Statistical Learning: Conformal prediction

Barabás Eszter
ELTE Eötvös Loránd Tudományegyetem
Supervisor: Csáji Balázs Csanád

Budapest, 2025.

Motivation

- quantifying uncertainty estimates for predictions
- produces prediction sets that are guaranteed to contain the true data with a preset probability
- range of likely outcomes instead of a single estimate

Theoretical Framework

Consider the data in the form of ordered pairs (x_i, y_i) called examples; $z_i = (x_i, y_i)$. We set

$$\mathcal{Z} := \mathcal{X} \times \mathcal{Y}.$$

Examples are chosen independently from some probability distribution P on $\mathcal{Z}$.

Definition

The exchangeability of P means that for every positive integer n, every permutation π of $\{1, \dots, n\}$, and every measurable set $E \subseteq \mathcal{Z}^n$,

$$\mathbb{P}\big((z_1, \dots, z_n) \in E\big) = \mathbb{P}\big((z_{\pi(1)}, \dots, z_{\pi(n)}) \in E\big).$$

Definition

A *simple predictor* is a measurable function

$$D : \mathcal{Z}^* \times \mathcal{X} \to \mathcal{Y}.$$

Theoratical Framework

additional inputs: $\alpha \in (0,1)$, called the *significance level*; $1-\alpha$ is the *confidence level*

algorithm Γ outputs a subset of $\mathcal{Y}$:

$$\Gamma^{\alpha}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \subseteq \mathcal{Y}.$$

If $\alpha_1 \geq \alpha_2$,

$$\Gamma^{\alpha_2}(x_1, \ldots, y_{n-1}, x_n) \supseteq \Gamma^{\alpha_1}(x_1, \ldots, y_{n-1}, x_n). \qquad (1)$$

Definition

A confidence predictor is a measurable function

$$\Gamma : (0,1) \times \mathcal{Z}^* \times \mathcal{X} \to 2^{\mathcal{Y}}$$

that satisfies the monotonicity condition (1) for all significance levels $\alpha_1 \geq \alpha_2$, all $n \in \mathbb{Z}^+$, and all incomplete data sequences $(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$.

Theoretical Framework

Definition

A bag or multiset of size $n \in \mathbb{N}$ is a collection $z_1, \dots, z_n$ of n elements from a measurable space $\mathcal{Z}$, where order is irrelevant and repetitions are allowed. We denote by $\mathcal{Z}^{(n)}$ the set of all bags of size n, and by $\mathcal{Z}^{(*)} = \bigcup_{n \geq 1} \mathcal{Z}^{(n)}$ the set of all finite bags.

Definition

A nonconformity measure is a measurable function

$$A : \mathcal{Z}^{(*)} \times \mathcal{Z} \to \overline{\mathbb{R}}$$

that assigns to each bag of old examples and each new example $z \in \mathcal{Z}$ a score $A(z_1, \dots, z_n, z)$ indicating how different z is from $z_1, \dots, z_n$.

For regression problems: $A(z_1, \dots, z_n, (x, y)) = |y - \hat{y}(x)|$

Quantiles

Suppose we have a sequence $Y_i \in \mathbb{R}$, $i = 1, \dots, n$ of real-valued response values and a significance level α. Our goal is to find a one-sided prediction interval $C_n = (-\infty, q_n]$ such that

$$\mathbb{P}\big(Y_{n+1} \leq q_n\big) \geq 1 - \alpha.$$

$$\mathbb{P}(Y_{n+1} \text{ is among the } (1-\alpha)(n+1) \text{ smallest of } Y_1, \dots, Y_{n+1}) \ \geq \ 1 - \alpha.$$

This is equivalent to

$$\mathbb{P}(Y_{n+1} \text{ is among the } (1-\alpha)(n+1) \text{ smallest of } Y_1, \dots, Y_n) \ \geq \ 1 - \alpha.$$

Accordingly, define

$$q_n = \lceil (1-\alpha)(n+1) \rceil\text{-th smallest of } Y_1, \dots, Y_n.$$

Regression

Let f_n be any point predictor trained on the n samples. Train a nonconformity score, for example, we define residuals:

$$R_i = |Y_i - f_n(X_i)|, \quad i = 1, \dots, n,$$

and the quantile

$$q_n = \lceil (1-\alpha)(n+1) \rceil\text{-th smallest of } R_1, \dots, R_n.$$

Then the naive prediction set is

$$C_n(x) = \left[\, f_n(x) - q_n,\ f_n(x) + q_n \,\right].$$

Split Conformal prediction algoritmn

- Split the training set into a proper training set D_1 of size n_1 and a calibration set D_2 of size n_2
- Fit predictor f_{n_1} using the data in D_1
- For each $i \in D_2$, compute

$$R_i = |Y_i - f_{n_1}(X_i)|$$

- Let q_{n_2} be the $\lceil (1-\alpha)(n_2+1) \rceil$-th smallest value among $\{R_i : i \in D_2\}$:

$$q_{n_2} = R_{(\lceil (1-\alpha)(n_2+1) \rceil)}$$

- For any new input x, define

$$C_n(x) = \left[\, f_{n_1}(x) - q_{n_2},\; f_{n_1}(x) + q_{n_2} \,\right]$$

Split conformal prediction

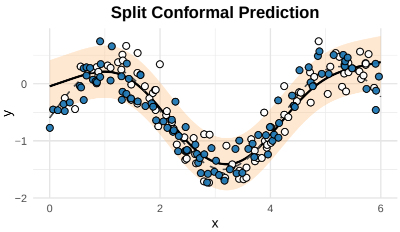


Figure: 90% coverage

$n = 200$ data points generated, $x_i \sim \mathrm{Uniform}(0, 6)$

$$f(x) = 0.9 \sin(1.5x) - 0.6$$

$\varepsilon_i \sim \mathcal{N}(0,\ 0.25^2)$

$y_i = f(x_i) + \varepsilon_i$

Method comparison

the **empirical coverage** over a test set $\{(X_j, Y_j)\}_{j=1}^{m}$ is

$$\frac{1}{m}\sum_{j=1}^{m} 1\{\,|Y_j - f_n(X_j)| \leq q_n\,\}.$$

the **width** of prediction interval for the i-th observation:

$$W_i = \hat{y}_i^{\text{upper}} - \hat{y}_i^{\text{lower}}$$

where

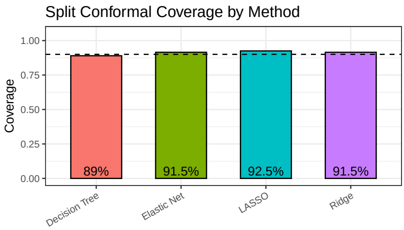
$$\hat{y}_i^{\text{lower}} = \hat{y}_i - q_{1-\alpha}, \quad \hat{y}_i^{\text{upper}} = \hat{y}_i + q_{1-\alpha}.$$

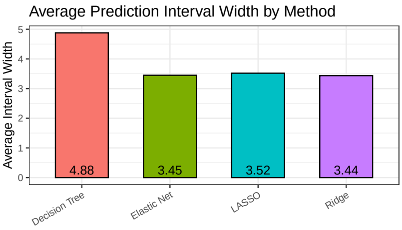
$$\mathcal{C}_i = [\hat{y}_i - q_{1-\alpha},\ \hat{y}_i + q_{1-\alpha}]$$

Thus, the (constant) interval width is:

$$W = 2q_{1-\alpha}$$

Method comparison





AI use

I used ChatGPT in my report

- to generate parts of the code for plots
- for grammatical checking

References

- Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Boston, MA: Springer US, 2005.
- Tibshirani, Ryan. "Conformal prediction." UC Berkeley (2023).
- Shafer, Glenn, and Vladimir Vovk. "A tutorial on conformal prediction." Journal of Machine Learning Research 9.3 (2008).
- Fontana, Matteo, Gianluca Zeni, and Simone Vantini. "Conformal prediction: a unified review of theory and new challenges." Bernoulli 29.1 (2023): 1-23.
- Marques, F., and C. Paulo. "Universal distribution of the empirical coverage in split conformal prediction." Statistics & Probability Letters 219.C (2025).