Conformal prediction

D. L. Razafindrakoto[1, 2]   A. Celisse[2]

[1]Safran Aircraft Engines

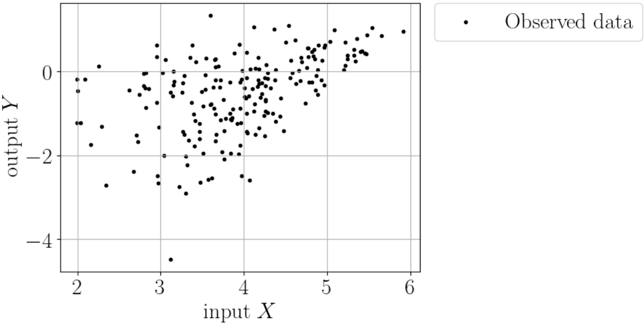
[2]SAMM, Université Paris 1 Panthéon-Sorbonne

M2 Tide, 9 Janvier 2026

Illustration

Data set

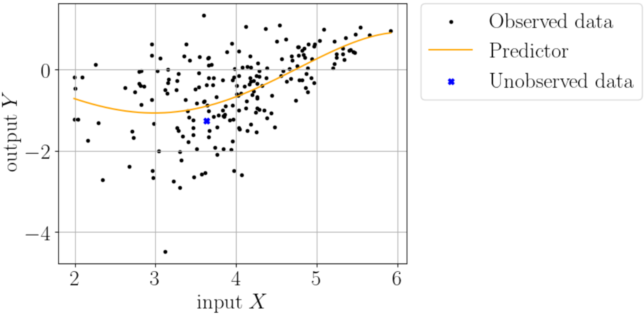
$(X_1, Y_1), \ldots, (X_n, Y_n)$, independent and identically distributed as (X, Y) random variables, where $X \sim \beta(6, 3)$ and $Y|X \sim \cos(X) + (1 - \cos(X))\mathcal{N}(0, 0.5)$.



Illustration

Prediction

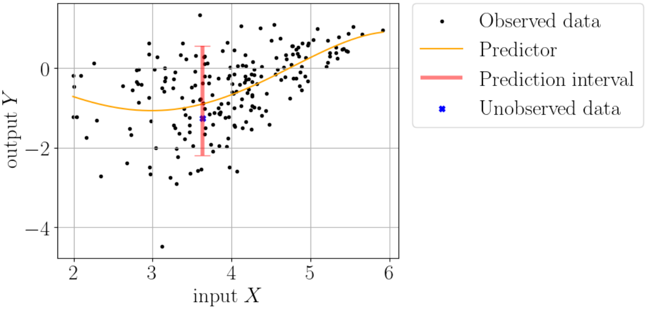
Given (X_{n+1}, Y_{n+1}) independent and identically distributed as $(X_1, Y_1), \ldots, (X_n, Y_n)$, a prediction $\hat{Y}_{n+1}$ approximates Y_{n+1}.



Illustration

Prediction region

Given (X_{n+1}, Y_{n+1}), independent and identically distributed as $(X_1, Y_1), \ldots, (X_n, Y_n)$, for a confidence control level α, a prediction region $\hat{C}_\alpha(X_{n+1})$ containts Y_{n+1} with probability greater that $1-\alpha$.



Setup

- Data set: $(X_1, Y_1), \ldots, (X_n, Y_n)$, (X_{n+1}, Y_{n+1}), $\mathcal{X} \times \mathcal{Y}$-valued independent and identically distributed random variables where $\mathcal{X} \subseteq \mathbb{R}$, and $\mathcal{Y} \subseteq \mathbb{R}$.
- Regression: Predict $\hat{Y}_{n+1} \approx Y_{n+1}$ provided X_{n+1} and the observed data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Confidence prediction region

For a confidence control level α, a confidence prediction region $C_\alpha(X_{n+1})$ fullfils the following

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

Solution

Conformal prediction (Vovk, Gammerman, and Shafer, 2005).

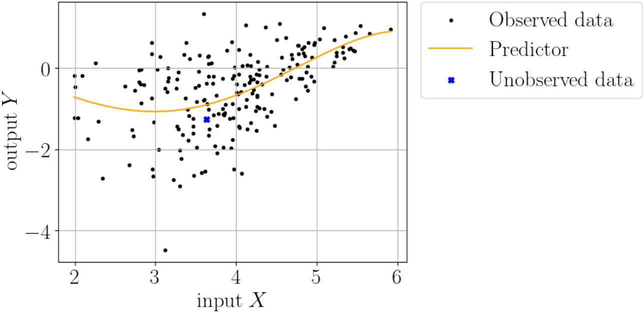
Predictor

- Feature map: $\phi(\cdot) : \mathcal{X} \mapsto \mathbb{R}^d$. For example, for every $x \in \mathbb{R}$, $\phi(x) = \left(1, x, x^2, x^3, x^4\right)^T$.

Ridge regression

For a data set D and a regularization parameter $\lambda \in (0, +\infty)$

$$\hat{\beta}_{\lambda;D} := \underset{\beta \in \mathbb{R}^d}{\arg\min} \frac{1}{|D|} \sum_{(x,y) \in D} \left(y - \phi(x)^T \beta\right)^2 + \lambda \|\beta\|_2^2 .$$



Full conformal prediction

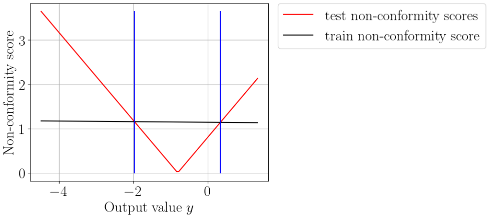
Non-conformity scores

For every $i \in \{1, \dots, n+1\}$ and every $y \in \mathcal{Y}$

$$S_{D^y}(X_i, Y_i) = \left| Y_i - X_i^T \hat{\beta}_{\lambda; D^y} \right|, \quad \text{if } 1 \le i \le n,$$

$$S_{D^y}(X_i, y) = \left| y - X_i^T \hat{\beta}_{\lambda; D^y} \right|, \quad \text{if } i = n+1,$$

where the data set D^y is defined as $D^y := \{(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)\}$.



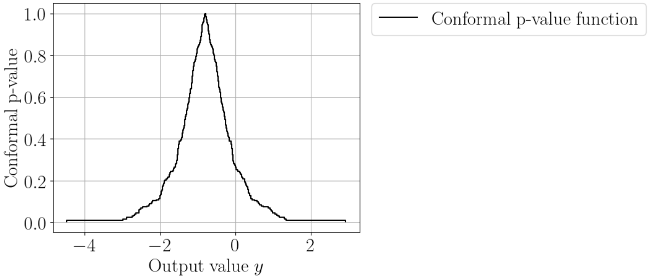
- It measures the "strangeness" each data points in D^y.
- For example, $S_{D^y}(X_{n+1}, y)$ is the least strange for a value y around -0.7 and gets more strange for values away from -0.7.

Full conformal prediction

Full conformal p-value function

For every output value $y \in \mathcal{Y}$

$$\widehat{\pi}_D^{\text{Full}}(X_{n+1}, y) := \frac{1 + \sum_{i=1}^{n} \mathbb{1}\{S_{D^y}(X_i, Y_i) \geq S_{D^y}(X_{n+1}, y)\}}{n+1}.$$



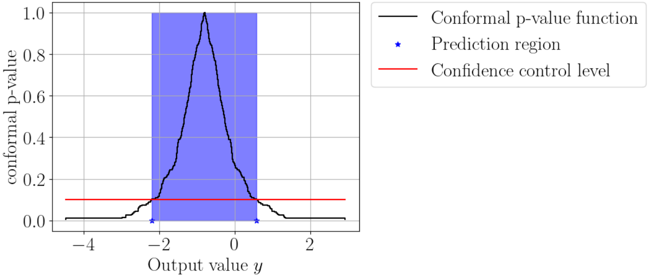
- It measures how strange is (X_{n+1}, y) relative the other data points in D^y.
- For example, it is relatively the least strange for a value y around -0.7, and relatively more strange for values y far from -0.7.

Full conformal prediction

Full conformal prediction region (FCPR)

For a confidence control level α, the FCPR $\hat{C}_{\alpha}^{\mathrm{Full}}(X_{n+1})$ is defined as

$$\hat{C}_{\alpha}^{\mathrm{Full}}(X_{n+1}) := \left\{ y \in \mathcal{Y} : \widehat{\pi}_{D}^{\mathrm{Full}}(X_{n+1}, y) > \alpha \right\}.$$



- It is the set of the values of y, such (X_{n+1}, y) is relatively not too strange.

Full conformal prediction

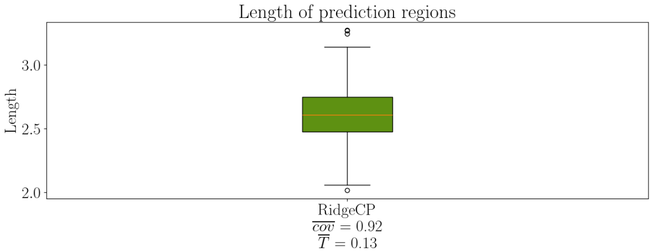
Coverage guarantee

If $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable, then the FCPR $\hat{C}_\alpha^{\mathrm{Full}}(X_{n+1})$ enjoys the following guarantee,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha^{\mathrm{Full}}(X_{n+1}) \right] \geq 1 - \alpha.$$

Moreover $S_{D^{Y_{n+1}}}(X_1, Y_1), \ldots, S_{D^{Y_{n+1}}}(X_{n+1}, Y_{n+1})$ are almost surely distinct, then

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha^{\mathrm{Full}}(X_{n+1}) \right] \leq 1 - \alpha + \frac{1}{n+1}.$$



- $\overline{cov}$ which is a empirical estimation of the coverage probability is indeed not far from $0.9 = 1 - \alpha$.

Proof (Part 1)

Since $\hat{f}_{D^{Y_{n+1}}}$ is invariant to permutation of the data points in $D^{Y_{n+1}}$, and $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable, it follows that $S_{D^{Y_{n+1}}}(X_1, Y_1), \ldots, S_{D^{Y_{n+1}}}(X_{n+1}, Y_{n+1})$ are exchangeable.

Let us note for every $i \in \{1, \ldots, n+1\}$, $S_i := S_{D^{Y_{n+1}}}(X_i, Y_i)$, and sort them $S_{(1)} \geq \ldots \geq S_{(n+1)}$. Let us consider the following set

$$\begin{aligned}
\mathrm{Strange}(\alpha) &:= \left\{ i \in \{1, \ldots, n+1\}, \sum_{j=1}^{n+1} \mathbb{1}\left\{S_j \geq S_i\right\} \leq \alpha(n+1) \right\} \\
&= \left\{ i \in \{1, \ldots, n+1\}, \sum_{j=1}^{n+1} \mathbb{1}\left\{S_{(j)} \geq S_{(i)}\right\} \leq \alpha(n+1) \right\} \\
&\subseteq \left\{ i \in \{1, \ldots, n+1\}, (i) \leq \alpha(n+1) \right\} \\
&\subseteq \left\{ i \in \{1, \ldots, n+1\}, (i) \leq \lfloor \alpha(n+1) \rfloor \right\}.
\end{aligned}$$

It follows that

$$\mathrm{Card}(\mathrm{Strange}(\alpha)) \leq \sum_{i=1}^{n+1} \mathbb{1}\left\{(i) \leq \lfloor \alpha(n+1) \rfloor\right\} \leq \lfloor \alpha(n+1) \rfloor.$$

Proof (Part 2)

Let us consider the probability of not covering

$$
\begin{aligned}
&\mathbb{P}\left[\, Y_{n+1} \notin \hat{C}_{\alpha}^{\mathrm{Full}}\left(X_{n+1}\right) \right] \\
&= \mathbb{P}\left[\, \hat{\pi}_D\left(X_{n+1}, Y_{n+1}\right) \leq \alpha \,\right] \\
&= \mathbb{P}\left[\frac{\sum_{i=1}^{n+1} \mathbb{1}\left\{ S_{D^{Y_{n+1}}}\left(X_i, Y_i\right) \geq S_{D^{Y_{n+1}}}\left(X_{n+1}, Y_{n+1}\right) \right\}}{n+1} \leq \alpha \right] \\
&= \mathbb{P}\left[\, n+1 \in \mathrm{Strange}\left(\alpha\right) \right],
\end{aligned}
$$

and since $S_1, \ldots, S_{n+1}$ are exchangeable

$$
\begin{aligned}
\mathbb{P}\left[\, n+1 \in \mathrm{Strange}\left(\alpha\right) \right] &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left[\, i \in \mathrm{Strange}\left(\alpha\right) \right] \\
&= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\left[\, \mathbb{1}\left\{ i \in \mathrm{Strange}\left(\alpha\right) \right\} \right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{ i \in \mathrm{Strange}\left(\alpha\right) \right\} \right] \\
&= \frac{\mathrm{Card}\left(\mathrm{Strange}\left(\alpha\right)\right)}{n+1} = \frac{\lfloor \alpha\left(n+1\right) \rfloor}{n+1} \leq \alpha.
\end{aligned}
$$

Proof (Part 3)

Let us consider the case where $S_1, \ldots, S_{n+1}$ are almost surely distinct. It follows that almost surely $S_{(1)} > \ldots > S_{(n+1)}$. Going back to set of strange points

$$
\begin{aligned}
\mathrm{Strange}(\alpha) &:= \left\{ i \in \{1, \ldots, n+1\}, \sum_{j=1}^{n+1} \mathbb{1}\left\{S_j \geq S_i\right\} \leq \alpha(n+1) \right\} \\
&= \left\{ i \in \{1, \ldots, n+1\}, \sum_{j=1}^{n+1} \mathbb{1}\left\{S_{(j)} \geq S_{(i)}\right\} \leq \alpha(n+1) \right\} \\
&= \left\{ i \in \{1, \ldots, n+1\}, (i) \leq \alpha(n+1) \right\} \\
&\supseteq \left\{ i \in \{1, \ldots, n+1\}, (i) \leq \lceil \alpha(n+1) \rceil \right\}.
\end{aligned}
$$

It follows that

$$
\mathrm{Card}(\mathrm{Strange}(\alpha)) \geq \sum_{i=1}^{n+1} \mathbb{1}\left\{(i) \leq \lceil \alpha(n+1) \rceil\right\} = \lceil \alpha(n+1) \rceil .
$$

Therefore, following a similar argument as before

$$
\mathbb{P}\left[Y_{n+1} \notin \hat{C}_{\alpha}^{\mathrm{Full}}(X_{n+1}) \right] \geq \frac{\lceil \alpha(n+1) \rceil}{n+1} \geq \frac{\alpha(n+1) - 1}{n+1} \geq \alpha - \frac{1}{n+1}.
$$

Practical considerations

Full conformal prediction region depends on

- the confidence control level $1 - \alpha$, and
- the quality of the predictor $\hat{f}_\lambda : \mathcal{X} \mapsto \mathcal{Y}$.

In general, computing the full conformal prediction exactly is too computationaly costly.

- In fact, a brute force approach requires training as many predictors as the cardinality of the space of output values $\mathcal{Y}$, which in regression is $+\infty$.

Commonly-used computationaly afordable approximation are

- Split conformal prediction (Papadopoulos, 2008),
- Cross conformal prediction (Barber et al., 2021), and
- Stable conformal prediction (Ndiaye, 2022).

We devised a new approximation with better guarantees (Razafindrakoto, Celisse, and Lacaille, 2026) for kernel regression with regularization.

Non-conformity scores (split conformal)

Data sets

For $n_{\mathrm{train}}, n_{\mathrm{cal}} \in \mathbb{N}$ such that $n_{\mathrm{train}} + n_{\mathrm{cal}} = n$,

- the calibration data set D_{cal} is defined as $D_{\mathrm{cal}} := \{(X_1, Y_1), \dots, (X_{n_{\mathrm{cal}}}, Y_{n_{\mathrm{cal}}})\}$,
- the training data set D_{train} is defined as $D_{\mathrm{train}} := \left\{\left(X_{n_{\mathrm{cal}}+1}, Y_{n_{\mathrm{cal}}+1}\right), \dots, (X_n, Y_n)\right\}$.

Non-conformity scores

For every $i \in \{1, \dots, n_{\mathrm{cal}}\}$ and every $y \in \mathcal{Y}$

$$
\begin{aligned}
S_{D_{\mathrm{train}}}(X_i, Y_i) &= \left| Y_i - X_i^T \hat{\beta}_{\lambda; D_{\mathrm{train}}} \right|, \quad \text{if } 1 \leq i \leq n_{\mathrm{cal}}, \\
S_{D_{\mathrm{train}}}(X_i, y) &= \left| y - X_i^T \hat{\beta}_{\lambda; D_{\mathrm{train}}} \right|, \quad \text{if } i = n + 1.
\end{aligned}
$$

- Pros: Only need to train one predictor.
- Cons: Less points for training and less non-conformity scores.

Split conformal prediction

Split conformal p-value function

For every output value $y \in \mathcal{Y}$

$$\widehat{\pi}_D^{\mathrm{Split}}(X_{n+1}, y) := \frac{1 + \sum_{i=1}^{n_{\mathrm{cal}}} \mathbb{1}\left\{S_{D_{\mathrm{train}}}(X_i, Y_i) \geq S_{D_{\mathrm{train}}}(X_{n+1}, y)\right\}}{n_{\mathrm{cal}} + 1}.$$

Split conformal prediction region (SCPR)

For a confidence control level α, the SCPR $\hat{C}_\alpha^{\mathrm{Split}}(X_{n+1})$ is defined as

$$\begin{aligned}
\hat{C}_\alpha^{\mathrm{Split}}(X_{n+1}) &:= \left\{y \in \mathcal{Y}, \widehat{\pi}_D^{\mathrm{Split}}(X_{n+1}, y) > \alpha\right\} \\
&= \Bigg[X_{n+1}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}} - \left|Y_{(i_{n,\alpha})} - X_{(i_{n,\alpha})}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}}\right|, \\
&\qquad X_{n+1}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}} + \left|Y_{(i_{n,\alpha})} - X_{(i_{n,\alpha})}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}}\right|\Bigg],
\end{aligned}$$

where $\left|Y_{(1)} - X_{(1)}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}}\right| \leq \ldots \leq \left|Y_{(n_{\mathrm{cal}})} - X_{(n_{\mathrm{cal}})}^T \hat{\beta}_{\lambda; D_{\mathrm{train}}}\right|$ and $i_{n,\alpha} = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$.

Recap

- Formulating confidence prediction regions with conformal prediction,
- Computing split conformal prediction regions.

Reference

[1] Rina Foygel Barber et al. "Predictive inference with the jackknife+". In: *The Annals of Statistics* 49.1 (2021), pp. 486–507.

[2] Eugene Ndiaye. "Stable conformal prediction sets". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 16462–16479.

[3] Harris Papadopoulos. "Inductive conformal prediction: Theory and application to neural networks". In: *Tools in artificial intelligence*. Citeseer, 2008.

[4] Davidson Lova Razafindrakoto, Alain Celisse, and Jérôme Lacaille. "Approximate full conformal prediction in RKHS". In: *arXiv preprint arXiv:2601.13102* (2026).

[5] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.