Federated Conformal prediction

Eric Moulines[1], Vincent Plassier[2],
[1] CMAP, Ecole polytechnique, Paris,
[2] Lagrange research center for mathematics and calculus

Elisabeth's 62 birthday

Outline

- 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.

Theory and practice of machine learning

- Machine learning algorithms have demonstrated substantial effectiveness in practical applications.
- A comprehensive theoretical foundation for machine learning is in place, primarily based on principles such as VC Dimension or Rademacher Complexity.
- However, there exists a significant gap between theory and practice.
  - While theoretical results provide performance guarantees, these often lose their potency when applied to practical, finite datasets, resulting in their limited use in real-world scenarios.

Assessing predictive uncertainty

In statistics, it is common practice to

- compute confidence intervals for parameters.
  - ◇ However, in machine learning, this is not straightforward since we typically do not have meaningful parameters to estimate..
  - ◇ Machine Learning is increasingly model-free and non-parametric !
- calculate prediction sets for future observations
  - ◇ Feasible assuming the independence of the training and test set and comes with validity guarantees;
- perform tests of hypothesis. Interestingly, even under sensible assumptions, testing proves feasible for complex observations.

Introduction to Conformal Prediction

- Conformal prediction extends rank tests, common in nonparametric statistics, to test the IID assumption.
- The linkage of testing with estimation, specifically with confidence intervals, traces back to the work of J. Neyman in 1934.
- In essence, prediction can be seen as the estimation of future data points.

Conformal predictors

- In the basic setting, successive values $z_1, z_2, \cdots \in \mathcal{X} \times \mathcal{Y}$, $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are observed.
- Before the $(n+1)$ th value z_{n+1} is observed, the training set consists of $(z_1, \ldots, z_n)$ and our goal is to predict the new response y_{n+1} given the features x_{n+1}.
- Objective (informal): Prediction algorithm that outputs a set of elements of $\mathcal{Y}$, implicitly meant to contain y_{n+1}.
- Objective (formal): A prediction set is a (measurable) function γ_n that maps a sequence $(z_1, \ldots, z_n) \in (\mathcal{X} \times \mathcal{Y})^n$ to a set $\gamma_n(z_1, \ldots, z_n) \subseteq \mathcal{Y}$.
- A trade-off between reliability and informativeness has to be faced by the algorithm while giving as output the prediction sets.
  - Giving as a prediction set $\mathcal{Y}$ is not useful !

Family of confidence predictors

- we often to deal with nested families of set predictors depending on a parameter $\alpha \in [0, 1]$, the significance level or miscoverage level, reflecting the required reliability of the prediction. The smaller α is, the bigger the reliability in our guess.
- The quantity $1 - \alpha$ is usually called the confidence level.
- As a consequence, we define a confidence predictor to be a nested family of set predictors (γ_n^α), such that, given α_1, α_2 and $0 \leq \alpha_1 \leq \alpha_1 \leq 1$

$$\gamma_n^{\alpha_1}(z_1, \ldots, z_n) \supseteq \gamma_n^{\alpha_2}(z_1, \ldots, z_n)$$

An illustration

Figure 1: Prediction set examples on Imagenet. *We show three progressively more difficult examples of the class* `fox squirrel` *and the prediction sets (i.e., $\mathcal{C}(X_{\text{test}})$) generated by conformal prediction.*

Figure: From Angelopoulos and Bates (2022)

Non-conformity measures

- Objective: estimates how unusual an example looks with respect to the previous ones. The order in which old examples $(z_1, \ldots, z_n)$ appear should not make any difference.
- To underline this point, we will use the term bag (in short, B) and the notation $z_{1:n} = [z_1, \ldots, z_n]$.
  - A bag is a multiset; $z_{1:n}$ is the bag we get from $(z_1, \ldots, z_n)$ when we ignore the order.
- A nonconformity measure $V(B, z) : \mathbf{Z}^n \times \mathbf{Z} \to \mathbb{R}$ is a way of scoring how different an example z is from a bag B.
  - There is not just one nonconformity measure !
- Informally !
  - A low value of $V(B, z)$ indicates that the point z conforms to bag B,
  - A high value indicates that z is atypical relative to B.

Nonconformity for regression

- Assume that $\mathcal{Y} = \mathbb{R}$ and $x \in \mathcal{X} = \mathbb{R}^d$.
- Let $\widehat{f} : \mathcal{X} \mapsto \mathcal{Y}$ a regression function fitted by running an algorithm $\mathcal{A}$ on z and $B = z_{1:n}$.
- A possible choice of nonconformity score:

$$V(B, (x, y)) = |y - \widehat{f}(x)|.$$

- But of course many variants can be considered !

Nonconformity for classification

- Assume that $y \in \mathcal{Y} = \{1, \ldots, K\}$ and denote $\mathcal{S}_{\mathcal{Y}}$ the probability simplex.
- Let $\widehat{f} : \mathcal{X} \mapsto \mathcal{S}_{\mathcal{Y}}$ be a soft classifier fitted by running an algorithm an algorithm $\mathcal{A}$ on z and $B = z_{1:n}$:

$$\widehat{f}(B, x) = (\widehat{f}_1(B, x), \ldots, \widehat{f}_K(B, x)) .$$

- A simple choice

$$V(B, (x, y)) = 1 - \widehat{f}_y(B, x)$$

- A more sophisticated choice (used to define Adaptive Prediction Set)

$$V(B, (x, y)) = \sum_{y' \in \mathcal{Y}} \widehat{f}_{y'}(B, x) \, \mathbb{1}_{\{\widehat{f}_{y'}(B,x) > \widehat{f}_y(B,x)\}}.$$

Full conformal prediction construction

At each $\mathbf{x} \in \mathcal{X}$, define the conformal prediction interval $\gamma_n^\alpha(Z_{1:n}, \mathbf{x})$ by repeating the following procedure

For each $y \in \mathcal{Y}$.

1. Calculate the nonconformity scores

$$V_i^{(\mathbf{x},y)} = V\left(Z_{-i} \cup \{(\mathbf{x}, y)\}, Z_i\right) \quad \text{and} \quad V_{n+1}^{(\mathbf{x},y)} = V\left(Z_{1:n}, (\mathbf{x}, y)\right)$$

2. Include y in the prediction interval $\gamma_n^\alpha(Z_{1:n}, \mathbf{x})$ if

$$V_{n+1}^{(\mathbf{x},y)} \leq Q_{1-\alpha}(\bar{\mu}_n^{(\mathbf{x},y)})$$
$$\bar{\mu}_n^{\mathbf{x},y} = (n+1)^{-1}\left\{\sum_{i=1}^n \delta_{V_i^{(\mathbf{x},y)}} + \delta_\infty\right\}.$$

where $Q_\beta(\nu)$ the level β quantile of a distribution ν on the real line

Quantile Lemma

Lemma

Let $V_1, \ldots, V_{n+1}$ be exchangeable random variables, and denote

$$\mu_n = (n+1)^{-1} \left\{ \sum_{i=1}^{n} \delta_{V^i} + \delta_\infty \right\} .$$

Then for any $\beta \in (0,1)$, we have

$$\mathbb{P}\left(V_{n+1} \leq Q_\beta(\mu_n)\right) \geq \beta$$

Furthermore, if ties between $V_1, \ldots, V_{n+1}$ occur with probability zero, then the above probability is upper bound by $\beta + 1/(n+1)$.

In words, consider n exchangeable observations of a scalar random variable, let's say $V_1, \ldots, V_n$. The rank of another observation V_{n+1} among $V_1, \ldots, V_{n+1}$ is uniformly distributed over the set $\{1, \ldots, n+1\}$, due to exchangeability.

Validity and efficiency

A set predictor γ_n^α is conservatively valid at a significance level $\alpha \in [0,1]$, if the probability of making an error namely the event $Y_{n+1} \notin \gamma_n^\alpha(Z_{1:n}, X_{n+1})-$ does not exceed α.

Theorem (After (Vovk et al., 2005))

Assume that $(X_i, Y_i) \in \mathbb{R}^d \times \mathcal{Y}, i = 1, \ldots, n+1$ are exchangeable. For any nonconformity score function, and any $\alpha \in (0,1)$, define the conformal prediction (based on the first n samples) at $\mathbf{x} \in \mathcal{X}$ by

$$\gamma_n^\alpha(Z_{1:n}, \mathbf{x}) = \left\{ y \in \mathcal{Y} : V_{n+1}^{(\mathbf{x},y)} \leq Q_{1-\alpha}(\bar{\mu}^{(\mathbf{x},y)}) \right\}$$

$$\bar{\mu}_n^{\mathbf{x},y} = (n+1)^{-1} \left\{ \sum_{i=1}^{n} \delta_{V_i^{(\mathbf{x},y)}} + \delta_\infty \right\}.$$

Then,

$$\mathbb{P}\left(Y_{n+1} \in \gamma_n^\alpha\left(Z_{1:n}, X_{n+1}\right)\right) \geq 1 - \alpha$$

Furthermore, if ties between $V_1^{(X_{n+1}, Y_{n+1})}, \ldots, V_{n+1}^{(X_{n+1}, Y_{n+1})}$ occur with probability zero, then this probability is upper bounded by $1 - \alpha + 1/(n+1)$.

Take-home message

The conformal prediction framework allows the construction of prediction sets with finite sample validity without assumptions on the generative model beyond exchangeability.

Formally, for $Z_i = \{Y_i, X_i\}_{i=1:n}, Z_i \overset{\text{iid}}{\sim} \mathbb{P}$ and miscoverage level α, conformal inference allows us to construct a confidence set $C_\alpha(X_{n+1})$ from $Z_{1:n}$ and X_{n+1} such that

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

Split conformal prediction

- Full conformal prediction is computationally intractable...
- In the split conformal prediction, the data sequence is split into two parts:
  - the training set $(Z_1^0, \ldots, Z_m^0)$ used for fitting the regression function $\widetilde{f}$
  - the calibration set $(Z_1, \ldots, Z_n)$ independent of the training set $(Z_1^0, \ldots, Z_m^0)$.
- We use the training set to feed the underlying algorithm, and, using the derived decision rule, we compute the non-conformity scores for each example in the calibration set.
- For every potential label y of the new unlabelled object X_{n+1}, its score $V_{n+1}(X_{n+1}, \cdot)$ is calculated and is compared to the ones of the calibration set.

Split conformal prediction construction

At each $\mathbf{x} \in \mathcal{X}$, define the conformal prediction interval $\tilde{\gamma}_n^\alpha(Z_{1:n}, \mathbf{x})$ by repeating the following procedure

For each $y \in \mathcal{Y}$.

1. Calculate the nonconformity scores, $i = 1, \ldots, n$,

$$V_i = V\left(Z_{1:m}^0, Z_i\right) \quad \text{and} \quad V_{n+1}^{(\mathbf{x},y)} = V\left(Z_{1:m}^0, (\mathbf{x}, y)\right)$$

2. Key point: Contrary to full conformal prediction, V_i, $i = 1, \ldots, n$ does not depend upon on $(\mathbf{x}, y)$ and need to be computed once for all on the calibration set.
3. Include y in the prediction interval $\gamma_n^\alpha(Z_{1:n}, \mathbf{x})$ if

$$V_{n+1}^{(\mathbf{x},y)} \leq Q_{1-\alpha}(\tilde{\mu}_n)$$

where

$$\tilde{\mu}_n = (n+1)^{-1} \left\{ \sum_{i=1}^{n} \delta_{V_i} + \delta_\infty \right\} .$$

All is working as before... the validity is now conditional to the training set - but is also valid unconditionally ! With this reduced computation cost, it is possible to combine easily conformal algorithms with computationally demanding estimators.

Split conformal prediction illustration

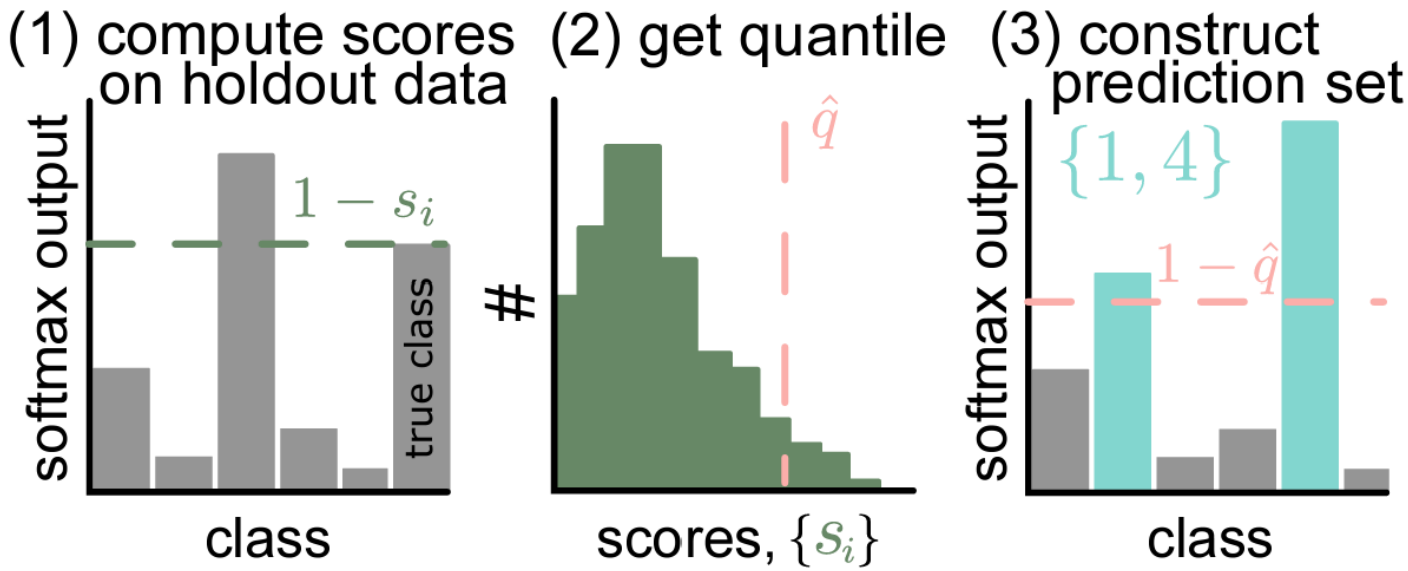


Figure: From (Angelopoulos and Bates, 2022)

Too good to be true (after Angelopoulos and Bates, 2022)

How is it possible to construct a statistically valid prediction set even if the heuristic notion of uncertainty of the underlying model is arbitrarily bad?

- if the scores V_i correctly rank the inputs from lowest to highest magnitude of model error, then the resulting sets will be smaller for easy inputs and bigger for hard ones.
- If the scores are inappropriate, in the sense that they do not approximate this ranking, then the sets will be useless.
  - For example, if the scores are random noise, then the sets will contain a random sample of the label space, where that random sample is large enough to provide valid marginal coverage.

Too good to be true (after Angelopoulos and Bates, 2022)

How is it possible to construct a statistically valid prediction set even if the heuristic notion of uncertainty of the underlying model is arbitrarily bad?

- if the scores V_i correctly rank the inputs from lowest to highest magnitude of model error, then the resulting sets will be smaller for easy inputs and bigger for hard ones.
- If the scores are inappropriate, in the sense that they do not approximate this ranking, then the sets will be useless.
  - For example, if the scores are random noise, then the sets will contain a random sample of the label space, where that random sample is large enough to provide valid marginal coverage.

Although the guarantee always holds, the usefulness of the prediction sets is primarily determined by the score function. This should be no surprise- the score function incorporates almost all

Weighted exchangeability

Definition

Random variables $V_1, \ldots, V_n$ are said to be weighted exchangeable, with weight functions $w_1, \ldots, w_n$, if the density f over a reference measure μ_n of their joint distribution can be factorized as

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} w_i(v_i) \cdot g(v_1, \ldots, v_n),$$

where g is any function that does not depend on the ordering of its inputs, i.e., $g\left(v_{\sigma(1)}, \ldots, v_{\sigma(n)}\right) = g(v_1, \ldots, v_n)$ for any permutation σ of $1, \ldots, n$.

Clearly, weighted exchangeability with weight functions $w_i \equiv 1$ for $i = 1, \ldots, n$ reduces to ordinary exchangeability.

Independent case

Lemma

Assume that $Z_i \sim \mathbb{P}_i$, $i = 1, \ldots, n$ are independent, where each $\mathbb{P}_i$ is absolutely continuous with respect to $\mathbb{P}_{\text{ref}}$, for $i \geq 2$. Then $Z_1, \ldots, Z_n$ are weighted exchangeable, with weight functions $w_i = \mathrm{d}\mathbb{P}_i / \mathrm{d}\mathbb{P}_{\text{ref}}$, $i \in \{1, \ldots, n\}$.

Of course, weighted exchangeability encompasses more than independent sampling, and allows for a nontrivial dependency structure between the variables, just as exchangeability is broader than the i.i.d. case.

Weighted quantile lemma

Lemma (After Tibshirani et al. (2020))

Let $Z_i, i = 1, \ldots, n+1$ be weighted exchangeable random variables, with weight functions $w_1, \ldots, w_{n+1}$. Let $V_i = \mathcal{S}(Z_i, Z_{-i})$, where $Z_{-i} = Z_{1:(n+1)} \setminus \{Z_i\}$, for $i = 1, \ldots, n+1$, and $\mathcal{S}$ is an arbitrary score function. Define

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j\left(z_{\sigma(j)}\right)}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j\left(z_{\sigma(j)}\right)}, i = 1, \ldots, n+1,$$

where the summations are taken over permutations σ of the numbers $1, \ldots, n+1$. Then for any $\beta \in (0,1)$,

$$\mathbb{P}\left(V_{n+1} \leq Q_\beta\left(\sum_{i=1}^{n} p_i^w(Z_1, \ldots, Z_{n+1})\,\delta_{V_i} + p_{n+1}^w(Z_1, \ldots, Z_{n+1})\delta_\infty\right)\right) \geq \beta$$

Weighted conformal prediction

Theorem

Assume that $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, n+1$ *are weighted exchangeable with weight functions* $w_1, \ldots, w_{n+1}$. *For any score function* $\mathcal{S}$, *and any* $\alpha \in (0,1)$, *define the weighted conformal band (based on the first n samples) at a point* $\mathbf{x} \in \mathcal{X}$ *by*

$$\gamma_n^\alpha(Z_{1:n}, \mathbf{x}) = \Big\{ y \in \mathbb{R} : V_{n+1}^{(\mathbf{x},y)} \leq Q_{1-\alpha}\Big(\sum_{i=1}^n p_i^w(Z_{1:n}, (\mathbf{x}, y)) \delta_{V_i^{(\mathbf{x},y)}} + p_{n+1}^w(Z_{1:n}, (\mathbf{x}, y)) \delta_\infty \Big) \Big\}.$$

Then

$$\mathbb{P}\left(Y_{n+1} \in \gamma_n^\alpha(Z_{1:n}, X_{n+1})\right) \geq 1 - \alpha\,.$$

The independent case

- The weights

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j\left(z_{\sigma(j)}\right)}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j\left(z_{\sigma(j)}\right)}, i = 1, \ldots, n+1,$$

  are in general intractable, unless much stronger structure allows major simplification
- This is in particular the case when
  1. the calibration data are i.i.d.
  2. the distribution of the query point (x, y) differs from the calibration set.
- Used in Tibshirani et al (2020) to address covariate shift [but it can also be used for label shift.

Independent case with a distribution shift

- Denote by $w(x, y) = \mathrm{d}\mathbb{P}_{\text{query}}/\mathrm{d}\mathbb{P}_{\text{cal}}(x, y)$ the pdf of the query w.r.t. the calibration distribution.
- Define the weights

$$p_i^w(Z_{1:n}, (x, y)) = \frac{w(Z_i)}{\sum_{j=1}^n w(Z_j) + w(x, y)}, \quad i = 1, \ldots, n,$$

$$p_{n+1}^w(Z_{1:n}, (x, y)) = \frac{w(x, y)}{\sum_{j=1}^n w(Z_j) + w(x, y)}$$

- Define the weighted empirical measure

$$\mu_n^{(x,y)} = \sum_{i=1}^n p_i^w(Z_{1:n}, (x, y))\delta_{V_i^{(x,y)}} + p_{n+1}^w(Z_{1:n}, (x, y))\delta_\infty \,.$$

- Special cases:
  - (Tibshirani et al., 2020) Covariate shift: $w(x, y) = w(x)$
  - (Podkopaev et al, 2021) Label shift: $w(x, y) = w(y)$.

Independent case with a distribution shift

- For any $\mathbf{x} \in \mathcal{X}$, and $\alpha \in (0,1)$ define the prediction set

$$\gamma_n^{\alpha}(Z_{1:n}, \mathbf{x}) = \left\{ y \in \mathcal{Y}, V_{n+1}^{(\mathbf{x},y)} \leq Q_{1-\alpha}(\mu_n^{(\mathbf{x},y)}) \right\}, .$$

- The weighted conformal prediction Theorem, shows that

$$\mathbb{P}\left(Y_{n+1} \in \gamma_n^{\alpha}(Z_{1:n}, X_{n+1})\right) \geq 1 - \alpha \, .$$

- In the special case of split conformal prediction, then

$$\mu_n^{(x,y)} = \sum_{i=1}^{n} p_i^w(Z_{1:n}, (x,y))\delta_{V_i} + p_{n+1}^w(Z_{1:n}, (x,y))\delta_{\infty}$$

the nonconformity scores can be computed once for all, it is only required to update the weights.

An illustration

- Consider the following toy classification task with 3 classes $\mathcal{Y} = \{1, 2, 3\}$ where class proportions are given as $p = (0.1, 0.6, 0.3)$ and $q = (0.3, 0.2, 0.5)$, and for each data point the covariates are sampled according to $X \mid Y = y \sim \mathcal{N}(m_y, \Sigma)$ where $m_1 = (-2; 0)^\top$, $m_2 = (2; 0)^\top$, $m_3 = (0; 2\sqrt{3})^\top$, $\Sigma = \text{diag}(4, 4)$.
- Perform split-conformal prediction sets for a single draw of data from the source and target distributions using the Bayes-optimal rule (not even learned) as an underlying predictor.

An illustration

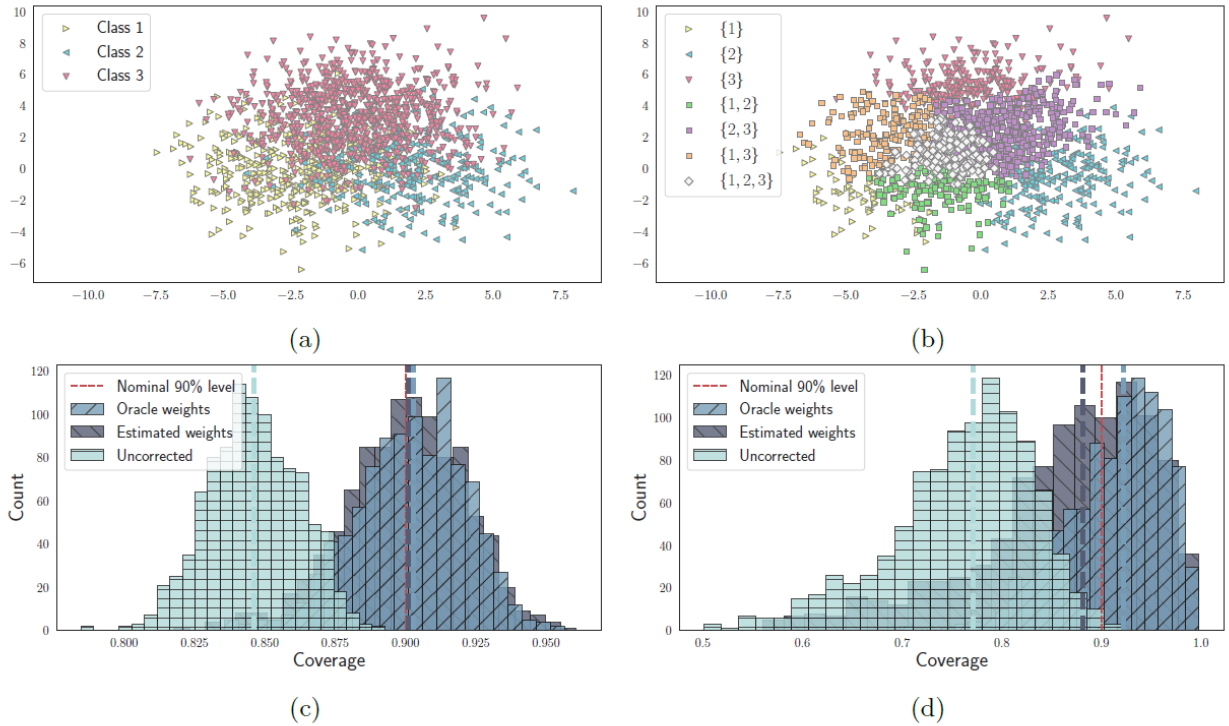


Figure 1: (a) Test data sample for the toy simulation in Section 2.2. (b) Corresponding conformal prediction sets when label shift is accounted for with oracle importance weights. (c) Empirical coverage on shifted data for the toy simulation in Section 2.2. (d): Empirical coverage on the `wine quality` dataset. Dashed vertical lines describe the median coverage values, which are significantly worse when label shift is not accounted for, while using estimated weights mimics the oracle reasonably well.

Figure: After (Podkopaev, Ramdas, 2021)

Bayesian predictions

Bayesian statistics provides optimal predictions for future observations provided the model is well-specified. Informally,

- The conditional distribution of the observations (likelihood) is known, $p(Z_1, \ldots, Z_n|\theta)$
- The prior distribution of the parameter $\pi(\theta)$ is known
- In such case, assuming $Z_{n+1} \perp\!\!\!\perp (Z_1, \ldots, Z_n) \mid \theta$, the posterior predictive distribution for the response at a new $X_{n+1} = x_{n+1}$ takes on the form

$$p\left(y \mid x_{n+1}, Z_{1:n}\right) = \int f_\theta\left(y \mid x_{n+1}\right) p\left(\theta \mid Z_{1:n}\right) d\theta$$

- Asymptotically exact samples from the posterior can be obtained through Markov chain Monte Carlo (MCMC) and the above density can be computed through Monte Carlo (MC), or by direct sampling from an approximate model.

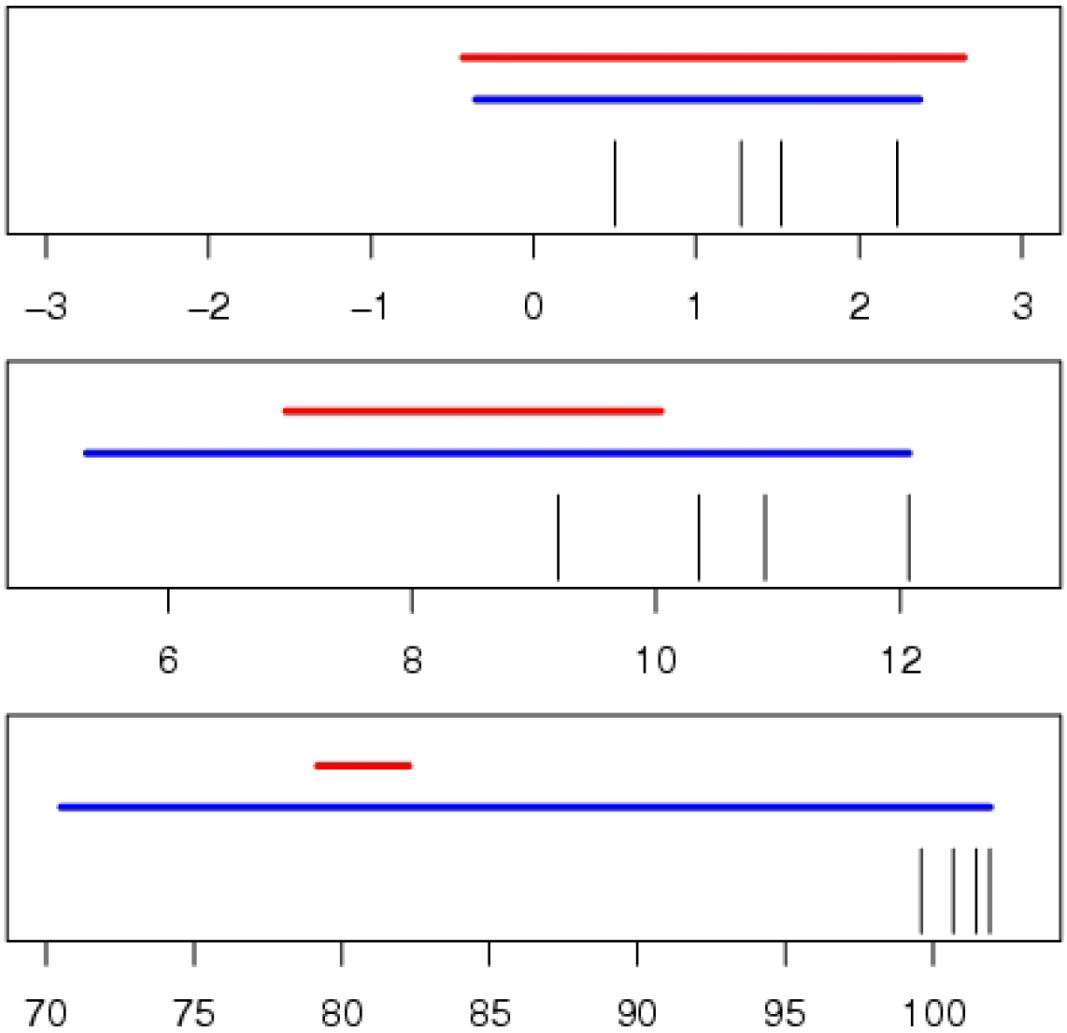
Bayesian predictive distributions

- Given a Bayesian predictive distribution, one can then construct the highest density $100 \times (1-\alpha)\%$ posterior predictive credible intervals, which are the shortest intervals to contain $(1-\alpha)$ of the predictive probability.
- Alternatively, the central $100 \times (1-\alpha)\%$ credible interval can be computed using the $\alpha/2$ and $1-\alpha/2$ quantiles. Posterior predictive distributions condition on the observed $Z_{1:n}$ and represent subjective and coherent beliefs.
- However, it is well known that model misspecification can lead Bayesian intervals to be poorly calibrated in the frequentist sense (Dawid, 1982; Fraser et al., 2011):
  - the long run proportion of the observed data lying in the $(1-\alpha)$ Bayes predictive interval is not necessarily equal to $(1-\alpha)$.
  - This has consequences for the robustness of such approaches and trust in using Bayesian models to aid decisions.
- 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 be misleading



$N(1, 1)$

$N(10, 1)$

$N(100, 1)$

Figure: From (Vovk,2021) lecture

Bayes prediction intervals can mislead

- 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 $\theta = 100$: 3.08vs 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.

Federated learning

- **Federated learning** is an increasingly important framework for large-scale learning.
  - FL allows many agents to train a model together under the coordination of a central server without ever transmitting the agents' data over the network, in an attempt to preserve privacy. There has been a considerable amount of FL work over the past 5 years.
- Compared to classical machine learning techniques, FL has two unique features.
  - the networked agents are massively distributed, communication bandwidth is limited, and agents are not always available (*system heterogeneity*).
  - the data distribution at different agents can vary greatly (*statistical heterogeneity*)

Federated inference procedures that allow to build prediction sets for each agent with a confidence level that can be guaranteed.

Setup

- Consider a federated learning system with n agents.
  - Each agent $i \in [n]$ owns a local calibration set $\mathcal{D}_i = \{(X_k^i, Y_k^i)_{k=1}^{N^i}\}$, where N^i is the number of calibration samples for the agent i.
  - The calibration data are i.i.d. and that the statistical heterogeneity is due to *label shifts*:

$$(X_k^i, Y_k^i) \sim P^i = P_{X|Y} \times P_Y^i,$$

    where $P_{X|Y}$, the conditional distribution of the feature given the label, is assumed identical among agents but P_Y^i, the prior label distribution, may differ across agents.
- A predictive model $\hat{f}$ has been learned by federated learning. The results are agnostic to the learning procedure.

Objective

- For an agent $\star \in [n]$, and each $\alpha \in (0, 1)$, we are willing to compute a set-valued predictor, $\mathcal{C}_\alpha$ with confidence level $1 - \alpha$, which depends on the calibration data of all the agents.
  - The goal is to construct informative conformal prediction sets for each agent, even when its calibration set is limited in size, by using the calibration data of all the agents participating in the FL.
  - The calibration data must always remain local to the networked agents.
- Objective: Attain both conformal and theoretical privacy guarantees – matched to the privacy guarantees that can be obtained in the FL training procedure.

Naive split conformal prediction

- Consider the calibration dataset $\{(X_k^i, Y_k^i)\colon k \in [N^i]\}_{i\in[n]}$ with data distributed according to $\{P^i\}_{i\in[n]}$.
- For $\{\pi_i\}_{i\in[n]} \in \Delta_n$ we define the mixture distribution of labels given for $y \in \mathcal{Y}$ by
$$P_Y^{\mathrm{cal}}(y) = \sum_{i=1}^n \pi_i P_Y^i(y).$$
- Our goal is to determine a set of likely outputs for a new data point $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star)$ drawn on agent $\star \in [n]$ from the distribution $P^\star$.
- The conformal approach relies on non-conformity scores $V_k^i = V(X_k^i, Y_k^i)$, $i \in [n]$, $k \in [N^i]$ to determine the prediction set.
- These non-conformity scores are uniformly weighted to generate the conventional prediction set
$$\mathcal{C}_{\alpha,\bar{\mu}}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \le Q_{1-\alpha}(\bar{\mu})\},$$
$$\bar{\mu} = (N+1)^{-1}\Big(\sum_{i=1}^n \sum_{k=1}^{N^i} \delta_{V_k^i} + \delta_1\Big).$$
- Naive approach leads to significant under-coverage in the presence of label shift.

A first solution

- Assumptions: for all $i \in [n]$ and $y \in \mathcal{Y}$, we have access to the likelihood ratios:
$$w_y^i = P_Y^i(y)/P_Y^{\text{cal}}(y).$$
- Define
$$p^\star_{y,\mathbf{y}} = \frac{W_{y,\mathbf{y}}}{W_{\mathbf{y},\mathbf{y}} + \sum_{\tilde{y}\in\mathcal{Y}} N_{\tilde{y}} W_{\tilde{y},\mathbf{y}}},$$
$$\mu^\star_{\mathbf{y}} = p^\star_{\mathbf{y},\mathbf{y}}\delta_1 + \sum_{i=1}^{n}\sum_{k=1}^{N^i} p^\star_{Y_k^i,\mathbf{y}}\delta_{V_k^i}.$$
where the weights $\{W_{y,\mathbf{y}}\}_{(y,\mathbf{y})\in\mathcal{Y}^2}$ - derived from (Tibshirani et al, 2020) are provided in (Plassier et al., 2022) too complex to be displayed and computed!
- For any covariate $\mathbf{x} \in \mathcal{X}$, define the $(1-\alpha)$-prediction set with oracle weights
$$\mathcal{C}_{\alpha,\mu^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}\left(\mu^\star_{\mathbf{y}}\right)\right\}.$$
- In contrast to the exchangeable setting, the quantile is calculated based on a weighted empirical distribution depending on $\mathbf{y}$.

Conformal guarantees

Theorem

For any $\alpha \in [0, 1)$*, we have*

$$1 - \alpha \leq \mathbb{P}\left(Y^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\mu^{\star}}(X^{\star}_{N^{\star}+1})\right) \leq 1 - \alpha + \mathbb{E}\left[\max_{(i,k)\in\mathcal{I}} \{p^{\star}_{Y^{i}_{k}, Y^{\star}_{N^{\star}+1}}\}\right],$$

- It is important to note that the lower bound holds even in the presence of ties between non-conformity scores.
- The prediction set requires the challenging computation of the weights $p^{\star}_{y,\mathbf{y}}$. Indeed, the calculation of $W_{y,\mathbf{y}}$ requires the summation over $N!$ elements.

Practical split conformal prediction

- Let $\bar{N} \leq N$, be randomly sampled accoring to a multinomial distribution with parameter $(\bar{N}, \{\pi_i\}_{i\in[n]})$.
  - ◇ We denote by $\bar{N}^i$ the multinomial count associated with agent i.
  - ◇ We sample $\bar{N}^i \wedge N^i$ calibration data from agent i and denote $V_k^i = V(X_k^i, Y_k^i)$.

  For any label $y \in \mathcal{Y}$, the weight $\bar{p}^\star_{y,\mathbf{y}}$ is given by:

$$\bar{p}^\star_{y,\mathbf{y}} = \frac{w^\star_y}{w^\star_{\mathbf{y}} + \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}^i} w^\star_{Y_k^i}}.$$

- We consider the prediction set

$$\bar{\mu}^\star_{\mathbf{y}} = \bar{p}^\star_{\mathbf{y},\mathbf{y}}\delta_1 + \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}^i} \bar{p}^\star_{Y_k^i,\mathbf{y}}\delta_{V_k^i},$$
$$\mathcal{C}_{\alpha,\bar{\mu}^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu}^\star_{\mathbf{y}})\right\}.$$

Theorem

Set $\bar{N} = \lfloor N/2 \rfloor$ *and* $\pi_i = N^i/N$, *for any* $i \in [n]$. *Then,*

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{6}{N} + \frac{36 + 6\log N}{N}\|w^\star\|_\infty^2 + \frac{14\log N}{N}\sum_{i:\ \frac{N^i}{12} < \log N} \sqrt{N^i}.$$

- If $n = 1$ and $N \geq 46$, the set $\{i \in [n] : N^i < 12 \log N\}$ is empty. In this case, the convergence rate reduces to $N^{-1} \log N$.
- If each agent has the same number of calibration data, the convergence rate $N^{-1} \log N$ is ensured when $N \geq 12n \log N$.