Conformal prediction

A not-so-gentle mashup of oh-so-gentle introductions

Disclaimer

! (NIH)

A tall order

About UQ...

- $X \mapsto \hat{Y} = f(X)$ not enough for *decision making*. Need CIs / ...
- Bootstrap methods. Cost! Asymptotics & assumptions
- Asymptotic estimates (e.g. CLT+BE). But real life has **fat tails**...
- So... go Bayesian or go home?
  - costly, approximations are usually unjustified
  - priors are basically arbitrary, so why trust the posterior?

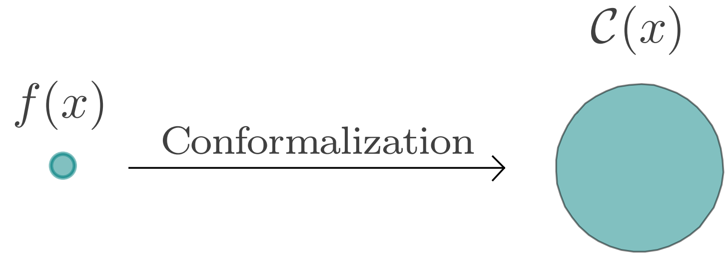
$\Rightarrow$ Need proper **"confidence sets"** with guarantees *based on data* and not on assumptions

The goal of conformal prediction

Given: **supervised trained model** f, and (unseen) **calibration dataset** $\mathcal{D}_{\mathrm{cal}}$ i.i.d.

Attach "**UQ / calibration layer**" to f which outputs "good" **prediction sets**

- Regression:
  intervals covering true values
- Classification:
  discrete sets containing the true class



with high probability

For new (X, Y) compute set $\mathcal{C}(X)$ s.t.

$$\mathbb{P}(Y \in \mathcal{C}(X)) \approx 1 - \alpha \qquad \text{(\textbf{coverage})}$$

What we get

Conformal predictor: $\mathcal{C}\colon \mathcal{X} \to 2^{\mathcal{Y}}$ with **guaranteed coverage**

- **Distribution-free**
- **Model-agnostic**: works with RFs, GBTs, NNs and any **black-box**
- **Efficiency**: $|\mathcal{C}(X)| \ll |\mathcal{Y}|$ (no trivial solution)
- **Adaptivity**: $|\mathcal{C}(X)|$ depends on the model's uncertainty
- $\mathcal{D}_{\mathrm{cal}}$ unseen by f, hence "*split CP*". Same distribution as $\mathcal{D}_{\mathrm{train}}$

What we get

Conformal predictor: $\mathcal{C}: \mathcal{X} \to 2^{\mathcal{Y}}$ with **guaranteed coverage**

- $\mathcal{C} = \mathcal{C}(f, \mathcal{D}_{\text{cal}}, \alpha)$
- **Distribution-free**
- **Model-agnostic**: works with RFs, GBTs, NNs and any **black-box**
- **Efficiency**: $|\mathcal{C}(X)| \ll |\mathcal{Y}|$ (no trivial solution)
- **Adaptivity**: $|\mathcal{C}(X)|$ depends on the model's uncertainty
- $\mathcal{D}_{\text{cal}}$ unseen by f, hence "*split CP*". Same distribution as $\mathcal{D}_{\text{train}}$

What this means

- Interpretable? ✓
  - Context for the outputs. Alternatives matter in high-stakes decisions!

    $$\{\text{defective ball-bearing}, \text{unbalanced wheel}\} \neq \{\text{defective ball-bearing}, \text{axle failure}\}$$

- Useful for automated decisions? ✓
  - Rigorous, calibrated uncertainty estimates
  - OOD detection
  - ...
- Criterion to select between algorithms ("best" prediction sets)
- And more!

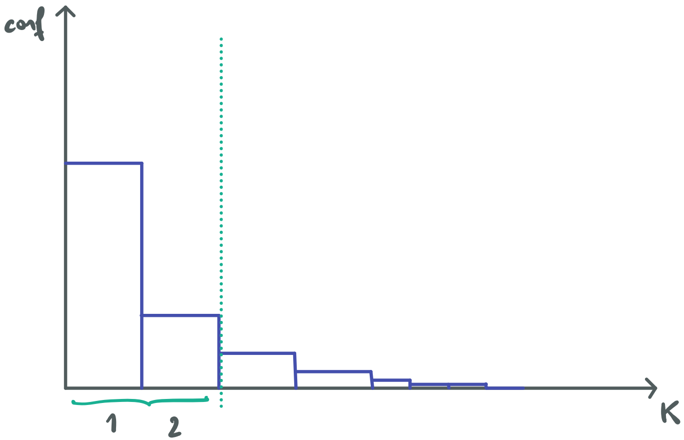
An algorithm for classification

A first idea

$f\colon \mathcal{X} \to [0,1]^K$ classifier. Desired cover. 90%

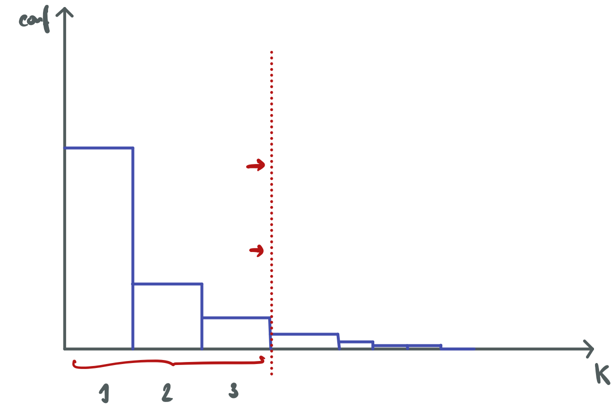
New x'. Sort confidences: $c'_{(1)}, \dots, c'_{(K)}$

($c_{(j)}$ is the j-th order statistic of c)



Include classes $\mathcal{C}(x') := \{f_{(1)}, \dots, f_{(m)}\}$ **while** $\sum_{j=1}^{m} c'_{(j)} < 0.9$

Poorly calibrated (NNs overconfident...)



For every $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ **calibration set**:

Sum confidences into $s_i := \sum_{j=1}^{m_i} c^i_{(j)}$

taking classes until **true** y_i, in position m_i

$\min \hat{q} \in [0,1]$ s.t. $\hat{\mathbb{P}}(S \leqslant \hat{q}) \geqslant 0.9$

$\mathcal{C}(x') := \{f(x')_{(1)}, \dots, f(x')_{(m)}\}$ where m is s.t. $\sum_{j=1}^{m} c'_{(j)} < \hat{q}$

Calibrated on unseen data!

Adaptive predictive sets for classification

For each $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ compute a **conformity score**: $s_i = s(x_i, y_i) := \sum_{j=1}^{m(y_i)} f(x_i)_{(j)}$,

s is the model's *up-to-true-label total confidence*

Look at $\hat{q}_{.9}$, the 90th percentile of s $\quad q_{.9} := F_S^{-1}(.9) := \min_q \{\mathbb{P}(S \leqslant q) \geqslant .9\}$

For $\sim 90\%$ of $\mathcal{D}_{\text{cal}}$'s samples f has up-to-true-label confidence below $\hat{q}_{.9}$

Define $\quad \mathcal{C}(x') := \{y_1, \ldots, y_{m'} : \sum_{j=1}^{m'} f(x')_{(j)} \leqslant \hat{q}_{.9}\}$

Adding classes until the total confidence reaches $\hat{q}_{.9}$ results in $\sim 90\%$ of the sets $\mathcal{C}(x_i)$ including the true class

Profit! $\quad \mathbb{P}(Y' \in \mathcal{C}(X')) \approx .9$

A simpler algorithm

1. For each $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ compute a **conformity score**: $s_i = s(x_i, y_i) := 1 - f(x_i)_{y_i}$,

   S is the *model's "uncertainty" for the* ***correct class***

   Look at the 90th percentile $\hat{q} = \hat{q}_{.9}$

   $\mathbb{P}(S \leqslant \hat{q}_{.9}) \geqslant .9$ means that:

   f has confidence $\geqslant 1 - \hat{q}$ for 90% of $\mathcal{D}_{\text{cal}}$'s labels

2. In other words:

   $\sim 90\%$ of $\mathcal{D}_{\text{cal}}$'s labels have **true-class uncertainty** below $\hat{q}_{.9}^{(n)}$

3. Define $$\mathcal{C}(x') := \{ y \in \mathcal{Y} \colon s(x', y) \leqslant \hat{q}_{.9}^{(n)} \} = \{ y \colon f(x')_y \geqslant 1 - \hat{q}_{.9}^{(n)} \}$$

4. **Profit!**

The general recipe

What we did:

1. Take a **heuristic notion** of uncertainty associated to f

   Ex: softmax outputs

2. Define a **conformal score** $s(x,y) \in \mathbb{R}$ for all (x,y) and compute over $\mathcal{D}_{\text{cal}}$

   Higher is worse

3. Compute a high **conformal quantile** $\hat{q} = \hat{q}_{1-\alpha}$

   $(1-\alpha)$ fraction of samples in $\mathcal{D}_{\text{cal}}$ have score $\leqslant \hat{q}$

4. Define $$\mathcal{C}(x') := \{y : s(x', y) \leqslant \hat{q}\}$$

5. Then $$\mathbb{P}(Y' \in \mathcal{C}(X')) \approx 1 - \alpha$$

   (theorem)

The fundamental theorem

Theorem. ([VGS05])

Let $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *be* ***exchangeable****. Define* $\hat{q} := \hat{q}_{1-\alpha}^{(n)} := \hat{F}_{s,n}^{-1}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$*. Then* $\mathcal{C} = \mathcal{C}_{\hat{q}} \colon \mathcal{X} \to 2^{\mathcal{Y}}$ *constructed as*

$$\mathcal{C}(X) := \{y \in \mathcal{Y} \colon s(X, y) \leqslant \hat{q}\}$$

fulfills

$$1 - \alpha \leqslant \mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \leqslant 1 - \alpha + \frac{1}{n+1},$$

for f, α, n *arbitrary.*

Key property: Exchangeability implies that S_{n+1} is **indistinguishable** from the other S_i. It has equal probability of falling between any two conformal scores.

$$\mathbb{P}(S_{n+1} \leqslant S_k) = \frac{k}{n+1}.$$

But… how good can this be?

It's not magic…

- Constant scores? Useless sets
- Random scores? Useless sets
- Ranking of scores does not reflect model error? Useless sets
- Informative scores? ✓
- Adapted loss functions? ✓
- Heavy tails? ✓ Predictive sets will be larger

Beyond the discrete

Can we do the same for regression?

1. $Y = f(X) + \varepsilon \in \mathbb{R}$ regression model. Uncertainty heuristic: **residuals** $|Y - \hat{f}(X)|$

2. Compute **conformal scores** $s(x_i, y_i) := |y_i - f(x_i)|$ over $\mathcal{D}_{\text{cal}}$

3. Form $$\hat{q}_{1-\alpha} = \min \left\{ q \colon \frac{|\{i \colon s_i \leqslant q\}|}{n} \geqslant 1 - \alpha \right\}$$

4. Define $$\mathcal{C}(x') := \{y \in \mathbb{R} \colon s(x', y) \leqslant \hat{q}_{1-\alpha}\} = [\hat{f}(x') - \hat{q}_{1-\alpha}, \hat{f}(x') + \hat{q}_{1-\alpha}]$$

5. **Profit?**

   Constant size for prediction interval

   Why do we learn the mean $\mathbb{E}[Y \mid X]$, when **we care about quantiles**?

An idea

- Recall: CDF of $Y|X$ is $\mathbb{P}(Y \leqslant y|X)$

  α-th **conditional quantile function** $t_\alpha(X) := \inf\{y \in \mathbb{R}: \mathbb{P}(Y \leqslant y|X) \geqslant \alpha\}$

- The **conditional prediction interval**

$$\mathcal{C}(X) = [t_{\alpha/2}(X), t_{1-\alpha/2}(X)]$$

  trivially satisfies

$$\mathbb{P}(Y \in \mathcal{C}(X)|X) = 1 - \alpha$$

- Alas... we don't have access to the CDF of $Y|X$

  So we can **learn the quantiles** instead **and conformalize** (finite sample guarantees)

Conformalized Quantile Regression

- Use **pinball loss** ρ_α to learn quantiles $(\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2})$

$$\rho_\alpha(y, \hat{y}) = \begin{cases} \alpha\,(y - \hat{y}) & \text{if } y > \hat{y} \\ (1-\alpha)\,(\hat{y} - y) & \text{otherwise} \end{cases}$$

- Use **signed distance to the closest quantile** as score

$$s(x, y) := \max\{\hat{t}_{\alpha/2}(x) - y, y - \hat{t}_{1-\alpha/2}(x)\}$$

- Compute $\hat{q} = \hat{q}_{1-\alpha}$

- Same prediction rule:

$$\begin{aligned} \mathcal{C}(x') &= \{y \in \mathbb{R} : s(x', y) \leqslant \hat{q}\} \\ &= [\hat{t}_{\alpha/2}(x') - \hat{q}, \hat{t}_{1-\alpha/2}(x') + \hat{q}] \\ &\supseteq [\hat{t}_{\alpha/2}(x'), \hat{t}_{1-\alpha/2}(x')] \end{aligned}$$

[RPC19]

Conformalized Quantile Regression

- Use **pinball loss** ρ_α to learn quantiles $(\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2})$

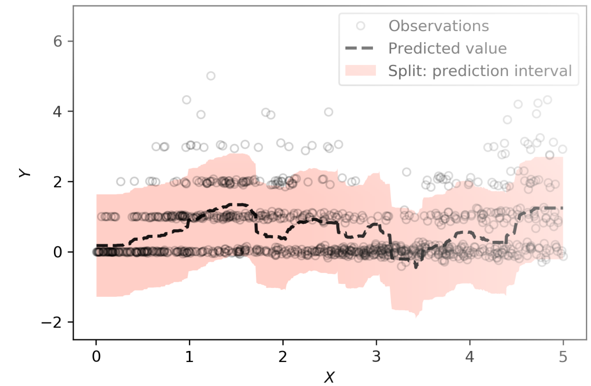
$$\rho_\alpha(y, \hat{y}) = \begin{cases} \alpha\,(y - \hat{y}) & \text{if } y > \hat{y} \\ (1-\alpha)\,(\hat{y} - y) & \text{otherwise} \end{cases}$$

- Use **signed distance to the closest quantile** as score

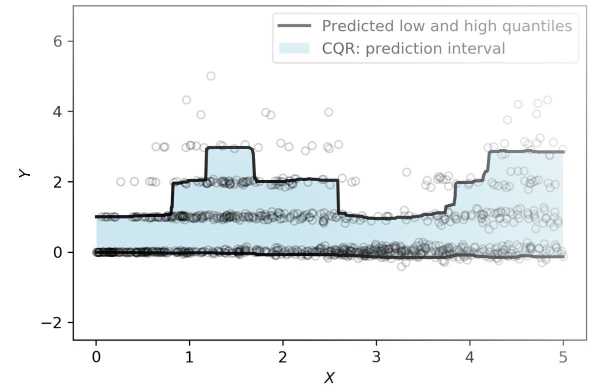
$$s(x, y) := \max\{\hat{t}_{\alpha/2}(x) - y, y - \hat{t}_{1-\alpha/2}(x)\}$$

- Compute $\hat{q} = \hat{q}_{1-\alpha}$
- Same prediction rule:

$$\begin{aligned} \mathcal{C}(x') &= \{y \in \mathbb{R} : s(x', y) \leqslant \hat{q}\} \\ &= [\hat{t}_{\alpha/2}(x') - \hat{q}, \hat{t}_{1-\alpha/2}(x') + \hat{q}] \\ &\supseteq [\hat{t}_{\alpha/2}(x'), \hat{t}_{1-\alpha/2}(x')] \end{aligned}$$



(a) Split: Avg. coverage 91.4%; Avg. length 2.91.



(c) CQR: Avg. coverage 91.06%; Avg. length 1.99.

[RPC19]

Potential stumbling blocks

Difficulty #1: Marginal vs conditional

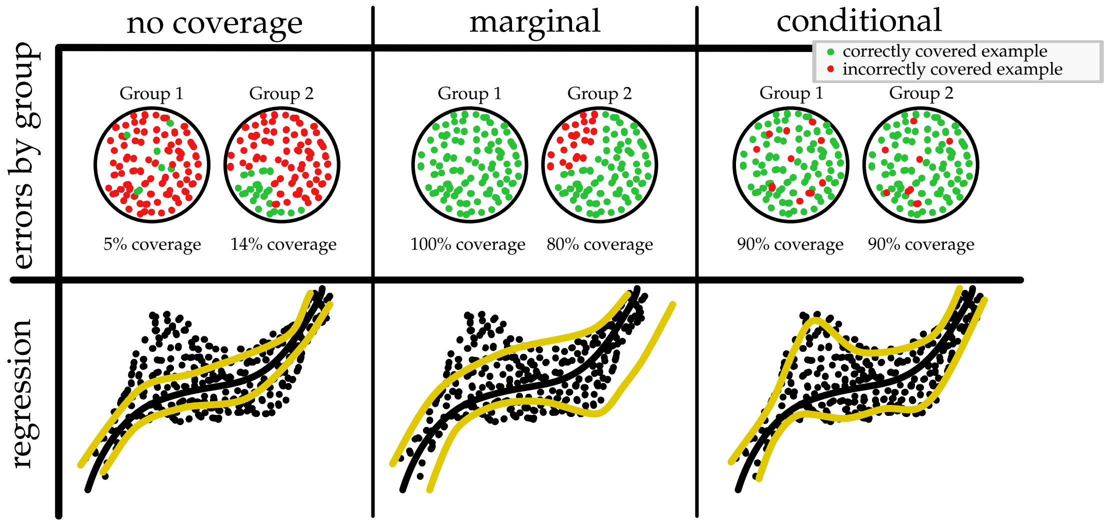
$\mathbb{P}(Y \in \mathcal{C}(X)) \approx 1 - \alpha$ is a *marginal* guarantee

- One usually wants a stronger **conditional guarantee** $\mathbb{P}(Y \in \mathcal{C}(X) | \mathcal{D})$

- Not conditional on $\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{train}}$
  - ⇨ Fluctuations wrt. $1 - \alpha$
  - ⇨ Need n large enough (see later, and [Vov12])

- Not conditional on groups
  - ⇨ **Coverage unbalanced** in $\mathcal{X}$ or $\mathcal{Y}$ (only "easy" samples) $C = \mathbb{E}[\mathbb{1}\{Y \in \mathcal{C}(X)\}]$
  - ⇨ Check coverage separately over a partition of $\mathcal{X}$ or $\mathcal{Y}$ $\hat{C} = \hat{\mathbb{E}}_{\mathcal{D}_{\text{val}}}[\mathbb{1}\{Y_i \in \mathcal{C}(X_i)\}]$
  - ⇨ Changes to the score, many techniques.

Conditional coverage



[AB22]

Measuring coverage

$$C = \mathbb{E}[\mathbb{1}\{Y \in \mathcal{C}(X)\}]$$

1. Cross-validation over $\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{val}}$ of **empirical coverage**

$$\hat{C} = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{x\,,\,y \in \mathcal{D}_{\text{val}}} \mathbb{1}\{y \in \mathcal{C}_{\mathcal{D}_{\text{cal}}}(x)\}$$

   Mean should concentrate around $1 - \alpha$

   Bad $\hat{C}$ is a good indicator of *distribution shift* (more later)

2. Distribution of $\hat{C}$ is known [AB22, App. C]

   Good checks available, formulas to look for errors. Use moments to verify implementation

3. Verify conditional coverage with **feature-stratified** or **size-stratified** coverage

Conditional guarantees

- **Feature-balanced CP** [Vov12, AB22]

  1. Want: $\mathbb{P}(Y' \in \mathcal{C}(X') | X_1' = g) \approx 1 - \alpha$ for all $g \in \{1, \ldots, G\} = \mathrm{range}(X_1)$
  2. Stratify by group: $s_i^{(g)}$, $\hat{q}^{(g)}$
  3. $\mathcal{C}(x) := \{y \colon s(x, y) \leqslant \hat{q}^{(x_1)}\}$ ✓

- **Class-conditional CP** [Vov12, AB22]

  1. Want: $\mathbb{P}(Y' \in \mathcal{C}(X') | Y = y) \approx 1 - \alpha$ for all $y \in \mathcal{Y}$
  2. Stratify by class: $s_i^{(k)}$, $\hat{q}^{(k)}$
  3. $\mathcal{C}(x) := \{y \colon s(x, y) \leqslant \hat{q}^{(y)}\}$ ✓

Difficulty #2: distribution shift

Distribution shift, $\{(X_i, Y_i)\}_{i=1}^{n+1}$ non exchangeable

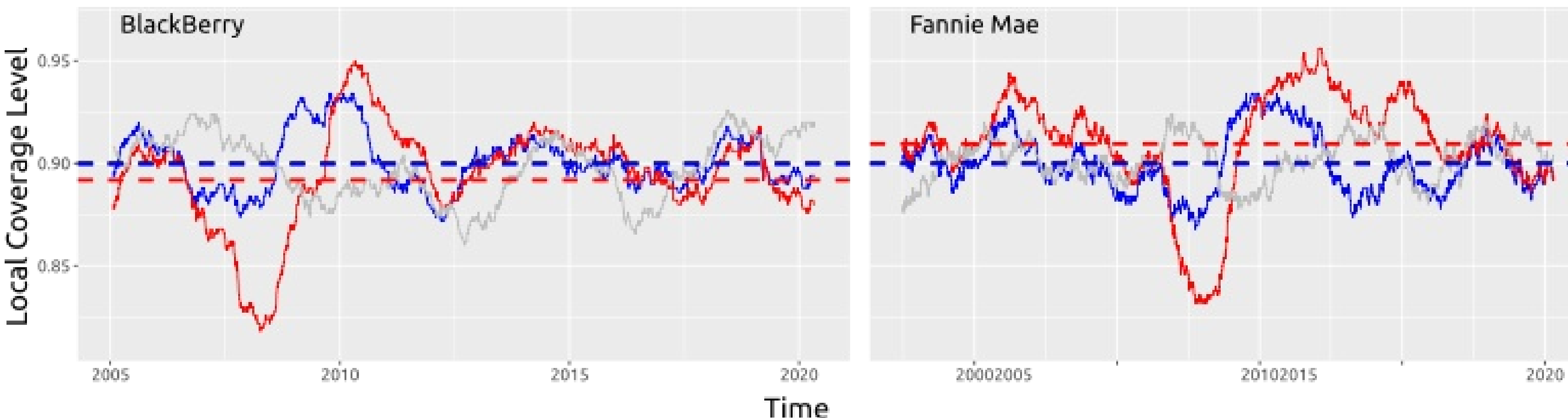
Time series, streaming data, finite data, interactive systems, ...

Adaptive CP [GC21]: Compute $\alpha_{n+1}, \alpha_{n+2}, \ldots,$ with $\begin{cases} \text{increase } \alpha_t & \text{if } \; Y_t \in \mathcal{C}(X_t) \\ \text{decrease } \alpha_t & \text{if } \; Y_t \notin \mathcal{C}(X_t) \end{cases}$

$$\alpha_{t+1} := \alpha_t + \gamma\,(\alpha - \mathrm{err}_t), \;\text{ where } \mathrm{err}_t := \mathbb{1}_{\mathcal{C}_t}(Y_t)$$

And reestimate $\hat{q}_{1-\alpha}$.

Also see [GC22, BCRT22]



Difficulty #2: distribution shift

Distribution shift, $\{(X_i, Y_i)\}_{i=1}^{n+1}$ non exchangeable

Time series, streaming data… $\textsc{nex}$CP [BCRT22]

- **Fixed, non-negative weights** $\sum w_i = 1$ for conformal scores (decay)
- Computes $\hat{q}_{1-\alpha}^{(n,w_i)}$ wrt. **weighted empirical distribution** $\hat{F}_{s,w}^{(n)} = \frac{1}{n+1} \sum w_i \, \delta_{s_i}$
- $\mathcal{C}(x_{n+1}) := \left\{ y \colon s(x_{n+1,y}) \leqslant \hat{q}_{1-\alpha}^{(n,w_i)} \right\}$

More difficulties

- In some domains, coverage is not the right notion!
  - ⇨ **Conformal Risk Control** [ABF+22]
- Data waste
  - ⇨ **Full** conformal prediction
  - ⇨ **jackknife+**

How good is my CP?

- Adaptivity: not guaranteed but essential. Smallest average $|\mathcal{C}(X)|$ not enough
- **Histogram** of $|\mathcal{C}(x_i)|$ informative but not conclusive
- Coverage checks: formulae to look for errors [AB22, §3]
  - Analytic expression for sample coverage
  - Use moments to verify implementation
  - Bad coverage is a good indicator of distribution shift
- Dependence on the calibration set:

$$\mathbb{P}(Y' \in \mathcal{C}(X') | \mathcal{D}_{\text{cal}}) \sim \text{Beta}(n+1-m, m), \quad m := \lfloor (n+1)\,\alpha \rfloor$$

Invert the CDF to compute n for δ, ε. See [Vov12] for this and more

Extensions

A long list

- Group-balanced CP: ensure per-group coverage
- Class-conditional CP: ensure per-class coverage
- Conformal risk control: minimise false negative rate, maximise "fairness",… [ABF+22]
- Outlier detection: p-values instead of the $3\,\sigma$ hack [BAL+21]
- CP under distribution shift: real data, streaming data, … [BCRT22, GC22, GC21]
- CP without exchangeability.
- Joint optimization: "smooth sorting", increases CW-efficiency [SDCD22]

Recap

The core idea

1. Take a **heuristic notion** of uncertainty associated to f
2. Define a **conformal score** $s(x, y) \in \mathbb{R}$ and compute over $\mathcal{D}_{\text{cal}}$
3. Compute a high **conformal quantile** $\hat{q} = \hat{q}^{(n)}_{1-\alpha}$
4. Define $$\mathcal{C}(x') := \{y : s(x', y) \leqslant \hat{q}\}.$$
5. Then: $$\mathbb{P}(Y' \in \mathcal{C}(X')) \approx 1 - \alpha.$$

Methods for classification and regression. Trivial to implement. Ready-to-use examples

Simple techniques to verify implementation

Adaptive and heuristic methods for time series

Applications to OOD and multi-task

Can optimize arbitrary risks

The core issues

1. Lack of conditional guarantees
2. Efficiency (class-wise and group-wise)
3. Distribution shift
4. ...

Happy conformalizing!

miguel@appliedai.de

Sources⇝

A learning path

1. Videos: **excellent** tutorials by Angelopolous & Bates (YouTube)
2. A&B's **easy and comprehensive introduction** [AB22]

   (Many of the references in this talk are introduced here)
3. NeurIPS 2022 talk by Candès on **distribution shift**, NEXCP and related papers
4. **Awesome Conformal Prediction** on github for ALL the pointers (too many)
5. Some of Vovk's work, e.g. conditional guarantees (and lack thereof) [Vov12]
6. Look for papers by Angelopoulos, Bates, Candès, Jordan, Lei, Tibshirani, Wasserman,...
7. aai TRANSFERLAB?

References

[AB22] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *ArXiv:2107.07511 [cs, math, stat]*, dec 2022.

[ABF+22] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. aug 2022.

[BAL+21] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *ArXiv:2101.02703 [cs, stat]*, aug 2021.

[BCRT22] Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. sep 2022.

[GC21] Isaac Gibbs and Emmanuel Candes. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021.

[GC22] Isaac Gibbs and Emmanuel Candès. Conformal Inference for Online Prediction with Arbitrary Distribution Shifts. oct 2022.

[RPC19] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SDCD22] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Optimal Conformal Classifiers. In *International Conference on Learning Representations*. May 2022.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, New York, mar 2005.

[Vov12] Vladimir Vovk. Conditional Validity of Inductive Conformal Predictors. In *Proceedings of the Asian Conference on Machine Learning*, pages 475–490. PMLR, nov 2012.