Trustworthy Machine Learning

PAC Conformal Prediction

Sangdon Park

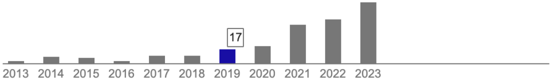
POSTECH

Motivation: Conditional Guarantee?

Vladimir Vovk

Conditional validity of inductive conformal predictors

Authors	Vladimir Vovk
Publication date	2012/11/17
Conference	Asian conference on machine learning
Pages	475-490
Publisher	PMLR
Description	Conformal predictors are set predictors that are automatically valid in the sense of having coverage probability equal to or exceeding a given confidence level. Inductive conformal predictors are a computationally efficient version of conformal predictors satisfying the same property of validity. However, inductive conformal predictors have been only known to control unconditional coverage probability. This paper explores various versions of conditional validity and various ways to achieve them using inductive conformal predictors and their modifications.
Total citations	Cited by 251



Scholar articles
Conditional validity of inductive conformal predictors
V Vovk - Asian conference on machine learning, 2012
Cited by 251 Related articles All 19 versions

Conditional Guarantees

Marginal Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(X_{n+1})\right\} \geq 1 - \alpha$$

Conditional Guarantees

Marginal Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(X_{n+1})\right\} \geq 1 - \alpha$$

X-conditional Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(x) \middle| X_{n+1} = x\right\} \geq 1 - \alpha$$

Conditional Guarantees

Marginal Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(X_{n+1})\right\} \geq 1 - \alpha$$

X-conditional Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(x) \middle| X_{n+1} = x\right\} \geq 1 - \alpha$$

- Hopeless :([Lei and Wasserman, 2014]

Conditional Guarantees

Marginal Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(X_{n+1})\right\} \geq 1-\alpha$$

X-conditional Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(x) \middle| X_{n+1} = x\right\} \geq 1-\alpha$$

- Hopeless :([Lei and Wasserman, 2014]

Training-conditional Guarantee (=PAC Guarantee):

$$\mathbb{P}\left\{\mathbb{P}\left\{y \in \hat{C}(x)\right\} \geq 1-\varepsilon\right\} \geq 1-\delta$$

Conditional Guarantees

Marginal Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(X_{n+1})\right\} \geq 1 - \alpha$$

X-conditional Guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}(x) \middle| X_{n+1} = x\right\} \geq 1 - \alpha$$

- Hopeless :([Lei and Wasserman, 2014]

Training-conditional Guarantee (=PAC Guarantee):

$$\mathbb{P}\left\{\mathbb{P}\left\{y \in \hat{C}(x)\right\} \geq 1 - \varepsilon\right\} \geq 1 - \delta$$

- We will explore this!

PAC Guarantee: A Goodness Metric

PAC-style coverage guarantee

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

PAC Guarantee: A Goodness Metric

PAC-style coverage guarantee

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This implies that we need the i.i.d. assumption.

PAC Guarantee: A Goodness Metric

PAC-style coverage guarantee

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This implies that we need the i.i.d. assumption.
- $\hat{C}$ is learned from a calibration set $Z_n \sim \mathcal{D}^n$.

PAC Guarantee: A Goodness Metric

PAC-style coverage guarantee

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This implies that we need the i.i.d. assumption.
- $\hat{C}$ is learned from a calibration set $Z_n \sim \mathcal{D}^n$.
- We will interpret conformal prediction to a learning problem [Valiant, 1984].
  - See tolerance region [Wilks, 1941] and training-conditional inductive conformal prediction [Vovk, 2013] for an equivalent result.

PAC Guarantee: A Goodness Metric

PAC-style coverage guarantee

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This implies that we need the i.i.d. assumption.
- $\hat{C}$ is learned from a calibration set $Z_n \sim \mathcal{D}^n$.
- We will interpret conformal prediction to a learning problem [Valiant, 1984].
  - See tolerance region [Wilks, 1941] and training-conditional inductive conformal prediction [Vovk, 2013] for an equivalent result.
- The main goal is to find a PAC learning algorithm for the set of conformal sets.

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm
  - Why not use $\tau = 0$?

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm
  - Why not use $\tau = 0$?
    - No! Produces trivial conformal sets.

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm
  - Why not use $\tau = 0$?
    - No! Produces trivial conformal sets.
- How to minimize the size of conformal sets?

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm
  - Why not use $\tau = 0$?
    - No! Produces trivial conformal sets.
- How to minimize the size of conformal sets?
  - Another objective of the PAC learning algorithm

Conformal Prediction with a PAC Guarantee

Learning-theoretic View [Park et al., 2020]

Parameterized conformal sets

$$C(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- How to find τ that satisfies the PAC guarantee?
  - Finding a PAC learning algorithm
  - Why not use $\tau = 0$?
    - No! Produces trivial conformal sets.
- How to minimize the size of conformal sets?
  - Another objective of the PAC learning algorithm
  - Minimize the size, while satisfying the PAC guarantee.

Minimizing Set Size

Secondary goal: minimizing set size

$$\text{maximizing } \tau \quad \Longrightarrow \quad \text{minimizing the expected set size}$$

Minimizing Set Size

Secondary goal: minimizing set size

maximizing $\tau \quad \Longrightarrow \quad$ minimizing the expected set size

- Recall the conformal set definition:

$$C_\tau(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

Minimizing Set Size

Secondary goal: minimizing set size

$$\text{maximizing } \tau \quad \Longrightarrow \quad \text{minimizing the expected set size}$$

- Recall the conformal set definition:

$$C_\tau(x) \coloneqq \{y \in \mathcal{Y} \mid f(x,y) \geq \tau\}$$

- We have

$$\tau_1 \leq \tau_2 \quad \Longrightarrow \quad \forall x, C_{\tau_2}(x) \subseteq C_{\tau_1}(x),$$

  i.e., size is monotonically decreasing in τ.

Minimizing Set Size

Secondary goal: minimizing set size

$$\text{maximizing } \tau \quad \Longrightarrow \quad \text{minimizing the expected set size}$$

- Recall the conformal set definition:

$$C_\tau(x) := \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$$

- We have

$$\tau_1 \leq \tau_2 \quad \Longrightarrow \quad \forall x, C_{\tau_2}(x) \subseteq C_{\tau_1}(x),$$

  i.e., size is monotonically decreasing in τ.
- Maximizing τ eventually minimizes the expected size, *i.e.*,

$$\mathbb{E}\left\{S(C(x))\right\} \leq \sup_x S(C(x))$$

  - $S(\cdot)$: a size metric

PAC Learning Algorithm

$$\mathcal{A}_{\mathsf{Binom}}: \qquad \hat{\tau} = \max_{\tau \in \mathbb{R}_{\geq 0}} \tau \qquad \text{subj. to} \qquad U_{\mathsf{Binom}}(C_\tau, Z_n, \delta) \leq \varepsilon$$

- $\mathcal{A}_{\mathsf{Binom}}$ returns $\hat{\tau} = 0$ if the constraint is infeasible.

PAC Learning Algorithm

$$\mathcal{A}_{\textsf{Binom}}: \qquad \hat{\tau} = \max_{\tau \in \mathbb{R}_{\geq 0}} \; \tau \qquad \text{subj. to} \qquad U_{\textsf{Binom}}(C_\tau, Z_n, \delta) \leq \varepsilon$$

- $\mathcal{A}_{\textsf{Binom}}$ returns $\hat{\tau} = 0$ if the constraint is infeasible.
- For the PAC guarantee, we need to bound $\mathbb{P}\{y \notin C_\tau(x)\}$

PAC Learning Algorithm

$$\mathcal{A}_{\textsf{Binom}}: \qquad \hat{\tau} = \max_{\tau \in \mathbb{R}_{\geq 0}} \tau \qquad \text{subj. to} \qquad U_{\textsf{Binom}}(C_\tau, Z_n, \delta) \leq \varepsilon$$

- $\mathcal{A}_{\textsf{Binom}}$ returns $\hat{\tau} = 0$ if the constraint is infeasible.
- For the PAC guarantee, we need to bound $\mathbb{P}\{y \notin C_\tau(x)\}$
  - Bound the expected error via a concentration inequality!

PAC Learning Algorithm

$$\mathcal{A}_{\textsf{Binom}}: \qquad \hat{\tau} = \max_{\tau \in \mathbb{R}_{\geq 0}} \tau \qquad \text{subj. to} \qquad U_{\textsf{Binom}}(C_\tau, Z_n, \delta) \leq \varepsilon$$

- $\mathcal{A}_{\textsf{Binom}}$ returns $\hat{\tau} = 0$ if the constraint is infeasible.
- For the PAC guarantee, we need to bound $\mathbb{P}\{y \notin C_\tau(x)\}$
  - Bound the expected error via a concentration inequality!
- Recall that $U_{\textsf{Binom}}(C_\tau, Z_n, \delta)$ is the binomial tail bound, *i.e.*,

$$U_{\textsf{Binom}}(C_\tau, Z_n, \delta) := \inf \left\{ \theta \in [0, 1] \;\middle|\; F(E_\tau; n, \theta) \leq \delta \right\}$$

  - $F(k; n, \varepsilon)$: the cumulative distribution function of the binomial distribution with n trials and success probability ε
  - $E_\tau := \sum_{i=1}^{n} \mathbb{1}\left(y_i \notin C_\tau(x_i)\right)$

PAC Guarantee

Theorem (Vovk [2013], Park et al. [2020])

The algorithm $\mathcal{A}_{\mathit{Binom}}$ is PAC, i.e., for any f, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and $n \in \mathbb{Z}_{\geq 0}$, we have

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta,$$

where the inner probability is taken over a labeled example $(x, y) \sim \mathcal{D}$, the outer probability is taken over i.i.d. labeled examples $Z_n \sim \mathcal{D}^n$, and $\hat{C} = \mathcal{A}_{\mathit{Binom}}(Z_n)$.

PAC Guarantee

Theorem (Vovk [2013], Park et al. [2020])

The algorithm $\mathcal{A}_{Binom}$ is PAC, i.e., for any f, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and $n \in \mathbb{Z}_{\geq 0}$, we have

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta,$$

where the inner probability is taken over a labeled example $(x, y) \sim \mathcal{D}$, the outer probability is taken over i.i.d. labeled examples $Z_n \sim \mathcal{D}^n$, and $\hat{C} = \mathcal{A}_{Binom}(Z_n)$.

- Vovk [2013] provides the original proof.

PAC Guarantee

Theorem (Vovk [2013], Park et al. [2020])

The algorithm $\mathcal{A}_{Binom}$ is PAC, i.e., for any f, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and $n \in \mathbb{Z}_{\geq 0}$, we have

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta,$$

where the inner probability is taken over a labeled example $(x, y) \sim \mathcal{D}$, the outer probability is taken over i.i.d. labeled examples $Z_n \sim \mathcal{D}^n$, and $\hat{C} = \mathcal{A}_{Binom}(Z_n)$.

- Vovk [2013] provides the original proof.
- Park et al. [2020] interprets it in a learning-theoretic view

PAC Guarantee

Theorem (Vovk [2013], Park et al. [2020])

The algorithm $\mathcal{A}_{\textit{Binom}}$ is PAC, i.e., for any f, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and $n \in \mathbb{Z}_{\geq 0}$, we have

$$\mathbb{P}\left\{\mathbb{P}\left\{y \notin \hat{C}(x)\right\} \leq \varepsilon\right\} \geq 1 - \delta,$$

where the inner probability is taken over a labeled example $(x, y) \sim \mathcal{D}$, the outer probability is taken over i.i.d. labeled examples $Z_n \sim \mathcal{D}^n$, and $\hat{C} = \mathcal{A}_{\textit{Binom}}(Z_n)$.

- Vovk [2013] provides the original proof.
- Park et al. [2020] interprets it in a learning-theoretic view
- Park and Kim [2023] provides a simplified proof.

PAC Guarantee: A Proof Sketch

Define:

- C_τ: a prediction set C with a parameter τ
- $L(C_\tau) := \mathbb{P}\{y \notin C_\tau(x)\}$
- $\mathcal{H}_\varepsilon := \{\tau \in \mathbb{R}_{\geq 0} \mid L(C_\tau) > \varepsilon\}$ – Suppose that $\mathbb{R}_{\geq 0}$ is a set of finely quantized real numbers.
- $\tau^* := \inf \mathcal{H}_\varepsilon$

We have:

$$
\begin{aligned}
\mathbb{P}\left\{L(C_{\mathcal{A}_{\text{Binom}}(Z)}) > \varepsilon\right\} &\leq \mathbb{P}\left\{\exists \tau \in \mathcal{H}_\varepsilon, U_{\text{Binom}}(C_\tau, Z, \delta) \leq \varepsilon\right\} \\
&\leq \mathbb{P}\left\{U_{\text{Binom}}(C_{\tau^*}, Z, \delta) \leq \varepsilon\right\} \qquad (1) \\
&\leq \mathbb{P}\left\{L(C_{\tau^*}) > \varepsilon \wedge U_{\text{Binom}}(C_{\tau^*}, Z, \delta) \leq \varepsilon\right\} \\
&\leq \mathbb{P}\left\{L(C_{\tau^*}) > U_{\text{Binom}}(C_{\tau^*}, Z, \delta)\right\} \\
&\leq \delta, \qquad (2)
\end{aligned}
$$

- (1): $\mathbb{1}\left(y \notin C_\tau(x)\right)$ and U_{Binom} are non-decreasing in τ (*i.e.*, Lemma 2 in [Park et al., 2022])
- (2): the property of the binomial tail bound U_{Binom}.

Application: Image Classification

Qualitative Results

Certain → Uncertain

$\{\widehat{\text{king penguin}}\}$

$\left\{\begin{array}{c}\widehat{\text{Chihuahua}}, \\ \text{toy terrier}, \\ \text{Italian greyhound}, \\ \text{Boston bull}, \\ \text{miniature pinscher}\end{array}\right\}$

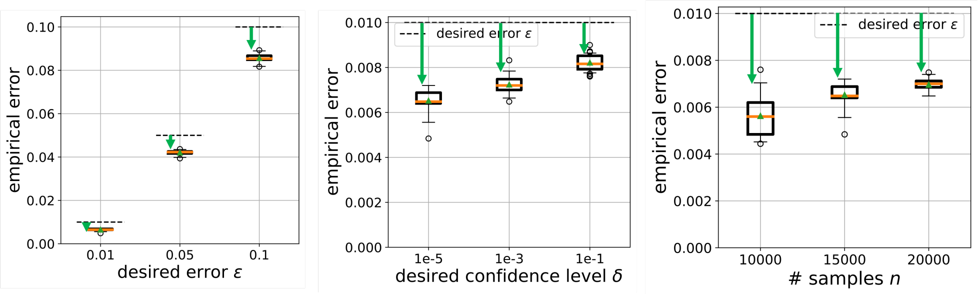
$\left\{\begin{array}{c}\text{barber chair}, \\ \text{hand blower}, \\ \text{medicine chest}, \\ \text{paper towel}, \\ \text{plunger}, \\ \text{shower curtain}, \\ \text{soap dispenser}, \\ \widehat{\text{toilet seat}}, \\ \text{tub, washbasin}, \\ \text{washer, toilet tissue}\end{array}\right\}$

$\widehat{\text{label}}$: predicted label, green: true label

- As an image (and a model's understanding) is uncertain, the set size gets larger.

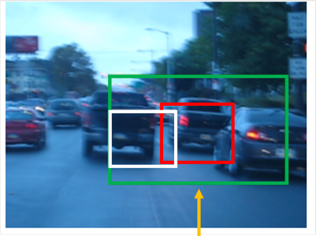
Application: Image Classification

Quantitative Results



$\delta = 10^{-5}, n = 20K$ | $\epsilon = 0.01, n = 20K$ | $\epsilon = 0.01, \delta = 10^{-5}$

Application: Regression



A point prediction fails, but a "conformal set" contains the true bounding box

White: Ground truth, Red: a point prediction, Green: Over-approximation of a conformal set

- The visualized conformal set is the bounding box that covers all boudning boxes in a conformal set.

Conclusion

- PAC conformal prediction constructs a conformal set with the PAC guarantee.
  - This is conformal prediction conditioned on a calibration set.
- Interesting questions:
  - Can we consider group-conditional conformal prediction?

Reference I

J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

S. Park and T. Kim. Pac neural prediction set learning to quantify the uncertainty of generative language models. *arXiv preprint arXiv:2307.09254*, 2023.

S. Park, O. Bastani, N. Matni, and I. Lee. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJxVI04YvB`.

S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=DhP9L8vIyLc`.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Reference II

V. Vovk. Conditional validity of inductive conformal predictors. *Machine learning*, 92(2-3): 349–376, 2013.

S. S. Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.