Conformal Prediction: an Introduction

Léo Andéol[1]

Institut de Mathématiques de Toulouse

October 2022

[1] leo.andeol@math.univ-toulouse.fr

Outline

1. Conformal Prediction
2. Split Conformal Prediction
3. Full Conformal Prediction
4. Conclusion
   - References

Motivation

- Machine learning models make pointwise predictions: a class label (classification), a real number (regression),
- In order to apply these models in real situations, we would rather like prediction **sets** that incorporate the uncertainty of the prediction,
- In particular, prediction sets with guarantees that are valid in practice (therefore finite-sample, distribution-free and model-free).

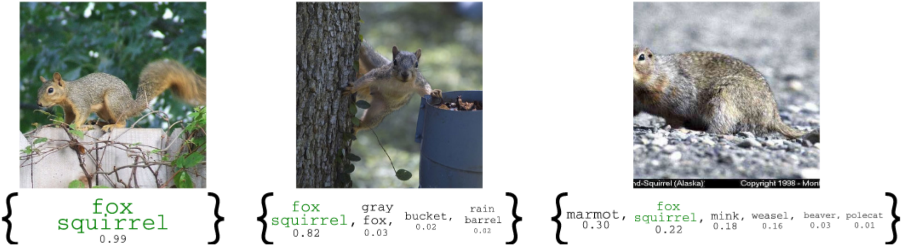
{ fox squirrel 0.99 }
{ fox squirrel, 0.82 gray fox, 0.03 bucket, 0.02 rain barrel 0.02 }
d-Squirrel (Alaska) Copyright 1998 - Mon
{ marmot, 0.30 fox squirrel, 0.22 mink, 0.18 weasel, 0.16 beaver, 0.03 polecat 0.01 }

Figure 1: *Illustration of prediction sets of variable size depending on the difficulty of the image (Angelopoulos and Bates, 2021)*

Notations and assumptions I

In order to formalize the problem, we introduce some notations and assumptions (as in (Angelopoulos and Bates, 2021; Romano et al., 2019; Gupta et al., 2022)):

- $(X_1, Y_1), ..., (X_{n+1}, Y_{n+1})$ a sequence of pairs of i.i.d (or exchangeable) random variables of an unknown distribution $\mathcal{P}$ with values in the measurable space $\mathcal{X} \times \mathcal{Y}$
- where $\mathcal{X}$ corresponds to the input space ($\mathbb{R}^d$, images, etc) and $\mathcal{Y}$ to the label space ($\mathbb{R}$ in regression often, $[\![K]\!]$ in classification with K the number of classes, etc)
- Consider $(X_1, Y_1), ..., (X_n, Y_n)$ as our training samples, and (X_{n+1}, Y_{n+1}) as our test sample.

Notations and assumptions II

We aim to construct a prediction set which contains the true label of the test sample with high probability.

- The prediction set is a measurable function of the training samples as well as the input of the test sample for a desired error rate $\alpha \in \left(0, \frac{1}{2}\right)$: $C_\alpha(X_{n+1}) \overset{\text{def}}{=} C_\alpha((X_1, Y_1), ..., (X_n, Y_n), X_{n+1}) \subseteq \mathcal{Y}$
- The measurability condition is: $\{(x_1, y_1), ..., (x_{n+1}, y_{n+1}) : y_{n+1} \in C_\alpha(x_{n+1})\}$ is measurable in $(\mathcal{X} \times \mathcal{Y})^{n+1}$
- We also consider a measurable function we call predictor (or model) $f : \mathcal{X} \to \mathcal{Y}$

Problem I

Our problem is to find $C_\alpha(X_{n+1})$ such that, for all distributions $\mathcal{P}$, for all $\alpha \in \left(0, \frac{1}{2}\right)$, we have

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

We refer to it as "marginal coverage" (Angelopoulos and Bates, 2021), but is also referred to as "conservative validity" (Vovk et al., 2005)

Problem II

- Of course there is a trivial solution, $C_\alpha(X_{n+1}) = \mathcal{Y}$
- Therefore the problem is not only to find C_α, but a non-trivial one and ideally the smallest
- There is multiple approaches to conformal prediction, we will introduce some classical ones!

Outline

1 Conformal Prediction

2 Split Conformal Prediction

3 Full Conformal Prediction

4 Conclusion
- References

Introduction

- Split conformal prediction is chronologically not the first type of conformal prediction, but the simplest
- It consists in partitioning (splitting) our training dataset into two disjoint subsets $\{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{train}}}$ for training our model and $\{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$ for building the prediction set.

A first approach I

Let us consider the case of regression where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, then we train our predictor $\hat{\mu} = f : \mathbb{R}^d \to \mathbb{R}$ given any algorithm $\mathcal{A}$:

$$\hat{\mu}(x) \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_{\text{train}}\}) \tag{2}$$

Now we must **conformalize** this predictor!

A first approach II

Given a tolerable error rate α,

- We define $C_\alpha(X_{n+1}) = [\hat{\mu}(x) - k, \hat{\mu}(x) + k]$ where k is the conformalization constant or margin
- This is one easy and intuitive way to define a prediction set but not the only one.
- We compute absolute residuals on the calibration set $R_i = |Y_i - \hat{\mu}(X_i)| \quad \forall i \in \mathcal{I}_{\text{cal}}$

$$Q_{1-\alpha}\left(R, \mathcal{I}_{\text{cal}}\right) := (1-\alpha)\left(1 + \frac{1}{|\mathcal{I}_{\text{cal}}|}\right)\text{-th empirical quantile of } \{R_i : i \in \mathcal{I}_{\text{cal}}]\} \tag{3}$$

- Finally, we can set $k = Q_{1-\alpha}\left(R, \mathcal{I}_{\text{cal}}\right)$

$$C_\alpha\left(X_{n+1}\right) = \left[\hat{\mu}\left(X_{n+1}\right) - Q_{1-\alpha}\left(R, \mathcal{I}_{\text{cal}}\right), \hat{\mu}\left(X_{n+1}\right) + Q_{1-\alpha}\left(R, \mathcal{I}_{\text{cal}}\right)\right]. \tag{4}$$

A first approach III

Theorem

If $(X_i, Y_i), i = 1, \ldots, n+1$ *are exchangeable, then the prediction interval* $C_\alpha(X_{n+1})$ *constructed by the aforementioned algorithm satisfies,*

$$P(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha. \tag{5}$$

We will look into the proof in another session.

A first approach IV

The scores $s(x, y)$ are a generalization of the absolute residuals

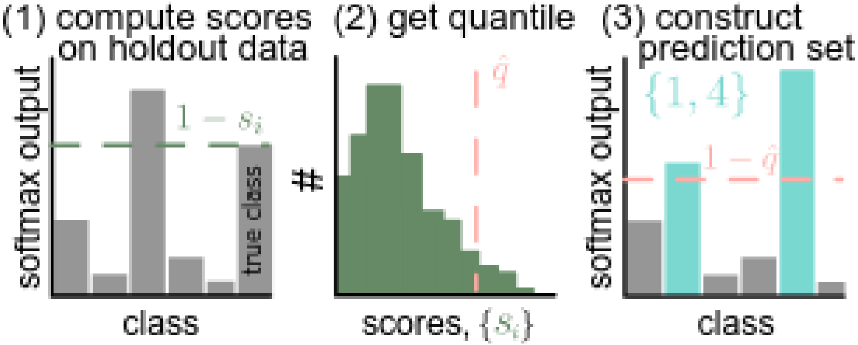

Figure 2: *Illustration of the general algorithm (for classification) to compute prediction sets (Angelopoulos and Bates, 2021)*

A first approach V

Empirically, we observe the following results,

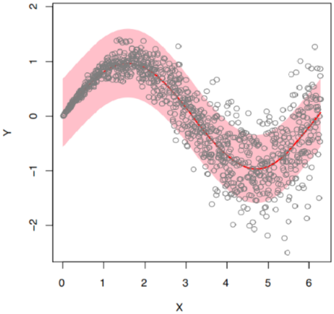

Figure 3: *Visualization of the resulting prediction intervals(Fontana et al., 2020)*

It is clear here that the prediction sets are fixed in size (even from the formulation with width $= 2Q_{1-\alpha}$, whereas ideally they would vary in x

Marginal vs Conditional Coverage I

We recall that the obtained coverage is marginal:

$$P(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha \tag{6}$$

Ideally, we would like conditional coverage:

$$P(Y_{n+1} \in C_\alpha(X_{n+1}) | X_{n+1}) \geq 1 - \alpha \tag{7}$$

However it was proved impossible by Vovk (2012) without supplementary assumptions. Therefore, we try to improve the empirical conditional coverage of our prediction sets.

Marginal vs Conditional Coverage II

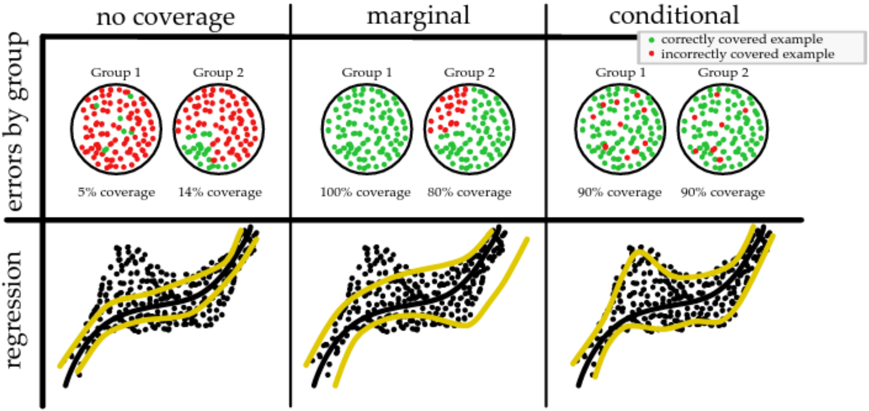

Figure 4: *Illustration of no, marginal and conditional coverage (Angelopoulos and Bates, 2021)*

Marginal vs Conditional Coverage III

Moreover, it has been proven (Vovk, 2012) that the coverage follow a Beta distribution,

$$P\left(Y_{n+1} \in C_\alpha\left(X_{n+1}\right) \mid \{(X_i, Y_i)\}_{i=1}^{n}\right) \sim \mathrm{Beta}(n+1-l, l), \tag{8}$$

where $l = \lfloor (n+1)\alpha \rfloor$. Notice that the conditioning is on the calibration set, not on the test sample X_{n+1}!

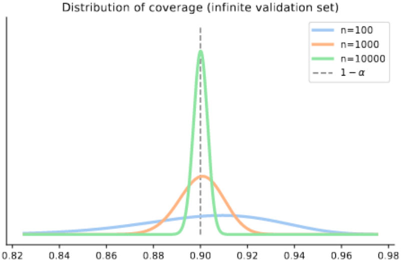

Figure 5: *Distribution of the coverage*

Non-conformity scores

In order to improve the conditional coverage of our prediction set, it needs to be more adaptive. There is two ways to consider this:

- We could replace the residual by a more general non-conformity score, involving for instance the conditional variance,
- We could formulate the prediction set itself more adaptively,

As shown by (Gupta et al., 2022), these two approaches are in fact equivalent. There is very few restrictions on which score function can be used, but guarantees remain valid. However, the value of the prediction set will depend on the scoring function.

CQR I

- We will now introduce the Conformalized Quantile regression(Romano et al., 2019).
- This approach, while using a different non-conformality score does not use an adaptive one, but has another approach to obtain adaptivity,
- We can replace our pointwise predictor (regressor in this case) by two quantile regressors, and therefore obtain already guaranty-free intervals as output of our model, which is adaptive (as long as the quantile function estimated is) and can be made valid with conformalization.

The naive prediction set, given $\alpha_{\mathrm{lo}} = \frac{\alpha}{2}$ and $\alpha_{\mathrm{hi}} = 1 - \frac{\alpha}{2}$ (for an equally split probability of error on both ends of the interval, but not necessary) is,

$$C(x) = \left[q_{\alpha_{\mathrm{lo}}}(x), q_{\alpha_{\mathrm{hi}}}(x)\right]. \tag{9}$$

CQR II

In order to conformalize, we follow the same pattern as the previous method, except we define residuals (non-conformity scores) in a different manner (notice that they can have a negative value):

$$R_i \overset{\text{def}}{=} \max \left\{ \hat{q}_{\alpha_{\mathrm{lo}}} \left(X_i \right) - Y_i, Y_i - \hat{q}_{\alpha_{\mathrm{hi}}} \left(X_i \right) \right\}. \tag{10}$$

Using these residuals, we compute the conformality constant in the same manner as the previous method and obtain the following prediction set:

$$C_\alpha^{\text{cqr}} = \left[\hat{q}_{\alpha_{\mathrm{lo}}} \left(X_{n+1} \right) - Q_{1-\alpha} \left(R, \mathcal{I}_{\text{cal}} \right), \hat{q}_{\alpha_{\mathrm{hi}}} \left(X_{n+1} \right) + Q_{1-\alpha} \left(R, \mathcal{I}_{\text{cal}} \right) \right]. \tag{11}$$

The Theorem 1 is also true for C_α^{cqr}.

CQR III

The intervals are much more adaptive, and we are closer to conditional coverage:

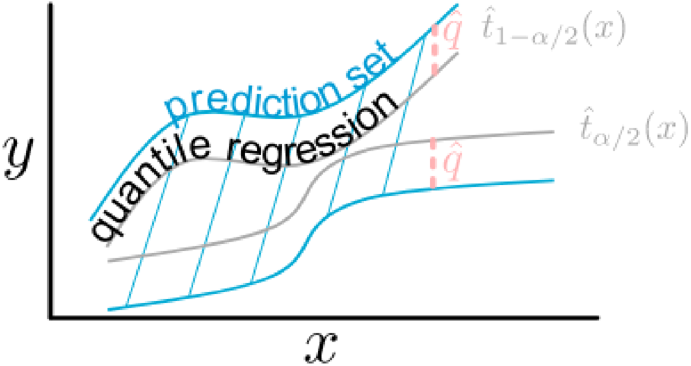

Figure 6: *Illustration of the CQR Method (Angelopoulos and Bates, 2021)*

Other types of prediction sets I

So far we have mentioned only simple types of prediction sets and a classical application (regression). However it is important to note that conformal prediction can be applied in a variety of situation such as segmentation (c.f. figure 7), or object detection (de Grancey et al., 2022) such as:

$$\begin{aligned} R^i_{x_{\min}} &= \hat{x}^i_{\min} - x^i_{\min} & R^i_{y_{\min}} &= \hat{y}^i_{\min} - y^i_{\min} \\ R^i_{x_{\max}} &= x^i_{\max} - \hat{x}^i_{\max} & R^i_{y_{\max}} &= y^i_{\max} - \hat{y}^i_{\max} \end{aligned} \tag{12}$$

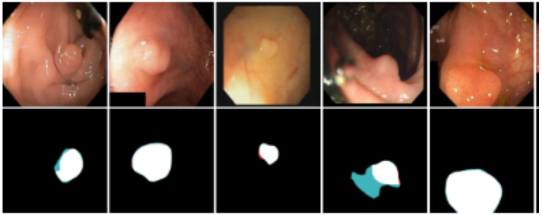

Figure 7: *Example of conformal prediction for segmentation (Angelopoulos and Bates,*

Other types of prediction sets II

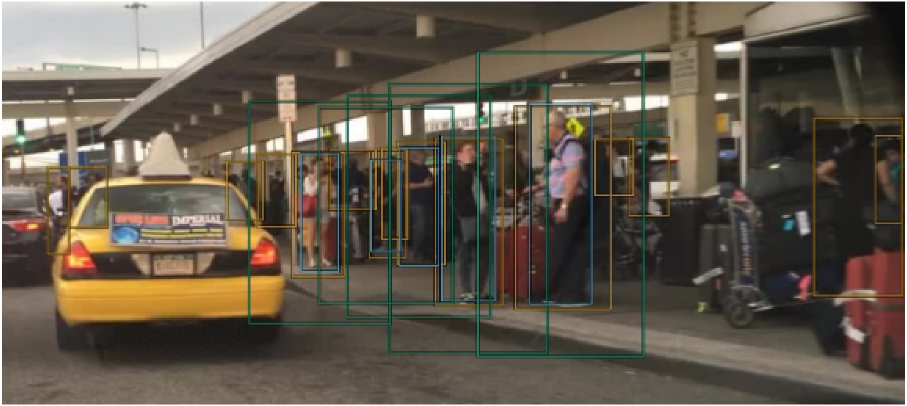

Figure 8: *Margins for object detection (de Grancey et al., 2022) (in yellow the ground truth, in blue the inference and in green the conformalized boxes)*

Outline

1 Conformal Prediction

2 Split Conformal Prediction

3 Full Conformal Prediction

4 Conclusion

- References

Introduction

- We now look at Full Conformal Prediction, the first approach chronologically
- The point of view here is quite different as, intuitively, for a given sequence $(X_1, Y_1), ..., (X_n, Y_n)$ and a given X_{n+1}, we will consider all $y \in \mathcal{Y}$ that would conform within the sequence $(X_1, Y_1), ..., (X_{n+1}, y)$.
- This conformity within the sequence is measured using non-conformity score on the sequence : $S : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}$, we set to $|y - f(X_{n+1})|$[2]

[2] We could use almost any S that is permutation invariant to the n first samples

Method

From this score function we can build the following prediction interval

$$\mathcal{C}\left(X_{n+1}\right)=\left\{y: S\left(\left(X_1, Y_1\right), \ldots,\left(X_n, Y_n\right),\left(X_{n+1}, y\right)\right) \leq Q_{1-\alpha}(y)\right\}, \tag{13}$$

where $Q_{1-\alpha}(y)$ is the $\lceil(1-\alpha)(n+1)\rceil/\mathrm{n}$ empirical quantile of the scores $(R_1, \ldots, R_n)$, where

$$R_i(y)=S\left(\left(\left(X_1, Y_1\right), \ldots,\left(X_n, Y_n\right),\left(X_{n+1}, y\right)\right) \setminus\left(X_i, Y_i\right),\left(X_i, Y_i\right)\right). \tag{14}$$

However, compared to split conformal prediction, the score does not only rely on the ith sample but on all of them, and therefore the quantile $Q_{1-\alpha}(y)$ depends on y. Moreover, the function f through the scores $R_i(y)$ depends on y and therefore need to be retrained for all $y \in \mathcal{Y}$.

Pros and Cons

- Unlike split conformal, all of the data is leveraged for both training and conformalization,
- However, is vastly more costly in classification ($\mathcal{Y}$ finite), and not computable for regression, except with assumptions on the predictor, or discretization of the space.
- Multiple methods in-between Split and Full conformal exist, and will be discussed in future sessions.

Conclusion

- Prediction sets with finite-sample, distribution-free, model-free guarantees,
- Split conformal prediction more common than full conformal prediction due to complexity and cost,
- Different non-conformity scores can be explored for better prediction sets,
- We will go into the main methods with more details in the coming sessions.

Thank you for attending!

References

A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021. [Online]. Available: https://arxiv.org/abs/2107.07511

Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf

C. Gupta, A. K. Kuchibhotla, and A. Ramdas, "Nested conformal prediction and quantile out-of-bag ensemble methods," *Pattern Recognition*, vol. 127, p. 108496, 2022.

V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: a unified review of theory and new challenges," 2020. [Online]. Available: https://arxiv.org/abs/2005.07972

V. Vovk, "Conditional validity of inductive conformal predictors," in *Proceedings of the Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. C. H. Hoi and W. Buntine, Eds., vol. 25. Singapore Management University, Singapore: PMLR, 04–06 Nov 2012, pp. 475–490. [Online]. Available: https://proceedings.mlr.press/v25/vovk12.html