Ch.7 Weighted Variants of Conformal Prediction

Conformal Prediction

Jikwang Kim

Seoul National University

November 3, 2025

Outline

1. Weighted quantiles and the weighted conformal algorithm

2. Conformal prediction under covariate and label shifts

3. Localized conformal prediction

4. Fixed-weight conformal prediction

5. A general outlook through weighted permutations

Outline

1. Weighted quantiles and the weighted conformal algorithm
2. Conformal prediction under covariate and label shifts
3. Localized conformal prediction
4. Fixed-weight conformal prediction
5. A general outlook through weighted permutations

Weighted Conformal Quantile

- (Unweighted) Full CP

$$\begin{aligned} \mathcal{C}(X_{n+1}) &= \{y : S^y_{n+1} \leq \hat{q}^y\}, \\ \text{where } \hat{q}^y =& \text{Quantile}(S^y_1, \ldots, S^y_n, S^y_{n+1}; 1-\alpha) \\ =& \text{Quantile}\left(\frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{S^y_i}; 1-\alpha\right) \end{aligned}$$

- Weighted Full CP

$$\begin{aligned} \mathcal{C}(X_{n+1}) &= \{y : S^y_{n+1} \leq \hat{q}^y_w\}, \\ \text{where } \hat{q}^y_w &= \text{Quantile}\left(\sum_{i=1}^{n+1} w_i \delta_{S^y_i}; 1-\alpha\right) \end{aligned}$$

- The weights can be fixed a-priori, or represent the ftn. of the data.

Weighted Conformal Algorithm

Algorithm 7.1: Weighted full CP

1. Input: Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test pt. X_{n+1}, target coverage level $1-\alpha$, conformal score ftn. s, and weight $w_1, \ldots w_{n+1}$
2. For each possible response value $y \in \mathcal{Y}$, 다음을 계산
    1. 각 sample 별 score $S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y), \quad i = 1, \ldots, n.$
    2. Test point score $S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y)$
    3. Quantile $\hat{q}_w^y = \text{Quantile}(\sum_{i=1}^{n+1} w_i \delta_{S_i^y}; 1-\alpha)$
3. Return the prediction set $\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^y \le \hat{q}_w^y\}$

WARNING! The test point has not been replaced yet.

Weighted Conformal Algorithm

Algorithm 7.2: Weighted split CP

1. Input: Pretraining data $\mathcal{D}_{\text{pre}}$, calibration data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test pt. X_{n+1}, target coverage level $1 - \alpha$, and weight $w_1, \ldots w_{n+1}$
2. Score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 학습, using $\mathcal{D}_{\text{pre}}$.
3. $S_i = s(X_i, Y_i)$ for $i \in [n]$ 계산, and $\hat{q}_w = \text{Quantile}(\sum_{i=1}^{n} w_i \delta_{S_i^y} + w_{n+1}\delta_{+\infty}; 1 - \alpha)$ 계산
4. Return the prediction set

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}\}$$

WARNING! The unreplaced test points are estimated conservatively.

Outline

1 Weighted quantiles and the weighted conformal algorithm

2 Conformal prediction under covariate and label shifts

3 Localized conformal prediction

4 Fixed-weight conformal prediction

5 A general outlook through weighted permutations

Setup

- Train data $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$
- Test point $(X_{n+1}, Y_{n+1}) \sim \hat{P}_{XY}$
  - Covariate-shifted : $(X_{n+1}, Y_{n+1}) \sim Q_X \times P_{Y|X}$

Ex) 소득에 따른 보유 차량 종류
Training - 주로 서울 (P_X) / Test - 주로 부산 (Q_X)
But 같은 소득에서는 같은 규칙 $(P_{Y|X})$

  - Label-shifted : $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times Q_Y$

Ex) (same)
Training - 주로 소형차 (P_Y) / Test - 주로 준중형차 (Q_Y)
But 각 클래스 내에서 동일한 규칙 $(P_{X|Y})$

$\Rightarrow$ $w_i =$ (the likelihood ratio relating these two distributions)

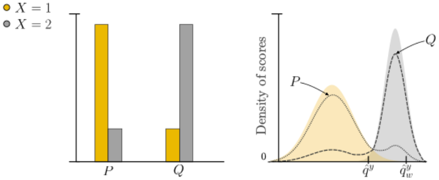
Covariate shift

- Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_X P_{Y|X}, \quad (X_{n+1}, Y_{n+1}) \sim Q_X P_{Y|X}$.
- If we know the LR for test versus train dist.s, we can be construct the weight:

$$w_i \propto \frac{dQ_X}{dP_X}(X_i) \quad \text{where } \frac{dQ_X}{dP_X} \text{ is the Radon-Nykodym derivative.}$$

- Then, we can guarantee the marginal coverage.

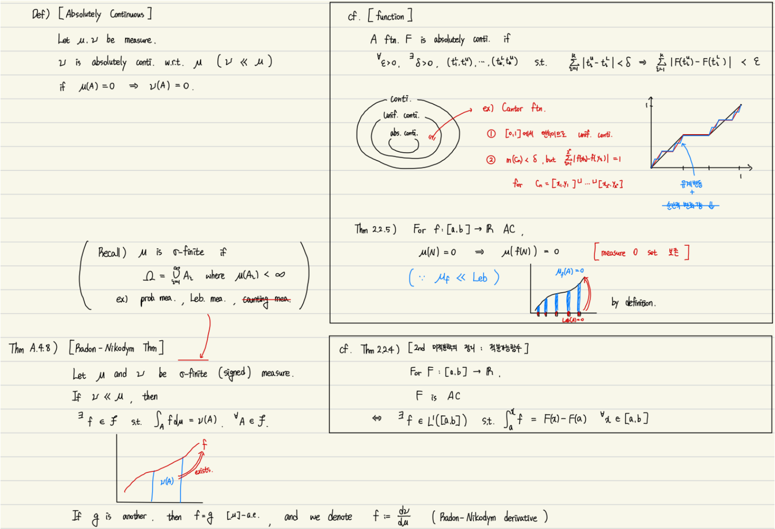
Covariate shift



- For simplicity, let $s(X, Y) = I(X = 1)$.
- In above, $P_{P_X}(X = 1) > P_{Q_X}(X = 1)$, i.e., $X = 1$ is over-represented in the training data.

$\Rightarrow$ The unweighted conformal quantile will be biased downwards.

Cf. Radon-Nikodym Derivative



conti.
unif. conti.
abs. conti.
1
1
유계변동
+
~~순간적 변화가능~~
$\mu_f(A)=0$
$Leb(A)=0$
f
exists.
$\nu(A)$

Def) [Absolutely Continuous]

Let μ, ν be measure.

ν is absolutely conti. w.r.t. μ $(\nu \ll \mu)$

if $\mu(A)=0 \Rightarrow \nu(A)=0$.

cf. [function]

A ftn. F is absolutely conti. if

$\forall \varepsilon>0$, $\exists \delta>0$, $(t_1^l, t_1^u), \cdots, (t_k^l, t_k^u)$ s.t. $\sum_{i=1}^{k} |t_i^u - t_i^l| < \delta \Rightarrow \sum_{i=1}^{k} |F(t_i^u) - F(t_i^l)| < \varepsilon$

ex) Cantor ftn.

① [0,1]에서 연속이므로 unif. conti.

② $m(C_n) < \delta$, but $\sum_{i=1}^{2^n} |f(x_i) - f(y_i)| = 1$

for $C_n = [x_1, y_1] \sqcup \cdots \sqcup [x_{2^n}, y_{2^n}]$

Thm 2.2.5) For $f:[a,b] \to \mathbb{R}$ AC,

$\mu(N)=0 \Rightarrow \mu(f(N))=0$ [measure 0 set 보존]

$(\because \mu_f \ll Leb)$

by definition.

(Recall) μ is σ-finite if

$\Omega = \bigcup_{i=1}^{\infty} A_i$ where $\mu(A_i) < \infty$

ex) prob. mea., Leb. mea., ~~counting mea.~~)

Thm A.4.8) [Radon-Nikodym Thm]

Let μ and ν be σ-finite (signed) measure.

If $\nu \ll \mu$, then

$\exists f \in \mathcal{F}$ s.t. $\int_A f d\mu = \nu(A)$. $\forall A \in \mathcal{F}$.

cf. Thm 2.2.4) [2nd 미적분학의 정리 : 적분가능함수]

For $F:[a,b] \to \mathbb{R}$,

F is AC

$\Leftrightarrow \exists f \in L^1([a,b])$ s.t. $\int_a^x f = F(x) - F(a)$ $\forall x \in [a,b]$

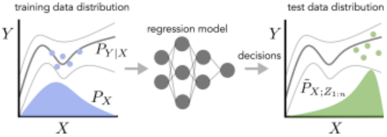
If g is another, then $f=g$ $[\mu]$-a.e., and we denote $f := \frac{d\nu}{d\mu}$ (Radon-Nikodym derivative)

Figure: Extra explanation of RN derivative

Cf. Radon-Nykodym Derivative

(Most Important Thing) We have to know P, and Q! i.e.,
we should have confidence that the data is sampled in P, and test point is in Q.
Example)

1. Casual Inference
   For $(X_i, T_i, Y_i) =$ (covariate, treatment, outcome), assume the model is fitted by all observed data. Also, for $\tau^{\text{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]$, and $\tau^{\text{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$,
   in ATE, $Q = P$, and in ATT, $Q_X = P_{X|T=1}$.
2. Adaptive Learning
   - X : the protein sequence ($P_X =$ Unif$(X; \mathcal{X})$), and Y : the fitness of protein
   - fit the regression model $\mu_{Z_{1:n}}$, and propose the sequence with the highest predicted fitness ($Q_X = \tilde{P}_{X;Z_{1:n}} = \exp(\lambda \cdot \mu_{Z_{1:n}}(X_{\text{test}}))$)



Covariate shift

Main Issue : Non-exchangeability for train and test.

(Recall) Prop 2.2: Conditioning on the empirical distribution

$$Z_i\text{'s are exchangeable} \quad \Longleftrightarrow \quad Z_i|\hat{P}_n \sim \hat{P}_n$$

- Now, $(X_{n+1}, Y_{n+1})|\hat{P}_{n+1} \not\sim \hat{P}_{n+1}$
- Instead,

$$(X_{n+1}, Y_{n+1})|\hat{P}_{n+1} \sim \sum_{i=1}^{n+1} \frac{dQ_X}{dP_X}(X_i) \cdot \delta_{(X_i,Y_i)}$$

- Hence, $S_{n+1} \mid \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i \delta_{S_i}$

Covariate shift

Thm 7.3: Weighted CP with Covariate shift

For $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_X \times P_{Y|X}$, and $(X_{n+1}, Y_{n+1}) \sim Q_X \times P_{Y|X}$ independently, let $Q_X \ll P_X$. (i.e., $\frac{dQ_X}{dP_X}(x) < \infty \; \forall x \in \mathcal{X}$)
Fix any symmetric score s, and define the prediction set

$$\mathcal{C}(X_{n+1}) = \{y : S_{n+1}^y \leq \hat{q}_w^y\}, \quad \text{where } \hat{q}_w^y = \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{S_i^y}; 1-\alpha\right)$$

where

$$w_i = \frac{\frac{dQ_X}{dP_X}(X_i)}{\sum_{j=1}^{n+1} \frac{dQ_X}{dP_X}(X_j)}$$

Then, it satisfies the marginal coverage guarantee,

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

WARNING! $w_{n+1}\delta_\infty$ can be omited.

Label shift

- Now, let
  $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X|Y} P_Y, \quad (X_{n+1}, Y_{n+1}) \sim P_{X|Y} Q_Y.$
- Similarly,

$$w_i \propto \frac{dQ_Y}{dP_Y}(Y_i) \quad \text{where } \frac{dQ_Y}{dP_Y} \text{ is the Radon-Nykodym derivative.}$$

- However, there are still some problems remaining:
  the test point Y_{n+1} is not fixed in CP.

$\Rightarrow$ The weights depend on the hypothesized test point y.

Label shift

Thm 7.4: Weighted CP with Label shift

For $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X|Y} \times P_Y$, and $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times Q_Y$ independently, let $Q_Y \ll P_Y$. (i.e., $\frac{dQ_Y}{dP_Y}(y) < \infty \; \forall y \in \mathcal{Y}$)
Fix any symmetric score s, and define the prediction set

$$\mathcal{C}(X_{n+1}) = \{y : S_{n+1}^y \leq \hat{q}_w^y\}, \quad \text{where } \hat{q}_w^y = \text{Quantile}\left(\sum_{i=1}^n w_i^y \delta_{S_i^y}; 1-\alpha\right)$$

where, for $i \in [n]$,

$$w_i^y = \frac{\frac{dQ_Y}{dP_Y}(Y_i)}{\sum_{j=1}^n \frac{dQ_Y}{dP_Y}(Y_j) + \frac{dQ_Y}{dP_Y}(y)}, \quad \text{and } w_{n+1}^y = \frac{\frac{dQ_Y}{dP_Y}(y)}{\sum_{j=1}^n \frac{dQ_Y}{dP_Y}(Y_j) + \frac{dQ_Y}{dP_Y}(y)}$$

Then, it satisfies the marginal coverage guarantee,

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

General shift between the training dist. and test dist.

Thm 7.5: Weighted CP with General shift

For $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P$, and $(X_{n+1}, Y_{n+1}) \sim Q$ independently, let $Q \ll P$. (i.e., $\frac{dQ}{dP}(y) < \infty \; \forall y \in \mathcal{Y}$)
Fix any symmetric score s, and define the prediction set

$$\mathcal{C}(X_{n+1}) = \{y : S_{n+1}^y \leq \hat{q}_w^y\}, \quad \text{where } \hat{q}_w^y = \text{Quantile}\left(\sum_{i=1}^n w_i^y \delta_{S_i^y}; 1-\alpha\right)$$

where, for $i \in [n]$,

$$w_i^y = \frac{\frac{dQ}{dP}(X_i, Y_i)}{\sum_{j=1}^n \frac{dQ}{dP}(X_j, Y_j) + \frac{dQ}{dP}(X_{n+1}, y)}, \quad \text{and } w_{n+1}^y = \frac{\frac{dQ}{dP}(X_{n+1}, y)}{\sum_{j=1}^n \frac{dQ}{dP}(X_j, Y_j) + \frac{dQ}{dP}(X_{n+1}, y)} \tag{7.3}$$

Then, it satisfies the marginal coverage guarantee,

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

Marginal coverage guarantee with General shift

Proposition 7.6: Conditioning on the empirical distribution

Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ and $Z_{n+1} \sim Q$ independently, for some dist.s P, Q on $\mathcal{Z}$. Also, $Q \ll P$. (i.e., $\frac{dQ}{dP}(z) < \infty \, \forall z \in \mathcal{Z}$)
Let the empirical dist. of the $n+1$ data points be

$$\hat{P}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{Z_i}$$

Then,

$$Z_{n+1} \mid \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i \delta_{Z_i}$$

where

$$w_i = \frac{\frac{dQ}{dP}(Z_i)}{\sum_{j=1}^{n+1} \frac{dQ}{dP}(Z_j)}, \quad i \in [n+1]$$

Marginal coverage guarantee with General shift

proof of Thm 7.5) Note that

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \le \hat{q}_w^{Y_{n+1}} \iff S_{n+1} \le \text{Quantile}(\sum_{i=1}^{n} w_i^{Y_{n+1}} \delta_{Z_i}; 1-\alpha)$$

Now. we write $Z_i = (X_i, Y_i)$, and

$$w_i = w_i^{Y_{n+1}} = \frac{\frac{dQ}{dP}(Z_i)}{\sum_{j=1}^{n+1} \frac{dQ}{dP}(Z_j)}, \quad i \in [n+1]$$

Then,

$$\begin{aligned} \mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) &= \mathbb{P}\left(S_{n+1} \le \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{S_i};\, 1-\alpha\right)\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(S_{n+1} \le \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{S_i};\, 1-\alpha\right) \middle| \hat{P}_{n+1}\right)\right] \end{aligned}$$

Marginal coverage guarantee with General shift

proof of Thm 7.5) (continue)
Since the score s is symmetric, $s(\cdot, \mathcal{D}_{n+1}) = s(\cdot, \hat{P}_{n+1})$.
Then, for $S_{n+1} = s(Z_{n+1}, \hat{P}_{n+1})$,

$$
\begin{aligned}
&\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \\
&= \mathbb{E}\left[\mathbb{P}\left(s(Z_{n+1}, \hat{P}_{n+1}) \leq \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{s(Z_{n+1}, \hat{P}_{n+1})};\, 1-\alpha\right)\middle|\hat{P}_{n+1}\right)\right] \\
&\overset{?}{=} \mathbb{E}\left[\sum_{j=1}^{n+1} w_j 1\left\{s(Z_j; \widehat{P}_{n+1}) \leq \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{s(Z_i; \widehat{P}_{n+1})}; 1-\alpha\right)\right\}\right]
\end{aligned}
$$

Marginal coverage guarantee with General shift

proof of Thm 7.5) (continue)
Why?

$$\text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{s(Z_i;\hat{P}_{n+1})}; 1-\alpha\right) \in \hat{P}_{n+1} \quad \text{(by prop. 7.6.)}$$

$$\Rightarrow \left\{ s(Z_{n+1}, \hat{P}_{n+1}) \leq \text{Quantile}\left(\sum_{i=1}^{n} w_i \delta_{s(Z_{n+1},\hat{P}_{n+1})};\ 1-\alpha\right)\right\} = f(Z_{n+1})$$

$$\text{for some } f,\ \text{given } \hat{P}_{n+1}$$

$$\Rightarrow f(Z_{n+1}) \mid \hat{P}_{n+1} \overset{d}{=} \sum_{j=1}^{n+1} w_j \delta_{f(Z_j)} \quad \text{(by prop. 7.6., again)}$$

Marginal coverage guarantee with General shift

proof of Thm 7.5) (continue)

(Recall) Fact 2.12: Quantiles and CDFs

For $z \in \mathbb{R}^n$, TFAE:

(i) $\hat{F}_z(v) = \sup\{\tau : \text{Quantile}(z;\tau) \leq v\}, \quad \forall v \in \mathbb{R}$

(iii) $\hat{F}_z(\text{Quantile}(z;\tau)) \geq \tau, \quad \forall \tau \in [0,1]$

Fact 7.7: Quantiles and CDFs

For $z \in \mathbb{R}^n$, and the weights $w_1 \ldots, w_n$,

$$\sum_{i=1}^{n} w_i 1 \left\{ z_i \leq \text{Quantile}\left(\sum_{j=1}^{n} w_j \delta_{z_j}; \tau \right) \right\} \geq \tau \quad \forall \tau \in [0,1]$$

Hence,

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq \mathbb{E}[1-\alpha] = 1-\alpha \quad \square$$

Marginal coverage guarantee with General shift

proof of Prop 7.6) For all meas'l sets A, and B,

$$\begin{aligned}
&\mathbb{P}_{(Z_1,\dots,Z_{n+1})\sim P^n\times Q}(Z_{n+1}\in A, \hat{P}_{n+1}\in B)\\
&=\mathbb{E}_{P^n\times Q}\big[\mathbf{1}\{Z_{n+1}\in A\}\mathbf{1}\{\hat{P}_{n+1}\in B\}\big]\\
&=\mathbb{E}_{P^{n+1}}\left[\frac{dQ}{dP}(Z_{n+1})\cdot\mathbf{1}\{Z_{n+1}\in A\}\mathbf{1}\{\hat{P}_{n+1}\in B\}\right]\\
&=\mathbb{E}_{P^{n+1}}\left[w_{n+1}\sum_{i=1}^{n+1}\frac{dQ}{dP}(Z_i)\cdot\mathbf{1}\{Z_{n+1}\in A\}\mathbf{1}\{\hat{P}_{n+1}\in B\}\right]\\
&=\sum_{i=1}^{n+1}\mathbb{E}_{P^{n+1}}\left[w_i\frac{dQ}{dP}(Z_{n+1})\cdot\mathbf{1}\{Z_i\in A\}\mathbf{1}\{\hat{P}_{n+1}\in B\}\right]\\
&=\sum_{i=1}^{n+1}\mathbb{E}_{P^n\times Q}\left[w_i\mathbf{1}\{Z_i\in A\}\mathbf{1}\{\hat{P}_{n+1}\in B\}\right]
\end{aligned}$$

Hence,

$$Z_{n+1}\mid\hat{P}_{n+1}\sim\sum_{i=1}^{n+1}w_i\delta_{Z_i}\qquad\square$$

Comparing distribution shift and conditional coverage

We have a difficulty about the conditional coverage:

- If it guarantees the test-conditional coverage, i.e.,

  $$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} = x) \geq 1 - \alpha \quad \forall x,$$

  it implies coverage for any covariate shift. (i.e. any distribution of X, Q_X)

- If it guarantees the label-conditional coverage, i.e.,

  $$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid Y_{n+1} = y) \geq 1 - \alpha \quad \forall y,$$

  it implies coverage for any label shift. (i.e. any distribution of Y, Q_Y)

- But it is harder, since $\mathcal{C}(X)$ will be very large unless we have a large number of obs. in each category. (ex. $\sum_{i \in [n]} \mathbf{1}\{Y_i = y\}$ is large, for each $y \in \mathcal{Y}$)
- Hence, we guarantee the marginal coverage relative to a particular distribution, (ex. $P_{X|Y} \times Q_Y$) which is more informative.

Outline

1. Weighted quantiles and the weighted conformal algorithm
2. Conformal prediction under covariate and label shifts
3. Localized conformal prediction
4. Fixed-weight conformal prediction
5. A general outlook through weighted permutations

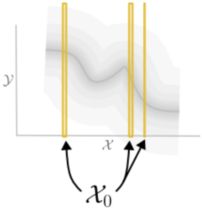
Improving the test-conditional coverage

- (Ch4) It's impossible in general to ensure test-conditional coverage.
- (Ch5) Certain score functions lead to approximate conditional coverage under additional assumptions.

- Consider the target

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

- The prediction set is determined by the conformal score of all training points, including the data points which are far away from the test point X_{n+1}.
- If the distribution of scores is very different across different regions of $\mathcal{X}$?



The distance between two feature vectors

- Hence, given X_{n+1}, we can think of the strategy that gives more weight to score calculations when data point is closer to X_{n+1}.

Definition : Localization Kernel

The **localization kernel** is the function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.
For example, $H(x,x') = \exp\{-\frac{\|x-x'\|_2^2}{2h^2}\}$ for $x, x' \in \mathbb{R}^d, h > 0$

- Let the weight w_i on the data point (X_i, Y_i) be

$$w_i \propto H(X_i, X_{n+1}),$$

and go back to Algorithm 7.1.

$\Rightarrow$ ~~$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}) \geq 1 - \alpha$~~, and also

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

Localized Conformal Algorithm

Algorithm 7.8: Localized CP

1. Input: Training data $(X_1, Y_1), \dots, (X_n, Y_n)$, test pt. X_{n+1}, target coverage level $1-\alpha$, conformal score ftn. s, and localization kernel H
2. For each possible response value $y \in \mathcal{Y}$, 다음을 계산
   1. 각 sample 별 score $S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y), \quad i = 1, \dots, n.$
   2. Test point score $S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y)$
   3. Weight $w_{i,j} = \frac{H(X_j, X_i)}{\sum_{j'=1}^{n+1} H(X_{j'}, X_i)}, \quad i, j = 1, \dots, n+1.$
   4. Weighted score $\tilde{S}_i^y = \sum_{j=1}^{n+1} w_{i,j} \mathbf{1}\{S_j^y < S_i^y\} \quad i = 1, \dots, n+1.$
   5. Quantile $\tilde{q}^y = \text{Quantile}(\tilde{S}_1^y, \dots, \tilde{S}_{n+1}^y; 1-\alpha)$
3. Return the prediction set $\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : \tilde{S}_{n+1}^y \le \tilde{q}^y\}$

This method is exactly same as the intuitive approach with Algorithm 7.1, but we use the data-dependent threshold $\tilde{q}^y$, rather than the usual threshold $1-\alpha$.

Localized Conformal Algorithm

Proposition 7.9: Localized CP is a version of Full CP

The prediction set defined in Algorithm 7.8 is equivalent to the full conformal prediction set, using the score function as

$$\tilde{s}((x,y);\mathcal{D}) = \sum_{j=1}^{m} \frac{H(x_j,x)}{\sum_{j'=1}^{m} H(x_{j'},x)} \mathbf{1}\{s(x_j,y_j;\mathcal{D}) < s(x,y;\mathcal{D})\}$$

for any (x,y) and any dataset $\mathcal{D} = ((x_1,y_1),\ldots,(x_m,y_m))$.

Also, since s and $\tilde{s}$ are symmetric in $\mathcal{D}$, by Thm 3.2, the **marginal coverage** must hold for localized CP, under the exchangeability assumption.

Approximate conditional coverage

- Now, we try to show the Localized CP can guarantee the approximate conditional coverage.
- Before that, we restrict the definition of the localization kernel for our algorithm.

Definition : Localization Kernel with Density

The **localization kernel** is the function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, such that $H(x, \cdot)$ is a density w.r.t. some measure ν for each $x \in \mathcal{X}$, i.e.,

$$\int_{\mathcal{X}} H(x, x')\, d\nu(x') = 1$$

For example, $H(x, x') = \exp\{-\frac{\|x-x'\|_2^2}{2h^2}\}$ for $x, x' \in \mathbb{R}^d, h > 0$

Randomly-localized Conformal Algorithm

Algorithm 7.10: Randomly-localized CP

1. Input: Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test pt. X_{n+1}, target coverage level $1 - \alpha$, conformal score ftn. s, and localization kernel H
2. Sample $\tilde{X}_{n+1} \sim H(X_{n+1}, \cdot)$
3. For each possible response value $y \in \mathcal{Y}$, 다음을 계산
    1. 각 sample 별 score $S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y), \quad i = 1, \ldots, n.$
    2. Test point score $S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y)$
    3. Weight $w_i = \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})}, \quad i, j = 1, \ldots, n+1.$
    4. Quantile $\hat{q}^y = \text{Quantile}(\sum_{i=1}^{n+1} w_i \delta_{S_i^y}; 1 - \alpha)$
4. Return the prediction set $\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y\}$

Conditional coverage guarantee with Localized CP

Thm 7.11: (Approximate) conditional coverage guarantee with Randomly-localized CP

For $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \overset{\text{i.i.d.}}{\sim} P$, for some P and s is a symmetric score ftn. Let $\mathcal{C}(X_{n+1})$ be the output of Algorithm 7.10., then

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid \tilde{X}_{n+1}) \geq 1 - \alpha \quad \text{a.s.}$$

proof) Note $(X_1, Y_1), \ldots, (X_n, Y_n) \perp\!\!\!\perp (X_{n+1}, Y_{n+1}, \tilde{X}_{n+1}).$, and

$$X_{n+1} \sim P_X,$$
$$Y_{n+1} \mid X_{n+1} \sim P_{Y|X},$$
$$\tilde{X}_{n+1} \mid (X_{n+1}, Y_{n+1}) \sim H(X_{n+1}, \cdot)$$

Then,

$$(X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1} \sim (P_X \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y|X}$$

$$\text{where } \frac{d(P_X \circ H(\cdot, \tilde{X}_{n+1}))(x)}{dP_X(x)} \propto H(x, \tilde{X}_{n+1})$$

Conditional coverage guarantee with Localized CP

proof) (continue)
Hence, given $\tilde{X}_{n+1}$, this is an instance of covariate shift-the distribution of X_{n+1} is no longer P_X, but the conditional distribution $P_{Y|X}$ is invariant.

We can apply Algorithm 7.1 (Weight full CP) to this, and the weight is

$$\tilde{w}_j \propto H(X_j, \tilde{X}_{n+1}) \propto \frac{d(P_X \circ H(\cdot, \tilde{X}_{n+1}))(X_j)}{dP_X(X_j)}.$$

Finally, above algorithm equals to Algorithm 7.10. □

Cf. Back to 4.7.2.

Original Test-coverage:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_0) \geq 1 - \alpha$$
$$\text{for all } P \text{ and all } \mathcal{X}_0 \subseteq \mathcal{X}, \text{ with } P_X(\mathcal{X}_0) \geq \delta$$

Relaxed version: (only certain special subsets $\mathcal{X}_0$)

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid \|X_{n+1} - x_*\| \leq r_*) \geq 1 - \alpha$$
$$\text{for all } P \text{ and all } x_* \in \mathcal{X}, r_* \geq 0 \text{ with } P_X(\|X - x_*\| \leq r_*) \geq \delta$$

It is same as only certain special kernel $H(\cdot, \cdot)$:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid x_*) \geq 1 - \alpha$$
$$\text{if } x_* \sim H(X_{n+1}, \cdot) \quad \text{with } \int_{B(x,r_*)} H(x, x') d\nu(x') \geq \delta$$

Outline

1. Weighted quantiles and the weighted conformal algorithm
2. Conformal prediction under covariate and label shifts
3. Localized conformal prediction
4. Fixed-weight conformal prediction
5. A general outlook through weighted permutations

Robustness for non-exchangeability

Example: Time series

- For time points $1, \dots, n+1$, if we want to get a prediction set of X_{n+1},
- we expect that the distribution of recent points is closer to that of the test point.

$\Rightarrow$ Set recent data higher weights relative to data from long ago.

- In this chapter, consider the fixed weights, i.e., data-independent weights,
- for a conservative approach with certain violations of exchangeability.

Fixed-weight CP

Thm 7.12: Lower bound of marginal coverage with fixed-weight CP

For the weights $w_1, \ldots, w_{n+1}$ with $w_{n+1} \geq w_i,\ \forall i \in [n]$, r.v.s $Z_1, \ldots Z_{n+1}$ with any joint distribution, and the symmetric score s, let

$$\mathcal{C}(X_{n+1}) = \{y : S^y_{n+1} \leq \hat{q}_w\}, \quad \text{where } \hat{q}^y_w = \text{Quantile}\left(\sum_{i=1}^{n+1} w_i \delta_{S^y_i}; 1-\alpha\right)$$

Then,

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1-\alpha-\sum_{i=1}^{n} w_i\, d_{\mathsf{TV}}\big((Z_1, \ldots, Z_{n+1}), (Z_1, \ldots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \ldots, Z_i)\big)$$

where d_{TV} means the total variation distance between distributions.

- Exchangeability Assumption X

$\Rightarrow$ Instead, introduce the TV distance of the swapped versions of the data.

1. If exchangeable, $d_{\mathsf{TV}} = 0$.
2. If not, set a smaller weight w_i when the distribution of Z_i is more different from that of Z_{n+1}.

Fixed-weight CP

proof) For the vector of whole scores $S = (S_1, \ldots, S_{n+1})$, and swapped scores $S^i = (S_1, \ldots, S_{i-1}, S_{n+1}, S_{i+1}, \ldots, S_n, S_i)$,

define a **score vector returning operator** $h : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}^{n+1}$ as

$$h(z_1, \ldots, z_{n+1}) = (s(z_1; (z_1, \ldots, z_{n+1})), \ldots, s(z_{n+1}; (z_1, \ldots, z_{n+1}))),$$

i.e., $S = h(Z_1, \ldots, Z_{n+1})$, and

$$S^i = h(Z_1, \ldots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \ldots, Z_n, Z_i) \quad (\because s \text{ is symmetric.})$$

Fixed-weight CP

proof) (continue)
Then, by Data Processing Inequality,

$$\begin{aligned} d_{\mathsf{TV}}(S, S^i) &= d_{\mathsf{TV}}(h(Z_1, \dots, Z_{n+1}), h(Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \dots, Z_n, Z_i)) \\ &\le d_{\mathsf{TV}}((Z_1, \dots, Z_{n+1}), (Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \dots, Z_n, Z_i)) \end{aligned}$$

and it suffices to show

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \ge 1 - \alpha - \sum_{i=1}^{n} w_i \cdot d_{\mathsf{TV}}(S, S^i).$$

(Important Fact) We use $d_{\mathsf{TV}}(S, S^i)$, rather than $d_{\mathsf{TV}}((\mathcal{D}_n, Z_{n+1}), (\mathcal{D}_n^{Z_{n+1}}, Z_i))$.
$\Rightarrow$ It is working better when the total variation distance between the data points Z_i and Z_{n+1} could be large, but the distributions of the corresponding scores S_i, S_{n+1} might be similar.

Fixed-weight CP

proof) (continue)
Recall the definition of Total Variation, for any measurable $A \subseteq (\mathcal{X} \times \mathcal{Y})^{n+1}$,

$$d_{\mathsf{TV}}(S, S^i) = \sup_A \|p_S(A) - p_{S'}(A)\|$$

Using it, for $S = S^{n+1}$,

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) &= \mathbb{P}\left(S_{n+1} \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{S_j}; 1-\alpha\right)\right) \\
&= \sum_{i=1}^{n+1} w_i \cdot \mathbb{P}\left(S_{n+1} \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{S_j}; 1-\alpha\right)\right) \\
&\geq \sum_{i=1}^{n+1} w_i \left[\mathbb{P}\left((S^i)_{n+1} \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{(S^i)_j}; 1-\alpha\right)\right) - d_{\mathsf{TV}}(S, S^i)\right. \\
&= \mathbb{E}\left[\sum_{i=1}^{n+1} w_i \cdot \mathbf{1}\left\{(S^i)_{n+1} \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{(S^i)_j}; 1-\alpha\right)\right\}\right] - \sum_{i=1}^{n} w_i \, d_{\mathsf{TV}}(S, S^i).
\end{aligned}$$

Fixed-weight CP

proof) (continue)
Claim:

$$S_i \leq \mathsf{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{S_j}; 1-\alpha\right) \Rightarrow (S^i)_{n+1} \leq \mathsf{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{(S^i)_j}; 1-\alpha\right) \qquad (7.8)$$

If holds,

$$\begin{aligned}
&\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \\
&\geq \mathbb{E}\left[\sum_{i=1}^{n+1} w_i \cdot \mathbf{1}\left\{S_i \leq \mathsf{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{S_j}; 1-\alpha\right)\right\}\right] - \sum_{i=1}^{n} w_i \, d_{\mathsf{TV}}(S, S^i) \\
&\geq 1-\alpha-\sum_{i=1}^{n} w_i \cdot d_{\mathsf{TV}}(S, S^i). \quad (\because \text{Fact. } 7.7)
\end{aligned}$$

Fixed-weight CP

proof) (continue)
To show (7.8), we only consider the cases $i \in [n]$. We will need following result:

(Recall) Lemma 3.4: Replacement Lemma

Let $v_1, \ldots, v_{n+1} \in \mathbb{R}$. For any $t \in [0,1]$,

$$v_{n+1} \le \text{Quantile}(v_1, \ldots, v_n; t) \iff v_{n+1} \le \text{Quantile}(v_1, \ldots, v_n, v_{n+1}; t(1+\frac{1}{n}))$$

Lemma 7.13: Replacement Lemma Generalization

For any distributions P_0, and P_1, and some $\epsilon \in [0,1]$, let
$P = (1-\epsilon)P_0 + \epsilon P_1$ be their mixture.
Then for any $x \in \mathbb{R}$, and $\tau \in [0,1]$,

$$x \le \text{Quantile}(P; \tau) \Rightarrow x \le \text{Quantile}\big((1-\epsilon) \cdot P_0 + \epsilon \cdot \delta_x; \tau\big)$$

Fixed-weight CP

proof) (continue)
For applying this, we setup $\epsilon = w_{n+1} - w_i \geq 0, \tau = 1 - \alpha$, and

$$P_1 = \delta_{S_{n+1}}, \quad P_0 = \frac{\sum_{j\in[n],\, j\neq i} w_j \cdot \delta_{S_j} + w_i \cdot (\delta_{S_i} + \delta_{S_{n+1}})}{1 - \epsilon}$$

Then,

$$\begin{aligned}
& S_i \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{S_j}; 1-\alpha\right) \\
\Leftrightarrow\ & S_i \leq \text{Quantile}\left((1-\epsilon)P_0 + \epsilon P_1; 1-\alpha\right) \\
\Rightarrow\ & S_i \leq \text{Quantile}\left((1-\epsilon)P_0 + \epsilon \delta_{S_i}; 1-\alpha\right) \\
\Leftrightarrow\ & S_i \leq \text{Quantile}\left(\sum_{j\in[n], j\neq i} w_j \cdot \delta_{S_j} + w_{n+1} \cdot \delta_{S_i} + w_i \cdot \delta_{S_{n+1}}; 1-\alpha\right) \\
\Leftrightarrow\ & (S^i)_{n+1} \leq \text{Quantile}\left(\sum_{j=1}^{n+1} w_j \delta_{(S^i)_j}; 1-\alpha\right). \quad \square
\end{aligned}$$

Outline

1. Weighted quantiles and the weighted conformal algorithm

2. Conformal prediction under covariate and label shifts

3. Localized conformal prediction

4. Fixed-weight conformal prediction

5. A general outlook through weighted permutations

Generalized CP

- In this chapter, we talk about reinterpretation CP through weighted distributions over permutations.
- We can generalize CP to work with any joint dist.s, even non-exchangeable.

- The main idea is if we know only **the probability of observing every ordering of the data points conditionally on their values**, this is enough to run CP.

Generalized CP

(Recall) Proposition 7.6: Conditioning on the empirical distribution

Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ and $Z_{n+1} \sim Q$ independently, then

$$Z_{n+1} \mid \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i \delta_{Z_i}, \text{ where } w_i = \frac{\frac{dQ}{dP}(Z_i)}{\sum_{j=1}^{n+1} \frac{dQ}{dP}(Z_j)}, \quad i \in [n+1]$$

Now, let the density $f : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}$ as $(Z_1, \ldots, Z_{n+1}) \sim f$, then

$$Z_{n+1} \mid \hat{P}_{n+1} \sim \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})\, \delta_{Z_{\sigma(n+1)}}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})}. \quad (\because \text{non-i.i.d.})$$

That is, once we fix the values of the data points, then we can explicitly describe the distribution of the test point in terms of their ordering, if we have knowledge of the joint density f.

Generalized CP

Thm 7.14: Generalized Weighted CP

For $(Z_1, \dots Z_n, Z_{n+1}) \sim P$, and the joint pdf f for any measure on $\mathcal{X} \times \mathcal{Y}$, fix any score s (not necessarily symmetric), and define the prediction set

$$\mathcal{C}(X_{n+1}) = \{y : s((X_{n+1}, y); \mathcal{D}_{n+1}) \leq \hat{q}_w^y\},$$

$$\text{where } \hat{q}_w^y = \text{Quantile}\left(\sum_{\sigma \in \mathcal{S}_{n+1}} w_\sigma^y \, \delta_{s(Z^y_{\sigma(n+1)}; \mathcal{D}^{y,\sigma}_{n+1})} ; 1 - \alpha \right)$$

where

$$w_\sigma^y = \frac{f(Z^y_{\sigma(1)}, \dots, Z^y_{\sigma(n+1)})}{\sum_{\sigma' \in \mathfrak{S}_{n+1}} f(Z^y_{\sigma'(1)}, \dots, Z^y_{\sigma'(n+1)})}$$

and the permuted dataset

$$(D^y_{n+1})_\sigma = (Z^y_{\sigma(1)}, \dots, Z^y_{\sigma(n+1)}) = ((X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n+1)}, y))$$

Then, it satisfies the marginal coverage guarantee,

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

Generalized CP

- Pros
  1. f 를 잘 몰라도 가능한 경우 존재: i.i.d. case $\left(w_\sigma^y = \frac{1}{(n+1)!}\right)$, etc.
  2. 가중치가 entire permutation σ 에 의존 (more general than covariate/label shift)
  3. nonexchangeable data, nonsymmetric algorithm 에 적용 가능
- Cons
  1. 대부분 f 를 알아야 함
  2. $(n+1)!$ 의 permutation 을 모두 계산
- Worth
  1. Unified version of existing method of Conformal Prediction
  2. It points toward an underlying structure that facilitates conformal prediction.