Predictive Inference Is Free with the Jackknife+-after-Bootstrap

Byol Kim, Chen Xu, and Rina Foygel Barber

Kiho Park
parkkiho@uchicago.edu

1 Introduction

While ensemble learning can provide substantially more stable and accurate estimates, relatively little is known about how to perform provably valid inference on the resulting output. Particular challenges arise when the data distribution is unknown, or when the base learner is difficult to analyze. We propose a method for constructing a predictive confidence interval for a new observation. Our method integrates ensemble learning with the recently proposed jackknife+ [Barber et al. (2019)]. Our main contributions are as follows.

- We propose the jackknife+-after-bootstrap (J+aB), a method for constructing predictive confidence intervals that can be efficiently wrapped around an ensemble learning algorithm chosen by the user.
- We prove that the coverage of a J+aB interval is at worst $1-2\alpha$ for the assumption-free theory. This lower bound is non-asymptotic, and holds for any sample size and any distribution of the data.
- We verify that the empirical coverage of a J+aB interval is actually close to $1-\alpha$.

2 Background and related work

Suppose $(X_1, Y_1), \cdots, (X_n, Y_n) \overset{i.i.d.}{\sim} \mathcal{P}$ on $\mathbb{R}^p \times \mathbb{R}$. We assume that (X_{n+1}, Y_{n+1}) is drawn from the same distribution $\mathcal{P}$. We would like to our method to output a predictive interval $\widehat{C} : \mathbb{R}^p \to \mathbb{R}^2$ with the property $\mathbb{P}\left[Y_{n+1} \in \widehat{C}(X_{n+1})\right] \geq 1-\alpha$. Ideally, we want $\widehat{C}$ satisfy distribution-free predictive coverage at level $1-\alpha$.

2.1 Jackknife and jackknife+

Let $\mathcal{R}$ denote any regression algorithm which outputs $\widehat{\mu} : \mathbb{R}^p \to \mathbb{R}$. We write $\widehat{\mu} = \mathcal{R}(\{(X_i, Y_i)\}_{i=1}^n)$ and $\widehat{\mu}_{\setminus i} = \mathcal{R}(\{(X_j, Y_j)\}_{j=1, j\neq i}^n)$. Let $\widehat{q}_{\alpha,n}^{+}\{v_i\}$ and $\widehat{q}_{\alpha,n}^{-}\{v_i\}$ denote the upper and the lower α-quantiles of a list of n values indexed by i. The jackknife prediction interval is given by

$$\widehat{C}_{\alpha,n}^{\mathrm{J}}(x) = \widehat{\mu}(x) \pm \widehat{q}_{\alpha,n}^{+}\{R_i\} = [\widehat{q}_{\alpha,n}^{-}\{\widehat{\mu}(x) - R_i\}, \widehat{q}_{\alpha,n}^{+}\{\widehat{\mu}(x) + R_i\}]$$

where $R_i = |Y_i - \widehat{\mu}_{\setminus i}(X_i)|$. It turns out that fully assumption-free theory is impossible for the interval. By contrast, it is achieved by the jackknife+, which modifies it by replacing $\widehat{\mu}$ with $\widehat{\mu}_{\setminus i}$'s:

$$\widehat{C}_{\alpha,n}^{\mathrm{J+}}(x) = [\widehat{q}_{\alpha,n}^{-}\{\widehat{\mu}_{\setminus i}(x) - R_i\}, \widehat{q}_{\alpha,n}^{+}\{\widehat{\mu}_{\setminus i}(x) + R_i\}]$$

Barber et al. (2019) showed that this interval satisfies distribution-free predictive coverage at level $1-2\alpha$.

2.2 Ensemble methods

The ensemble method starts by creating multiple training data sets of size m from the available training data points $\{1, \cdots, n\}$. We may choose the sets by bootstrapping (with replacement—a typical choice is $m = n$), or by subsampling (without replacement, for instance with $m = n/2$).

Algorithm 1 Ensemble learning

```
Input: Data $\{(X_i, Y_i)\}_{i=1}^n$
for $b = 1, \cdots, B$ do
    Draw $S_b = (i_{b,1}, \cdots, i_{b,m})$ by sampling with or without replacement from $\{1, \cdots, n\}$.
    Compute $\widehat{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \cdots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
end for
Define $\widehat{\mu}_\varphi = \varphi(\widehat{\mu}_1, \cdots, \widehat{\mu}_B)$.
Output: Ensembled regression function $\widehat{\mu}_\varphi$.
```

In the case of the jackknife+, if each ensembled model makes B many calls to the base regression method $\mathcal{R}$, the jackknife+ would require a total of Bn calls to $\mathcal{R}$. By contrast, our method will require only $O(B)$ many calls to $\mathcal{R}$, making the computational burden comparable to obtaining a single ensemble prediction.

2.3 Related work

Distribution-free predictive inference has garnered attention in recent years for providing valid inferential tools that can work with complex machine learning algorithms. This is because many of the methods proposed are wrapper algorithms that can be used in conjunction with an arbitrary learning procedures.

The method we propose can also be viewed as a wrapper designed specifically for use with ensemble learners. Many authors came up with cost efficient wrappers for use in the ensemble prediction setting. However fully distribution-free coverage cannot be guaranteed for previous methods. By contrast, out method preserves exchangeability, and hence is able to maintain assumption-free and finite-sample validity.

3 Jackknife+-after-bootstrap (J+aB)

We present out method, the jackknife+-after-bootstrap (J+aB). To design a cost efficient wrapper method suited to the ensemble prediction setting, we make use of the "out-of-bag" estimates.

Algorithm 2 Jackknife+-after-bootstrap (J+aB)

```
Input: Data $\{(X_i, Y_i)\}_{i=1}^n$
for $b = 1, \cdots, B$ do
    Draw $S_b = (i_{b,1}, \cdots, i_{b,m})$ by sampling with or without replacement from $\{1, \cdots, n\}$.
    Compute $\widehat{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \cdots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
end for
for $i = 1, \cdots, n$ do
    Aggregate $\widehat{\mu}_{\varphi\setminus i} = \varphi(\{\widehat{\mu}_b : b = 1, \cdots, B, S_b \not\ni i\})$
    Compute the residual, $R_i = |Y_i - \widehat{\mu}_{\varphi\setminus i}(X_i)|$.
end for
Compute the J+aB prediction intveral: at each $x \in \mathbb{R}$,

    $\widehat{C}^{\text{J+aB}}_{\alpha,n,B}(x) = \left[\widehat{q}^{-}_{\alpha,n}\{\widehat{\mu}_{\varphi\setminus i}(x) - R_i\}, \widehat{q}^{+}_{\alpha,n}\{\widehat{\mu}_{\varphi\setminus i}(x) + R_i\}\right]$

Output: Predictive interval $\widehat{C}^{\text{J+aB}}_{\alpha,n,B}$
```

In settings where both model evaluations and aggregations remain relatively cheap, our J+aB algorithm is able to output a more informative confidence interval at virtually no extra cost beyond what is already necessary to produce a single ensemble point prediction. Thus, one can view the J+aB as offering predictive inference "free of charge."

Table: Comparison of computational costs for obtaining n_{test} predictions

	#calls to $\mathcal{R}$	#evaltuations	#calls to φ
Ensemble	B	Bn_{test}	n_{test}
J+ with Ensemble	Bn	$Bn(1+n_{\text{test}})$	$n(1+n_{\text{test}})$
J+aB	B	$B(1+n_{\text{test}})$	$n(1+n_{\text{test}})$

4 Theory

Assumption 1. *$(X_1,Y_1),\cdots,(X_n,Y_n),(X_{n+1},Y_{n+1}) \overset{i.i.d.}{\sim} \mathcal{P}$, where $\mathcal{P}$ is any distribution on $\mathbb{R}^p \times \mathbb{R}$.*

Assumption 2. *For $k \geq 1$, any fixed k-tuple $((x_1,y_1),\cdots,(x_k,y_k)) \in \mathbb{R}^p \times \mathbb{R}$, and any permutation σ on $\{1,\cdots,k\}$, it holds that $\mathcal{R}((x_1,y_1),\cdots,(x_k,y_k)) = \mathcal{R}((x_{\sigma(1)},y_{\sigma(1)}),\cdots,(x_{\sigma(k)},y_{\sigma(k)}))$ and $\varphi(y_1,\cdots,y_k) = \varphi(y_{\sigma(1)},\cdots,y_{\sigma(k)}))$. In other words, the base regression algorithm and the aggregation φ are both invariant to the ordering of the input arguments.*

Theorem 1 gives the distribution-free coverage guarantee for the J+aB prediction interval with one intriguing twist: the total number of base models, B, must be drawn at random rather than chosen in advance. This is because Algorithm 2 as given subtly violates symmetry. However, we shall see that requiring B to be Binomial is enough to restore symmetry, after which assumption-free theory is possible.

Theorem 1. *Fix any integers $\widetilde{B} \geq 1$ and $m \geq 1$, any base algorithm $\mathcal{R}$, and any aggregation function φ. Suppose the jackknife+-after-bootstrap method (Algorithm 2) is run with (i) $B \sim \text{Binomial}\left(\widetilde{B}, \left(1-\frac{1}{n+1}\right)^m\right)$ in the case of sampling with replacement or (ii) $B \sim \text{Binomial}\left(\widetilde{B}, \left(1-\frac{m}{n+1}\right)\right)$ in the case of sampling without replacement. Then, under Assumptions 1 and 2, the jackknife+-after-bootstrap prediction interval satisfies*

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})\right] \geq 1-2\alpha$$

where the probability holds with respect to the random draw of the training data $(X_1,Y_1),\cdots,(X_n,Y_n)$, the test data point (X_{n+1},Y_{n+1}), and the Binomial B.

In most settings where a large number of models are being aggregated, we would not expect the distinction of random vs fixed B to make a meaningful difference to the final output. In Appendix B, we formalize this intuition and give a stability condition on the aggregating map φ under which the J+aB has valid coverage for any choice of B.

Finally, we remark that although we have exclusively used the regression residuals $|Y_i - \widehat{\mu}_{\setminus i}(X_i)|$ in our exposition for concreteness, our method can also accommodate alternative measures of conformity. If $\widehat{c}_{\varphi\setminus i}$ is the trained conformity measure aggregated from the S_b's that did not use the i-th point, then the corresponding J+aB set is given by

$$\widehat{C}^{c-J+aB}_{\alpha,n,B}(x) = \left\{y : \sum_{i=1}^{n} \mathbb{1}[\widehat{c}_{\varphi\setminus i}(x,y) > \widehat{c}_{\varphi\setminus i}(X_i,Y_i)] < (1-\alpha)(n+1)\right\}$$

5 Experiments

In this section, we demonstrate that the J+aB intervals enjoy coverage near the nominal level of $1 - \alpha$ numerically, using three real data sets and different ensemble prediction methods. In addition, we also look at the results for the jackknife+, combined either with the same ensemble method (J+ENSEMBLE) or with the non-ensembled base method (J+NON-ENSEMBLE). We used three real data sets, which were also used in Barber et al. (2019). For the base regression method $\mathcal{R}$, we used either the ridge regression (RIDGE), the random forest (RF), or a neural network (NN). We fixed $\alpha = 0.1$ and used $n = 40$ observations for training. We varied the size m of each bootstrap replicate as $m/n = 0.2, 0.4, \cdots, 1.0$. We used $B = 20$ for $J + \text{ENSEMBLE}$, and we drew $B \sim \text{Binomial}(\widetilde{B}, (1 - \frac{1}{n+1})^m)$ with $\widetilde{B} = [20/\{(1 - \frac{1}{n+1})^m (1 - \frac{1}{n})^m\}]$.

From Figure 1, it is clear that the coverage of the J+aB, J+ENSEMBLE, and J+NON-ENSEMBLE is near the nominal level. In Figure 2, we observe no evidence of a consistent trend of one method always outperforming the other in terms of the precision of the intervals. Thus, we prefer the computationally efficient J+aB to the costly J+ENSEMBLE. In settings where the base learner is expected to benefit from ensembling, J+aB is a practical method for obtaining informative prediction intervals that requires a level of computational resources on par with the ensemble algorithm itself.

6 conclusion

We propose the jackknife+-after-bootstrap (J+aB), a computationally efficient wrapper method tailored to the setting of ensemble learning, where by a simple modification to the aggregation stage, the method outputs a predictive interval with fully assumption-free coverage guarantee in place of a point prediction. The J+aB provides a mechanism for quantifying uncertainty in ensemble predictions that is both straightforward to implement and easy to interpret, and can therefore be easily integrated into existing ensemble models.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+, 2019. arXiv preprint.

7 Question

Why $(1 - 1/n)^m$ is there? page 8.