Lecture 1
Distribution-free inference: aims and algorithms

Rina Foygel Barber

http://www.stat.uchicago.edu/~rina/

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)
- So, we run a test to check if the assumptions hold
  (goodness of fit / overdispersion / calibration ...)
- But, what if this test is only guaranteed to detect violations,
  under some other assumptions?

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)
- So, we run a test to check if the assumptions hold
  (goodness of fit / overdispersion / calibration ...)
- But, what if this test is only guaranteed to detect violations,
  under some other assumptions?

The goal of distribution-free inference is to provide guarantees
that are valid universally over all data distributions.

Distribution-free inference: questions

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y_{n+1} will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

Distribution-free inference: questions

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y_{n+1} will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

Lecture 1 — algorithms for prediction
Lecture 2 — challenges with other inference problems

Holdout methods

If we fit a regression model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ on training data,
how to build a prediction interval $\widehat{\mu}(X_{n+1}) \pm [...]$ for Y_{n+1}?

- If we use the training residuals $R_i = |Y_i - \widehat{\mu}(X_i)|$,
  these are too small due to overfitting

Holdout methods

If we fit a regression model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ on training data,
how to build a prediction interval $\widehat{\mu}(X_{n+1}) \pm [...]$ for Y_{n+1}?

- If we use the training residuals $R_i = |Y_i - \widehat{\mu}(X_i)|$,
  these are too small due to overfitting
- Holdout method: fit $\widehat{\mu}$ on training set $i = 1, \ldots, \frac{n}{2}$,
  then use residuals $\{R_i = |Y_i - \widehat{\mu}(X_i)|\}_{i=\frac{n}{2}+1,\ldots,n}$ to choose width

Holdout methods

Prediction interval:

$$\widehat{C}_n(X_{n+1}) = \overset{\text{fitted on training data } i=1,\dots,\frac{n}{2}}{\widehat{\mu}(X_{n+1})} \quad \pm \quad \underbrace{Q_{1-\alpha}(R_{\frac{n}{2}+1},\dots,R_n)}_{\substack{\text{the } (1-\alpha)(\frac{n}{2}+1)\text{-th smallest resid.} \\ \text{in the holdout set } i=\frac{n}{2}+1,\dots,n}}$$

Theorem:[1] if data points are i.i.d., $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha$

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

Prediction interval:

$$\widehat{C}_n(X_{n+1}) = \underbrace{\widehat{\mu}}_{\text{fitted on training data } i=1,\dots,\frac{n}{2}}(X_{n+1}) \quad \pm \quad \underbrace{Q_{1-\alpha}(R_{\frac{n}{2}+1},\dots,R_n)}_{\substack{\text{the } (1-\alpha)(\frac{n}{2}+1)\text{-th smallest resid.} \\ \text{in the holdout set } i=\frac{n}{2}+1,\dots,n}}$$

Theorem:[1] if data points are i.i.d., $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha$

Proof: $\left\{(X_i, Y_i)\right\}_{i=\frac{n}{2}+1,\dots,n+1}$ are i.i.d. conditional on training data
$\Rightarrow$ residuals $\left\{R_i\right\}_{i=\frac{n}{2}+1,\dots,n+1}$ are i.i.d. cond. on training data

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

Prediction interval: fitted on training data $i = 1, \ldots, \frac{n}{2}$

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \quad \pm \quad \underbrace{Q_{1-\alpha}(R_{\frac{n}{2}+1}, \ldots, R_n)}_{\substack{\text{the } (1-\alpha)(\frac{n}{2}+1)\text{-th smallest resid.} \\ \text{in the holdout set } i=\frac{n}{2}+1,\ldots,n}}$$

Theorem:[1] if data points are i.i.d., $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$

Proof: $\left\{(X_i, Y_i)\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are i.i.d. conditional on training data
$\Rightarrow$ residuals $\left\{R_i\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are i.i.d. cond. on training data
$\Rightarrow$ residuals $\left\{R_i\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are exchangeable

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

Prediction interval: fitted on training data $i = 1, \ldots, \frac{n}{2}$

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \underbrace{Q_{1-\alpha}(R_{\frac{n}{2}+1}, \ldots, R_n)}_{\substack{\text{the } (1-\alpha)(\frac{n}{2}+1)\text{-th smallest resid.} \\ \text{in the holdout set } i = \frac{n}{2}+1, \ldots, n}}$$

Theorem:[1] if data points are i.i.d., $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha$

Proof: $\left\{(X_i, Y_i)\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are i.i.d. conditional on training data
$\Rightarrow$ residuals $\left\{R_i\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are i.i.d. cond. on training data
$\Rightarrow$ residuals $\left\{R_i\right\}_{i=\frac{n}{2}+1,\ldots,n+1}$ are exchangeable
$\Rightarrow \mathbb{P}\left\{R_{n+1} \leq \text{ the } (1-\alpha)(\frac{n}{2}+1)\text{-th smallest of } R_{\frac{n}{2}+1}, \ldots, R_{n+1}\right\} \geq 1-\alpha$

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

- In the above construction,

  $$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [\ldots] \; = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [\ldots] \; \}$$

- We can generalize to any score function:[2]

  $$\widehat{C}_n(X_{n+1}) = \{ \text{ all } y \text{ values with } S(X_{n+1}, y) \leq [\ldots] \; \}$$

  where $S(x, y)$ measures "nonconformity" of the data point (x, y)

[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

- In the above construction,

  $\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [...] = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [...] \ \}$

- We can generalize to any score function:[2]

  $$\widehat{C}_n(X_{n+1}) = \{ \text{ all } y \text{ values with } S(X_{n+1}, y) \leq [...] \ \}$$

  where $S(x, y)$ measures "nonconformity" of the data point (x, y)

The method:

- Using data $i = 1, \ldots, n/2$, fit "nonconformity score" $S(x, y)$
- Compute $S_i = S(X_i, Y_i)$ for $i = n/2 + 1, \ldots, n$,
  & $Q_{1-\alpha}(S_{n/2} + 1, \ldots, S_n)$
- Prediction interval:
  $\widehat{C}_n(X_{n+1}) = \{y : S(X_{n+1}, y) \leq Q_{1-\alpha}(S_{n/2} + 1, \ldots, S_n)\}$

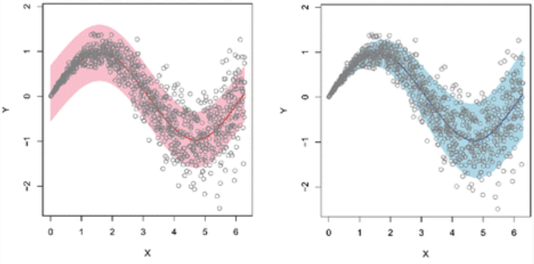
[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Holdout methods

Why any score function?

- If noise level varies with X, may want varying interval width:[3]

$$S(x,y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)} \quad \Rightarrow \quad \widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{\sigma}(X_{n+1}) \cdot Q_{1-\alpha}(\ldots)$$



(figure from Lei et al 2018)

[3] Lei et al 2018, *Distribution-Free Predictive Inference for Regression*

Holdout methods

Why any score function?

- If Y is discrete (or even categorical),[4]
  not natural to take $\widehat{C}_n(X_{n+1}) = $ [some continuous interval]
- Instead, take $\widehat{C}_n(X_{n+1}) = \{$all categories y that are likely given $X_{n+1}\}$

[4] Sadinle et al 2019, *Least ambiguous set-valued classifiers with bounded error levels*

Holdout methods

Why any score function?

- If Y is discrete (or even categorical),[4]
  not natural to take $\widehat{C}_n(X_{n+1}) = $ [some continuous interval]
- Instead, take $\widehat{C}_n(X_{n+1}) = \{$all categories y that are likely given $X_{n+1}\}$
- Compute estimate $\widehat{p}(y|x) \approx \mathbb{P}\{Y = y | X = x\}$,

then use $S(x, y) = 1/\widehat{p}(y|x) \;\Rightarrow\; \widehat{C}_n(X_{n+1}) = \{y : \widehat{p}(y|X_{n+1}) \geq [...]\}$

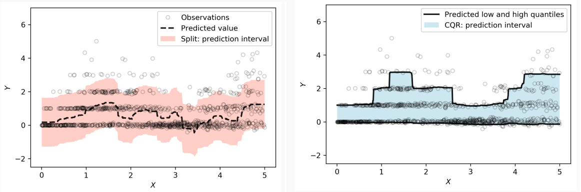
[4] Sadinle et al 2019, *Least ambiguous set-valued classifiers with bounded error levels*

Holdout methods

Why any score function?

- If the shape of distrib. of $Y|X$ varies with X,
  $\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [\dots]$ is not an optimal form for the interval
- Instead, can estimate conditional quantiles directly:[5]

$$S(x,y) = \max\{\widehat{q}_{\alpha/2}(x) - y,\ y - \widehat{q}_{1-\alpha/2}(x)\} \ \text{ or } \ \max\left\{\frac{\widehat{q}_{0.5}(x) - y}{\widehat{q}_{0.5}(x) - \widehat{q}_{\alpha/2}},\ \frac{y - \widehat{q}_{0.5}(x)}{\widehat{q}_{1-\alpha/2}(x) - \widehat{q}_{0.5}(x)}\right\}$$



(figure from Romano et al 2019)

[5] Romano et al 2019, *Conformalized quantile regression*
Kivaranovic et al 2019 *Adaptive, distribution-free prediction intervals for deep neural networks*

Conformal prediction

All methods so far rely on data splitting:

- Training: use $n/2$ data points to develop a score function $S(x, y)$
- Validation: use $n/2$ data points to learn the distrib. of $S(X, Y)$
- Then we can predict $S(X_{n+1}, Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

[6] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Conformal prediction

All methods so far rely on data splitting:

- Training: use $n/2$ data points to develop a score function $S(x, y)$
- Validation: use $n/2$ data points to learn the distrib. of $S(X, Y)$
- Then we can predict $S(X_{n+1}, Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

The drawback: sample splitting means that we only use $n/2$ data points to fit the model / the score function

Conformal prediction: use all data points for training & validation[6]

[6] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Conformal prediction

All methods so far rely on data splitting:

- Training: use $n/2$ data points to develop a score function $S(x, y)$
- Validation: use $n/2$ data points to learn the distrib. of $S(X, Y)$
- Then we can predict $S(X_{n+1}, Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

The drawback: sample splitting means that we only use $n/2$ data points to fit the model / the score function

Conformal prediction: use all data points for training & validation[6]

Terminology: $\begin{cases} \text{holdout method / split conformal / inductive conformal} \\ \text{conformal / full conformal / transductive conformal} \end{cases}$

[6] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Conformal prediction

- Suppose we observe training + test data:

$$(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i),\ i = 1, \dots, n, \quad R_{n+1} = Y_{n+1} - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \dots, |R_n|, |R_{n+1}|\Big]$

By exchangeability of $R_1, \dots, R_{n+1}$,
this event has $\geq 1 - \alpha$ probability

Conformal prediction

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \quad y \quad)$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i),\ i = 1, \ldots, n, \quad R_{n+1} = \quad y \quad - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

By exchangeability of $R_1, \ldots, R_{n+1}$,
this event has $\geq 1-\alpha$ probability if we plug in $y = Y_{n+1}$

Conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals
$$R_i = Y_i - \widehat{\mu}(X_i), \; i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$
- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

Conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals
$$R_i = Y_i - \widehat{\mu}(X_i),\ i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$
- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

Conformal prediction

Test value $y \in \mathbb{R}$

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals
$$R_i = Y_i - \widehat{\mu}(X_i),\ i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$
- Check if $|R_{n+1}| \leq \Big[(1-\alpha)$ quantile of $|R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}_n(X_{n+1})$

if No: discard y

Conformal prediction

Test value $y \in \mathbb{R}$

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get residuals
$$R_i = Y_i - \widehat{\mu}(X_i), \ i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$
- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}_n(X_{n+1})$

if No: discard y

$$\boxed{\mathbb{P}\Big\{Y \in \widehat{C}_n(X_{n+1})\Big\} = \mathbb{P}\{\text{for test value } y = Y_{n+1}, \text{ answer is Yes}\} \geq 1 - \alpha}$$

Conformal prediction

"Full" conformal can be run with any score function *on data sets*:

$$\big((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\big) \mapsto (S_1, \ldots, S_{n+1})$$

S_i = "nonconformity score" of data point i, relative to rest of the data

- Regression: $S_i = |y_i - \widehat{\mu}(x_i)|$ where $\widehat{\mu}$ is fitted on all data
- Quantile regression: S_i measures gap between y_i and $\widehat{q}(x_i)$
- Classification: $S_i = 1/\widehat{p}(y_i|x_i)$
- & many more

Conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get nonconformity scores
$$S_1, \ldots, S_{n+1}$$
- Check if $|S_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |S_1|, \ldots, |S_n|, |S_{n+1}|\Big]$

Conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get nonconformity scores
$$S_1, \ldots, S_{n+1}$$
- Check if $|S_{n+1}| \leq \Big[(1-\alpha)$ quantile of $|S_1|, \ldots, |S_n|, |S_{n+1}|\Big]$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

Conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get nonconformity scores
$$S_1, \ldots, S_{n+1}$$
- Check if $|S_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |S_1|, \ldots, |S_n|, |S_{n+1}|\Big]$

Test value $y \in \mathbb{R}$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}_n(X_{n+1})$

if No: discard y

Conformal prediction

Test value $y \in \mathbb{R}$

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model $\widehat{\mu}$ to all $n+1$ data points, & get nonconformity scores
$$S_1, \ldots, S_{n+1}$$
- Check if $|S_{n+1}| \leq \left[(1-\alpha) \text{ quantile of } |S_1|, \ldots, |S_n|, |S_{n+1}|\right]$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}_n(X_{n+1})$

if No: discard y

$$\boxed{\mathbb{P}\left\{Y \in \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}\left\{\text{for test value } y = Y_{n+1}, \text{ answer is Yes}\right\} \geq 1-\alpha}$$

Weighted conformal prediction

Conformal prediction (or holdout method) assumes:
training & test data are from the same distribution

One possible violation: covariate shift

- Marginal distribution of X is different in training vs. test data (e.g., some subpopulations are over- or under-represented in the training data)
- But, distribution of $Y|X$ is the same

Weighted conformal prediction

Assuming we know the shift (i.e., $\mathrm{d}P_X^{\text{test}}(x) \propto \overbrace{w(x)}^{\text{known fn.}} \cdot \ \mathrm{d}P_X^{\text{train}}(x)$), conformal can adjust for the shift by using *weighted exchangeability*[7]

[7] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*
Hu & Lei 2020, *A distribution-free test of covariate shift using conformal prediction*

Weighted conformal prediction

Assuming we know the shift (i.e., $\mathrm{d}P_X^{\text{test}}(x) \propto \overbrace{w(x)}^{\text{known fn.}} \cdot \; \mathrm{d}P_X^{\text{train}}(x)$),
conformal can adjust for the shift by using *weighted exchangeability*[7]

Given $n+1$ data points...

- With (unweighted) exchangeability,
  each one is equally likely to be the test point
  $\rightsquigarrow R_{n+1} \leq Q_{1-\alpha}(R_1, \ldots, R_{n+1})$ with prob. $1-\alpha$
- With weighted exchangeability, the distribution is nonuniform:

$$\mathbb{P}\left\{(x, y) \text{ is the test point}\right\} \propto w(x)$$

$\rightsquigarrow$ need to compute a weighted quantile: $Q_{1-\alpha}\left(\sum_i \frac{w(X_i)}{\sum_{i'} w(X_{i'})} \delta_{R_i}\right)$

[7] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*
Hu & Lei 2020, *A distribution-free test of covariate shift using conformal prediction*

Weighted conformal prediction

Application: survival analysis & censored data[8]

- "Clean" data $(X_i, Y_i) =$ (features, survival time)
- Censored observations $(X_i, \tilde{Y}_i)$ where $\tilde{Y}_i = \min\{C_i, Y_i\}$
- Main idea: choose a cutoff c_0 so that "usually" $Y_i \leq c_0$,
  & keep only data with $C_i \geq c_0$ (i.e., most Y's are not censored)

[8] Candès, Lei, Ren 2021, *Conformalized survival analysis*

Weighted conformal prediction

Application: survival analysis & censored data[8]

- "Clean" data $(X_i, Y_i) =$ (features, survival time)
- Censored observations $(X_i, \tilde{Y}_i)$ where $\tilde{Y}_i = \min\{C_i, Y_i\}$
- Main idea: choose a cutoff c_0 so that "usually" $Y_i \leq c_0$,
  & keep only data with $C_i \geq c_0$ (i.e., most Y's are not censored)
  - On this data set, can use CP to predict survival time Y
  - But, this may be a different distribution
    (population with $C_i \geq c_0 \quad \neq \quad$ general population)
  - If distrib. of $C|X$ known,
    can use weighted CP to correct for distribution shift

[8] Candès, Lei, Ren 2021, *Conformalized survival analysis*

Weighted conformal prediction

Application: estimating individual treatment effects[9]

- Data $(X_i, T_i, Y_i) =$ (features, treatment group $= 0$ or 1, outcome)
- $\text{ITE}_i =$ (value of Y_i, if $T_i = 1$) − (value of Y_i, if $T_i = 0$)
- Challenge: treatment assignment may depend on X
- Main idea: if propensity score $\mathbb{P}\{T = 1 \mid X = x\}$ is known, can use weighted CP to adjust for $X|T = 1$ versus $X|T = 0$

[9] Lei & Candès 2020, *Conformal inference of counterfactuals and individual treatment effects*

Weighted conformal prediction

A related problem — label shift (for categorical Y / classification)[10]

- Marginal distribution of Y is different in training vs. test data (e.g., some subpopulations are over- or under-represented in the training data)
- But, distribution of $X|Y$ is the same

If the label shift is known (i.e., the weight function $w(y) = \frac{P_{Y,\text{test}}(y)}{P_{Y,\text{train}}(y)}$), can use weighted exchangeability to guarantee coverage

[10] Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Full conformal: computational challenges

"Full" conformal prediction requires that the model fitting algorithm is re-run:

- For each test value X_{n+1} of interest
- For every possible value $y \in \mathbb{R}$

Approaches:

- In practice — use a grid of y values (but no theory)
- Specialized methods for specific algorithms e.g. Lasso[11]
- Discretized CP —
  "discretize the data" or "discretize the model"[12]

[11] Lei 2017, *Fast Exact Conformalization of Lasso using Piecewise Linear Homotopy*

[12] Chen, Chun, & B. 2017, *Discretized conformal prediction for efficient distribution-free inference*

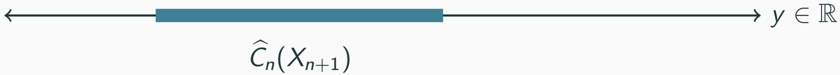
Full conformal: computational challenges

A preliminary observation—
It is valid to run CP on the interval $[\min_{1\le i\le n} Y_i,\ \max_{1\le i\le n} Y_i]$

- With prob. $\geq 1 - \frac{2}{n+1}$, including Y_{n+1} doesn't change the endpoints
- So, coverage is $\geq 1 - \alpha - \frac{2}{n+1}$

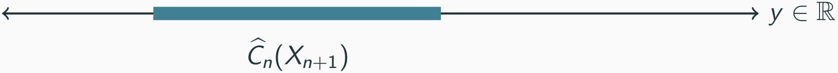
Full conformal: computational challenges

Conformal prediction:



Full conformal: computational challenges

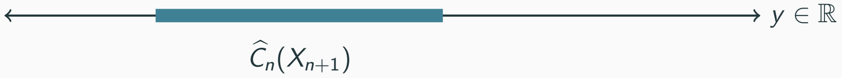
Conformal prediction:



Conformal prediction with rounding (informal version):

Full conformal: computational challenges

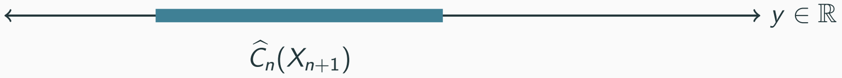
Conformal prediction:



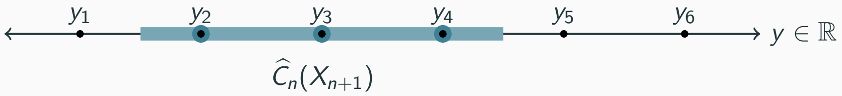
Conformal prediction with rounding (informal version):

Full conformal: computational challenges

Conformal prediction:

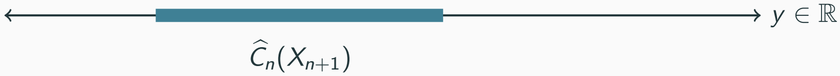


Conformal prediction with rounding (informal version):

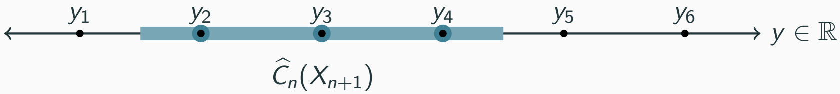


Full conformal: computational challenges

Conformal prediction:



Conformal prediction with rounding (informal version):



Problems to solve:

- Theory to guarantee coverage rate $1 - \alpha$?
- Avoid wider intervals due to discretized grid?

Full conformal: computational challenges

Why do we lose the coverage guarantee?

- If only fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for y in a grid...
- Equivalent to: fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, [Y_{n+1}])$

Y_{n+1} rounded to grid

Full conformal: computational challenges

Why do we lose the coverage guarantee?

- If only fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for y in a grid...
- Equivalent to: fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, [Y_{n+1}])$

Y_{n+1} rounded to grid

To maintain exchangeability:

- Fit $\widehat{\mu}$ on $(X_1, [Y_1]), \ldots, (X_n, [Y_n]), (X_{n+1}, y)$
- Residuals $R_i = [Y_i] - \widehat{\mu}(X_i)$, $R_{n+1} = y - \widehat{\mu}(X_{n+1})$

Full conformal: computational challenges

Why do we lose the coverage guarantee?

- If only fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for y in a grid...
- Equivalent to: fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, [Y_{n+1}])$

Y_{n+1} rounded to grid

To maintain exchangeability:

- Fit $\widehat{\mu}$ on $(X_1, [Y_1]), \ldots, (X_n, [Y_n]), (X_{n+1}, y)$
- Residuals $R_i = [Y_i] - \widehat{\mu}(X_i)$, $R_{n+1} = y - \widehat{\mu}(X_{n+1})$
- Prediction set on the grid:

$$[\widehat{C}_n](X_{n+1}) = \{y \text{ values on the grid such that } |R_{n+1}| \leq \ldots\}$$

- Final prediction set:

$$\widehat{C}_n(X_{n+1}) = \{y \in \mathbb{R} : \ y \text{ rounds to any value in } [\widehat{C}_n](X_{n+1})\}$$

Full conformal: computational challenges

Alternate approach—discretize the model, not the data

Full conformal: computational challenges

Alternate approach—discretize the model, not the data

- Run a discretized algorithm for model fitting:

$$(X_1, Y_1), \ldots, (X, y) \xrightarrow{[\;]} (X_1, [Y_1]), \ldots, (X, [y]) \xrightarrow{\text{fit model } \widehat{\mu}} [\widehat{\mu}]$$

- Calculate residuals

$$R_i = Y_i - [\widehat{\mu}](X_i),\ i = 1, \ldots, n, \quad R_{n+1} = y - [\widehat{\mu}](X)$$

- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } |R_1|, \ldots, |R_n|, |R_{n+1}|\Big]$

- Model is fitted once for each possible value of $[y] \in$ grid
- Residual is calculated for each possible value of $y \in \mathbb{R}$

Full conformal: computational challenges

If we run conformal prediction w/ discretized data:

✓ • Theory to guarantee coverage rate $1 - \alpha$?

? • Avoid wider intervals due to discretized grid?

If we run conformal prediction w/ discretized model:

✓ • Theory to guarantee coverage rate $1 - \alpha$?

✓ • Avoid wider intervals due to discretized grid?

Jackknife & CV / Jackknife+ & CV+

Holdout methods	vs	full conformal
(lose sample size)		(high computational cost)

Jackknife & CV / Jackknife+ & CV+

Holdout methods vs full conformal
(lose sample size) (high computational cost)

To avoid this tradeoff, can we use cross-validation?

Split data into $S_1 \cup \cdots \cup S_K$

For each $i \in S_k$, $R_i^{\mathrm{CV}} = |Y_i - \widehat{\mu}_{-S_k}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(R_i^{\mathrm{CV}}\right)$$

— Computational cost: $K + 1$ regressions

— Problem: theory from holdout setting no longer holds

Jackknife & CV / Jackknife+ & CV+

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)

Residuals $R_i^{\mathrm{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(R_i^{\mathrm{LOO}}\right)$$

[13] Steinberger & Leeb 2018, *Conditional predictive inference for high-dimensional stable algorithms*

Jackknife & CV / Jackknife+ & CV+

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)

Residuals $R_i^{\text{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(R_i^{\text{LOO}}\right)$$

— Predictive coverage under algorithmic stability assumption:[13]

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| \leq \epsilon\right\} \geq 1 - \nu$$

[13] Steinberger & Leeb 2018, *Conditional predictive inference for high-dimensional stable algorithms*

Jackknife & CV / Jackknife+ & CV+

Jackknife+:[14]

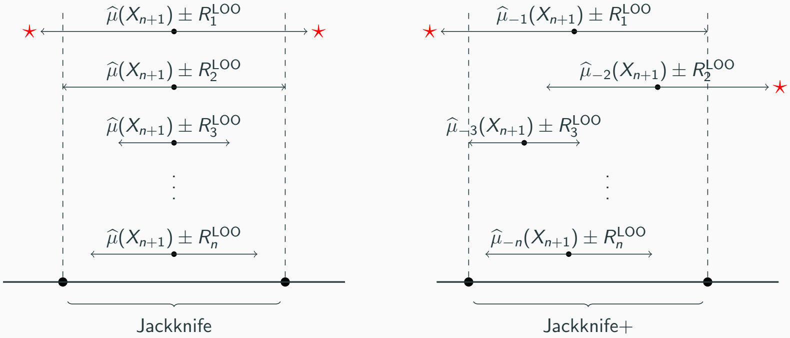
$$C(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left(\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}\right),\ \mathrm{Q}_{1-\alpha}\left(\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}\right)\right]$$

Compare to jackknife:

$$C(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left(\widehat{\mu}(X_{n+1}) - R_i^{\mathrm{LOO}}\right),\ \mathrm{Q}_{1-\alpha}\left(\widehat{\mu}(X_{n+1}) + R_i^{\mathrm{LOO}}\right)\right]$$

[14] B., Candès, Ramdas, Tibshirani 2019

Jackknife & CV / Jackknife+ & CV+



Jackknife & CV / Jackknife+ & CV+

Extension to CV+:[15]

$$C(X_{n+1}) = \left[\mathrm{Q}_\alpha \left(\widehat{\mu}_{-S_k}(X_{n+1}) - R_i^{\mathrm{CV}} \right),\ \mathrm{Q}_{1-\alpha} \left(\widehat{\mu}_{-S_k}(X_{n+1}) + R_i^{\mathrm{CV}} \right) \right]$$

Closely related to the cross-conformal prediction method[16]

[15] B., Candès, Ramdas, Tibshirani 2019

[16] Vovk 2015, Vovk et al 2018

Theorem: For any distrib. P and any $\mathcal{A}$, jackknife+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - 2\alpha.$$

Jackknife & CV / Jackknife+ & CV+

Theorem: For any distrib. P and any $\mathcal{A}$, jackknife+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - 2\alpha.$$

Theorem: For any distrib. P and any $\mathcal{A}$, K-fold CV+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq \begin{cases} 1 - 2\alpha - 1/K & \text{(new)} \\ 1 - 2\alpha - 2K/n & \text{(Vovk et al)} \end{cases}$$

$$\rightsquigarrow \quad \geq 1 - 2\alpha - \sqrt{2/n}.$$

Jackknife & CV / Jackknife+ & CV+

Proof idea: embed jackknife+ into a larger <u>exchangeable</u> problem

Jackknife & CV / Jackknife+ & CV+

Proof idea: embed jackknife+ into a larger exchangeable problem

- Exchangeable data $\{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\}$
- $\binom{n+1}{2}$ leave-two-out regressions: $\widetilde{\mu}_{-\{i,j\}}$ for $1 \leq i, j \leq n+1$
- We can observe n of these, i.e., $\widetilde{\mu}_{-\{i,n+1\}} = \widehat{\mu}_{-i}$ for $1 \leq i \leq n$

Online / streaming / time series data

Conformal prediction can also be applied to an online setting...

- If data points are iid,

  conformal p-values are valid (and $\perp\!\!\!\perp$) at each time t

  $\Rightarrow$ can use conformal to predict / to test for changepoints[17]

[17] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[18] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[19] Xu & Xie 2021, *Conformal prediction interval for dynamic time-series*

Online / streaming / time series data

Conformal prediction can also be applied to an online setting...

- If data points are iid,
  conformal p-values are valid (and $\perp\!\!\!\perp$) at each time t
  $\Rightarrow$ can use conformal to predict / to test for changepoints[17]
- If data points are independent but there is distribution drift,
  can bound cumulative error[18]

[17] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[18] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[19] Xu & Xie 2021, *Conformal prediction interval for dynamic time-series*

Online / streaming / time series data

Conformal prediction can also be applied to an online setting...

- If data points are iid,
  conformal p-values are valid (and $\perp\!\!\!\perp$) at each time t
  $\Rightarrow$ can use conformal to predict / to test for changepoints[17]
- If data points are independent but there is distribution drift,
  can bound cumulative error[18]
- If data points form a time series,
  some standard assumptions ensure asymptotic coverage[19]

[17] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[18] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[19] Xu & Xie 2021, *Conformal prediction interval for dynamic time-series*

Looking ahead...

Lecture 2 will discuss:

- Limitations of the predictive validity guarantee
- Challenges for inference on problems other than prediction
- Is "distribution free" the right framework?