, post-processing $\mathcal{T}_\lambda : \mathcal{X} \to \mathcal{Y}'$ take on values in any space $\mathcal{Y}'$. In practice, $\mathcal{Y}' = \mathcal{Y}$ for predictions, or $\mathcal{Y}' = 2^{\mathcal{Y}}$ for prediction sets. For $\mathcal{T}_\lambda$, the risk $R\left(\mathcal{T}_\lambda\right) \in \mathbb{R}$, denoted $R(\lambda)$, is defined to capture a problem-specific notion of the statistical error.

Objective: Train a function $\mathcal{T}_{\hat{\lambda}}$ based on $\hat{f}$ and the calibration data s.t. it achieves the following error-control property:

Definition 1 (Risk-controlling prediction)**.** Let $\hat{\lambda} \in \Lambda$ be a random variable. We say that $\mathcal{T}_{\hat{\lambda}}$ is an (α, δ)-risk-controlling prediction (RCP) if $\mathbb{P}\left(R\left(\mathcal{T}_{\hat{\lambda}}\right) \leq \alpha\right) \geq 1 - \delta$.

The risk tolerance α and error level δ are chosen by the user. $\hat{\lambda}$ is a function of the calibration data, so the probability in the above definition will be over the randomness in the sampling of $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)$.

2 Risk Control in Prediction

Goal: Find a function $\mathcal{T}_{\hat{\lambda}}$ whose risk is less than some user-specified threshold α.

Algorithm Outline: Search across the collection of functions $\{\mathcal{T}_\lambda\}_{\lambda\in\Lambda}$ and estimate their risk on the calibration data $(X_i, Y_i)_{i=1,\ldots,n}$. The output of the procedure will be a set of λ values, $\widehat{\Lambda} \subseteq \Lambda$ which are all guaranteed to control the risk, $R(\lambda)$.

1. For each λ_j in a discrete set $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$, define the null hypothesis $\mathcal{H}_j : R(\lambda_j) > \alpha$. Thus, rejecting $\mathcal{H}_j$ corresponds to selecting λ_j as a point where the risk is controlled.
2. For each null hypothesis, compute a finite-sample valid p-value using a concentration inequality.
3. Return $\widehat{\Lambda} = \mathcal{A}\left(\{p_j\}_{j\in\{1,\ldots,|\Lambda|\}}\right) \subset \Lambda$, where $\mathcal{A}$ is an algorithm that controls the family-wise error rate (FWER).

Result: Except with probability δ, each $\hat{\lambda} \in \widehat{\Lambda}$ yields an RCP $\mathcal{T}_{\hat{\lambda}}$.

Theorem 1. *Suppose p_j has a distribution stochastically dominating the uniform distribution for all j under $\mathcal{H}_j$. Let $\mathcal{A}$ be an FWER-controlling algorithm at level δ. Then $\hat{\Lambda} = \mathcal{A}(p_1, \ldots, p_N)$ satisfies the following:*

$$\mathbb{P}\left(\sup_{\lambda\in\widehat{\Lambda}}\{R(\lambda)\} \leq \alpha\right) \geq 1 - \delta,$$

where the supremum over an empty set is defined as $-\infty$. Thus, selecting any $\lambda \in \widehat{\Lambda}, \mathcal{T}_\lambda$ is an $(\alpha, \delta) - RCP$.

Theorem 1 reduces the problem of risk control into two subproblems:

1. Generate a p-value for each hypothesis.
2. Combine the hypotheses to discover the least conservative prediction that controls the risk at level α.

Result: any FWER-controlling procedure can be used to find Λ, then pick any $\lambda \in \Lambda$ as the chosen RCP. In the FDR case, choose $\hat{\lambda} = \min\widehat{\Lambda}$, which yields the most discoveries, thus the lowest FNR.

2.1 Calculating Valid p-Values

Calculate a valid p-value p_j for each null hypothesis $\mathcal{H}_j$, i.e., one satisfying $u \in [0, 1], \mathbb{P}(p_j \leq u) \leq u$ under $\mathcal{H}_j$.

Idea: Calculate the empirical risk of $\mathcal{T}_{\lambda_j}$ for each j then use a concentration inequality to get the p-value for $\mathcal{H}_j : R(\lambda_j) > \alpha$. If $\mathcal{H}_j$ is rejected, this implies the risk is controlled.

Define the risk function as the expectation of a loss function $L : R(\mathcal{T}) = \mathbb{E}[L(\mathcal{T}(X), Y)]$.

Consider the bounded case where $L(\mathcal{T}(X), Y) \in [0, 1]$, apply the hybridized Hoeffding-Bentkus (HB) inequality, which uses the empirical risk on the calibration set, $\widehat{R}_j = \frac{1}{n}\sum_{i=1}^{n} L\left(T_{\lambda_j}(X_i), Y_i\right)$, as the test statistic.

Proposition 1 (Hoeffding-Bentkus inequality p-values)**.** *The following is a valid p-value for $\mathcal{H}_j$:*

$$p_j^{\mathrm{HB}} = \min\left(\exp\left\{-nh_1\left(\widehat{R}_j \wedge \alpha, \alpha\right)\right\}, e\mathbb{P}\left(\mathrm{Bin}(n, \alpha) \leq \left\lceil n\widehat{R}_j \right\rceil\right)\right),$$

where $h_1(a, b) = a\log(\frac{a}{b}) + (1 - a)\log(\frac{1-a}{1-b})$.

Note: In the unbounded case, the HB inequality no longer applies, but asymptotically valid p-values can be obtained from the CLT.

2.2 Multiple Hypothesis Testing

Combine p-values to form the rejection set $\widehat{\Lambda}$ using multiple hypothesis testing.

Consider a list of null hypotheses, $\mathcal{H}_j$, $j = 1, \ldots, N$, with associated p-values p_j that stochastically dominate the uniform distribution on $[0, 1]$ under the null. Let the indices of the true nulls be $J_0 \subset \{1, \ldots, N\}$ and those of the non-nulls be $J_1 = \{1, \ldots, N\} \backslash J_0$.

Goal: FWER control uses the p-values to reject a subset of the $\mathcal{H}_j$ while limiting the probability of making any false rejections at a level δ.

Definition 2 (FWER-controlling algorithm)**.** An algorithm $\mathcal{A} : [0, 1]^N \to 2^{\{1,\ldots,N\}}$ is an FWER-controlling algorithm at level δ if $\mathbb{P}(\mathcal{A}(p_1, \ldots, p_N) \subseteq J_1) \geq 1 - \delta$.

Note: p-values in the above definition may be dependent, thus form $\widehat{\Lambda}$ with the Bonferroni correction, which satisfies Definition 2.

Proposition 2 (Bonferroni controls FWER)**.** *Let* $\mathcal{A}^{(\mathrm{Bf})}(p_1, \ldots, p_N) = \left\{\lambda_j : p_j \leq \frac{\delta}{|\Lambda|}\right\}$*. Then,* $\mathcal{A}^{(\mathrm{Bf})}$ *is an FWER-controlling algorithm.*

Note: For large N the multiplicity correction of Bonferroni correction degrades performance.

Thus, consider multiple testing methods that take advantage of problem structure to efficiently search the hypothesis space, mitigating this issue. This is possible because, adjacent p-values will be highly dependent for nearby λ, and non-nulls will generally cluster together. Thus, eventually only need one value $\hat{\lambda}$ with reasonably good performance that guarantees the error control.

2.2.1 Fixed Sequence Testing

The multiple testing method is designed for settings where a-priori which hypotheses are more or less likely to control the risk is known. For example, as in the case of the FDR, the risk function may be nearly monotonedecreasing in λ, making large λ much more promising than small ones.

Fixed sequence testing: sequentially test the hypotheses - e.g. from large λ to small λ-and stop upon the first acceptance. More generally,the fixed sequence test can be initialized at several different points along the ordering, provided the significance level is adjusted accordingly.

Algorithm 1 Fixed sequence testing

Input: error level $\delta \in (0, 1)$, parameter grid $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$, p-values $(p_1, \ldots, p_N)$, initializations $\mathcal{J} \subset \{1, \ldots, N\}$ (e.g., a coarse equi-spaced grid, with, say, 20 elements)
$\widehat{\Lambda} \leftarrow \emptyset$
for $j \in \mathcal{J}$ **do**
  if $\lambda_j \notin \widehat{\Lambda}$ **then** ▷ Avoid repeating values of j.
    while $p_j \leq \delta/|\mathcal{J}|$ **do**
      $\widehat{\Lambda} \leftarrow \widehat{\Lambda} \cup \{\lambda_j\}$
      $j \leftarrow j + 1$
Return: rejection set $\widehat{\Lambda}$

Proposition 3 (Fixed sequence testing controls FWER)**.** *Algorithm 1 is an FWER-controlling algorithm, i.e. it satisfies Definition 2.*

Proof. Consider $|\mathcal{J}| = 1$, at the first index where the null is encountered, the probability of making a false discovery is bounded by δ. Thus, the probability of making any false discoveries is bounded by δ. When $|\mathcal{J}| > 1$, the procedure is equivalent to running multiple instances of $|\mathcal{J}| = 1$ in parallel, at level $\delta/|\mathcal{J}|$. By the union bound, the probability of any false rejections is then bounded by δ. □

Fixed sequence testing yields the desired FWER level:

Proposition 4. *Let j^* be the index of the first null in the sequence. Then, for Algorithm 1 with $|\mathcal{J}| = 1$, FWER $= \mathbb{P}(p_{j^*} \leq \delta)$. As a result, if the null p-values are (asymptotically) uniform, the FWER is (asymptotically) δ as well.*

Proof. Since the procedure is sequential, the null $\mathcal{H}_{j^*}$ is rejected iff $p_{j^*} \leq \delta$. □

Note: By contrast, Bonferroni typically yields a FWER much smaller than δ

2.2.2 A General Recipe for FWER Control

Sequential graphical testing (SGT), is a more general and powerful framework for FWER. The SGT procedure is parameterized by a directed graph $\mathcal{G}$ comprising a node set Λ, and edge weights $g_{i,j} \in [0,1]$ for each pair $i, j \in \Lambda$ obeying $g_{i,i} = 0$ and $\sum_{j=1}^{n} g_{i,j} \leq 1$. In addition, each node i is allocated an initial error budget δ_i such that $\sum_i \delta_i = \delta$. From here, the algorithm iteratively tests each hypothesis $i \in \Lambda$ at the iteratively updated significance level δ_i (i.e., checks if $p_i \leq \delta_i$). If any hypothesis i is rejected, the procedure reallocates the error budget from node i to the rest of the nodes according to the edge weights, allowing them to be rejected more easily.

Algorithm 2 Sequential graphical testing [47]

Input: error level $\delta \in (0,1)$, parameter grid $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$, p-values $(p_1, \ldots, p_N)$, graph $\mathcal{G}$, initial error budget δ_i such that $\sum_i \delta_i = \delta$

$\widehat{\Lambda} \leftarrow \emptyset$

while $\exists i : p_i \leq \delta_i$ **do**

  Choose any i such that $p_i \leq \delta_i$

  $\widehat{\Lambda} \leftarrow \widehat{\Lambda} \cup \{\lambda_i\}$ ▷ Reject hypothesis i

  Update the error levels and the graph:

$$\delta_j \leftarrow \begin{cases} \delta_j + \delta_i g_{i,j} & \lambda_j \in \Lambda \setminus \widehat{\Lambda} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad g_{k,j} \leftarrow \begin{cases} \frac{g_{k,j} + g_{k,i} g_{i,j}}{1 - g_{k,i} g_{i,k}} & \lambda_k, \lambda_j \in \Lambda \setminus \widehat{\Lambda}, \quad k \neq j \\ 0 & \text{otherwise} \end{cases}$$

Return: rejection set $\widehat{\Lambda}$

Proposition 5 (SGT controls FWER)**.** *Algorithm 2 is an FWER-controlling algorithm, i.e. it satisfies Definition 2.*

The choices of the graph $\mathcal{G}$ and initial error budget $\{\delta_i\}_{i \in \Lambda}$ are critical for the power of the procedure. The general principle is to concentrate the initial error budget on hypotheses likely to reject. If these promising hypotheses are indeed rejected, then the error budget should accrue to adjacent hypotheses, giving them a higher chance of rejected.

2.3 Multiple Risks and Multi-Dimensional λ

Allow Λ with multiple dimensions and seek to control m risks $R_1, \ldots, R_m$ at levels $\alpha_1, \ldots, \alpha_m$ simultaneously. Define the null hypothesis

$$\mathcal{H}_j : R_l(\lambda_j) > \alpha_l, \text{ for some } l \in 1, \ldots, m.$$

To test this null hypothesis, examine the finer null hypotheses, $\mathcal{H}_{j,l} : R_l(\lambda_j) > \alpha_l$: $\mathcal{H}_j$ holds iff there exists a $l \in 1, \ldots, m$ such that $\mathcal{H}_{j,l}$ holds. Then apply an FWER-controlling procedure to test $\mathcal{H}_j$.

Proposition 6. *Let $p_{j,l}$ be a p-value for $\mathcal{H}_{j,l}$, for each $l = 1, \ldots, m$. Define $p_j := \max_l p_{j,l}$. Then, for all j such that $\mathcal{H}_j$ holds and for all $u \in [0,1]$, we have $\mathbb{P}(p_j \leq u) \leq u$.*

Thus, valid p-values are calculated for each λ_j, then apply techniques from the previous section to select a set $\widehat{\Lambda}$ that controls the FWER.