Conformal prediction

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2018) Distribution-free predictive inference for regression. JASA,113:1094–1111
- Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.
- A Tutorial on Conformal Prediction `https://www.youtube.com/watch?v=nql000Lu_iE` (Part 1); `https://www.youtube.com/watch?v=TRx4a2u-j7M` (Part 2); `https://www.youtube.com/watch?v=37HKrmA5gJE` (Part 3)

Table of Contents

Suppose we have fitted a Gaussian linear model based on the training data $(\mathbf{y}, \mathbf{X})$, obtaining the estimates

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}, \quad \hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2/(n-p)$$

There are (at least) two levels at which we can make predictions

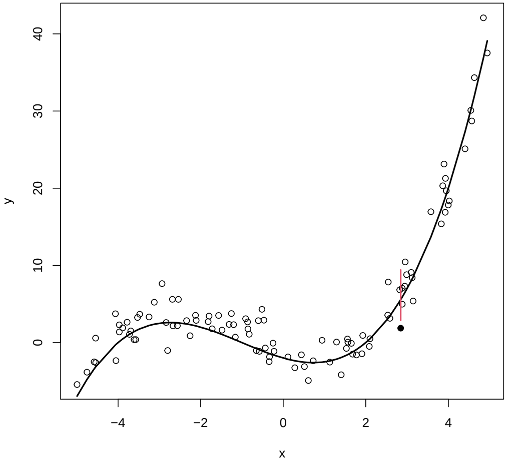
1. A *point prediction* is a single best guess about what a new Y will be when $X = x$
2. A *prediction interval*

$$C_\alpha(x) = x^t\hat{\beta} \pm t_{n-p}^{1-\alpha/2}\hat{\sigma}\sqrt{x^t(\mathbf{X}^t\mathbf{X})^{-1}x + 1}$$

for $Y|X = x$ with $(1-\alpha)$ *conditional coverage* guarantee, i.e.

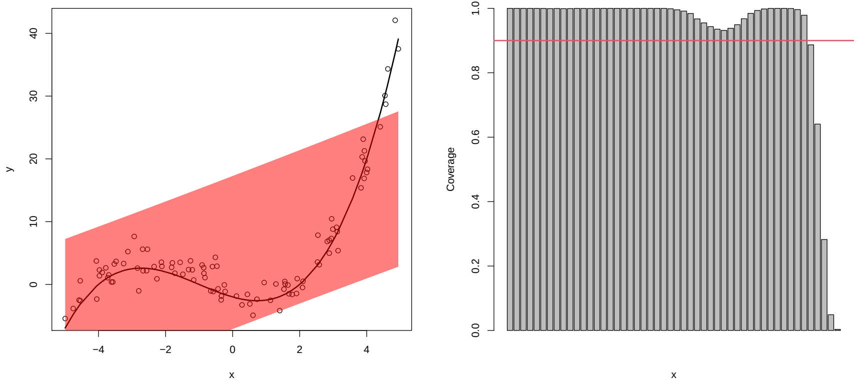
$$\mathrm{P}(Y \in C_\alpha(x)|X = x) = 1 - \alpha$$

where the probability is with respect to the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the new response Y at a fixed test point $X = x$



$$f(x) = \frac{1}{4}(x+4)(x+1)(x-2)$$

Model miss-specification



$1-\alpha = 90\%$, marginal coverage $\approx 93\%$

Table of Contents

Marginal and conditional coverage

- $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ follows some *unknown* joint distribution P_{XY}
- Training $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test (X_{n+1}, Y_{n+1}) i.i.d. (X, Y)
- C_α satisfies *distribution-free marginal coverage* at level $1 - \alpha$ if

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

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X_{n+1}, Y_{n+1})
- C_α satisfies *distribution-free conditional coverage* at level $1 - \alpha$ if

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

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$, and Y_{n+1} at a fixed test point $X_{n+1} = x$

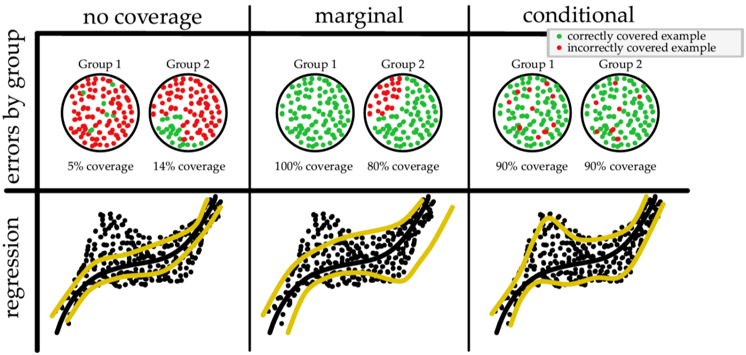


Figure 10: Prediction sets with various notions of coverage: *no coverage, marginal coverage, or conditional coverage (at a level of 90%). In the marginal case, all the errors happen in the same groups and regions in X-space. Conditional coverage disallows this behavior, and errors are evenly distributed.*

From: Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Table of Contents

Conformal prediction

Conformal prediction (Vovk, Gammerman, Saunders, Vapnik, 1996-1999) is a general framework for constructing prediction intervals by using *any* algorithm with finite sample and distribution-free *exact* marginal coverage, i.e.

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha \qquad \forall\ P_{XY}$$

Two main versions:

- *Full* conformal prediction
- *Split* conformal prediction

Algorithm 1 Full conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, test x_{n+1}, algorithm $\hat{\mu}$, level α, grid of values $\mathcal{Y} = \{y, y', y'', \ldots\}$

```
for y ∈ 𝒴 do
    Train μ̂^y(x) = μ̂(x; (x_1, y_1), ..., (x_n, y_n), (x_{n+1}, y))
    Compute R_i^y = |y_i − μ̂^y(x_i)| for i = 1, ..., n
    Sort R_1^y, ..., R_n^y in increasing order: R_(1)^y ≤ ... ≤ R_(n)^y
    Compute R_α^y = R_(k)^y with k = ⌈(1 − α)(n + 1)⌉
    Compute R^y = |y − μ̂^y(x_{n+1})|
end for
C_α(x_{n+1}) = {y ∈ 𝒴 : R^y ≤ R_α^y}
```

for $y \in \mathcal{Y}$ **do**

Train $\hat{\mu}^y(x) = \hat{\mu}(x; (x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y))$

Compute $R_i^y = |y_i - \hat{\mu}^y(x_i)|$ for $i = 1, \ldots, n$

Sort $R_1^y, \ldots, R_n^y$ in increasing order: $R_{(1)}^y \leq \ldots \leq R_{(n)}^y$

Compute $R_\alpha^y = R_{(k)}^y$ with $k = \lceil (1-\alpha)(n+1) \rceil$

Compute $R^y = |y - \hat{\mu}^y(x_{n+1})|$

end for

$C_\alpha(x_{n+1}) = \{y \in \mathcal{Y} : R^y \leq R_\alpha^y\}$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$. This is the only assumption of the method
- The prediction interval

  $$C_\alpha(x_{n+1}) = \{y \in \mathbb{R} : R^y \leq R^y_\alpha\},$$

  satisfies

  $$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

  if and only if $\alpha \in \{1/(n+1), 2/(n+1), \ldots, n/(n+1)\}$
- Informally, the null hypothesis that the random variable Y_{n+1} will have the outcome y, i.e.

  $$H_y : Y_{n+1} = y$$

  is rejected when $R^y > R^y_\alpha$

Nonparametric Statistics

- Machine Learning has strong historical roots in Nonparametric Statistics
- K-Nearest Neighbors was introduced by two statisticians (students of Jerzy Neyman), Evelyn Fix and Joseph Hodges (Fix and Hodges, 1951)
- Conformal Prediction turns out to have roots in Permutation Testing (Fisher, 1925; Efron, 2021)

Prediction interval for Y_{n+1} (Vovk et al., 2005)	Confidence interval for Δ (Lehmann, 1963)
Supervised learning Training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ Test point (X_{n+1}, Y_{n+1})	Two-sample location shift model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F(x)$ $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} F(y - \Delta)$
$H_y : Y_{n+1} = y$	$H_d : \Delta = d$
$(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y)$	$x_1, \ldots, x_n, y_1 - d, \ldots, y_m - d$
$\hat{C} = \{y : p_y^* > \alpha\}$	$\hat{C} = \{d : p_d^* > \alpha\}$

Table of Contents

Algorithm 2 Split conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, x_{n+1}, algorithm $\hat{\mu}$, validation sample size m, level α

1: Split $\{1, \ldots, n\}$ into L of size w and I of size $m = n - w$

2: Train $\hat{\mu}_L(x) = \hat{\mu}(x; (x_l, y_l), l \in L)$

3: Compute $R_i = |y_i - \hat{\mu}_L(x_i)|$ for $i \in I$

4: Sort $\{R_i, i \in I\}$ in increasing order: $R_{(1)} \leq \ldots \leq R_{(m)}$

5: Compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : |y - \hat{\mu}_L(x_{n+1})| \leq R_\alpha\} \\ &= [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha] \end{aligned}$$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$
- The prediction interval

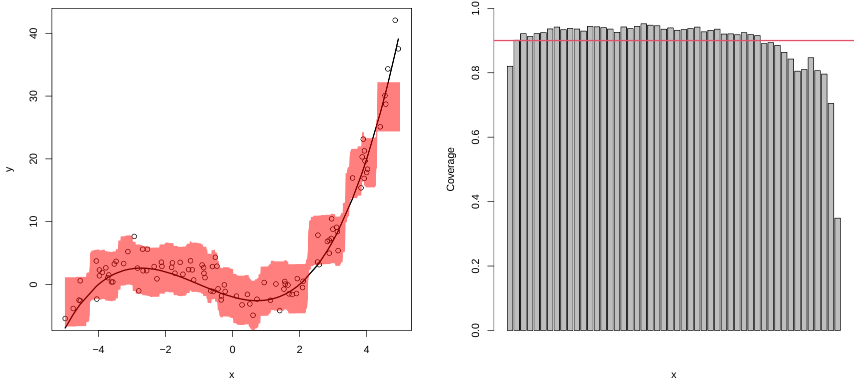
$$C_\alpha(x_{n+1}) = [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha]$$

  satisfies

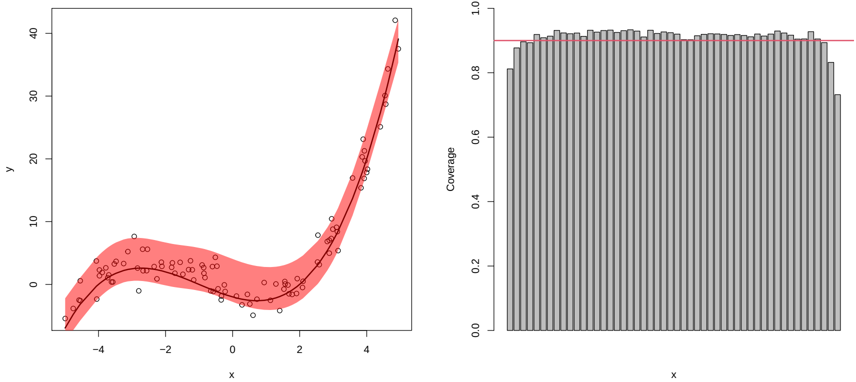
$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

  if and only if $\alpha \in \{1/(m+1), 2/(m+1), \ldots, m/(m+1)\}$
- Note that in computing the critical value $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$, we need to have $k \leq m$, which happens if $\alpha \geq 1/(m+1)$ (otherwise if $k > m$ we need to set $R_\alpha = +\infty$)

Random Forest



Smoothing splines



Conformity scores

- In the previous algorithm we used a statistic, called *conformity score*, which is the absolute value of the residual

$$R_i = |y_i - \hat{\mu}_L(x_i)|, \quad i \in I$$

  where $\hat{\mu}_L$ is an estimator of $\mathbb{E}(Y \mid X)$ based on $\{(X_i, Y_i), i \in L\}$
- The oracle knows the conditional distribution of $Y \mid X$. The oracle prediction interval

$$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$$

  where $q^\gamma(x)$ is the γ-quantile of $Y \mid X = x$, guarantees exact conditional coverage

$$\mathrm{P}(Y \in C^*_\alpha(X) | X = x) = 1 - \alpha \quad \forall\ x$$

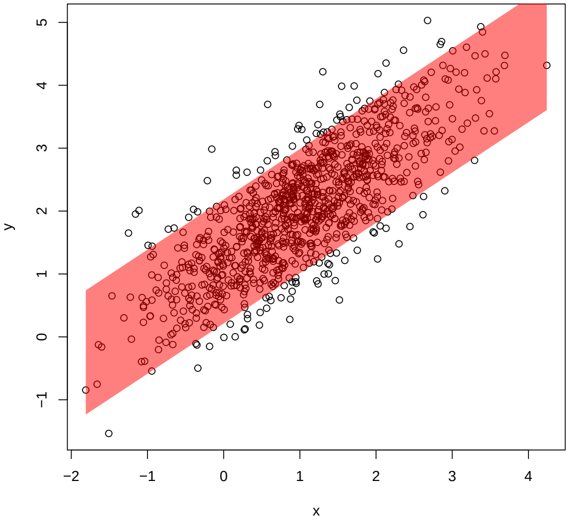
Suppose that

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right)$$

then the conditional distribution of $Y \mid X = x$ is

$$(Y|X = x) \sim N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1 - \rho^2)\right)$$

from which we can compute the quantile $q^\gamma(x)$



$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$ as a function of x

Table of Contents

Conformal quantile regression

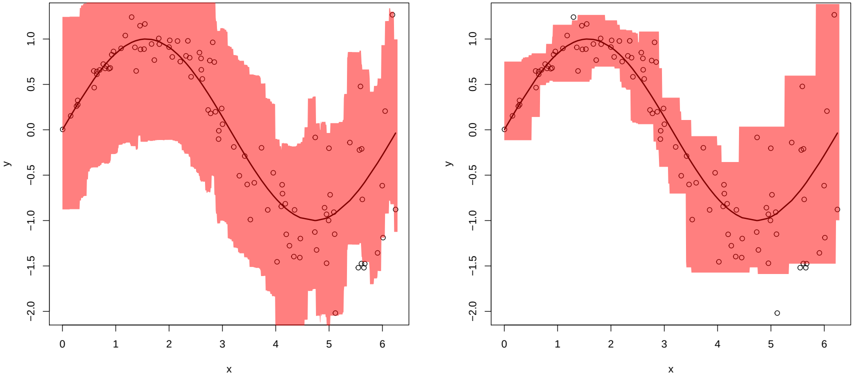
- Compute conformity scores

$$R_i = \max\left\{\hat{q}_L^{\gamma}(X_i) - Y_i,\, Y_i - \hat{q}_L^{1-\gamma}(X_i)\right\}, \quad i \in I$$

  where $\hat{q}_L^{\gamma}$ is an estimator of the γ-quantile of $Y \mid X$ based on $\{(X_i, Y_i), i \in L\}$
- Sort $\{R_i, i \in I\}$ in increasing order, obtaining $R_{(1)} \leq \ldots \leq R_{(m)}$, and compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$
- Compute the prediction interval

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : \max\left\{\hat{q}_L^{\gamma}(x_{n+1}) - y, y - \hat{q}_L^{1-\gamma}(x_{n+i})\right\} \leq R_\alpha\} \\ &= [\hat{q}_L^{\gamma}(x_{n+1}) - R_\alpha, \hat{q}_L^{1-\gamma}(x_{n+1}) + R_\alpha] \end{aligned}$$

  or $C_\alpha(x_{n+1}) = \emptyset$ if $R_\alpha < (1/2)(\hat{q}_L^{\gamma}(x_{n+1}) - \hat{q}_L^{1-\gamma}(x_{n+1}))$



$$X_i \sim U(0, 2\pi), \epsilon_i \sim N(0, 1), Y_i = \sin(X_i) + \frac{\pi|X_i|}{20}\epsilon_i$$