

Lecture 3: Conformal prediction

Vladimir Vovk

Centre for Reliable Machine Learning
Department of Computer Science
Royal Holloway, University of London

Mahalanobis Memorial Lectures 2020–21
19 March, 2021

Main points of this lecture

- Conformal prediction combines some advantages of machine learning and statistics.
- It can output prediction sets or predictive distributions.
- Efficiency of conformal prediction is an interesting research programme.

Plan

- 1 Machine learning
- 2 Conformal prediction sets
- 3 Conformal predictive distributions
- 4 Efficiency of conformal prediction

Theory and practice of machine learning

- Machine learning algorithms often work very well in practice.
- A theory of machine learning also exists (mostly based on VC dimension or Rademacher complexity).
- But there is a chasm between theory and practice. Theoretical results about performance guarantees for machine learning algorithms look fine asymptotically but become very weak when applied to practical datasets; not used in practice.

The two cultures



Leo Breiman (2001).

Statistical modeling: the two cultures (with discussion).

Statistical Science, 16:199–231.

There is a big difference between two strands of data analysis, old (statistics) and new (machine learning).

- **Data modelling culture** (98% of statisticians) often leads to irrelevant theory and questionable scientific conclusions.
- **Algorithmic modelling culture**, which concentrates on **predictive accuracy**, is the right way forward.

Assumptions in machine learning and statistics

- It is not really true that machine learning does not use data models.
- Mainstream machine learning uses one data model (which people often do not even bother to state): the observations are independent and identically distributed (the **IID model**, the **randomness model**, or the **exchangeability model**).
- Statistics: their models are usually more narrow, often parametric (e.g., Gaussian).
- An attempt at reconciliation: conformal prediction.

What do we lose under the IID model?

In statistics, people routinely

- compute **confidence intervals** for parameters (machine learning: we can't really do that; we do not have parameters);
- compute **prediction sets** for future observations (turns out to be feasible under the IID model, with guarantees of validity);
- do **testing** (testing the IID model turns out to be feasible, even for very complex observations).

Conformal prediction

- Conformal prediction adapts rank tests (standard in nonparametric statistics) to testing the IID assumption.
- Connection of testing with estimation (namely, with confidence intervals) goes back to at least Jerzy Neyman (1934).
- And prediction is just estimation of future data.

Plan

- 1 Machine learning
- 2 **Conformal prediction sets**
- 3 Conformal predictive distributions
- 4 Efficiency of conformal prediction

Conformity measures

- A **conformity measure** maps a finite multiset ζ of observations and another observation (x, y) to a **conformity score** $A(\zeta, (x, y))$.
- Intuitively, $A(\zeta, (x, y))$ shows how well (x, y) conforms to ζ .
- Sometimes $A(\zeta, (x, y))$ shows how strange (x, y) is compared with ζ , and then I will say “**nonconformity measure**”.
- Examples of nonconformity measures:

$$A(\zeta, (x, y)) := |y - \hat{y}|, \quad A(\zeta, (x, y)) := \left| \frac{y - \hat{y}}{\hat{\sigma}} \right|,$$

where \hat{y} is a prediction for the label y of the test object x computed from ζ as training set, and $\hat{\sigma}$ is an estimate of its accuracy.

Conformal prediction

Let $z_1 = (x_1, y_1), \dots, z_n = (x_n, y_n)$ be the training set and x_{n+1} be a new object. To find the prediction set for its label y_{n+1} at a significance level $\epsilon \in (0, 1)$:

- For each possible label y :
 - compute the conformity scores

$$\alpha_i := A(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n, (x_{n+1}, y)\}, z_i), \quad i = 1, \dots, n,$$

$$\alpha_{n+1} := A(\{z_1, \dots, z_n\}, (x_{n+1}, y))$$

($\{\cdot\}$ is a multiset); **all** of them depend on y ;

- compute the rank-based **p-value**

$$p(y) := \frac{|\{i : \alpha_i \leq \alpha_{n+1}\}|}{n+1}.$$

- Output the **prediction set** $\Gamma_{n+1} := \{y : p(y) > \epsilon\}$.

Validity result

Theorem

The probability that the conformal predictor makes an error (i.e., Γ_{n+1} fails to include y_{n+1}) does not exceed ϵ .

- This is the property of **validity**.
- Validity being achieved automatically, only **efficiency** is in question (we want the prediction sets to be small).

Stronger versions

- If we treat ties between (non)conformity scores more carefully, the p-values will be distributed uniformly on $[0, 1]$.
- Then the probability of error will be exactly ϵ .
- In the **online protocol** (observations arrive sequentially and after making each prediction we add them to the training set), the consecutive p-values are independent (and so the consecutive prediction sets make errors independently).
- The difference is tiny for training set of a reasonable size.

Regression

- Suppose $y \in \mathbb{R}$ (**regression** problem).
- Let me use **ridge regression** as a simple example of an “underlying algorithm”.
- The prediction for the label y of a new object x given a training set is

$$\hat{y} := x'(X'X + aI)^{-1}X'Y,$$

where X is the data matrix and Y is the vector of labels for the training set.

- Ridge regression has a Bayesian justification.

Ridge Regression Confidence Machine

- Let us take $A(\zeta, (x, y)) := |y - \hat{y}|$.
- How can we go over all $y \in \mathbb{R}$ to apply conformal prediction?
- We can solve the linear equations $\alpha_i(y) = \alpha_{n+1}(y)$ in y and compute Γ_{n+1} for the conformal predictor based on RR.
- As a result, Γ_{n+1} can be computed in time $O(n \log n)$.
- For **kernel ridge regression** (nonparametric version of ridge regression obtained using the “kernel trick”) the computation time is $O(n^2)$.
- We call this **Ridge Regression Confidence Machine (RRCM)**.

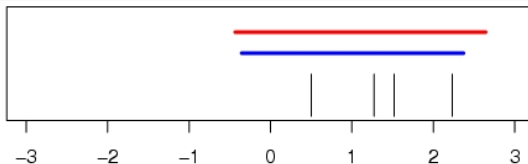
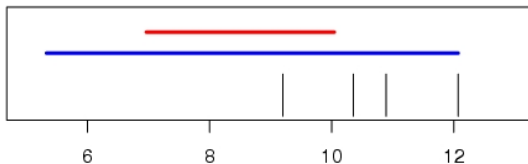
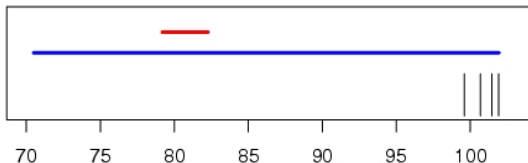
Bayesian predictions

- Bayesian statistics provides ideal predictions for future observations **provided we know the data-generating distribution**.
- These predictions are both **valid** (e.g., the right coverage probability for prediction intervals) and **efficient** (we can make them as short as possible, e.g., in expectation).
- Conformal prediction: validity is guaranteed (under a nonparametric assumption), and we try to achieve efficiency.

Breaking the Bayesian assumption

- Now let's see what happens when the Bayesian assumption is violated (but the IID assumption still holds).
- Suppose $y_i \sim N(\theta, 1)$, where $\theta \sim N(0, 1)$.
- Next slide: a version of Larry Wasserman's picture;
 $\epsilon := 20\%$; four observations are generated from $N(\theta, 1)$ for different θ .
- The blue lines are the CP prediction intervals and the red lines are the Bayes prediction intervals.

Bayes prediction intervals can mislead

 $N(1, 1)$  $N(10, 1)$  $N(100, 1)$

Efficiency

- The observations are generated from $N(\theta, 1)$.
- When $\theta = 1$ (and so the Bayesian assumption can be regarded as satisfied), the Bayes prediction intervals are on average only slightly shorter than RRCM's (3.08 vs 3.36; Bayes intervals are shorter in 54% of cases).
- But as θ grows, RRCM's intervals also grow (in order to cover the observations), whereas the width of the Bayes prediction intervals is constant. (For $\theta = 100$: 3.08 vs 31.2.)

Parametric vs nonparametric statistics

- No matter how carefully you choose your prior, you may be wrong.
- In parametric statistics, it is widely believed that, at least asymptotically, the choice of the prior does not matter much: the data will swamp the prior.
- However, even in parametric statistics the model (such as $N(\theta, 1)$) itself may be wrong.
- In nonparametric statistics, the situation is much worse:
the prior can swamp the data, no matter how much data you have

(Diaconis and Freedman, 1986). In this case, using Bayes prediction intervals becomes problematic.

Computational efficiency of full conformal prediction

- For a fairly narrow (but important) class of prediction algorithms, we can compute the conformal prediction intervals efficiently.
- Besides Ridge Regression and Kernel Ridge Regression, this class includes Nearest Neighbours.
- A more advanced result (Jing Lei): this is true for the Lasso as well.
- In general (especially when we need to normalize the features and tune some parameters), full conformal prediction can be very slow: we have to retrain the underlying predictor for each new test object and each postulated label for it.

Split conformal prediction

- To speed up computations, we can split the training set into two parts: the **training set proper** ζ and the **calibration set** C .
- Compute the conformity score $\alpha(z) := A(\zeta, z)$ for each $z \in C$ and for $z := (x, y)$, the test object x with a postulated label y .
- Compute the p-value from the calibration set:

$$p(y) := \frac{|\{z \in C \cup \{(x, y)\} : \alpha(z) \leq \alpha((x, y))\}|}{|C| + 1}.$$

Advantages of split conformal prediction

- In interesting cases, we can preprocess ζ so that computing $A(\zeta, z)$ for each z becomes quick (think of, e.g., $A(\zeta, (x, y)) := |y - \hat{y}| / \hat{\sigma}$).
- Computing p-values $p(y)$ becomes efficient.

Cross-conformal prediction (1)

- Split conformal predictors are computationally efficient, but they use only part of the training set for training (training set proper) and another part for calibrating conformity scores (calibration set).
- Full conformal predictors use the full training set both for calibration and training (which shows in their predictive efficiency).
- Can't we have the best of both worlds?

Cross-conformal prediction (2)

- My simple proposal (**cross-conformal prediction**, like cross validation): divide the training set into a number of folds, use each fold in turn as calibration set (training on the remaining folds), and average the resulting p-values. Only approximate validity.
- Barber et al.: jackknife+, a provably valid version of cross-conformal prediction.

Prediction in an open world

- Conformal prediction can be used “in an open world” (situation where we do not have an exhaustive list of labels).
- Suppose that we want to classify Android malware samples into families, given a training set of classified malware samples.
- When using conformal prediction, we compute a p-value for each possible family.
- What if the p-values for **all** families are small?
- This means that either we are witnessing a rare event, or this is a new family (“zero-day attack”).

Plan

- 1 Machine learning
- 2 Conformal prediction sets
- 3 Conformal predictive distributions**
- 4 Efficiency of conformal prediction

- Statisticians are fond of pointing out that p-values are **not** probabilities.
- p-values need some discipline to become probabilities.
- The definition of **p-values** that treats ties properly is

$$p(y) = p(y, \tau) := \frac{|\{i : \alpha_i < \alpha_{n+1}\}| + \tau |\{i : \alpha_i = \alpha_{n+1}\}|}{n + 1},$$

where $i = 1, \dots, n + 1$ and $\tau \in [0, 1]$ is a random number.

- What if $p(y)$ ranges from 0 to 1 (monotonically increasing) as y ranges from $-\infty$ to ∞ ?
- It starts looking like a distribution function.

Examples

- Examples of suitable conformity measures:

$$A(\zeta, (x, y)) := y - \hat{y}, \quad A(\zeta, (x, y)) := \frac{y - \hat{y}}{\hat{\sigma}}$$

(dropping absolute value).

- Let's consider the first one. There are still three versions:
 - **deleted**, as defined earlier (\hat{y} is computed from x and ζ);
 - **ordinary** (\hat{y} is computed from x and $\zeta \cup \{(x, y)\}$);
 - **studentized** (intermediate).

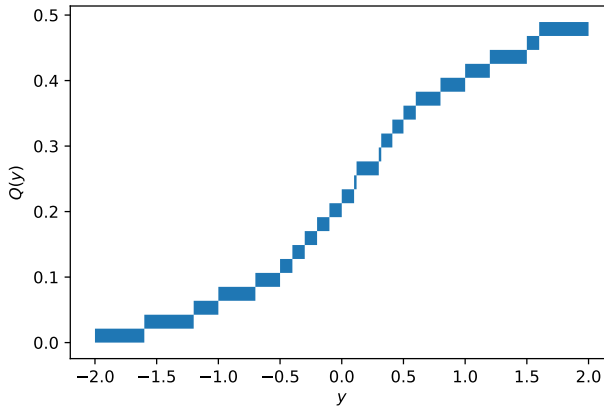
Getting a distribution function

- We say that $p(y, \tau)$, as defined earlier, is a **conformal predictive system (CPS)** if $p(y, \tau)$ is increasing in y (we know it's increasing in τ) and

$$\lim_{y \rightarrow -\infty} p(y, 0) = 0, \quad \lim_{y \rightarrow \infty} p(y, 1) = 1.$$

- The studentized version of RR is a CPS.
- In theory (and probably only in theory) the deleted and ordinary versions can violate the monotonicity of $p(y, \tau)$ in y ; this can only happen in the presence of extremely “high-leverage” objects.

What a conformal predictive distribution may look like



Calibration in probability

- Previous slide: we let τ vary between 0 and 1.
- We get a “thick distribution function”; RR gives ones of thickness $1/(n+1)$ apart from at most n points (usually very thin).
- But their thickness allows us to achieve the property of validity known in probability forecasting as **calibration in probability** (Dawid, Gneiting): the distribution of $p(y_n, \tau_n)$ is uniform on $[0, 1]$.

Calibration and sharpness

- Gneiting et al.'s motto (“paradigm”):
maximizing the sharpness of the predictive distributions subject to calibration.
- Calibration (in probability): agreement between the predictions and observations.
- Sharpness: property of predictions only.



Tilman Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery (2007).

Probabilistic forecasts, calibration and sharpness.
Journal of the Royal Statistical Society B, 69:243–268.

Computational efficiency

- As in the case of RRCM, we can compute the CPS based on RR explicitly.
- As for RRCM, this involves solving linear equations.
- The prediction can be computed in time $O(n^2)$.

Split conformal prediction

- For general prediction algorithms, in addition to full conformal prediction being slow, it is difficult to guarantee that $p(y, \tau)$ is increasing in y (this can be violated even for deleted and ordinary residuals!).
- If we use split conformal prediction, not only computing p-values becomes efficient, it also becomes easy to ensure that $p(y, \tau)$ is increasing in y :
 - Think again of $A(\zeta, (x, y)) := (y - \hat{y})/\hat{\sigma}$.
 - Alternatively, we can take any predictive system (such as Bayesian) $A(\zeta, (x, y))$ (increasing in y) without validity properties under the IID assumption and turn it into $p(y, \tau)$ (calibrated in probability).

Plan

- 1 Machine learning
- 2 Conformal prediction sets
- 3 Conformal predictive distributions
- 4 Efficiency of conformal prediction

De-Bayesing Bayesian algorithms

- Suppose the Bayesian assumption happens to be true.
- Do we lose much when we conformalize the Bayesian algorithm? (Asked independently by Evgeny Burnaev and Larry Wasserman.)
- We might expect (Larry Wasserman): when we do, the Bayesian assumption is very fragile. It's safer not to use it.

BRR vs conformalized RR

- Suppose that, in addition to the assumptions of BRR (=Bayesian ridge regression), x_n are IID and $\Sigma := \mathbb{E}(x_n x_n')$ exists and is non-singular.
- Then

$$\sqrt{n}(B^* - C^*) \xrightarrow{\text{law}} N(0, A),$$

$$\sqrt{n}(B_* - C_*) \xrightarrow{\text{law}} N(0, A),$$

where

- (B_*, B^*) is the Bayes prediction interval,
- (C_*, C^*) is the conformal prediction interval,
- $A = A(\epsilon, \mathbb{E}(x_n), \Sigma)$.

Further details



Evgeny Burnaev and Vladimir Vovk (2014).

Efficiency of conformalized ridge regression.

Proceedings of Machine Learning Research **35**:1–18
(COLT 2014).

This paper gives an explicit expression for the asymptotic variance A .

Predictive systems

- There are similar results for predictive distributions: how fast conformal predictive distributions and oracular predictive distributions converge to each other?
- An unusual notion of convergence: in Yuri Belyaev's sense.




Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie (2019).

Nonparametric predictive distributions based on conformal prediction.

Machine Learning **108**:445–474.

References: basics of conformal prediction

 Vladimir Vovk, Alex Gammerman, and Glenn Shafer (2005).

Algorithmic learning in a random world.
New York: Springer.

 Glenn Shafer and Vladimir Vovk (2008).

A tutorial on conformal prediction.
Journal of Machine Learning Research **9**:371–421.

References: more advanced things



Jing Lei (2019).

Fast exact conformalization of the Lasso using piecewise linear homotopy.

Biometrika, 106:749–764.

RR was easy; Lasso is much more difficult.



Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani (2021).

Predictive inference with the jackknife+.

Annals of Statistics 49:486–507.

A conformal-type version of jackknife with guarantees of validity.

Thank you for your attention!