Anytime Conformal Prediction

Arun Kumar Kuchibhotla

30 June, 2022

Carnegie Mellon University

Joint work

This is joint work with Kayla Scharfstein, CMU.

Very much a work in progress, all comments welcome.

Table of contents

Introduction to Conformal Prediction

Conformal Prediction

- Conformal prediction, in recent times, has received a lot of attention as the go-to method for constructing valid prediction sets under "weak" assumptions.
- The problem conformal prediction solves can be stated as follows. Suppose $Z_1, \ldots, Z_n, Z_{n+1}$ are exchangeable random variables in some measurable space, of which we only observe $Z_1, \ldots, Z_n$. For any $\alpha \in [0, 1]$, construct a set $\widehat{C}_{n,\alpha}$ such that $\mathbb{P}(Z_{n+1} \in \widehat{C}_{n,\alpha}) \geq 1 - \alpha$.
- There are several existing conformal methods that can achieve this goal without any more distributional assumptions. There are even extensions for dependent data and arbitrary individual sequences.
- The problem we want to study is

"can we stop at a sample size n at will?
Can n be a random time?"

Different Conformal Methods

- One of the simplest conformal methods to solve this problem is called the *split conformal prediction*; Vovk, Lei and Wasserman, and so on.
- There are more complicated methods such as full conformal, jackknife+, CV+, bootstrap-after-jackknife+, and so on.
- All these methods are shown to work under the weaker assumption of exchangeable data.
- Here we consider the assumption of IID data and require the stronger anytime prediction guarantee.
- We first review the split conformal method and mention a few reasons why IID assumption provides a great insight into the prediction problem.

Split Conformal Method

- The idea of the split conformal method is to split the data into two parts and obtain a transform to reduce the problem to 1-d.
- Obtain a real-valued transformation $\widehat{R}(\cdot)$ based on the first part and let $\widehat{q}_{n_2,\alpha}$ denote the $\lceil (n_2+1)(1-\alpha) \rceil$-th largest value of $\widehat{R}(Z_j)$ for Z_j's in the second part.
- Just under exchangeability, it can be proved that

$$\mathbb{P}(\widehat{R}(Z_{n+1}) \le \widehat{q}_{n_2,\alpha}) \ge 1-\alpha \quad \text{for all} \quad n \ge 1.$$

- Equivalently, if $\widehat{C}_{n,\alpha} = \{z : \widehat{R}(z) \le \widehat{q}_{n_2,\alpha}\}$, then

$$\mathbb{P}(Z_{n+1} \in \widehat{C}_{n,\alpha}) \ge 1-\alpha.$$

- In the IID setting, this guarantee can be written in terms of the common probability measure $\mu(\cdot)$ of Z_i's: $\mathbb{E}[\mu(\widehat{C}_{n,\alpha})] \ge 1-\alpha$.

Exchangeability to IID

- There is more to gain from strengthening the exchangeability assumption to IID random variables.
- Conditional on the first split, $\widehat{R}(Z_j)$ for Z_j's in the second split are IID observations and the goal of constructing a valid prediction now becomes finding a $\widehat{q}_{n_2,\alpha}$ such that

$$\mathbb{P}(\widehat{R}(Z_{n+1}) \leq \widehat{q}_{n_2,\alpha}|\widehat{R}) = \mathbb{E}[F_{\widehat{R}}(\widehat{q}_{n_2,\alpha})|\widehat{R}] \geq 1-\alpha,$$

  where $F_{\widehat{R}}(\cdot)$ is the cumulative distribution function of $\widehat{R}(Z)$ conditional on $\widehat{R}$.
- This reformulation already shows a great advantage of the IID assumption. Note that the population $1-\alpha$ quantile already satisfies $F(q_\alpha) \geq 1-\alpha$.
- Note that if $\widehat{q}_{n_2,\alpha}$ is some quantile of $\widehat{R}(Z_j)$ for Z_j's in the second split, then $F_{\widehat{R}}(\widehat{q}_{n_2,\alpha})$ is a uniform order statistics, the properties of which are understood well.

Exchangeability to IID

- In addition to asking for $\mathbb{E}[F_{\widehat{R}}(\widehat{q}_{n_2,\alpha})] \geq 1-\alpha$, one can also consider a PAC guarantee:

$$\mathbb{P}(F_{\widehat{R}}(\widehat{q}_{n_2,\alpha,\delta}) \geq 1-\alpha) \geq 1-\delta.$$

  Conditional coverage is at least $1-\alpha$ with probability at least $1-\delta$.
- By strengthening the exchangeability assumption to IID, we can understand at a conformal prediction much more, still without distributional assumptions.
- For example, we can say how the conformal method behaves for the coverage guarantee uniform over all $\alpha \in [0,1]$. From DKW, we know

$$\mathbb{E}\left[\sup_{1\leq k\leq n}\left|U_{k:n} - \frac{k}{n}\right|\right] \leq \frac{C}{\sqrt{n}} \quad \text{for all} \quad n \geq 1.$$

- We can also consider tail bounds for $\sup_k |U_{k:n} - k/n|$ so that a uniform PAC guarantee can be obtained.
- Such deviation inequalities also help aggregate several split conformal prediction sets based on different transformations $\widehat{R}(\cdot)$ to obtain a smaller prediction sets (Yang and Kuchibhotla, 2021).

Anytime Conformal Prediction: The Problem

The Problem

- Recall that under the IID assumption, the classical goal of conformal prediction is to create a set $\widehat{C}_{n,\alpha}$ given n observations $Z_1, \ldots, Z_n$ such that $\mathbb{E}[\mu(\widehat{C}_{n,\alpha})] \geq 1 - \alpha$.
- Here n is a fixed, given sample size.
- Consider the scenario where we observe (IID) data sequentially: we observe Z_1, report a prediction set $\widehat{C}_2$, observe Z_2, report prediction set Z_3, and so on.
- The analyst decides to stop at time T for some reason. Can we guarantee that $\widehat{C}_T$ still contains $1 - \alpha$ proportion of all future independent random variables?
- Formally, can we guarantee $\mathbb{E}[\mu(\widehat{C}_T)] \geq 1 - \alpha$? To see this, let $Z_1^*, Z_2^*, \ldots$ denote an independent sequence of random variables from $\mu(\cdot)$. Covering $1 - \alpha$ proportion of $Z_1^*, Z_2^*, \ldots$ is same as

$$\lim_{s \to \infty} \frac{1}{s} \sum_{i=1}^{s} \mathbf{1}\{Z_i^* \in \widehat{C}_T\} \geq 1 - \alpha.$$

The Problem and Desiderata

- The problem of anytime conformal prediction is to construct a sequence of sets such that

$$\mathbb{E}[\mu(\widehat{C}_{T,\alpha})] \geq 1 - \alpha,$$

for any random time T with $\widehat{C}_{T,\alpha}$ depending only on IID observations $Z_1, \dots, Z_{T-1}$.

- This goal can be shown to be equivalent to

$$\mathbb{E}\left[\min_{t\geq 1} \mu(\widehat{C}_{t,\alpha})\right] \geq 1 - \alpha. \quad (1)$$

- This equivalence implies that one may not have access to independent hold-out data for obtaining a transformation $\widehat{R}(\cdot)$.
- Similar to (1), one can also consider anytime PAC guarantee:

$$\mathbb{P}\left(\min_{t\geq 1} \mu(\widehat{C}_{t,\alpha,\delta}) \geq 1 - \alpha\right) \geq 1 - \delta.$$

The Problem and Desiderata

- Let us consider for now the anytime PAC guarantee:

$$\mathbb{P}\left(\min_{t\geq 1} \mu(\widehat{C}_{t,\alpha,\delta}) \geq 1-\alpha\right) \geq 1-\delta.$$

- This goal is readily possible in two "trivial" ways:
    - Irrespective of data, return a set that is $\mathcal{Z}$ with probability $1-\alpha$ and $\emptyset$ with probability α. Then $\min_{t\geq 1} \mu(\widehat{C}_{t,\alpha,\delta}) = 1-\alpha$ almost surely.
    - Take $\widehat{C}_{t,\alpha,\delta} = \mathcal{Z}$ for $t < n$ (for some n) and return the classical split conformal prediction set $\widehat{C}_{n,\alpha,\delta}$ for $t > n$. Then

$$\min_{1\leq t\leq n} \mu(\widehat{C}_{t,\alpha,\delta}) = 1 \quad \text{and} \quad \min_{t>n} \mu(\widehat{C}_{t,\alpha,\delta}) = \mu(\widehat{C}_{n,\alpha,\delta}).$$

- The problem with both these sets is that they do not converge to the "optimal" prediction set as $t \to \infty$. Note that unlike confidence regions, prediction sets do not shrink to a singleton as $t \to \infty$.
- We do expect that $\widehat{C}_{t,\alpha,\delta}$ becomes the optimal prediction set as t becomes ∞.

The Problem and Desiderata

- For a good choice of transformation $\widehat{R}(\cdot)$, the classical split conformal set $\widehat{C}_{n,\alpha}$ can be shown to satisfy

$$\text{Leb}(\widehat{C}_{n,\alpha} \Delta C_{\alpha}^{\text{opt}}) \leq O_p(r_n),$$

for $r_n \to 0$ as $n \to \infty$ at a "good" rate.

- A reasonable desiderata for the anytime conformal prediction problem is to construct $\widehat{C}_{t,\alpha,\delta}$ such that

$$\mathbb{P}\left(\min_{t \geq 1} \mu(\widehat{C}_{t,\alpha,\delta}) \geq 1 - \alpha\right) \geq 1 - \delta.$$

and

$$\text{Leb}(\widehat{C}_{t,\alpha,\delta} \Delta C_{\alpha}^{\text{opt}}) \leq \widetilde{O}_p(r_t),$$

where $\widetilde{O}(\cdot)$ holds $\log(t)$ or $\log\log(t)$ factors.

Anytime Conformal Prediction: A Solution

Recall

- The anytime conformal prediction problem is to construct $\widehat{C}_{t,\alpha,\delta}$ such that
$$\mathbb{P}\left(\min_{t\geq 1} \mu(\widehat{C}_{t,\alpha,\delta}) \geq 1-\alpha\right) \geq 1-\delta.$$
and
$$\text{Leb}(\widehat{C}_{t,\alpha,\delta} \Delta C_{\alpha}^{\text{opt}}) + \widetilde{O}_p(r_t),$$
where $\widetilde{O}(\cdot)$ holds $\log(t)$ or $\log\log(t)$ factors.
- **Example**: If $Z = (X, Y)$ and $Y = m_0(X) + \xi$, $\xi|X \sim N(0, \sigma^2)$, then the optimal prediction set for Y is $[m_0(X) \pm \sigma z_{\alpha/2}]$, the width of which is $\sigma z_{\alpha/2}$.
The conformal prediction set $[\widehat{m}_{n_1}(X) \pm \widehat{q}_{n_2,\alpha}]$. Here $\widehat{m}_{n_1}(\cdot)$ is obtained from the hold-out training data and $\widehat{q}_{n_2,\alpha}$ is obtained from the calibration data. With n_1 fixed and $n_2 \to \infty$, the prediction set becomes $[\widehat{m}_{n_1}(X) \pm \sigma z_{\alpha/2}]$.
Only with $n_1 \to \infty$, we get to the optimal set: $[m_0(X) \pm \sigma z_{\alpha/2}]$.

Towards the Solution: Part 1

- Recall that we do not have access to a hold-out data set for anytime conformal prediction as we want coverage guarantee from sample size 1. As a first step, let us assume we do have such a hold-out.
- Use the hold-out data to obtain a real-valued transformation $\widehat{R}_{n_1}(\cdot)$.
- As we get samples $Z_{n_1+1}, Z_{n_1+2}, \ldots$, we want to report $\widehat{q}_{n_1+j,\alpha,\delta}$ such that

$$\mathbb{P}\left(\min_{j\geq 1} \mu(\widehat{R}_{n_1}^{-1}(\widehat{q}_{n_1+j})) \geq 1-\alpha\right) \geq 1-\delta.$$

- Note that $\mu(\widehat{R}_{n_1}^{-1}(\widehat{q}_{n_1+j})) = F_{\widehat{R}_{n_1}}(\widehat{q}_{n_1+j})$.
- Hence, it suffices to find $\widehat{q}_{n_1+j}$ such that

$$\mathbb{P}\left(\min_{j\geq 1} \widehat{q}_{n_1+j} \geq F_{\widehat{R}_{n_1}}^{-1}(1-\alpha)\right) \geq 1-\delta.$$

  This means, we want an uniform in sample size upper bound on a particular quantile of $F_{\widehat{R}_{n_1}}(\cdot)$, which is possible via a distribution-free confidence band sequence for a CDF.

Towards the Solution: Part 1 (contd.)

- Confidence sequence for CDF: with observations $Z_{n_1+1}, \dots, Z_{n_1+j}$, we have the empirical cdf

$$\widehat{F}_{\widehat{R}_{n_1},j}(x) = \frac{1}{j}\sum_{s=1}^{j} \mathbf{1}\{\widehat{R}_{n_1}(Z_{n_1+s}) \le x\}.$$

- Howard and Ramdas (2021, Eq. (1)) implies that with probability at least $1-\delta$,

$$F^{-1}_{\widehat{R}_{n_1}}(1-\alpha) \le \min_{j\ge 1} \widehat{F}^{-1}_{\widehat{R}_{n_1},j}\left(1-\alpha+u_{j,\alpha,\delta}\right),$$

where $u_{j,\alpha,\delta} = 1.5\sqrt{\alpha(1-\alpha)\ell_\alpha(j)} + 0.8\ell_\alpha(j)$ with

$$\ell_\alpha(j) = \frac{1.4\log\log(2.1j)+\log(10/\delta)}{j}.$$

- Hence, the anytime conformal prediction set (with hold-out data) is

$$\widehat{C}_{n_1+j,\alpha,\delta} = \{z:\ \widehat{R}_{n_1}(z) \le \widehat{F}^{-1}_{\widehat{R}_{n_1},j}(1-\alpha+u_{j,\delta})\}.$$

The problem is that this converges to $[\widehat{m}_{n_1}(X) \pm \sigma z_{\alpha/2}]$ as $j\to\infty$. We need $\widehat{R}_{n_1}$ also to converge for optimality.

Final Solution

- In order to improve on the previous solution with hold-out data, we need to somehow sequentially update.
- Consider a sequence of transformations $\widehat{R}_0(\cdot), \widehat{R}_1(\cdot), \ldots$ with $\widehat{R}_t(\cdot)$ computed based on $Z_1, \ldots, Z_{t-1}$.

  Eg: $\widehat{R}_t(\cdot)$ obtained from SGD. If $Z = (X, Y)$, then one can consider $\widehat{R}_t(x) = \widehat{\beta}_t^\top x$ with $\widehat{\beta}_t$ obtained via

  $$\widehat{\beta}_t = \widehat{\beta}_{t-1} - \xi_t X_{t-1}(Y_{t-1} - X_{t-1}^\top \widehat{\beta}_{t-1}),$$

  for $t \geq 1$. Here ξ_t is some step size.
- Ideally, we would like to find the quantile $q_{t,\alpha}$ of $\widehat{R}_t(Z)$ and use the set $\{z : \widehat{R}_t(z) \leq q_{t,\alpha}\}$. But, we do not have any observations with the same distribution as $\widehat{R}_t(Z)$.

Idea: At time t, use $\widehat{R}_s(\cdot)$ for some $s < t$, say, $s = \eta^{\lfloor \log_\eta t \rfloor}$ $(\eta > 1)$.

Final Solution (Contd.)

- Formally, split the time into geometric epochs

$$\{t \geq 1\} = \bigcup_{k \geq 0} \left\{s : \eta^k \leq s < \eta^{k+1}\right\}.$$

- For $s \in [\eta^k, \eta^{k+1}]$, apply the step 1 with $\widehat{R}_{\lceil \eta^k \rceil}(\cdot)$.
- This means that at time n, this method is using a fraction n/η of the observations as hold-out data and the resulting prediction sets converge to the optimal one as the sample size tends to infinity.
- All this optimality hinges on using the "right" transformation $\widehat{R}(\cdot)$, which has been discussed by several authors. See, e.g., Gupta et al. (2022, Pattern Recognition) and Sesia & Candes (2020, Stat).
- The benefit of this method is that it is valid for any sequence of transformations $\widehat{R}_s(\cdot), s \geq 1$ and is a purely online method in that there is no need to store past data.

Conclusions

Conclusions

- Strengthening the exchangeability assumption to IID yields a better understanding of conformal problem and methods. It also paves way for using concentration inequalities.
- We have introduced an online version of conformal prediction problem.
- Making use of confidence sequences, we provide a solution and proved that the resulting prediction sets are asymptotically optimal under regularity assumptions.
- The proposed solution is purely online and does not require storing past data.
- It remains to be seen how well the proposed solution performs in practice.

Conclusions

- Strengthening the exchangeability assumption to IID yields a better understanding of conformal problem and methods. It also paves way for using concentration inequalities.
- We have introduced an online version of conformal prediction problem.
- Making use of confidence sequences, we provide a solution and proved that the resulting prediction sets are asymptotically optimal under regularity assumptions.
- The proposed solution is purely online and does not require storing past data.
- It remains to be seen how well the proposed solution performs in practice.

Thank You!