Split Conformal Prediction for Regression

A Tutorial

Mikolaj Mazurczyk & Christian Igel

UNIVERSITY OF COPENHAGEN

Outline

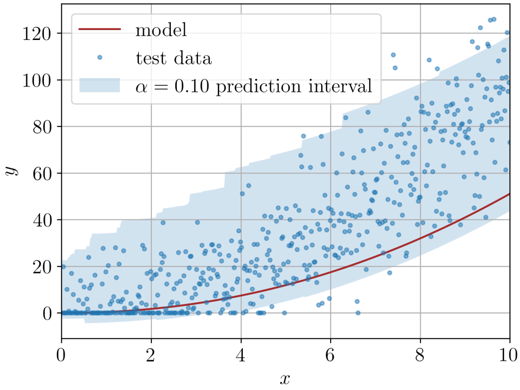
Download

Python notebook for tutorial

Lecture notes

Outline

What is the goal?



- Given a statistical model $\hat{f} : x \mapsto y$ learned from data, we want to quantify the probability that the true value y for an input x is within a certain prediction set $\mathcal{C}(x)$ with high probability.
- What would be your favourite method to address this task?
- Does this approach give you rigorous guarantees? That is, can you guarantee that the predictions will be in the prediction set with a probability of $(1 - \alpha)$?

Problem formulation

Given

- distribution P over $\mathcal{X} \times \mathcal{Y}$,
- fitted model $\hat{f} : \mathcal{X} \to \mathcal{Y}$,
- miscoverage level $\alpha \in \mathbb{R}$,
- calibration data $\{X_i, Y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P$,

construct a function $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that generates a prediction interval $\mathcal{C}\left(X_{\text{test}}\right)$ such that

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) \geq 1 - \alpha$$

for any test sample $\left(X_{\text{test}}, Y_{\text{test}}\right) \sim P$.

Coverage and conditional coverage

- There is an important difference between (marginal) coverage

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) \geq 1-\alpha$$

  and conditional coverage:

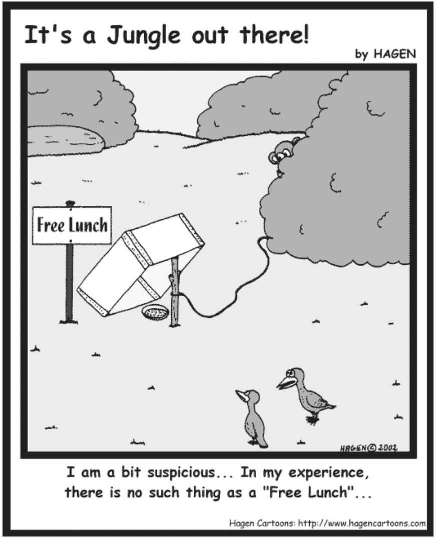
$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid X_{\text{test}}\right) \geq 1-\alpha$$

- Conditional coverage is difficult to achieve, in the continuous case in general impossible.
- Important: Do guarantees hold in expectation over all possible calibration data sets $\mathcal{D}_{\text{cal}}$ or for a concrete finite $\mathcal{D}_{\text{cal}}$?

No free lunch

It is not cheap to get rigorous uncertainty estimates.
You have to pay either by

- strong accurate assumption on the hypothesis class, the data generating distribution and/or the algorithm, or
- having i.i.d. calibration data.



Outline

Empirical cumulative distribution and quantiles

- Random variables and their realizations are denoted by uppercase and lowercase letters, respectively.
- $F_X(x) = \mathbb{P}(X \leq x)$: cumulative distribution function (CDF) of real-valued X.
- $\hat{F}_X(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq z\}$: empirical CDF based on realizations $\{x_i\}_{i=1}^{n}$ of X.
- $\mathcal{Q}_\tau(F_X) = \inf\{x : \tau \leq F_X(x)\}$: τ-quantile of a CDF F_X.
- Empirical τ-quantile of a collection $\{z_i\}_{i=1}^{n}$ using nearest-rank definition:

$$\widehat{\mathcal{Q}}_\tau\left(\{z_i\}_{i=1}^{n}\right) = \inf\left\{z' : \tau \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{z_i \leq z'\}\right\} ,$$

which is the $\lceil \tau \cdot n \rceil$-th value in the list of $\{z_i\}_{i=1}^{n}$ values sorted in increasing order:

$$\widehat{\mathcal{Q}}_{0.5}(\{1,2,3,4,5\}) = 3 \quad , \quad \widehat{\mathcal{Q}}_{0.5}(\{1,2,3,4\}) = 2 \quad , \quad \widehat{\mathcal{Q}}_{0.25}(\{1,2,3,\dots,100\}) = 25$$

Basic idea of split conformal prediction I

A standard (machine learning) way to estimate the accuracy of a predictive model:

- To estimate the expected accuracy of a model, we estimate its accuracy on i.i.d. validation data not used during training.
- In expectation over all draws of the validation data set, the mean error on the validation data set equals the expected error.
- To account for finite sample effects, we apply finite sample concentration bounds.

What about the following for estimating the uncertainty of a model:

- To estimate the uncertainty of a model, we estimate its uncertainty on an i.i.d. calibration data set not used during training.
- In expectation over all draws of the calibration data set, the α-quantile of uncertainties on the calibration data should be the expected α-quantile of uncertainty.
- To account for finite sample effects, we apply finite sample concentration bounds.

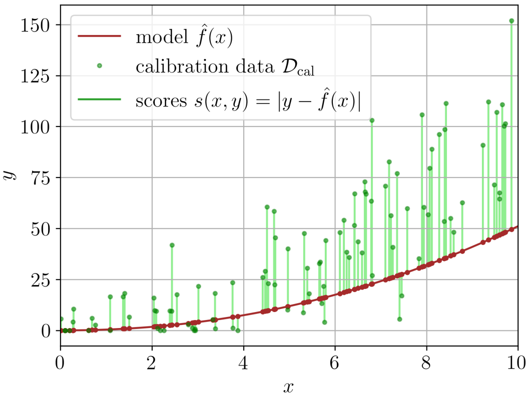
Basic idea of split conformal prediction II

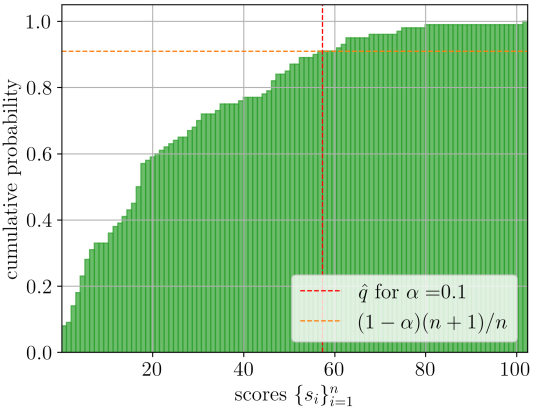
Let's consider regression with $\mathcal{Y} = \mathbb{R}$. The basic split conformal prediction (SCP) approach is rather simple:

1. For each of n points $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ compute a score function, say, the absolute error $s_i = s(x_i, y_i) = \left|y_i - \hat{f}(x_i)\right|$.
2. Sort $s_1, \dots, s_n$ in increasing order and create an empirical cumulative distribution function (CDF) of errors.
3. Pick an empirical quantile $\hat{q}$ corresponding to the $(1 - \alpha)$ quantile of the distribution.
4. Construct a prediction interval:

$$\mathcal{C}(x_{\text{test}}) = \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) \leq \hat{q}\} = \left[\hat{f}(x_{\text{test}}) - \hat{q}, \hat{f}(x_{\text{test}}) + \hat{q}\right]$$

Score function and its empirical cumulative distribution





Split conformal prediction

Algorithm 1: Split Conformal Prediction

Input: non-conformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; calibration set $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^{n}$ of size $n \in \mathbb{Z}^+$; test point x_{test} such that $\forall_{i \in \{1,\dots,n\}} (x_i, y_i), (x_{\text{test}}, y_{\text{test}}) \in \mathcal{X} \times \mathcal{Y}$; target miscoverage level $\alpha \in (0, 1)$

Output: prediction set $\mathcal{C}(x_{\text{test}})$ with coverage $\geq 1 - \alpha$, where $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$

for $i = 1$ **to** n **do** $s_i \leftarrow s(x_i, y_i)$ `// Compute scores on calibration set`

order scores $s_{(1)} \leq s_{(2)} \leq \dots \leq s_{(n)}$ `// Calculate conformal quantile`

set rank $k \leftarrow \lceil (1 - \alpha)(n + 1) \rceil$

if $k > n$ **then** set threshold $\hat{q} \leftarrow \infty$ `// Handle edge case`

else set threshold $\hat{q} \leftarrow s_{(k)}$

return $\mathcal{C}(x_{\text{test}}) \leftarrow \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) \leq \hat{q}\}$ `// Construct prediction set`

Main theorem: Marginal coverage guarantee

Theorem (marginal coverage guarantee [VGS99])

Assume that $\mathcal{C}$ *is defined as in Algorithm (1) for* $\alpha \in (0,1)$, $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{n}$, *and* $\forall_{i\in\{1,\dots,n\}} \left(X_i, Y_i\right), \left(X_{test}, Y_{test}\right) \overset{\text{i.i.d.}}{\sim} P$, *where* P *is some data distribution. Then*

$$1-\alpha \leq \mathbb{E}_{\mathcal{D}_{cal}}\left[\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right) \mid \mathcal{D}_{cal}\right)\right] \leq 1-\alpha+\frac{1}{n+1}+\epsilon_{tie} \ ,$$

where ϵ_{tie} *captures the likelihood that the nonconformity score of the test data point* S_{test} *is the same as another calibration score* S_i,

$$\epsilon_{tie} := \mathbb{P}\left(\exists_{i\in\{1,\dots,n\}}\ S_i = S_{test}\right) \ .$$

Comments on marginal coverage guarantee

- Bound is pretty tight.
- I.i.d. property can be replaced by a weaker exchangebility property.
- No assumption on loss function.
- Guarantee only holds in expectation over the calibration and test samples.
- We consider $\hat{q} := \widehat{\mathcal{Q}}_{(1-\alpha)((n+1)/n)}\left(\{s_i\}_{i=1}^n\right)$, which is the k-th value when sorting $\{s_i\}_{i=1}^n$ in increasing order with $k := \lceil (1-\alpha)(n+1) \rceil$; this k can be higher than $\lceil (1-\alpha)n \rceil$, that is, $\widehat{\mathcal{Q}}_{(1-\alpha)}\left(\{s_i\}_{i=1}^n\right) \leq \widehat{\mathcal{Q}}_{(1-\alpha)((n+1)/n)}\left(\{s_i\}_{i=1}^n\right)$. This choice allows us to prove the main marginal coverage guarantee later.

Properties of the residuals

The distribution of the residuals (i.e., the errors of the model) can be categorized as

- Homoscedastic, not dependent on input x, or
- Heteroscedastic, dependent on input x.

If $\mathcal{Y} = \mathbb{R}$ and the prediction set has the form $[\hat{f}(x) - c_{\mathrm{lo}}, \hat{f}(x) + c_{\mathrm{hi}}]$ then we distinguish (extends canonically to multi-dimensional outputs):

- Symmetric uncertainties, where upper and lower confidence bound are the same.
- Asymmetric uncertainties, where upper and lower confidence bound can differ.

Conformal prediction can be tailored to these settings.

Examples of score functions for regression

- Absolute error score (homoscedastic): $s(x, y) = |y - \hat{f}(x)|$, resulting in

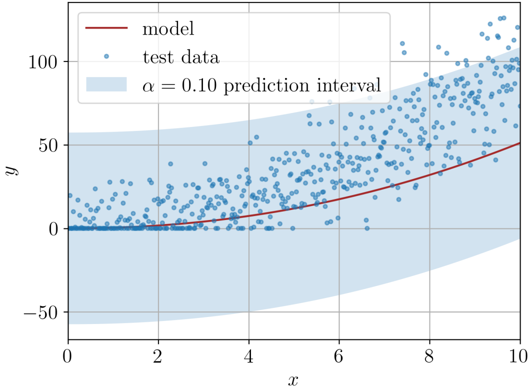
$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \;\middle|\; |y - \hat{f}(x)| \leq \hat{q} \right\} \iff y \in \left[\hat{f}(x) - \hat{q}, \hat{f}(x) + \hat{q} \right] \;.$$

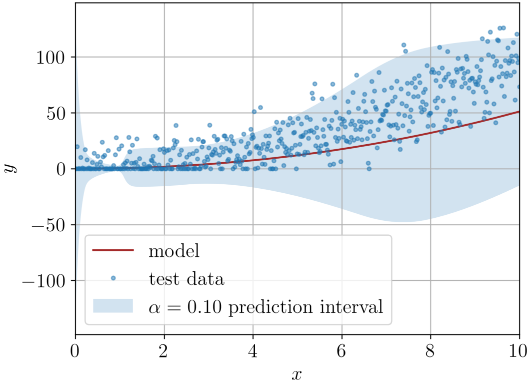
- Scaled absolute error score (heteroscedastic): Model also predicts a symmetric uncertainty estimate $u(x)$, thus $s(x, y) = |y - \hat{f}(x)|/u(x)$, resulting in

$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \;\middle|\; \frac{|y - \hat{f}(x)|}{u(x)} \leq \hat{q} \right\} \iff y \in \left[\hat{f}(x) - \hat{q} \cdot u(x), \hat{f}(x) + \hat{q} \cdot u(x) \right] \;.$$

  Uncertainty estimates could be
  - Predicted variances
  - Ensemble variances
  - Variances from Bayesian modelling
  - ...

Results for symmetric score functions





Left: Absolute error score

Right: Scaled absolute error score (using MLP neural network to estimate variance)

Examples of score functions for asymmetric uncertainties

- Signed-error Split Conformal Regression [LJL14]: We use

$$s_{i,\text{lo}} = s_{\text{lo}}\left(x_i, y_i\right) = \hat{f}(x_i) - y_i, \quad s_{i,\text{hi}} = s_{\text{hi}}\left(x_i, y_i\right) = y_i - \hat{f}(x_i),$$

$$\hat{q}_{\alpha_{\text{lo}}} := \widehat{Q}_{(1-\alpha_{\text{lo}})((n+1)/n)}\left(\left\{s_{i,\text{lo}}\right\}_{i=1}^{n}\right) \quad \text{and} \quad \hat{q}_{\alpha_{\text{hi}}} := \widehat{Q}_{(1-\alpha_{\text{lo}})((n+1)/n)}\left(\left\{s_{i,\text{hi}}\right\}_{i=1}^{n}\right),$$

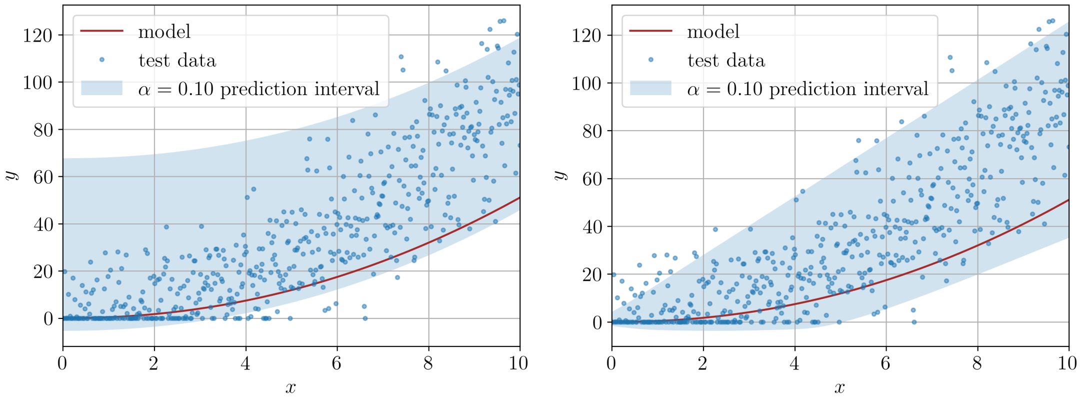
with $\alpha = \alpha_{\text{lo}} + \alpha_{\text{hi}}$, giving $\mathcal{C}\left(X\right) = \left[f\left(X\right) - \hat{q}_{\alpha_{\text{lo}}}, f\left(X_{\text{test}}\right) + \hat{q}_{\alpha_{\text{hi}}}\right]$.

- Conformalized Quantile Regression (CQR, [RPC19]): Assuming estimates of lower and upper quantiles $\hat{q}_{\tau_{\text{lo}}}\left(x\right)$ and $\hat{q}_{\tau_{\text{hi}}}\left(x\right)$ at levels τ_{lo} and τ_{hi}, the score function $s\left(x, y\right) = \max\left\{\hat{q}_{\tau_{\text{lo}}}\left(x\right) - y, y - \hat{q}_{\tau_{\text{hi}}}\left(x\right)\right\}$ corrects both:

$$\mathcal{C}\left(x\right) = \left\{y \in \mathbb{R} \mid \max\left\{\hat{q}_{\tau_{\text{lo}}}\left(x\right) - y, y - \hat{q}_{\tau_{\text{hi}}}\left(x\right)\right\} \leq \hat{q}\right\}$$
$$\iff y \in \left[\hat{q}_{\tau_{\text{lo}}}\left(x\right) - \hat{q}, \hat{q}_{\tau_{\text{hi}}}\left(x\right) + \hat{q}\right]$$

The lower and upper quantiles estimates could stem from quantile regression.

Results of different score functions with asymmetric uncertainties



Left: Signed absolute error score

Right: Conformalized Quantile Regression (using MLPs)

Properties of common score functions for regression

Score function	heteroscedastic	asymmetric	only requires $\hat{f}(x)$
Absolute error score	no	no	yes
Scaled absolute error score	yes	no	no
Conformalized Quantile Regression	yes	yes	no
Signed-error Split Conformal Regression	no	yes	yes

Outline

Order statistics

For a collection of n random variables $Z_1, \dots, Z_n$ the k-th order statistic $Z_{(k)}$ is the k-th value when these random variables are arranged in non-decreasing (ascending) order:

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(k)} \leq \cdots \leq Z_{(n)} \quad .$$

Lemma

For a list of n i.i.d. random variables $\mathcal{Z} = (Z_1, \dots, Z_n)$ and any $k \in \{1, \dots, n\}$ we have

$$\mathbb{P}\left(Z_i \leq Z_{(k)}\right) \geq \frac{k}{n} \quad \text{and} \quad \mathbb{P}\left(Z_i < Z_{(k)}\right) \leq \frac{k-1}{n} \quad .$$

Proof of helper lemma

Proof: From the definition of order statistics, we have

$$k \leq \sum_{i=1}^{n} \mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}$$

and because expectation preserves inequalities we can rewrite

$$\begin{aligned} k &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}\right] \\ &= \sum_{i=1}^{n} \mathbb{P}\left(Z_i \leq Z_{(k)}\right) \\ &= n \cdot \mathbb{P}\left(Z_i \leq Z_{(k)}\right) \quad \text{for any } i, \end{aligned}$$

where we used that the Z_i are i.i.d.; dividing both sides by n proves the first part of the lemma, the second part can be proven analogously.

Replacement lemma

Lemma (Replacement lemma)

Let $S_{(n;k)}$ denote the k-th order statistic computed over $\{S_i\}_{i=1}^n$ and $S_{(n+1;k)}$ the k-th order statistic computed over $\{S_i\}_{i=1}^n \cup \{S_{\text{test}}\}$. Then we have

$$S_{\text{test}} \leq S_{(n;k)} \iff S_{\text{test}} \leq S_{(n+1;k)} \ .$$

Proof:

$S_{\text{test}} \leq S_{(n;k)} \implies S_{\text{test}} \leq S_{(n+1;k)}$: Suppose $S_{\text{test}} > S_{(n+1;k)}$. Then $S_{(n+1;k)} = S_{(n;k)}$, because if we add S_{test} and it is strictly greater than $S_{(n+1;k)}$ it means that its addition does not change the order of first k lowest values. Therefore, $S_{\text{test}} > S_{(n+1;k)} \implies S_{\text{test}} > S_{(n;k)}$

$S_{\text{test}} \leq S_{(n+1;k)} \implies S_{\text{test}} \leq S_{(n;k)}$: k-th smallest entry in the list cannot increase if we add a new value to the list, so $S_{(n+1;k)} \leq S_{(n;k)}$.

□

Proof of marginal coverage guarantee: Lower bound

$\hat{Q} := \widehat{Q}_{(1-\alpha)((n+1)/n)}\left(\{S_i\}_{i=1}^{n}\right)$ is k-th largest score S_i computed over n samples from $\mathcal{D}_{\text{cal}}$, where $k := \lceil(1-\alpha)(n+1)\rceil$, i.e., $\hat{Q} = S_{(k)}$, and thus

$$\{Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\} = \left\{s(X_{\text{test}}, Y_{\text{test}}) \leq \hat{Q}\right\} = \left\{S_{\text{test}} \leq S_{(k)}\right\} \quad ,$$

where $S_{\text{test}} := s(X_{\text{test}}, Y_{\text{test}})$.

$\mathcal{D}_{\text{cal}}$ and $(X_{\text{test}}, Y_{\text{test}})$ being i.i.d. implies that the scores S_i and S_{test} are i.i.d.

Combining both lemmata and choosing $k = \lceil(1-\alpha)(n+1)\rceil$ gives

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) = \mathbb{P}\left(S_{\text{test}} \leq S_{(n+1;k)}\right) \geq \frac{k}{n+1} = \frac{\lceil(1-\alpha)(n+1)\rceil}{n+1} \geq 1-\alpha \quad ,$$

where we used simplified notation $\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) := \mathbb{E}_{\mathcal{D}_{\text{cal}}}[\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}) \mid \mathcal{D}_{\text{cal}})]$.

Proof of marginal coverage guarantee: Upper bound

From

$$S_{\text{test}} \leq S_{(n+1;k)} \iff \text{either } S_{\text{test}} < S_{(n+1;k+1)} \text{ or } S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}$$

we have

$$\begin{aligned}
\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) &= \mathbb{P}\left(S_{\text{test}} \leq S_{(n+1;k)}\right) \\
&= \mathbb{P}\left(S_{\text{test}} < S_{(n+1;k+1)}\right) + \mathbb{P}\left(S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}\right) \\
&\leq \mathbb{P}\left(S_{\text{test}} < S_{(n+1;k+1)}\right) + \epsilon_{\text{tie}} \\
&\leq \frac{(k+1)-1}{n+1} + \epsilon_{\text{tie}} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} + \epsilon_{\text{tie}} \\
&\leq 1-\alpha+\frac{1}{n+1}+\epsilon_{\text{tie}} \quad ,
\end{aligned}$$

where the second equality holds since the two events are mutually exclusive. □

Outline

Calibration-conditional coverage theorem

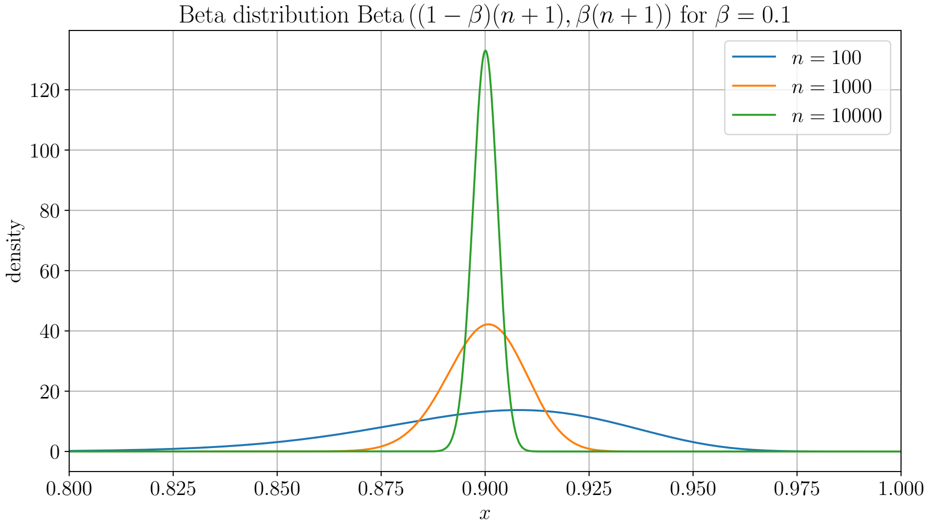
Theorem (Calibration-conditional coverage)

Suppose the data points $\left\{(X_i, Y_i)_i\right\}_{i=1}^{n} = \mathcal{D}_{\mathit{cal}}$ *and* $(X_{\mathit{test}}, Y_{\mathit{test}})$ *are i.i.d., and let* $\mathcal{C}$ *be constructed via split conformal prediction (Algorithm (1)) using any pre-trained non-conformity score function* s *at a miscoverage level* β*. Then the calibration-conditional coverage stochastically dominates the beta distribution*

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\mathit{test}} \in \mathcal{C}\left(X_{\mathit{test}}\right) \mid \mathcal{D}_{\mathit{cal}}\right) \leq 1-\alpha\right) \leq F_{\mathrm{Beta}((1-\beta)(n+1),\beta(n+1))}\left(1-\alpha\right) \quad ,$$

where $F_{\mathrm{Beta}(a,b)}$ *denotes the CDF of the* $\mathrm{Beta}\left(a, b\right)$ *distribution.*

Beta distribution



Calibration-conditional coverage theorem

Proof: Let F_S be the CDF of the distribution of scores $s(X, Y)$. Define $S_i = s(X_i, Y_i)$, which are i.i.d. draws from the distribution with CDF F_S. Let $S_{(1)} \leq \cdots \leq S_{(n)}$ be the order statistics of $S_1, \ldots, S_n$. For $k = \lceil (1-\beta)(n+1) \rceil$, we established

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) = \mathbb{P}\left(S_{\text{test}} \leq S_{(k)} \mid \mathcal{D}_{\text{cal}}\right) = F_S\left(S_{(k)}\right) \quad ,$$

where the last step holds since S_{test} is independent of $\mathcal{D}_{\text{cal}}$ and has CDF F_S. Therefore

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq 1-\alpha\right) = \mathbb{P}\left(F_S\left(S_{(k)}\right) \leq 1-\alpha\right) \quad .$$

Let $Z_i := F_S\left(S_i\right)$. Since F_S is non-decreasing, the order statistics $Z_{(1)} \leq \cdots \leq Z_{(n)}$ of $Z_1, \ldots, Z_n$ satisfy $Z_{(k)} = F_S\left(S_{(k)}\right)$.

Calibration-conditional coverage theorem

By the probability integral transform, if F_S is continuous, each Z_i is an i.i.d. sample from the uniform distribution $\mathcal{U}_{[0,1]}$. Otherwise, if F_S has discontinuities, the resulting Z_i stochastically dominate $U_1, \ldots, U_n \sim \mathcal{U}_{[0,1]}$, meaning

$$\forall_{i\in[n],x\in[0,1]}\ F_{Z_i}(x) \leq F_{U_i}(x)\,,$$

we get

$$\mathbb{P}\left(F_S\left(S_{(k)}\right) \leq 1-\alpha\right) = \mathbb{P}\left(Z_{(k)} \leq 1-\alpha\right) \leq \mathbb{P}\left(U_{(k)} \leq 1-\alpha\right) \ .$$

By the definition of the beta distribution, the k-th order statistic of a sample of size n from $\mathcal{U}_{[0,1]}$ is a beta random variable

$$U_{(k)} \sim \mathrm{Beta}\left(k, n+1-k\right) \ ,$$

and since $k \geq (1-\beta)(n+1)$, $F_S\left(S_{(k)}\right)$ stochastically dominates $\mathrm{Beta}\left((1-\beta)(n+1), \beta(n+1)\right)$. $\square$

Calibration-conditional coverage upper bound

Lemma (Calibration-conditional coverage upper bound)

Under the conditions of the calibration-conditional coverage theorem, we have that

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\mathit{test}} \in \mathcal{C}\left(X_{\mathit{test}}\right) \mid \mathcal{D}_{\mathit{cal}}\right) \leq 1-\alpha\right) \leq e^{-2n\Delta^2} \;,$$

where $\beta := \alpha - \Delta$.

Proof of calibration-conditional coverage upper bound

Proof: In the proof of the calibration-conditional coverage theorem, we established that

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq 1-\alpha\right) \leq \mathbb{P}\left(U_{(k)} \leq 1-\alpha\right) \quad ,$$

where $U_{(k)}$ is the k-th order statistic of n uniformly distributed random variables $U_1, \ldots, U_n \sim \mathcal{U}_{[0,1]}$ with $k = \lceil (1-\beta)(n+1) \rceil = \lceil (1-\alpha+\Delta)(n+1) \rceil$. This is equivalent to saying that at least k realizations of U_i are $\leq 1-\alpha$. Thus, by using $V_i = \mathbb{1}\left\{U_i \leq 1-\alpha\right\}$ the probability can be rewritten as

$$\mathbb{P}\left(U_{(k)} \leq 1-\alpha\right) = \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq k\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)(n+1)\right) \quad .$$

Proof of calibration-conditional coverage upper bound

Since $V_1, \dots, V_n$ are i.i.d. with expected values

$$\mathbb{E}\left[V_i\right] = \mathbb{E}\left[\mathbb{1}\left\{U_i \leq 1-\alpha\right\}\right] = \mathbb{P}\left(U_i \leq 1-\alpha\right) = 1-\alpha$$

we can apply Hoeffding's inequality:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)(n+1)\right) &\leq \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)\, n\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} V_i - n\mathbb{E}\left[V\right] \geq (1-\alpha+\Delta)\, n - n(1-\alpha)\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} V_i - n\mathbb{E}\left[V\right] \geq n\Delta\right) \\
&\leq e^{-2n\Delta^2}
\end{aligned}$$

$\square$

Outline

Bin-wise conditional coverage (poor man's conditional coverage)

Theorem (Bin-wise conditional coverage)

Assuming the setup of the marginal coverage guarantee theorem, partition $\mathcal{X}$ into B bins $\mathcal{X}_1, \dots, \mathcal{X}_B$. Further, let $\mathcal{I}_b = \{i \in \{1, \dots, n\} \mid X_i \in \mathcal{X}_b\}$ be the index set of features X_i belonging to bin $\mathcal{X}_b$, $n_b = |\mathcal{I}_b|$ and $\hat{q}_b := \widehat{Q}_{(1-\alpha)((n_b+1)n_b)}\left(\{S_i\}_{i\in\mathcal{I}_b}\right)$ be the empirical quantile computed according to Algorithm (1) for each bin $b \in \{1, \dots, B\}$ separately by partitioning $\mathcal{D}_{\text{cal}}$ into $\{(X_i, Y_i)\}_{i\in\mathcal{I}_b}$. By defining

$$\mathcal{C}\left(X_{\text{test}}\right) = \{y \ : \ X_{\text{test}} \in \mathcal{X}_m \wedge S_{\text{test}} \leq \hat{q}_b\}$$

we get the bin-wise conditional coverage guarantee

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(Y_{\text{test}}\right) \mid X_{\text{test}} \in \mathcal{X}_b\right) \geq 1 - \alpha$$

for all $b \in \{1, \dots, B\}$ with $\mathbb{P}\left(X_{\text{test}} \in \mathcal{X}_b\right) > 0$.

Split Localized Conformal Prediction (SLCP)

- To get test-conditional coverage, we would ideally like to know the conditional CDF

$$q^*_{1-\alpha}(x) := \mathcal{Q}_{1-\alpha}\left(F_{S|X}(S|x)\right)$$

with

$$\mathbb{P}\left(S \leq q^*_{1-\alpha}(x) \mid X = x\right) = 1 - \alpha \ .$$

- Idea [HTGL23]: Approximate $q^*_{1-\alpha}(x)$ by $\tilde{q}_{1-\alpha}(x) \approx q^*_{1-\alpha}(x)$, consider

$$\mathbb{P}\left(S - \tilde{q}_{1-\alpha}(x) \leq q^*_{1-\alpha}(x) - \tilde{q}_{1-\alpha}(x) \mid X = x\right) = 1 - \alpha \ ,$$

and apply conformal prediction on the residual of ideal and approximated quantiles

$$\mathbb{P}\left(S' \leq \hat{q}\right) \geq 1 - \alpha \ ,$$

with the new score $S' = S - \tilde{q}_{1-\alpha}(x)$.

Prediction interval using absolute error score

Let us assume that the original score function is the absolute error:

$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \;\middle|\; \overbrace{\overbrace{|y - \hat{f}(x)|}^{s(x,y)} - \tilde{q}_{1-\alpha}(x)}^{s'(x,y)} \leq \hat{q} \right\}$$

$$\iff y \in \left\{ \hat{f}(x) - \underbrace{\hat{q}}_{\text{CP correction}} - \underbrace{\tilde{q}_{1-\alpha}(x)}_{\text{estimated quantile}}, \hat{f}(x) + \hat{q} + \tilde{q}_{1-\alpha}(x) \right\}$$

Example using Nadaraya–Watson estimator and signed absolute error

- Nadaraya–Watson estimator for conditional CDF: Given $\{(x_i, y_i)\}_{i=1}^m$, kernel $K : \mathbb{R} \to \mathbb{R}$ with zero mean, the Nadaraya–Watson estimator of the CDF $F_{Y|X}$ is

$$\hat{F}_{Y|X}(y|x) = \sum_{i=1}^{m} w(x_i|x)\, \mathbb{1}\{y \leq y_i\}$$

with

$$w(x_i|x) = \frac{K(\|x_i - x\|)}{\sum_{j=1}^m K(\|x_j - x\|)} = \frac{K(\|x_i - x\|)}{\sum_{j=1}^m K(\|x_j - x\|)} \,.$$

- We estimate the localized conditional CDF on the training data ($\mathcal{D}_{\text{cal}}$ and test data remain exchangeable):

$$\tilde{q}_{1-\alpha}(x) = \widetilde{Q}_{1-\alpha}\left(\hat{F}_{S|X}(S|x)\right) = \inf\left\{ s : 1-\alpha \leq \sum_{(x',y')\in\mathcal{D}_{\text{train}}} w(x'|x)\, \mathbb{1}\{s \leq s(x', y')\} \right\}$$

- We use Signed-error Split Conformal Regression.

SLCP with kernel density estimation

Algorithm 2: Split Localized Conformal Prediction (SLCP)

Input: $\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{train}} \in 2^{\mathcal{X}\times\mathcal{Y}}$; model $\hat{f} : \mathcal{X} \to \mathcal{Y}$; test input $x_{\text{test}} \in \mathcal{X}$; $\alpha_{\text{lo}}, \alpha_{\text{hi}} \in (0,1)$; kernel $K : \mathbb{R} \to \mathbb{R}$

Output: Prediction set $\mathcal{C}(x_{\text{test}}) \in 2^{\mathcal{Y}}$ with coverage $\geq 1 - (\alpha_{\text{lo}} + \alpha_{\text{hi}})$ and $\forall_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \lim_{|\mathcal{D}_{\text{cal}}|\to\infty} \mathbb{P}\left(y \in \mathcal{C}\left(x\right) | X_{\text{test}} = x\right) \geq 1 - \alpha$.

$s_{\text{lo}}(x,y) := -s_{\text{hi}}(x,y) := \hat{f}(x) - y$

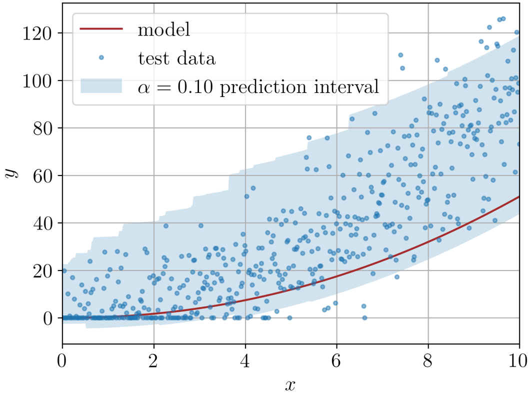
for $l \in \{lo, hi\}$ **do**

- $w(x_i|x) := \frac{K(\|x_i - x\|)}{\sum_{(x_j,y_j)\in\mathcal{D}_{\text{train}}} K(\|x_j - x\|)}$
- $\tilde{q}_{1-\alpha_l}(x) := \inf\left\{s : 1 - \alpha_l \leq \sum_{(x_i,y_i)\in\mathcal{D}_{\text{train}}} w(x_i|x)\mathbb{1}[s \leq s(x_i,y_i)]\right\}$
- $\hat{q}_l \leftarrow$ run SCP on $\mathcal{D}_{\text{cal}}$ with miscoverage α_l and score $s(x,y) = s_l(x,y) - \tilde{q}_{1-\alpha_l}(x)$

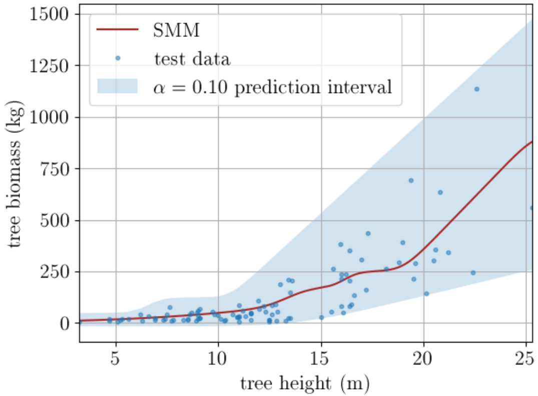
return

$$\mathcal{C}(x_{test}) := \left\{y \in \mathcal{Y} \;\middle|\; y \in \left[\hat{f}(x_{test}) - \tilde{q}_{1-\alpha_{lo}}(x_{test}) - \hat{q}_{lo}, \hat{f}(x_{test}) + \tilde{q}_{1-\alpha_{hi}}(x_{test}) + \hat{q}_{hi}\right]\right\}$$

SLCP results



Real-world example



Data: Allometric data from trees in southern China (Qiue et al., in preparation)
Model: Smooth min-max (SMM) neural network (Igel, *ICML*, 2024)
CP method: CQR with MLPs

Summary

Conformal prediction

- gives prediction sets with rigorous coverage guarantees,
- requires no assumption on model, data generating distribution, and algorithm except calibration data and test data being i.i.d.,
- can deal with heteroscedastic, asymmetric noise/uncertainty,
- can turn your favourite uncertainty estimates into proper prediction sets.

Conditional coverage remains difficult.

Conformal prediction goes back to Vovk and Gammerman (e.g., [VGS99]), there is an excellent review by Angelopoulos and Bates [AB23], and a new textbook draft by Angelopoulos, Barber, and Bates [ABB25] – we based this tutorial on the latter.

Thanks

Danmarks Grundforskningsfond
Danish National Research Foundation

novo nordisk foundation

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

References

Anastasios N. Angelopoulos and Stephen Bates.
Conformal prediction: A gentle introduction.
Foundations and Trends® in Machine Learning, 16(4):494–591, 2023.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates.
Theoretical foundations of conformal prediction.
arXiv:2411.11824, 2025.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu.
Split localized conformal prediction.
arXiv:2206.13092, 2023.

Henrik Linusson, Ulf Johansson, and Tuve Löfström.
Signed-error conformal regression.
In *Advances in Knowledge Discovery and Data Mining (PKDD)*, pages 224–236. Springer, 2014.

Yaniv Romano, Evan Patterson, and Emmanuel Candes.
Conformalized quantile regression.
In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

Volodya Vovk, Alexander Gammerman, and Craig Saunders.
Machine-learning applications of algorithmic randomness.
In *International Conference on Machine Learning (ICML)*, page 444–453. Morgan Kaufmann Publishers Inc., 1999.

Outline

Hoeffding's inequality

Theorem (Hoeffding's inequality)

Let $X_1, \dots, X_n$ *be independent real-valued random variables, such that for each* $i \in \{1, \dots, n\}$ *there exist* $a_i \leq b_i$, *such that* $X_i \in [a_i, b_i]$. *Then for every* $\epsilon > 0$*:*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \epsilon\right) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)}$$

and

$$\mathbb{P}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] - \sum_{i=1}^{n} X_i \geq \epsilon\right) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)} .$$

Probability integral transform

Theorem (Probability integral transform)

Suppose that a random variable X has a continuous distribution for which the cumulative distribution function (CDF) is F_X. Then the random variable Y defined as

$$Y := F_X(X)$$

has a standard uniform distribution $\mathcal{U}_{[0,1]}$. If F_X is not continuous then Y is a superuniform random variable, meaning that

$$\forall_{u \in [0,1]} \ \mathbb{P}(Y \leq u) \leq u \ ,$$

or equivalently, Y dominates the uniform distribution:

$$\forall_{u \in [0,1]} \ F_Y(u) \leq F_U(u) \, , \quad \textit{where} \quad U \sim \mathcal{U}_{[0,1]} \ .$$