Conformal prediction and its extensions

1 Conformal prediction

Methods for distribution-free prediction.

Assumption: exchangeability

- The **data** $Z_i = (X_i, Y_i)$ are assumed to be exchangeable (for example, i.i.d.).
- The **algorithm** which maps data to a fitted model $\hat{\mu} : \mathcal{X} \to \mathbb{R}$ is assumed to treat the data points symmetrically.

1.1 Split conformal prediction

(inductive conformal prediction)

1. initial training data set: pre-trained model $\hat{\mu} : \mathcal{X} \to \mathbb{R}$.
2. holdout/calibration set: nonconformity scores $R_i = \left| Y_i - \hat{\mu}\left(X_i\right) \right|, \quad i = 1, \dots, n$.
3. prediction set: $\widehat{C}_n\left(X_{n+1}\right) = \hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1} \cdot \delta_{R_i} + \frac{1}{n+1} \cdot \delta_{+\infty}\right)$.

   (the $\lceil (n+1)(1-\alpha) \rceil$th smallest of $R_1, \dots, R_n$)

Drawback: the loss of accuracy due to sample splitting.

1.2 Full conformal prediction

(transductive conformal prediction)

1. training data & a hypothesized test point: $\hat{\mu}^y = \mathcal{A}\left(\left(X_1, Y_1\right), \dots, \left(X_n, Y_n\right), \left(X_{n+1}, y\right)\right)$ for each $y \in \mathbb{R}$.
2. residuals: $R_i^y = \begin{cases} \left| Y_i - \hat{\mu}^y\left(X_i\right) \right|, & i = 1, \dots, n \\ \left| y - \hat{\mu}^y\left(X_{n+1}\right) \right|, & i = n+1 \end{cases}$.

3. prediction set: $\widehat{C}_n\left(X_{n+1}\right)=\left\{y \in \mathbb{R}: R_{n+1}^y \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_i^y}\right)\right\}$.

Drawback: a steep computational cost.

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1-\alpha$ holds true for both split conformal and full conformal.

1.3 Jackknife+

(close to cross-conformal prediction in Vovk (2013), offering a compromise between the computational and statistical costs)

1. training data with ith point removed: $\hat{\mu}_{-i}=\mathcal{A}\left(\left(X_1, Y_1\right), \ldots,\left(X_{i-1}, Y_{i-1}\right),\left(X_{i+1}, Y_{i+1}\right), \ldots,\left(X_n, Y_n\right)\right)$.
2. residuals: $R_i^{\mathrm{LOO}}=\left|Y_i-\hat{\mu}_{-i}\left(X_i\right)\right|$.
3. prediction set:
$\left[\mathrm{Q}_\alpha\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{\widehat{\mu}_{-i}\left(X_{n+1}\right)-R_i^{\mathrm{LOO}}}+\frac{1}{n+1} \cdot \delta_{-\infty}\right),\ \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{\widehat{\mu}_{-i}\left(X_{n+1}\right)+R_i^{\mathrm{LOO}}}+\frac{1}{n+1} \cdot \delta_{+\infty}\right)\right]$

Drawback: while in practice the it generally provides coverage close to the target level $1-\alpha$, its theoretical guarantee only ensures $1-2\alpha$ probability of coverage in the worst case.

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1-2\alpha$ holds true for jackknife+.

2 Conformal time-series forecasting

Stankeviciute, M Alaa, and Schaar (2021): CF-RNNs

Multi-horizon time-series forecasting problem

Notation:

- the ith data point: $Z_i=(y_{1:t}^{(i)}, y_{t+1:t+H}^{(i)})$. Note that the label $y_{t+1:t+H}^{(i)}$ is now an H-dimensional value, in contrast with the scalar y value from before.

Assumption:

- exchangeable time-series observations

2.1 Methodology

(Split conformal prediction)

1. training set: train the underlying (auxiliary) model $\hat{\mu} : \mathbb{R}^t \to \mathbb{R}^H$, which produces multi-horizon forecasts **directly** (conditionally independent predictions).
2. calibration set: obtain the H-dimensional nonconformity scores

   $R_i = \left[\left|y_{t+1}^{(i)} - \hat{y}_{t+1}^{(i)}\right|, \dots, \left|y_{t+H}^{(i)} - \hat{y}_{t+H}^{(i)}\right|\right]^\top.$
3. prediction set: $\Gamma_1^\alpha\left(y_{(1:t)}^{(n+1)}\right), \dots, \Gamma_H^\alpha\left(y_{(1:t)}^{(n+1)}\right)$, where $\Gamma_h^\alpha\left(y_{(1:t)}^{(n+1)}\right) = \left[\hat{y}_{t+h}^{(n+1)} - \hat{\varepsilon}_h, \hat{y}_{t+h}^{(n+1)} + \hat{\varepsilon}_h\right]$, $\forall h \in \{1, \dots, H\}$ with the critical nonconformity scores $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_H$ become the $\lceil (n+1)(1-\alpha/H) \rceil$-th smallest residuals in the corresponding nonconformity score distributions. (Bonferroni correction)

THEOREM:

- $\mathcal{D} = \left\{\left(y_{1:t}^{(i)}, y_{t+1:t+H}^{(i)}\right)\right\}_{i=1}^{n}$: **exchangeable** time-series observations.
- $\hat{\mu}$: model predicting H-step forecasts using **the direct strategy**.

$$\mathbb{P}\left(\forall h \in \{1, \dots, H\} \cdot y_{t+h} \in [\hat{y}_{t+h} - \hat{\varepsilon}_h, \hat{y}_{t+h} + \hat{\varepsilon}_h]\right) \geq 1 - \alpha.$$

3 Conformal prediction beyond exchangeability

Barber et al. (2023)

1. Nonexchangeable conformal with a **symmetric** algorithm (weights)
2. Nonexchangeable conformal with **nonsymmetric** algorithms (weights & swap)

3.1 Notation

- the ith data point $Z_i = (X_i, Y_i)$
- the full data sequence $Z = (Z_1, \dots, Z_{n+1})$
- the sequence after swap $Z^i = (Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, Z_n, Z_i)$

3.2 Methodology

1. Choose **fixed (non-data-dependent)** weights $w_1, \ldots, w_n \in [0,1]$ with the intuition that a higher weight should be assigned to a data point that is "trusted" more.
2. Normalize weights $\tilde{w}_i = \frac{w_i}{w_1+\cdots+w_n+1}, i = 1, \ldots, n$ and $\tilde{w}_{n+1} = \frac{1}{w_1+\cdots+w_n+1}$.
3. Generate "tagged" data points $(X_i, Y_i, t_i) \in \mathcal{X} \times \mathbb{R} \times \mathcal{T}$.
4. Swap data set, resulting in Z^K with $K \sim \sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_i$, i.e., two data points have swapped tags.
5. Apply algorithm $\mathcal{A}$ to Z^K in place of Z.

3.3 Split conformal prediction

- prediction set: $\widehat{C}_n(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{R_i} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right)$

3.4 Full conformal prediction

1. training data & a hypothesized test point: $\hat{\mu}^{y,k} = \mathcal{A}\left(\left(X_{\pi_k(i)}, Y^y_{\pi_k(i)}, t_i\right) : i \in [n+1]\right)$ for any $y \in \mathbb{R}$ and $k \in [n+1]$, where π_k is the permutation on $[n+1]$ swapping indices k and $n+1$, and $Y_i^y = \begin{cases} Y_i, & i = 1, \ldots, n \\ y, & i = n+1 \end{cases}$.
2. residuals: $R_i^{y,k} = \begin{cases} \left|Y_i - \hat{\mu}^{y,k}(X_i)\right|, & i = 1, \ldots, n \\ \left|y - \hat{\mu}^{y,k}(X_{n+1})\right|, & i = n+1 \end{cases}$.
3. prediction set: $\widehat{C}_n(X_{n+1}) = \left\{y : R_{n+1}^{y,K} \le Q_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R_i^{y,K}}\right)\right\}$.

THEOREM:

- Lower bounds on coverage.

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \ge 1 - \alpha - \sum_{i=1}^n \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R(Z), R\left(Z^i\right)\right)$$

- Upper bounds on coverage.

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} < 1 - \alpha + \tilde{w}_{n+1} + \sum_{i=1}^n \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R(Z), R\left(Z^i\right)\right),$$

if $R_1^{Y_{n+1},K}, \ldots, R_n^{Y_{n+1},K}, R_{n+1}^{Y_{n+1},K}$ are distinct with probability 1.

The results hold true for both nonexchangeable split conformal and full conformal.

So, if $\tilde{w}_{n+1} = \frac{1}{w_1 + \cdots + w_n + 1}$ is small (the effective sample size is large), then mild violations of exchangeability can only lead to mild undercoverage or to mild overcoverage.

3.5 Jackknife+

1. training data with ith point removed and data tag swapped:

   $\hat{\mu}_{-i}^{k} = \mathcal{A}\left(\left(X_{\pi_k(j)}, Y_{\pi_k(j)}, t_j\right) : j \in [n+1], \pi_k(j) \notin \{i, n+1\}\right)$.

2. residuals: $R_i^{k,\mathrm{LOO}} = \left|Y_i - \hat{\mu}_{-i}^{k}\left(X_i\right)\right|$.

3. prediction set:

   $\left[\mathrm{Q}_{\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{\hat{\mu}_{-i}^{K}(X_{n+1}) - R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{-\infty}\right),\ \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{\hat{\mu}_{-i}^{K}(X_{n+1}) + R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right)\right]$

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1 - 2\alpha - \sum_{i=1}^{n} \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R_{\text{jack}+}(Z), R_{\text{jack}+}(Z^i)\right)$

4 References

Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. "Conformal Prediction Beyond Exchangeability." *The Annals of Statistics* 51 (2). https://doi.org/10.1214/23-aos2276.

Stankeviciute, Kamile, Ahmed M Alaa, and Mihaela van der Schaar. 2021. "Conformal Time-Series Forecasting." *Advances in Neural Information Processing Systems* 34: 6216–28.

Vovk, Vladimir. 2013. "Cross-Conformal Predictors." *Annals of Mathematics and Artificial Intelligence* 74 (1-2): 9–28. https://doi.org/10.1007/s10472-013-9368-4.