Statistical Learning - module III

Aldo Solari aldo.solari@unimib.it

Course webpage: https://aldosolari.github.io/SL3/

Contents

Chapter 1

Conformal prediction

As a motivating example, suppose that each data point i corresponds to a patient, with a X_i encoding relevant covariates (age, family history, current symptoms, etc.), while the response Y_i measures a quantitative outcome (e.g., reduction in blood pressure after treatment with a drug). When a new patient arrives at the doctor's office with covariate values X_{n+1}, the doctor would like to be able to predict their eventual outcome Y_{n+1} with a range, making a statement along the lines of: "Based on your age, family history, and current symptoms, you can expect your blood pressure to go down by 10–15mmHg".

1.1 Marginal coverage vs conditional coverage

Throughout, P_{XY} will denote a joint distribution on $(X, Y) \in \mathbb{R} \times \mathbb{R}^d$, and we will write P_X to denote the induced marginal on X, and $P_{Y|X}$ for the conditional distribution of $Y|X$.

We would like that the predictive interval C_α *covers* the true response value with high probability for any distribution P_{XY}. Do we require that coverage holds with high probability on average over X_{n+1} or point-wise at any value $X_{n+1} = x$?

We say that C_α satisfies *distribution-free marginal coverage* at level $1 - \alpha$ if

$$\text{pr}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha$$

where the probability is w.r.t. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test point (X_{n+1}, Y_{n+1}), all drawn i.i.d. from P_{XY}. In other words, the probability that C_α covers the true test value Y_{n+1} is at least $1 - \alpha$, on average over a random draw of the training and test data from any distribution P_{XY}.

We say that C_α satisfies *distribution-free conditional coverage* at

$$\text{pr}(Y_{n+1} \in C_\alpha(X_{n+1}) | X_{n+1} = x) \geq 1 - \alpha$$

for all distributions P_{XY} and almost all x, where the probability is w.r.t. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the test response Y_{n+1} at a fixed test point $X_{n+1} = x$.

The difference is that for marginal coverage, the probability is taken over both X_{n+1} and Y_{n+1}, while for conditional coverage, X_{n+1} is fixed and the probability is taken over Y_{n+1} only (and over all the training data in both situations).

For practical purposes, marginal coverage does not seem to be sufficient. On the other hand, the problem of conditional inference is statistically very challenging, and is known to be incompatible with the distribution-free setting: see Foygel Barber et al. (2019).

Now, how should we interpret the difference between marginal and conditional coverage? With $\alpha = 0.05$, we expect that the doctor's statement ("...you can expect your blood pressure to go down by 10–15mmHg") should hold with 95% probability. For marginal coverage, the probability is taken over both X_{n+1} and Y_{n+1}, while for conditional coverage, X_{n+1} is fixed and the probability is taken over Y_{n+1} only (and over all the training data in both situations). This means that for marginal coverage, the doctor's statements have a 95% chance of being accurate on average over all possible patients that might arrive at the clinic (marginalizing over X_{n+1}), but might for example have 0% chance of being accurate for patients under the age of 25, as long as this is averaged out by a higher-than-95% chance of coverage for patients older than 25. The stronger definition of conditional coverage, on the other hand, removes this possibility, and requires that whatever statement the doctor makes (different for each patient) has a 95% chance of being true for every individual patient, regardless of the patient's age, family history, etc.

Chapter 2

Full conformal prediction

Conformal prediction, also known as *full conformal prediction*, is a general framework for constructing prediction sets in regression problems with finite sample and distribution free marginal coverage.

The main reference is the 2005 book Algorithmic Learning in a Random World (Vovk et al., 2005), although the subject was pioneered by Vladimir Vovk and collegues already in the 90s (Gammerman et al., 2013). Under very mild assumptions, conformal predictions sets provide exact coverage.

These appealing theoretical properties are contrasted by very high computational cost, which hinders its practical application.

2.1 Algorithm

Algorithm 2.1 describes how to compute the full conformal prediction interval.

Algorithm 1 Full conformal prediction

Require: (x_i, y_i), $i = 1, \ldots, n$, α, regression algorithm $\hat{\mu}$, new test point x_{n+1}, tentative values $\mathcal{Y} = \{y, y', y'', \ldots\}$

1: **for** $y \in \mathcal{Y}$ **do**
2: $\quad \hat{\mu}^y(x) = \hat{\mu}(x; (x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y))$
3: $\quad$ **for** $i = 1, \ldots, n$ **do**
4: $\quad\quad T_i^y = |y_i - \hat{\mu}^y(x_i)|$
5: $\quad$ **end for**
6: $\quad T_\alpha^y = T_{(k)}^y$ is the kth ordered statistic $T_{(1)}^y \leq \ldots \leq T_{(n)}^y$ with $k = \lceil (1-\alpha)(n+1) \rceil$.
7: $\quad R^y = |y - \hat{\mu}^y(x_{n+1})|$
8: **end for**
9: $\hat{C}_\alpha(x_{n+1}) = \{y \in \mathcal{Y} : R^y \leq T_\alpha^y\rceil\}$

2.2 Theory

Assume that $Z_i = (X_i, Y_i)$, $i \in [n+1]$ are $n+1$ independent identically distributed random vectors from a probability distribution P_{XY} on the sample space $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, where $[n]$ denotes $\{1, \ldots, n\}$. Suppose that the realizations $z_i = (x_i, y_i)$, $i \in [n]$ and x_{n+1} are available, and we want to predict Y_{n+1} based on x_{n+1}. We aim to construct a prediction set $C_\alpha(x) = C_\alpha(x; Z_i, i \in [n]) \subseteq \mathbb{R}$ such that its marginal coverage is at least $1 - \alpha$, i.e.

$$\text{pr}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha, \tag{2.1}$$

where the probability is taken over all Z_i, $i \in [n+1]$.

Let $\phi_\alpha = \phi_\alpha(Z) \in \{0, 1\}$ be a Bernoulli random variable, where $Z = (Z_i, i \in [n+1])$. Denote by $\phi_\alpha^y = \phi_\alpha(Z^y)$ with $Z^y = (Z_1, \ldots, Z_n, Z_{n+1}^y)$ and $Z_{n+1}^y = (X_{n+1}, y)$.

Theorem 2.2.1. *Assume that ϕ_α is a Bernoulli random variable such that $\text{E}(\phi_\alpha) \leq \alpha$. Then the prediction set*

$$C_\alpha(x) = \{y \in \mathbb{R} : \phi_\alpha^y = 0\},$$

satisfies (3.1). Exact coverage $\text{pr}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$ is obtained if and only if $\text{E}(\phi_\alpha) = \alpha$.

Proof. $\text{pr}(Y_{n+1} \notin C_\alpha(X_{n+1})) = \text{E}(\phi_\alpha) \leq \alpha$. □

Informally, ϕ_α^y can be thought of as a test for the null hypothesis that Y_{n+1} assumes the value of y, that is $H_y : Y_{n+1} = y$. Theorem 2.2.1 states that a valid prediction set can be obtained by inverting a collection of such tests.

In the next Theorem we provide an example of ϕ_α that can be used in Theorem 2.2.1. For ease of exposition, we will restrict our attention to the residual statistic

$$R = R(Z) = |Y_{n+1} - \hat{\mu}(X_{n+1})| \tag{2.2}$$

where $\hat{\mu}(x) = \hat{\mu}(x; Z)$ is an estimator of the regression function $\text{E}(Y|X)$ (Papadopoulos et al., 2002; Lei et al., 2018). In conformal prediction, the statistic R serves as a measure of conformity of the new observation of Y_{n+1} with the previously observed data, and the results presented here extend naturally to alternative choices of R.

For $i \in [n]$, define the i-th residual

$$T_i = T_i(Z) = |Y_i - \hat{\mu}(X_i)|. \tag{2.3}$$

Denote by $T_{(1)} \leq \ldots \leq T_{(n)}$ the sorted values of $T_1, \ldots, T_n$, with ties broken arbitrarily, and let

$$T_\alpha = T_{(k)} \tag{2.4}$$

with $k = \lceil (1 - \alpha)(n + 1) \rceil$, be the kth ordered statistic.

Theorem 2.2.2. *Assume that $\hat{\mu} = \hat{\mu}(\cdot, Z)$ is a function symmetric in Z. Then the Bernoulli variable $\phi_\alpha = \mathbb{1}\{R > T_\alpha\}$ satisfies $\mathbb{E}(\phi_\alpha) \leq \alpha$, where R and T_α are defined in (2.2) and (2.4), respectively. If $T_1, \ldots, T_n, R$ are almost surely distinct, then $\mathbb{E}(\phi_\alpha) = \alpha$ if and only if $\alpha \in \{1/(n+1), 2/(n+1), \ldots, n/(n+1)\}$, where T_i is defined in (2.3).*

Proof. Proof omitted. □

In spite of its elegance and theoretical appeal, the computational cost of the general method is still rather prohibitive: Theorem 2.2.2 with $\phi_\alpha^y = \mathbb{1}\{R^y > T_\alpha^y\}$ involves re-computing $R^y = |y - \hat{\mu}^y(x_{n+1})|$, $T_i^y = |y_i - \hat{\mu}^y(x_i)|$ and T_α^y at each tentative value of y since in general $\hat{\mu}^y(x) = \hat{\mu}^y(x; Z_1, \ldots, Z_n, Z_{n+1}^y)$ varies as y varies.

Chapter 3

Split conformal prediction

In spite of its elegance and theoretical appeal, the computational cost of full conformal prediction proved to be rather prohibitive in practical applications.

To address this issue, Papadopoulos et al. (2002) and Lei et al. (2018) have proposed inductive or *split conformal prediction* which successfully addresses the issue of computational efficiency, but at the cost of introducing extra randomness due to a one-time random split of the data.

3.1 Algorithm

Algorithm 2 describes how to compute the split conformal prediction set.

Algorithm 2 Split Conformal

Require: data $(x_1, y_1), \ldots, (x_n, y_n)$, x_{n+1}, validation sample size m, statistic R, level $\alpha \in (0, 1)$

1: split $[n]$ into L of size w and I of size $m = n - w$

2: compute $\{R_i\}_{i=1}^{m}$ and $R_\alpha = R_{(\lceil (1-\alpha)(m+1) \rceil)}$

return split conformal prediction set $C_\alpha(x_{n+1}) = \{y \in \mathbb{R} : R \leq R_\alpha\}$

3.2 Theory

Assume that $Z_i = (X_i, Y_i)$, $i \in [n+1]$ are $n+1$ independent identically distributed random vectors from a probability distribution P_{XY} on the sample space $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, where $[n]$ denotes $\{1, \ldots, n\}$. Suppose that the realizations $z_i = (x_i, y_i)$, $i \in [n]$ and x_{n+1} are available, and we want to predict Y_{n+1} based on x_{n+1}. We aim

to construct a prediction set $C_\alpha(x) = C_\alpha(x; Z_i, i \in [n]) \subseteq \mathbb{R}$ such that its marginal coverage is at least $1 - \alpha$, i.e.

$$\text{pr}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha, \tag{3.1}$$

where the probability is taken over all Z_i, $i \in [n+1]$.

Consider a partition of $[n]$ into a calibration set L of size w and a validation set I of size $m = n - w$, independently of the observed data values. Define a statistic

$$R = R(Z_L, Z_{n+1})$$

referred to as conformity score in conformal inference, to serve as a measure of plausibility of the value y as a realization of Y_{n+1} for the observed value of X_{n+1}, where $Z_L = (Z_i, i \in L)$. Examples include

$$R = |Y_{n+1} - \hat{\mu}_L(X_{n+1})|, \tag{3.2}$$

where $\hat{\mu}_L$ is an estimator of $\mathbb{E}(Y \mid X)$ based on Z_L (Papadopoulos et al., 2002; Lei et al., 2018) and

$$R = \max\left\{\hat{q}_L^{\gamma}(X_{n+1}) - Y_{n+1}, Y_{n+1} - \hat{q}_L^{1-\gamma}(X_{n+1})\right\}, \tag{3.3}$$

where $\hat{q}_L^{\gamma}$ is an estimator of the γ-quantile of $Y \mid X$ based on Z_L (Romano et al., 2019; Sesia and Candès, 2020). Denote the validation set by $I = \{j_1, \ldots, j_m\}$ and let

$$R_i = R(Z_L, Z_{j_i}), \quad i \in [m]. \tag{3.4}$$

For $\alpha \in (0, 1)$, define a quantile $R_\alpha = R_{\lceil (1-\alpha)(m+1) \rceil}$, where $R_1 \leq \ldots \leq R_m$ are ordered statistics obtained by sorting $R_1, \ldots, R_m$ in non-decreasing order with ties broken arbitrarily.

Lemma 3.2.1. *The Bernoulli variable $\phi_\alpha = \mathbb{1}\{R > R_\alpha\}$ satisfies $\mathrm{E}(\phi_\alpha) \leq \alpha$. If $R_1, \ldots, R_m, R$ are almost surely distinct, then $\mathrm{E}(\phi_\alpha) = \alpha$ if and only if $\alpha \in \{1/(m+1), 2/(m+1), \ldots, m/(m+1)\}$.*

In particular, for R defined as in (3.2) and (3.3), Algorithm 2 returns $C_\alpha(x_{n+1}) = [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha]$ and $C_\alpha(x_{n+1}) = [\hat{q}_L^{\gamma}(x_{n+1}) - R_\alpha, \hat{q}_L^{1-\gamma}(x_{n+1}) + R_\alpha]$, respectively. The former is always an interval, whereas the latter is either an interval or an empty set, i.e. $C_\alpha(x_{n+1}) = \emptyset$ if and only if $R_\alpha < (1/2)[\hat{q}_L^{\gamma}(x_{n+1}) - \hat{q}_L^{1-\gamma}(x_{n+1})]$ (Gupta et al., 2019).

3.3 Proofs

We provide an explicit formulation of split conformal prediction within the permutation framework (Fisher, 1925).

Consider the group of transformations $\Sigma = \{\sigma_1, \ldots, \sigma_{m+1}\}$ whose $m+1$ elements are restricted permutations consisting of swapping the index $n+1$ with another index of $I \cup \{n+1\} = (j_1, \ldots, j_m, n+1)$, i.e. for $i \in [m]$, $\sigma_i = (\sigma_i(j_1), \ldots, \sigma_i(j_m), \sigma_i(n+1))$ is such that $\sigma_i(n+1) = j_i$, $\sigma_i(j_i) = n+1$ and $\sigma_i(j_k) = j_k$ for $j_k \neq j_i$. Here σ_{m+1} denotes the identity permutation. Note that Σ is a group with respect to the operation of composition of transformations: Σ contains an identity element; every element of Σ has an inverse in Σ; for all $\sigma, \tilde{\sigma} \in \Sigma$, $\sigma \circ \tilde{\sigma} \in \Sigma$.

For any $\sigma \in \Sigma$, let $\sigma Z = (Z_1^*, \ldots, Z_{n+1}^*)$ be the transformed vector with $Z_i^* = Z_{\sigma(i)}$ if $i \in I \cup \{n+1\}$ and $Z_i^* = Z_i$ otherwise, and let

$$R(\sigma Z) = R(Z_L, Z_{\sigma(n+1)})$$

be the statistic R calculated on σZ.

Since $Z_1, \ldots, Z_{n+1}$ are independent and identically distributed by assumption, $Z \overset{d}{=} \sigma Z$ holds for every $\sigma \in \Sigma$. This implies the group invariance condition (Hoeffding, 1952; Lehmann and Romano, 2006; Hemerik and Goeman, 2018, 2020):

$$(R(\sigma_1 Z), \ldots, R(\sigma_{m+1} Z)) \overset{d}{=} (R(\sigma_1 \circ \sigma Z), \ldots, R(\sigma_{m+1} \circ \sigma Z)) \tag{3.5}$$

for every $\sigma \in \Sigma$, where $\overset{d}{=}$ denotes equality in distribution. Note that (3.5) is implied by exchangeability of $(R_1, \ldots, R_m, R)$ (Commenges, 2003; Kuchibhotla, 2020), where $R_1 = R(\sigma_1 Z)$, …, $R_m = R(\sigma_m Z)$, $R = R(\sigma_{m+1} Z)$.

For $\alpha \in (0, 1)$, let $\tilde{R}_{(1)} \leq \ldots \leq \tilde{R}_{(m+1)}$ be the sorted values of $R_1, \ldots, R_m, R$, with ties broken arbitrarily, and let

$$\tilde{R}_\alpha = \tilde{R}_{(k)}$$

with $k = \lceil (1-\alpha)(m+1) \rceil$, be the the kth ordered statistic. We have $\tilde{R}_\alpha > R_\alpha$ if $R \leq \tilde{R}_\alpha$ and $\tilde{R}_\alpha = R_\alpha$ if $R > \tilde{R}_\alpha$. Then $\phi_\alpha = \mathbb{1}\{R > \tilde{R}_\alpha\} = \mathbb{1}\{R > R_\alpha\}$.

From Theorem 1 in Hemerik and Goeman (2018), we obtain that the Bernoulli variable $\phi_\alpha = \mathbb{1}\{R > \tilde{R}_\alpha\}$ satisfies $\mathrm{E}(\phi_\alpha) \leq \alpha$.

Finally, Condition 1 and Proposition 1 in Hemerik and Goeman (2018) ensure that if $R_1, \ldots, R_m, R$ are almost surely distinct, then $\mathrm{E}(\phi_\alpha) = \alpha$ if and only if $\alpha \in \{1/(m+1), 2/(m+1), \ldots, m/(m+1)\}$.

Chapter 4

Aggregated conformal prediction

Split conformal prediction is a computationally efficient method for performing distribution-free predictive inference in regression. It involves, however, a one-time random split of the data, and the result can strongly depend on the particular split. This kind of randomness of the prediction interval parallels the "p-value lottery" discussed in Meinshausen et al. (2009).

An illustration of the potential impact of a single random split is shown in Figure 4 featuring 10 prediction intervals obtained from 10 different data splits.

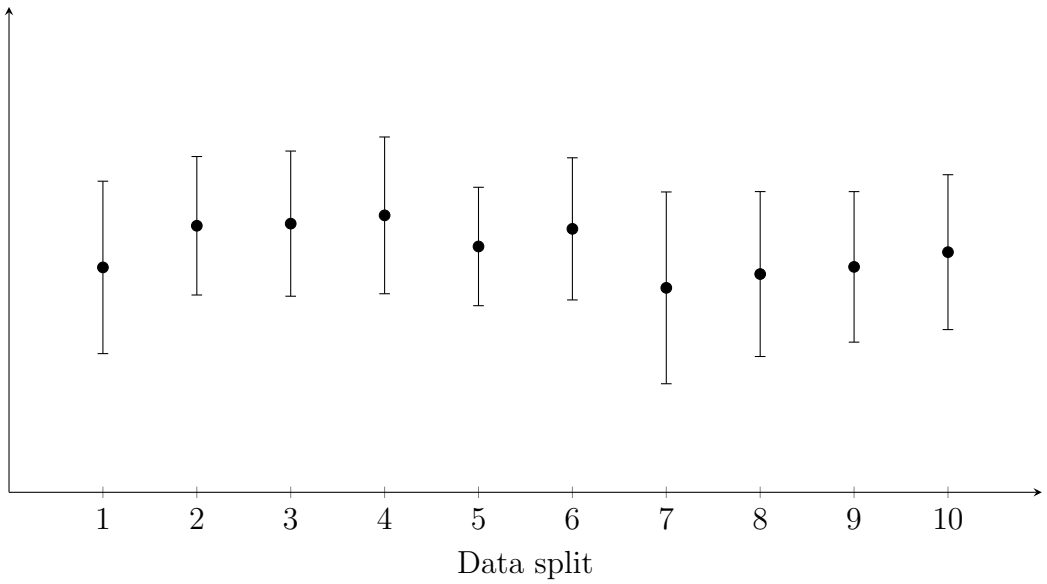

Figure 4.1: Realizations of 10 split conformal prediction intervals $C_\alpha(x)$ for the same test point x.

A straightforward strategy for overcoming this problem is to aggregate results obtained from multiple data splits. The idea of *aggregated conformal prediction* was introduced by Carlsson et al. (2014). Methods with proven coverage guarantees include K-fold cross-conformal prediction of Vovk (2015), jackknife+ and K-fold CV+ of Barber et al. (2021), and K-subsample conformal of Gupta et al. (2019). See Table 4 for an overview.

Method	*Coverage*	*Reference*
Cross-conformal	$\geq 1 - 2\alpha - a(n, K)$	Vovk (2015)
Jackknife+/CV+	$\geq 1 - 2\alpha - \min\{a(n, K), b(n, K)\}$	Barber et al. (2021)
Subsampling conformal	$\geq \min\{2, K\}\alpha$	Gupta et al. (2019)

Table 4.1: Aggregated conformal prediction methods with proven coverage guarantees, where $a(n, K) = (2 - 2/K)/(n/K + 1)$ and $b(n, K) = (1 - K/n)/(K + 1)$.

The coverage guarantees of all methods listed in Table 4 are based on the fact that double of the average p-value is a valid p-value, a result established by Rüschendorf (1982) and discussed in Vovk and Wang (2020). Only the factor $b(n, K)$ derived in Theorem 4 of Barber et al. (2021) is based on a different argument that makes use of Landau's theorem for tournaments (Landau, 1953).

4.1 Multi split conformal prediction

The multi split approach consists in constructing single split prediction sets multiple times, and proceeds with aggregating the results by including those points that are included in single split prediction intervals with frequency greater than a threshold.

We proceed as follows: choose the number of splits $B \in \mathbb{N}$. Partition $[n]$ into $L^{[b]}$ of size $w^{[b]}$ and $I^{[b]}$ of size $m^{[b]} = n - w^{[b]}$, independently of the observed data values, and choose a statistic $R^{[b]}$, for $b = 1, \ldots, B$. For $\beta \in (0, 1)$, variable $\phi_\beta^{[b]} = \mathbb{1}\{R^{[b]} > R_\beta^{[b]}\}$ has expected value $\mathbb{E}(\phi_\beta^{[b]}) \leq \beta$ by Lemma 3.2.1. Let

$$V_\beta = \sum_{b=1}^{B} \phi_\beta^{[b]} \tag{4.1}$$

be the number of successes (1s). The following Theorem provides an upper bound for $\mathbb{P}(V_\beta \geq k)$, the probability of at least k successes out of B trials.

By Markov's inequality,

$$\mathbb{P}(V_\beta \geq k) \leq \frac{\mathrm{E}(V_\beta)}{k} = \frac{B\beta}{k}. \tag{4.2}$$

This result can be used to aggregate results of split conformal inference performed over a number of different data splits. Let

$$\Pi_\beta = 1 - \frac{V_\beta}{B} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{Y \in C_\beta^{[b]}(X)\}$$

be the proportion of prediction sets $C_\beta^{[b]}(X)$ that include Y. For $\alpha \in (0,1)$ and a threshold $\tau = 1 - k/B$, the multi split conformal prediction set defined as

$$C_\alpha^\tau(x) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\} \tag{4.3}$$

has coverage at least $1-\alpha$ by Theorem 2.2.1 with $\phi_\alpha = \mathbb{1}(\Pi_\beta \leq \tau)$, where $\beta = \alpha(1-\tau)$.

Algorithm 3 describes how to compute the multi split conformal prediction set.

Algorithm 3 Multi split conformal prediction

Require: data $(x_1, y_1), \ldots, (x_n, y_n)$, x, number of splits $B \in \mathbb{N}$, inference sample sizes $(m^{[b]})_{b=1}^B$, statistics $(R^{[b]})_{b=1}^B$, threshold $\tau \in [0, (B-1)/B]$, level $\alpha \in (0,1)$.

1: **for** $b \leftarrow 1$ to B **do**
2: compute $C_\beta^{[b]}(x)$ using Algorithm 2 with $m^{[b]}$, $R^{[b]}$ and level $\beta = \alpha(1-\tau)$
3: **end for**
4: compute

$$C_\alpha^\tau(x) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\}$$

return multi split conformal prediction set $C_\alpha^\tau(x)$

In general, C_α^τ is not guaranteed to be an interval, even when single split prediction sets $C_\beta^{[b]}$ are all intervals. To compute C_α^τ efficiently, one can use Algorithm 1 in Gupta et al. (2019). If the computation of each single split interval $C_\beta^{[b]}$ takes time $\leq T$, the overall time to compute C_α^τ is $O(B \log B) + BT$.

The parameter τ can be regarded as a tuning parameter, and proper choice of τ is essential for good performance. Consider the case without assumption, i.e. $\lambda = 0$. On the one hand, setting $\tau = 1 - 1/B$ gives the Bonferroni-intersection method of Lei et al. (2018) with $C_\alpha^{1-1/B} = \bigcap_b C_{\alpha/B}^{[b]}$. On the other hand, setting $\tau = 0$ gives an unadjusted-union $C_\alpha^0 = \bigcup_b C_\alpha^{[b]}$.

An intermediate choice $\tau = 1/2$ amounts to constructing B single split confidence intervals at level $\alpha/2$, that is $C_{\alpha/2}^{[b]}$, which is a small but not negligible price to pay for using multiple splits rather than just one split.

Notice that the multi split prediction set C_α^τ is very flexible because it allows to use splits of different proportions m/n and possibly different statistics R across splits.

Bibliography

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *Annals of Statistics*, 49(1):486–507.

Carlsson, L., Eklund, M., and Norinder, U. (2014). Aggregated conformal prediction. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 231–240. Springer.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of nonparametric statistics*, 15(2):171–185.

Fisher, R. A. (1925). Statistical methods for research workers. *Statistical methods for research workers.*

Foygel Barber, R., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*.

Gammerman, A., Vovk, V., and Vapnik, V. (2013). Learning by transduction. *arXiv preprint arXiv:1301.7375*.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. (2019). Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562*.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *Test*, 27(4):811–825.

Hemerik, J. and Goeman, J. J. (2020). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192.

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

Landau, H. (1953). On dominance relations and the structure of animal societies: Iii the condition for a score structure. *The bulletin of mathematical biophysics*, 15(2):143–148.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

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

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

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

Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression. *arXiv preprint arXiv:1905.03222*.

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability*, pages 623–632.

Sesia, M. and Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28.

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

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.