

Federated Conformal Prediction General

Min, Xia

June 27, 2024

1 Conformal Prediction General[1]

Definition 1.1 (Exchangeability). [6] For any r.v. x_1, \dots, x_k , we say they are exchangeable if for any permutation $\sigma : [k] \rightarrow [k]$ (bijection), $(x_1, \dots, x_k) \stackrel{d.}{=} (x_{\sigma(1)}, \dots, x_{\sigma(k)})$.

Definition 1.2 (Weighted Exchangeability). [7] For any r.v. x_1, \dots, x_k , we say they are weighted exchangeable if their joint density can be factorized as

$$f(x_1, \dots, x_k) = \prod_{i=1}^k w_i(x_i) \cdot g(x_1, \dots, x_k),$$

where g is exchangeable, i.e., $g(x_1, \dots, x_k) = g(x_{\sigma(1)}, \dots, x_{\sigma(k)})$.

For conformal prediction two classes of targets are studied.

Definition 1.3 (Marginal Coverage). $(X, Y) \in \mathbb{R}^p \times \mathbb{R} \sim P_{XY}$ which is unknown. Given training set $Tr = \{(X_i, Y_i)\}_{i=1}^n$, and test on (X_{n+1}, Y_{n+1}) , both i.i.d.

C_α satisfies distribution-free marginal coverage at level $1 - \alpha$ if

$$P(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha, \quad \forall P_{XY}$$

The probability is with respect to $\{(X_i, Y_i)\}_{i=1}^{n+1}$.

Definition 1.4 (Conditional Coverage). $(X, Y) \in \mathbb{R}^p \times \mathbb{R} \sim P_{XY}$ which is unknown. Given training set $Tr = \{(X_i, Y_i)\}_{i=1}^n$, and test on (X_{n+1}, Y_{n+1}) , both i.i.d.

C_α satisfies distribution-free marginal coverage at level $1 - \alpha$ if

$$P\left(Y_{n+1} \in C_\alpha(X_{n+1}) \middle| X_{n+1} = x\right) \geq 1 - \alpha, \forall P_{XY}$$

The probability is with respect to $\{(X_i, Y_i)\}_{i=1}^n$ and Y_{n+1} .

Definition 1.5 (Conditional Validity). [8] Call a conformal set C_α is (ε, δ) valid if

$$P(\{X : P(Y \in C_\alpha(X)) \geq 1 - \varepsilon\}) \geq 1 - \delta,$$

which means we have enough, probability $1 - \delta$, X makes conditional coverage is guaranteed.

Definition 1.6 (Conformal Score Function). For data pair (X, Y) and point predictor and any loss function $V(\cdot, \cdot)$, call $R = S(X, Y) = V(Y, \hat{f}(X))$ be the conformal score (or residual).

Definition 1.7 (Efficiency). X is some r.v. following the testing distribution and C_α is efficient if $\mathbb{E}[|C_\alpha(X)|]$ is small. Define $\text{Size}(C_\alpha) = \frac{1}{n} \sum_{i=1}^n |C_\alpha(X_i)|$.

2 Standard Split Conformal Prediction

- First divide training set D into two sets: D_1 for proper training set and D_2 for calibration set. And let $n_i = |D_i|$, fit point predictor \hat{f}_1 on D_1 .
- Calculate residuals on D_2 : $R_i = |Y_i - \hat{f}_1(X_i)|$, $i \in D_2$.
- Find quantile on calibration residuals: $\hat{q}_2 = \lceil (1 - \alpha)n_2 \rceil$ smallest of R_i , $i \in D_2$.
- Construct a conformal set: $C_\alpha(x) = [\hat{f}_1(x) - \hat{q}_2, \hat{f}_1(x) + \hat{q}_2]$.

Let $R_{n+1} = |Y_{n+1} - \hat{f}_1(X_{n+1})|$. Let rank statistic $R_{(j)}$ be the j -th smallest in R_i , $i \in D_2$, and $k_\alpha = \lceil (1 - \alpha)(n_2 + 1) \rceil$. As

$$\{Y_{n+1} \in C_\alpha(X_{n+1})\} = \{R_{n+1} \leq \hat{q}_2\} = \{R_{n+1} \leq R_{(k_\alpha)}\},$$

and R_i , $i \in D_2$, R_{n+1} are exchangeable, we have

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha(X_{n+1}) \middle| D_1\right) \in \left[1 - \alpha, 1 - \alpha + \frac{1}{n_2 + 1}\right].$$

Assume a more general score function $V(x, y) = V((x, y); \hat{f}_1)$, define $R_i = V(X_i, Y_i)$ and change the conformal set to

$$C_\alpha(x) = \{y : S(x, y) = V(y, f(x)) \leq R_{(k_\alpha)}\}.$$

Remark 2.1. *Further condition on calibration set, which means conditioning on entire training set D and assume $R = V(x, y)$ has distribution F . As*

$$\{Y_{n+1} \in C_\alpha(X_{n+1})\} = \{R_{n+1} \leq R_{(k_\alpha)}\},$$

Assume the distribution function of $R_{(j)}$ is $F_{(j)}$, and we have

$$\mathbb{P}\left(\mathbb{P}\left(Y_{n+1} \in C_\alpha(X_{n+1}) \middle| D\right) \leq t\right) = \mathbb{P}\left(\mathbb{P}\left(R_{n+1} \leq R_{(k_\alpha)} \middle| D\right) \leq t\right)$$

$$\text{condition on } D \text{ randomness comes from } R_{n+1}, = \mathbb{P}\left(F(R_{(k_\alpha)}) \leq t\right)$$

$$= \mathbb{P}\left(R_{(k_\alpha)} \leq F^{-1}(t)\right)$$

$$= F_{(k_\alpha)}(F^{-1}(t)) \quad (1)$$

rank statistic has density $F'_{(j)}(x) = jC_{n_2}^j x^{j-1}(1-x)^{n-j}f(x)$, thus take derivative on formula (1), and $\mathbb{P}\left(Y_{n+1} \in C_\alpha(X_{n+1}) \middle| D\right)$ has density

$$k_\alpha C_{n_2}^{k_\alpha} t^{k_\alpha-1} (1-t)^{n-k_\alpha}.$$

3 Standard Full Conformal Prediction

Full CP has similar steps as split CP. It uses all data points for training.

- Fix any x and trial data y to construct training set $\{(X_1, Y_1), \dots, (X_n, Y_n), (x, y)\}$.
- Train point predictor \hat{f} on training set and define residuals $R_i = |Y_i - \hat{f}(X_i)|$, $i \in [n]$, $R_{n+1} = |y - \hat{f}(x)|$.
- Define j -th rank statistic of R_i , $i \in [n]$ as $R_{(j)}$, $k_\alpha = \lceil (1-\alpha)(n_2+1) \rceil$, and conformal set

$$C_\alpha(x) = \{y : R_{n+1} \leq R_{(k_\alpha)}\}.$$

As $\{Y_{n+1} \in C_\alpha(X_{n+1})\} = \{R_{n+1} \leq R_{(k_\alpha)}\}$, and the exchangeability of data

$$\mathbb{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \in \left[1 - \alpha, 1 - \alpha + \frac{1}{n+1}\right].$$

4 Standard CP under covariate shift

Follow the procedure of split CP and heterogeneity between training and test data[7].

Assume

$$Z_i = (X_i, Y_i) \sim P = P_X \times P_{Y|X}, i = 1, \dots, n,$$

$$Z_{n+1} = (X_{n+1}, Y_{n+1}) \sim P' = P'_X \times P_{Y|X}.$$

- Fix any trial data y to construct training set $\{(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)\}$. Train point predictor \hat{f} on new training set.
- Calculate nonconformity scores $R_i = S(X_i, Y_i)$, $i \in \{1, \dots, n\}$, $R_{n+1} = S(X_{n+1}, y)$ based on \hat{f} .
- Calculate importance weights p_i based on likelihood ratio w :

$$w(x) = \frac{dP'_X(x)}{dP_X(x)},$$

$$p_i = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}, i = 1, \dots, n+1.$$

- Calculate $1 - \alpha$ quantile of distribution $\sum_{i=1}^n p_i \delta_{R_i} + p_{n+1} \delta_\infty$ as q_α . Define conformal set $C_\alpha(x) = \{y : R_{n+1} \leq q_\alpha\}$.

All independent variables are weighted exchangeable. Let E_Z be $\{Z_1, \dots, Z_{n+1}\} = \{z_1, \dots, z_{n+1}\}$. Assume joint density is $f(z_1, \dots, z_{n+1}) = \prod_{i=1}^{n+1} dP(z_i) \cdot w(x_{n+1})$. Condition on E_Z , calculate R_i based on \hat{f} and z_i , for all permutation σ

$$\mathbb{P}(R_{n+1} = r_i | E_Z) = \mathbb{P}(Z_{n+1} = z_i | E_Z) = \frac{\sum_{\sigma(n+1)=i} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}{\sum_{\sigma} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})} = p_i,$$

which leads to $R_{n+1} | E_Z \sim \sum_{i=1}^{n+1} p_i \delta_{r_i}$. Let $Q(1 - \alpha, F)$ be the quantile function,

$$\mathbb{P}\left(R_{n+1} \leq Q(1 - \alpha, \sum_{i=1}^{n+1} p_i \delta_{r_i}) | E_Z\right) \geq 1 - \alpha,$$

means

$$\mathbb{P} \left(R_{n+1} \leq Q(1 - \alpha, \sum_{i=1}^n p_i \delta_{r_i} + p_{n+1} \delta_{\infty}) \middle| E_Z \right) \geq 1 - \alpha,$$

as condition on E_Z , $\sum_{i=1}^n p_i \delta_{r_i} + p_{n+1} \delta_{\infty} = \sum_{i=1}^n p_i \delta_{R_i} + p_{n+1} \delta_{\infty}$ (left p is based on z and right based on Z). The p_i in following formula is different from previous one.

$$\mathbb{P} \left(R_{n+1} \leq Q(1 - \alpha, \sum_{i=1}^n p_i \delta_{R_i} + p_{n+1} \delta_{\infty}) \middle| E_Z \right) \geq 1 - \alpha,$$

thus taking expectation on all E_Z ,

$$\mathbb{P}(Y_{n+1} \in C_{\alpha}(X_{n+1})) = \mathbb{P}(R_{n+1} \leq q_{\alpha}) \geq 1 - \alpha$$

5 Standard CP under Nonexchangeability

Conformal Prediction Beyond Exchangeability[2]

When data is not exchangeable, like covariate shift setting, standard CP is not valid. This article gives a more general method under unknown distribution nonexchangeability.

Definition 5.1 (Total Variation of Distribution). *Given two distribution with density $p(x)$, $q(x)$ the total variation is defined as*

$$d_{TV}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx$$

Assume data $Z = \{(Z_1 = (X_1, Y_1), t_1), \dots, (Z_{n+1} = (X_{n+1}, Y_{n+1}), t_{n+1})\}$ and Z^k be data sequence swap Z_k and Z_{n+1} in Z (t remains the same), $Z^{n+1} = Z$. Train a asymmetric point estimator based on Z^k call \hat{f}^k . And let the residual of sample i based on \hat{f}^k be $R_i^k = |Y_i - \hat{f}^k(X_i)|$. Given weights w_i , $i = 1, \dots, n+1$ and choose $K = i$ with probability w_i . Define conformal set

$$C_{\alpha}(X_{n+1}) = \left\{ y : R_{n+1}^K \leq Q(1 - \alpha, \sum_{i=1}^{n+1} w_i \delta_{R_i^K}) \right\}.$$

First calculate

$$\begin{aligned} \{Y_{n+1} \notin C_{\alpha}(X_{n+1})\} &= \left\{ R_{n+1}^K > Q(1 - \alpha, \sum_{i=1}^{n+1} w_i \delta_{R_i^K}) \right\} \\ &= \left\{ R_{n+1}^K > Q(1 - \alpha, \sum_{i=1}^n w_i \delta_{R_i^K} + w_{n+1} \delta_{\infty}) \right\} \end{aligned}$$

notice

$$\begin{aligned} \sum_{i=1}^n w_i \delta_{R_i^K} + w_{n+1} \delta_\infty &= \sum_{i \neq K} w_i \delta_{R_i^K} + w_K (\delta_{R_K^K} + \delta_\infty) + (w_{n+1} - w_K) \delta_\infty \\ \text{As } \delta_\infty \leq \delta_x, \forall x \in \mathbb{R}, &\leq \sum_{i \neq K} w_i \delta_{R_i^K} + w_K (\delta_{R_K^K} + \delta_{R_{n+1}^K}) + (w_{n+1} - w_K) \delta_\infty \\ \text{As } w_{n+1} \geq w_K, &\leq \sum_{i \neq K} w_i \delta_{R_i^K} + w_K (\delta_{R_K^K} + \delta_{R_{n+1}^K}) + (w_{n+1} - w_K) \delta_{R_K^K} \\ &= \sum_{i \neq K} w_i \delta_{R_i^K} + w_K \delta_{R_{n+1}^K} + w_{n+1} \delta_{R_K^K}, \end{aligned}$$

thus

$$Q(1 - \alpha, \sum_{i=1}^n w_i \delta_{R_i^K} + w_{n+1} \delta_\infty) \geq Q(1 - \alpha, \sum_{i \neq K} w_i \delta_{R_i^K} + w_K \delta_{R_{n+1}^K} + w_{n+1} \delta_{R_K^K}),$$

and

$$\{Y_{n+1} \notin C_\alpha(X_{n+1})\} \subset \left\{ R_{n+1}^K \geq Q(1 - \alpha, \sum_{i \neq K} w_i \delta_{R_i^K} + w_K \delta_{R_{n+1}^K} + w_{n+1} \delta_{R_K^K}) \right\}.$$

Let $R^k = (R_1^k, \dots, R_{k-1}^k, R_{n+1}^k, R_{k+1}^k, \dots, R_n^k, R_k^k)$, $r = (r_1, \dots, r_{n+1})$ and $S(r) = \{j \in [n+1] : r_j > Q(1 - \alpha, \sum_{i=1}^{n+1} w_i \delta_{r_i})\}$. As assume $r = R^k$, $r_k = R_{n+1}^k$

$$\left\{ R_{n+1}^K \geq Q(1 - \alpha, \sum_{i \neq K} w_i \delta_{R_i^K} + w_K \delta_{R_{n+1}^K} + w_{n+1} \delta_{R_K^K}) \right\} = \{K \in S(R^K)\}.$$

Finally we have

$$\begin{aligned} P(\{Y_{n+1} \notin C_\alpha(X_{n+1})\}) &\leq P(K \in S(R^K)) \\ &= \sum_{i=1}^{n+1} w_i P(i \in S(R^i)), \end{aligned}$$

notice

$$P(i \in S(R^i)) = \int_{\Omega_i} dP(R^i) \leq \int_{\Omega_i} dP(R^{n+1}) + \int \left| dP(R^i) - dP(R^{n+1}) \right|,$$

thus

$$\begin{aligned} P(\{Y_{n+1} \notin C_\alpha(X_{n+1})\}) &\leq P(K \in S(R^K)) \\ &= \sum_{i=1}^{n+1} w_i P(i \in S(R^{n+1})) + \sum_{i=1}^n w_i d_{TV}(R^i, R^{n+1}). \end{aligned}$$

6 Federated Conformal Prediction Article1

Efficient Conformal Prediction under Data Heterogeneity[5]

Idea: The marginal coverage is measured over all training data and test points. However, if there is a high variability in the coverage probability as a function of the training data, the test coverage probability may be substantially below $1 - \alpha$ for a particular training set.

Definition 6.1 (empirical miscoverage rate). $\alpha(Tr) = P(Y_{n+1} \notin C_\alpha(X_{n+1}) | Tr)$

In this article, assume n agents each has calibration data $(X_k^i, Y_k^i) \sim P_X^i P_{Y|X}$, $k = 1, \dots, n^i$, $i = 1, \dots, n$, and calibration set $D_i = \{(X_k^i, Y_k^i)\}_{k=1}^{n^i}$, $i = 1, \dots, n$. Let calibration distribution be $P^{cal} = \sum_{i=1}^n \pi_i P_X^i P_{Y|X}$, where $\pi_i = n_i / \left(\sum_{j=1}^n n_j \right)$, and the test distribution $P^{test} = P_X^{n+1} P_{Y|X}$. Let the general density ratio be $w(x, y) = \frac{dP_X^{n+1}(x)}{\sum_{i=1}^n \pi_i dP_X^i(x)}$.

- Utilize the GMM to compute parameters $\{\pi_y^i, \mu_y^i, \Sigma_y^i\}_{y \in \mathcal{Y}^i}$ on D_i . Note that P_X^i is approximated by $|\mathcal{Y}^i|$ centers mixed GMM, $P_X^i = \sum_{y \in \mathcal{Y}^i} \pi_y^i N(\phi(x); \mu_y^i, \Sigma_y^i)$, $i = 1, \dots, n+1$, where $\phi()$ be some latent map used while training \hat{f} . Further $w(x, y)$ can be calculated.
- Fix any trial data y , similar to covariate shift setting, a common idea should be calculate importance weight $p_k^i = w(X_k^i, Y_k^i)/W$, $k = 1, \dots, n^i$, $i = 1, \dots, n$, where $W = \sum_{i=1}^n \sum_{k=1}^{n^i} w(X_k^i, Y_k^i) + w(X_{n+1}, y)$, and $p_{n+1} = w(X_{n+1}, y)/W$. Similarly define residuals R_k^i , R_{n+1} .
- Conformal set: $C_\alpha(X_{n+1}) = \left\{ y : R_{n+1} \leq Q(1 - \alpha, \sum_{i=1}^n \sum_{k=1}^{n^i} p_k^i \delta_{R_k^i} + p_{n+1} \delta_\infty) \right\}$.

7 Federated Conformal Prediction Article2

Federated Conformal Predictors for Distributed Uncertainty Quantification[4]

The federated learning setting is similar to article1, but specific with classification setting. Data $(X_k^i, Y_k^i) \sim P^i$, $k = 1, \dots, n_i$, $i = 1, \dots, n$ be calibration distribution for

each i -th agent. The test distribution is $P^{test} = \sum_{i=1}^n \lambda_i P^i$. And $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, where \mathcal{Y} has finite items.

Definition 7.1 (FL Exchangeability). *Follow previous setting and $(X, Y) \sim P^{test}$. The scores on i -th agent $S(X_1^i, Y_1^i), \dots, S(X_{n_i}^i, Y_{n_i}^i), S(X, Y)$ are exchangeable with probability λ_i .*

Remark 7.2. *FL exchangeability means the test distribution is the mixture of all agent distribution and with probability λ_i , (X, Y) has distribution P^i .*

- Write $N = \sum_{i=1}^n n_i$, $\lambda_i = (n_i + 1)/(N + n)$.
- Fix trial data y for new test X . Calculate scores on each agent with given global classifier f , $\{R_k^i = S(X_k^i, Y_k^i)\}_{i \in [n], k \in [n_i]}$ and $R = S(X, y)$.
- Define conformal set $C_\alpha(X) = \{y : R \leq \hat{q}_\alpha\}$, where \hat{q}_α is the $\lceil (1 - \alpha)(N + n) \rceil$ smallest of agent scores.

Let $n_i(q) = |\{k \leq n_i : R_k^i \leq q\}|$ and $\sum_{i=1}^n n_i(\hat{q}_\alpha) = \lceil (1 - \alpha)(N + n) \rceil$. Define event

$$E = \{\forall i \in [n], \{R_k^i\}_{k=1}^{n_i} = \{r_k^i\}_{k=1}^{n_i}\}.$$

Then first follow FL exchangeability the whole space can be divided into n disjoint subspace $\Omega_i = \{R_1^i, \dots, R_{n_i}^i, R \text{ are exchangeable}\}$. And (X, Y) belongs to which P^i is independent of all other things.

$$P(R \leq \hat{q}_\alpha | E) = \sum_{i=1}^n \lambda_i P(R \leq \hat{q}_\alpha | E, \Omega_i).$$

Similar to results in split CP, $P(R \leq \hat{q}_\alpha | E, \Omega_i) \geq n_i(\hat{q}_\alpha)/(n_i + 1)$ as $n_i(\hat{q}_\alpha)$ scores are smaller than \hat{q}_α in $R_1^i, \dots, R_{n_i}^i$ which are exchangeable with R . Thus,

$$P(R \leq \hat{q}_\alpha | E) \geq \sum_{i=1}^n \lambda_i \frac{n_i(\hat{q}_\alpha)}{n_i + 1} = \frac{\sum_{i=1}^n n_i(\hat{q}_\alpha)}{N + n} = \frac{\lceil (1 - \alpha)(N + n) \rceil}{N + n} \geq 1 - \alpha.$$

8 Federated Conformal Prediction Article3

One-Shot Federated Conformal Prediction[3]

Still assume a similar setting, data $(X_k^i, Y_k^i) \sim P^i$, $k = 1, \dots, m$, $i = 1, \dots, n$ be calibration distribution for each i -th agent. And scores $R^i = (R_1^i, \dots, R_m^i)$ for i -th agent.

Core idea of this article is: if each agent gives a quantile \hat{q}_α^i , further find a quantile of these quantiles to generate the conformal set.

- Each agent compute scores R^i , given α calculate k', l'
- Each agent returns k' -th smallest score to the central server $R_{(k')}^i$
- Central server find l' -th smallest of $\{R_{(k')}^i\}_{i=1}^n$ \hat{q} . Define conformal set $C_\alpha(X) = \{y : R = S(X, y) \leq \hat{q}\}$

The decision of k', l' is easy to understand.

$$\{Y \in C_\alpha(X)\} = \{R \leq R_{(k')}^{(l')}\},$$

where $R_{(k')}^{(l')}$ is the l' -th smallest of $R_{(k')}^i$. Order statistic has explicit distribution, assume $R_k^i \sim G$, then $R_{(k)}^i \sim \sum_{j=k}^m C_m^j G^j (1-G)^{m-j}$, and further calculate the quantile of quantile can get the distribution of \hat{q} . This implies k', l' can be find through simple calculation.

References

- [1] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.
- [2] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.
- [3] Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR, 2023.
- [4] Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.
- [5] Vincent Plassier, Nikita Kotelevskii, Aleksandr Rubashevskii, Fedor Noskov, Maksim Velikanov, Alexander Fishkov, Samuel Horvath, Martin Takac, Eric Moulines, and Maxim Panov. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 4879–4887. PMLR, 2024.
- [6] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- [7] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- [8] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.