Trustworthy Machine Learning

Conformal Prediction

Sangdon Park

POSTECH

Conformal Prediction

Vladimir Vovk

FOLLOW

Department of Computer Science, Royal Holloway, University of London
Verified email at rhul.ac.uk - Homepage

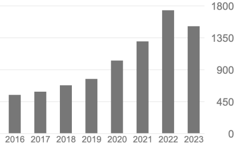
Machine Learning · Conformal Prediction · Foundations of Probability · Foundations of Statistics · Mathematical Finance

TITLE	CITED BY	YEAR
Algorithmic learning in a random world V Vovk, A Gammerman, G Shafer Springer	1435	2005
Ridge regression learning algorithm in dual variables C Saunders, A Gammerman, V Vovk	1044	1998
Aggregating strategies V Vovk Proceedings of 3rd Annu. Workshop on Comput. Learning Theory, 371-383	956	1990
A Tutorial on Conformal Prediction. G Shafer, V Vovk Journal of Machine Learning Research 9 (3)	853	2008

GET MY OWN PROFILE

Cited by — VIEW ALL

	All	Since 2018
Citations	14499	7052
h-index	58	41
i10-index	140	91

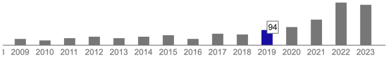


Conformal Prediction

Vladimir Vovk

Algorithmic learning in a random world

Authors	Vladimir Vovk, Alexander Gammerman, Glenn Shafer
Publication date	2005/3/1
Volume	29
Publisher	Springer
Description	This book is about conformal prediction, an approach to prediction that originated in machine learning in the late 1990s. The main feature of conformal prediction is the principled treatment of the reliability of predictions. The prediction algorithms described—conformal predictors—are provably valid in the sense that they evaluate the reliability of their own predictions in a way that is neither over-pessimistic nor over-optimistic (the latter being especially dangerous). The approach is still flexible enough to incorporate most of the existing powerful methods of machine learning. The book covers both key conformal predictors and the mathematical analysis of their properties. Algorithmic Learning in a Random World contains, in addition to proofs of validity, results about the efficiency of conformal predictors. The only assumption required for validity is that of" randomness"(the prediction algorithm is presented with …
Total citations	Cited by 1435



Scholar articles	Algorithmic learning in a random world V Vovk, A Gammerman, G Shafer - 2005 Cited by 1435 Related articles All 10 versions

> …we are *hedging* the prediction — we are adding to it a statement about how strongly we believe it.
> – Vovk et al., 2005

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

- Conventional prediction is a "point" prediction.

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

- Conventional prediction is a "point" prediction.
- Conformal prediction is a set-valued prediction.

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

- Conventional prediction is a "point" prediction.
- Conformal prediction is a set-valued prediction.
- The set contains "likely-correct" alternative options.
  - The set size measures "uncertainty"!

Motivation

Conventional prediction:

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

- Conventional prediction is a "point" prediction.
- Conformal prediction is a set-valued prediction.
- The set contains "likely-correct" alternative options.
  - The set size measures "uncertainty"!
- Why not confidence prediction? User-friendly?

Why "Conformal"?

mathoverflow

Home

PUBLIC

How is the "conformal prediction" conformal?

Asked 6 years, 5 months ago Modified 6 months ago Viewed 2k times

16

+100

Thanks for your interest. The term "conformal prediction" was suggested by Glenn Shafer, and at first I did not like it exactly for the reason that you mention: it has nothing (or very little) to do with conformal mappings in complex analysis. But then I discovered other meanings, even in maths; e.g., Wikipedia has five on its disambiguation page for "conformal":

- Conformal film on a surface (same thickness)
- Conformal fuel tanks on military aircraft
- Conformal coating in electronics
- Conformal hypergraph, in mathematics
- Conformal software, in ASIC Software

So the word did not look taken to me anymore. The expression that we had used before Glenn proposed "conformal prediction" was even worse ("transductive confidence machine").

Thanks to Hengrui Luo for drawing my attention to this question.

As for question (2), the answer depends on which robust predictors you have in mind. The predictors with most similar properties are the ones in classical statistics (such as the standard prediction intervals in linear regression based on Student's t distribution); the main difference is that they are parametric. There is a predictive version of tolerance intervals in nonparametric statistics, but their treatment of objects (x parts of observations (x,y), where y are labels) is limited. Upper bounds on the probability of error given by standard PAC predictors are often too high to be useful.

Share Cite Improve this answer Follow

answered Apr 13, 2017 at 7:35

Vladimir Vovk

Conformal (Prediction) Sets

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

Conformal (Prediction) Sets

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

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from Lei et al. [2018], Vovk et al. [2005], Tibshirani et al. [2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].

Conformal (Prediction) Sets

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

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from Lei et al. [2018], Vovk et al. [2005], Tibshirani et al. [2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].
- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: a conformity scoring function
  - Measures how well (x, y) *conforms* to a trained model f (via a proper training set)
  - $f(x, y)$ is a likelihood of x for being y

Conformal (Prediction) Sets

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

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from Lei et al. [2018], Vovk et al. [2005], Tibshirani et al. [2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].
- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: a conformity scoring function
  - Measures how well (x, y) *conforms* to a trained model f (via a proper training set)
  - $f(x, y)$ is a likelihood of x for being y
- q: A parameter to be chosen by an algorithm.

Conformity Scoring Functions I

Conformity scoring functions can be (almost) any model!

- classification (*e.g.*, resnet):

$$f(x,y) \coloneqq f_{\mathsf{cls}}(x,y)$$

  - f_{cls}: a classification model

Conformity Scoring Functions II

Conformity scoring functions can be (almost) any model!

- Standard regression (in 1-dimension):

$$f(x, y) := -|\mu(x) - y|$$

  - μ: a regressor

Conformity Scoring Functions III

Conformity scoring functions can be (almost) any model!

- Probabilistic regression (via a Gaussian model with a diagonal covariance matrix [Nix and Weigend, 1994]):

$$f(x, y) := \mathcal{N}(y; \mu(x), \sigma^2_{1:d}(x))$$

  - d: The dimension of $\mathcal{Y}$.
  - Implementation: $\mu(x) = f_{\mathsf{mu}}(x)$ and $\ln \sigma^2 = f_{\mathsf{var}}(x)$
    - ★ $f_{\mathsf{mu}}(x)$: a neural network
    - ★ $f_{\mathsf{var}}(x)$: a neural network

Back to Conformal Sets

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

- A conformity scoring function f is given.
- f is a target to measure uncertainty.
- How to choose q?

Assumption: Exchangeability

Assumption

A sequence of random variables $X_1, X_2, \dots$ is exchangeable if for any permutation σ, the following holds:

$$\mathbb{P}\left\{X_1, X_2, \dots\right\} = \mathbb{P}\left\{X_{\sigma(1)}, X_{\sigma(2)}, \dots\right\}.$$

Assumption: Exchangeability

Assumption

A sequence of random variables $X_1, X_2, \ldots$ is exchangeable if for any permutation σ, the following holds:

$$\mathbb{P}\left\{X_1, X_2, \ldots\right\} = \mathbb{P}\left\{X_{\sigma(1)}, X_{\sigma(2)}, \ldots\right\}.$$

- The i.i.d. assumption implies the exchangeability assumption (why?).

A Goodness Metric: Coverage Guarantee

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

A Goodness Metric: Coverage Guarantee

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set

A Goodness Metric: Coverage Guarantee

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n + 1$.

A Goodness Metric: Coverage Guarantee

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)

A Goodness Metric: Coverage Guarantee

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)
- $\hat{C}$: A conformal set constructed by the training set

A Goodness Metric: Coverage Guarantee

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)
- $\hat{C}$: A conformal set constructed by the training set
- $1 - \alpha \in (0, 1)$: A desired coverage rate

Quantile

Quantile of a Distribution

The level β quantile of a distribution F:

$$\text{Quantile}(\beta; F) := \inf\{z \mid \mathbb{P}\{Z \leq z\} \geq \beta\}$$

- F: a distribution over the augmented real line, $\mathbb{R} \cup \{\infty\}$
- $Z \sim F$
  - allows multiple instances of the same element

Quantile

Quantile of an Empirical Distribution

The level β quatile of an empirical distribution of the values $v_{1:n}$:

$$\text{Quantile}(\beta; v_{1:n}) := \text{Quantile}\left(\beta; \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}\right)$$

- $v_{1:n} := \{v_1, \ldots, v_n\}$: an unordered multiset
- δ_a: a δ-distribution (*i.e.*, a point mass at a)

Quantile

Quantile of an Empirical Distribution

The level β quatile of an empirical distribution of the values $v_{1:n}$:

$$\text{Quantile}(\beta; v_{1:n}) := \text{Quantile}\left(\beta; \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}\right)$$

- $v_{1:n} := \{v_1, \ldots, v_n\}$: an unordered multiset
- δ_a: a δ-distribution (*i.e.*, a point mass at a)

Lemma (Tibshirani et al. [2019])

If $V_1, \ldots, V_{n+1}$ are exchangeable random variables, then for any $\beta \in (0, 1)$, we have

$$\mathbb{P}\left\{V_{n+1} \leq \textit{Quantile}(\beta; V_{1:n} \cup \{\infty\})\right\} \geq \beta.$$

Quantile Lemma: A Proof Sketch I

- One fact about quantiles of a discrete distribution F with support points $a_1, \ldots, a_k \in \mathbb{R}$:
  - Denote $q := \text{Quantile}(\beta; F)$
  - Reassign the points a_i strictly larger than q to arbitrary values also strictly larger than q, yielding a new distribution $\tilde{F}$
  - Still we have $\text{Quantile}(\beta; F) = \text{Quantile}(\beta; \tilde{F})$
- Thus, we have

$$V_{n+1} > \text{Quantile}(\beta; V_{1:n} \cup \{\infty\}) \Longleftrightarrow V_{n+1} > \text{Quantile}(\beta; V_{1:n+1}).$$

- This implies

$$\begin{aligned} \mathbb{P}\Big\{V_{n+1} \leq \text{Quantile}(\beta; V_{1:n} \cup \{\infty\})\Big\} &= \mathbb{P}\Big\{V_{n+1} \leq \text{Quantile}(\beta; V_{1:n+1})\Big\} \\ &= \frac{\lceil \beta(n+1) \rceil}{n+1} \\ &\geq \beta. \end{aligned} \tag{1}$$

Quantile Lemma: A Proof Sketch II

- Why (1)? By exchangeability, we have for any integer $k \in \{1, \ldots, n+1\}$,

$$\mathbb{P}\{V_{n+1} \leq V_{[k]}\} = \frac{k}{n+1},$$

where $[k]$ is the k-th smallest value of $V_1, \ldots, V_{n+1}$.

  - Suppose there is no tie (see Kuchibhotla [2020] for a general proof). We have

$$\begin{aligned} \mathbb{P}\left\{V_{n+1} \leq V_{[k]}\right\} &= \mathbb{P}\left\{\bigvee_{i=1}^{k} V_{n+1} = V_{[i]}\right\} \\ &= \sum_{i=1}^{k} \mathbb{P}\left\{V_{n+1} = V_{[i]}\right\} \\ &= \sum_{i=1}^{k} \frac{n!}{(n+1)!} \\ &= \frac{k}{n+1}. \end{aligned} \tag{2}$$

Quantile Lemma: A Proof Sketch III

- Why (2)? For each permutation π, due to the exchangeability,

$$\begin{aligned}\mathbb{P}\{V_1 \leq \cdots \leq V_{n+1}\} &= \mathbb{P}\{(V_1, \ldots, V_{n+1}) \in A\} \\ &= \mathbb{P}\{(V_{\pi(1)}, \ldots, V_{\pi(n+1)}) \in A\} \\ &= \mathbb{P}\{V_{\pi(1)} \leq \cdots \leq V_{\pi(n+1)}\},\end{aligned}$$

where $A \coloneqq \{(x_1, \ldots, x_{n+1}) \mid x_1 \leq \cdots \leq x_{n+1}\}$.

Quantile Algorithm

Given $(X_1, Y_1), \ldots, (X_n, Y_n)$,

$$\hat{q}_{1-\alpha} \coloneqq \text{Quantile}(1-\alpha, V_{1:n} \cup \{\infty\}),$$

where $V_i \coloneqq -f(X_i, Y_i)$.

Coverage Guarantee of the Quantile Algorithm

Theorem (Vovk et al. [2005], Lei et al. [2018])

Assume that (X_i, Y_i) *for* $i \in \{1, \ldots, n+1\}$ *are exchangeable. For any scoring function* f *and any* $\alpha \in (0,1)$*, denote the conformal set by*

$$\hat{C}(x) \coloneqq \left\{ y \in \mathcal{Y} \middle| -f(x,y) \leq \hat{q}_{1-\alpha} \right\}.$$

Then, we have

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

where the probability is taken over (X_i, Y_i).

- This is a marginal coverage guarantee.

Coverage Guarantee of the Quantile Algorithm: A Proof Sketch

- Observe that

$$Y_{n+1} \in \hat{C}(X_{n+1}) \iff V_{n+1} \leq \text{Quantile}(1-\alpha, V_{1:n} \cup \{\infty\})$$

- Due to the symmetric construction of scores (using the same scoring function f), for any permutation π we have

$$(Z_1, \ldots, Z_{n+1}) \stackrel{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(n+1)}) \iff (V_1, \ldots, V_{n+1}) \stackrel{d}{=} (V_{\pi(1)}, \ldots, V_{\pi(n+1)})$$

  where $Z_i := (X_i, Y_i)$.
- As $(Z_1, \ldots, Z_{n+1})$ are exchangeable, so are $(V_1, \ldots, V_{n+1})$.
- Use the quantile lemma.

Power of Conformal Prediction

The coverage guarantee is drawn with minimal assumptions.

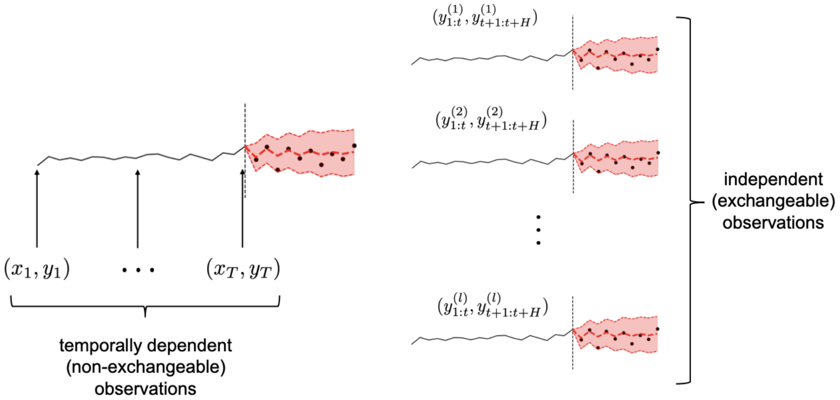
- It does not make assumptions on a distribution except for the exchangeability.
- The guarantee holds for any conformity scoring function.

Size of Conformal Sets

- Application-dependent issues
  - classification: set size
  - 1-D regression: interval length
  - multi-dimentional regression: *e.g.*, volume
- Larger set: uncertain (*e.g.*, the entire set)
- Smaller set: more certain (*e.g.*, a singleton)
- We will see some analysis in PAC conformal prediction.

Interesting Variation: Time-series Forecasting

Conformal Time-series Forecasting [Stankeviciute et al., 2021]



- Conformal prediction for independent time-series data
  - *e.g.*, temperature change for each year

Conformal Time-series Forecasting I

Problem

Setup:

- $y_{t:t'} := (y_t, y_{t+1}, \ldots, y_{t'}) \in \mathbb{R}^d \times \ldots \mathbb{R}^d$: a time-series of d-dimensional observation
  - Let $d = 1$
- H: a prediction horizon
- $\hat{y}_{t'+1:t'+H}$: predicted future observations (*e.g.*, the output of a RNN)
- $C_{t+h}(y_{1:t})$: a prediction interval at time $t + h$
  - $C_{t+h}(y_{1:t}) := [\hat{y}^L_{t+h}, \hat{y}^U_{t+h}]$

Conformal Time-series Forecasting II

Problem

Desired coverage guarantee:

$$\mathbb{P}\left\{\forall h \in \{1, \ldots, H\},\ y_{t+h} \in C_{t+h}(y_{1:t})\right\} \geq 1 - \alpha$$

- The probability is taken over $y_{1:t+H}$.
- $1 - \alpha$: a desired coverage rate

Goal: Find C_{t+h} for all $h \in \{1, \ldots, H\}$.

Conformal Time-series Forecasting

Approach

- $\mathcal{D} := \{(y_{1:T}^{(i)}, y_{T+1:T+H}^{(i)})\}_{i=1}^{m}$: a calibration set

Conformal Time-series Forecasting

Approach

- $\mathcal{D} := \{(y^{(i)}_{1:T}, y^{(i)}_{T+1:T+H})\}_{i=1}^m$: a calibration set
- Observe that

$$\mathbb{P}\left\{\exists h \in \{1,\ldots,H\},\ y_{t+h} \notin C_{t+h}(y_{1:t})\right\} \leq \sum_{h\in\{1,\ldots,H\}} \mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \quad (3)$$

$$\leq \alpha \quad (4)$$

Conformal Time-series Forecasting

Approach

- $\mathcal{D} := \{(y_{1:T}^{(i)}, y_{T+1:T+H}^{(i)})\}_{i=1}^{m}$: a calibration set
- Observe that

$$\mathbb{P}\left\{\exists h \in \{1,\ldots,H\},\ y_{t+h} \notin C_{t+h}(y_{1:t})\right\} \leq \sum_{h\in\{1,\ldots,H\}} \mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \tag{3}$$

$$\leq \alpha \tag{4}$$

  - (3) holds due to the union bound.

Conformal Time-series Forecasting

Approach

- $\mathcal{D} := \{(y_{1:T}^{(i)}, y_{T+1:T+H}^{(i)})\}_{i=1}^{m}$: a calibration set
- Observe that

$$\mathbb{P}\left\{\exists h \in \{1,\ldots,H\},\ y_{t+h} \notin C_{t+h}(y_{1:t})\right\} \le \sum_{h\in\{1,\ldots,H\}} \mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \tag{3}$$

$$\le \alpha \tag{4}$$

  - (3) holds due to the union bound.
  - (4) holds if $\mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \le \frac{\alpha}{H}$

Conformal Time-series Forecasting

Approach

- $\mathcal{D} := \{(y_{1:T}^{(i)}, y_{T+1:T+H}^{(i)})\}_{i=1}^{m}$: a calibration set
- Observe that

$$\mathbb{P}\left\{\exists h \in \{1, \ldots, H\},\ y_{t+h} \notin C_{t+h}(y_{1:t})\right\} \leq \sum_{h \in \{1,\ldots,H\}} \mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \tag{3}$$

$$\leq \alpha \tag{4}$$

  - (3) holds due to the union bound.
  - (4) holds if $\mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \leq \frac{\alpha}{H}$
- Due to the standard conformal prediction, we can find C_{t+h} such that

$$\mathbb{P}\{y_{t+h} \notin C_{t+h}(y_{1:t})\} \leq \frac{\alpha}{H}.$$

Conclusion

- Conformal prediction is a powerful tool to construct a prediction set (for measuring uncertainty) with correctness guarantees.
- Conformal prediction has many applications due to its "distribution-free" and "scoring-function-free" nature.
- The original conformal prediction framework can be extended to "conditional" cases (*e.g.*, PAC conformal prediction).

Reference I

A. K. Kuchibhotla. Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*, 2020.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

K. Stankeviciute, A. M Alaa, and M. van der Schaar. Conformal time-series forecasting. *Advances in neural information processing systems*, 34:6216–6228, 2021.

Reference II

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32:2530–2540, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.