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 $R(\mathcal{T}_\lambda) \in \mathbb{R}$, 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2 Risk Control in Prediction

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

Algorithm Outlin: 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.

2.1 Calculating Valid p-Values

2.2 Multiple Hypothesis Testing

2.3 Multiple Risks and Multi-Dimensional λ

2.4 An Alternative Approach: Uniform Concentration