An introduction to conformal prediction and distribution-free inference

(Lecture 1)

Rina Foygel Barber (University of Chicago)

Columbia Statistics short course May/June 2024

http://rinafb.github.io/

Topics — Lecture 1

Introduction

- Motivation for distribution-free inference
- Data-driven methods: holdout sets & CV

Conformal prediction theory & methodology

- Intro to exchangeability
- Split conformal prediction

Summary

- Summary & preview

Introduction

Motivation for distribution-free inference

Regression & prediction

Supervised learning setting:

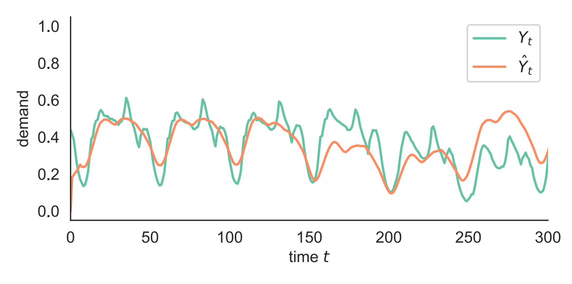
$$\text{Training data } (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$$

Goals:

- Inference on the regression — model distribution of Y given X
- Predictive inference — predict value of Y given X
  for test points $(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \ldots$

Regression & prediction

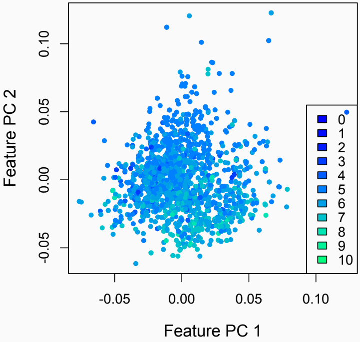
Example: the `Elec` data set[1] —
tracking electricity demand in Australia

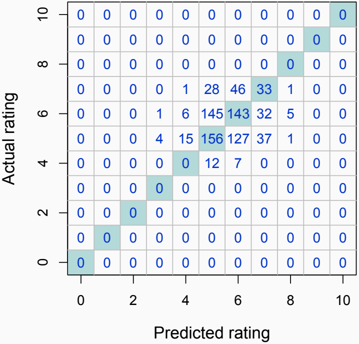


[1] Harries 1999, *Splice-2 comparative evaluation: Electricity pricing*

Regression & prediction

Example: the `Wine Quality` data set[2]





[2] Cortez et al 2009, `Wine Quality` data set, UCI Machine Learning Repository

Regression & prediction — classical approach

- We assume a parametric model on (X, Y) or on $Y \mid X$

  e.g., for linear regression, $Y = X^\top \beta + \mathcal{N}(0, \sigma^2)$

- We perform estimation & inference on the parameters....

  e.g., for linear regression, distribution of $\widehat{\beta}$ and $\widehat{\sigma}^2$

- ....& then we can provide prediction intervals:

  e.g., for linear regression, $Y_{n+1} \in X_{n+1}^\top \widehat{\beta} \pm \ldots\ldots.$

Regression & prediction — nonparametric approach

- We allow a nonparametric model for (X, Y) or $Y \mid X$,
  with assumptions/constraints

  e.g., assume $\mathbb{E}[Y \mid X]$ is smooth

- We perform estimation & inference on the model....

  e.g., $\widehat{\mu}(x) =$ estimate of $\mathbb{E}[Y \mid X = x]$, via a Gaussian kernel

- ....& then we can provide prediction intervals:

  e.g., $Y_{n+1} \in \widehat{\mu}(X_{n+1}) \pm \ldots\ldots$

Regression & prediction — ML approach

- Train an overparametrized model for $Y \mid X$

  e.g., train a neural net on $\{(X_i, Y_i)\}$

- Provide predictions for new feature vectors

  e.g., $\widehat{Y}_{n+i}$ = neural net's prediction for feature X_{n+i}

- Use a data-driven strategy for uncertainty quantification

  e.g., holdout data / cross-validation / bootstrapping / etc

Regression & prediction — challenges

What can go wrong?

- For the parametric approach — our model may be wrong
- For the nonparametric approach — our assumptions (e.g., smoothness) may not hold
- For the ML approach — is data-driven inference guaranteed to give valid answers?

Regression & prediction — challenges

Our choices:

- Rely on assumptions being correct
- Or, test empirically whether our assumptions hold
- Or, use inference methods that don't rely on assumptions (or, only rely on weaker assumptions)

Introduction

Data-driven methods: holdout sets & CV

Regression & prediction — data-driven predictive inference

Setting:

- Features $X \in \mathcal{X}$, response $Y \in \mathbb{R}$ (or $Y \in \mathcal{Y}$)
- Available training data $(X_1, Y_1), \ldots, (X_n, Y_n) \rightsquigarrow$ fit model $\widehat{\mu}$
- Goal: given $X_{n+1}, X_{n+2}, \ldots$, predict $Y_{n+1}, Y_{n+2}, \ldots$

Prediction? or predictive inference?

$\widehat{Y}_{n+i} = \widehat{\mu}(X_{n+i})$ $\qquad$ $Y_{n+i} \in \widehat{\mu}(X_{n+i}) \pm \underbrace{\text{(margin of error)}}_{\text{how to calculate?}}$

Using the training set for inference

Using the training loss:

If fitted model $\widehat{\mu}$ overfits to training data, generally

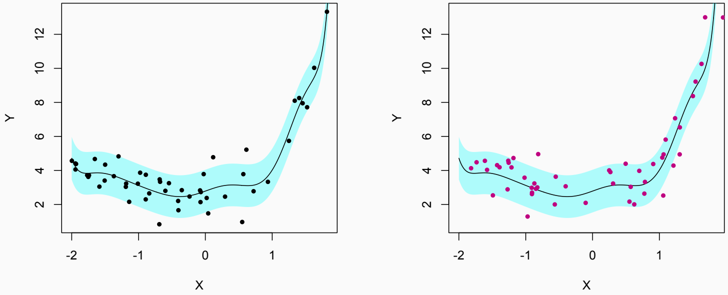
$$\underbrace{|Y_{n+i} - \widehat{\mu}(X_{n+i})|}_{\text{test error}} \gg \underbrace{\frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{\mu}(X_i)|}_{\text{avg. training error}}$$

even if training & test data are from the same distribution

Regression & prediction — data-driven predictive inference

Simulation: suppose we construct prediction intervals as

$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \mathrm{Quantile}_{1-\alpha}\underbrace{(|Y_1 - \widehat{\mu}(X_1)|, \ldots, |Y_n - \widehat{\mu}(X_n)|)}_{\text{residuals on training data}}$$



Train: 90% coverage

Test: 78% coverage

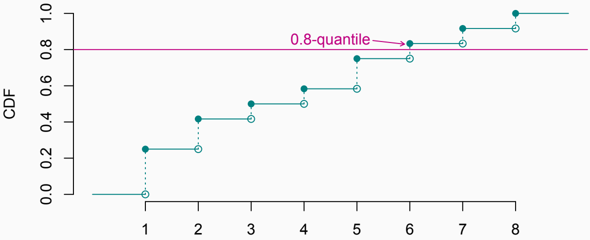
Using the training set for inference

Definition: quantile

Let $\tau \in [0,1]$. The τ-quantile of the list $z_1, \ldots, z_m \in \mathbb{R}$ is

$$\text{Quantile}_\tau(z_1, \ldots, z_m) = \inf\left\{ t \in \mathbb{R} : \underbrace{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{z_i \le t}}_{\text{what fraction of values are } \le t?} \ge \tau \right\}$$

Data: 1 1 1 2 2 3 4 5 5 6 7 8



Using a holdout set for inference

Holdout set approach—split the training data, $n = n_0 + n_1$

- Fit model $\widehat{\mu}$ on pretraining set $\{(X_i, Y_i)\}_{1 \le i \le n_0}$
- Compute residuals on calibration set, $\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \le n}$
- Prediction interval:

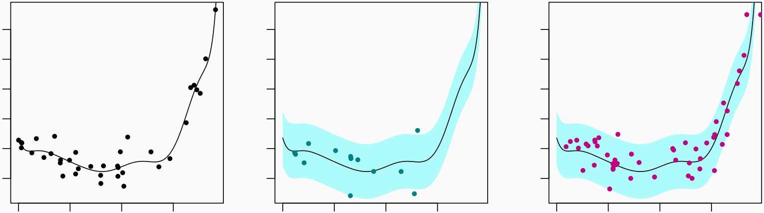
$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \text{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \le n}\big)$$

fitted on pretraining data

computed on calibration data

Using a holdout set for inference

Simulation:



Pretrain | Calibration: 90% coverage | Test: 89% coverage

Note: lower sample size $\rightsquigarrow$ $\widehat{\mu}$ is less accurate $\rightsquigarrow$ intervals are wider

Using a holdout set for inference

- The naive method fits a more accurate $\widehat{\mu}$,
  but the margin of error is too small due to overfitting
- The holdout set method fits a less accurate $\widehat{\mu}$,
  but the margin of error is correctly calibrated
- Can we use cross-validation (CV) to get the best of both?

Using cross-validation for inference

Leave-one-out CV (the "jackknife"):

$$\mathcal{C}(X_{n+j}) = \widehat{\mu}(X_{n+j}) \pm \text{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}_{-i}(X_i)|\}_{i=1,\ldots,n}\big)$$

fitted on all data

$\widehat{\mu}_{-i}$ fitted on $\{(X_\ell, Y_\ell)\}_{\ell\neq i}$
(the leave-one-out model)

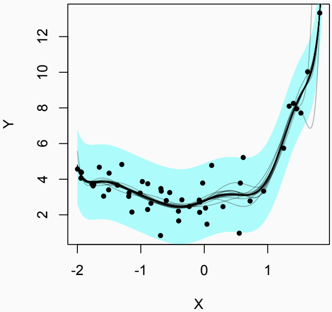
More computationally efficient: K-fold CV

- Partition $\{1,\ldots,n\} = A_1 \cup \cdots \cup A_K$, with $|A_k| = n/K$
- Fit models $\widehat{\mu}_{-A_k}$ to data $\{(X_i, Y_i)\}_{i\notin A_k}$
- Compute the margin of error using residuals

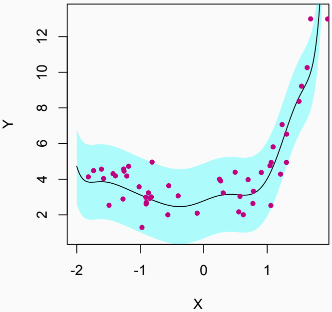
$$\big\{|Y_i - \widehat{\mu}_{-A_k}(X_i)|\big\}_{k=1,\ldots,K;i\in A_k}$$

Using cross-validation for inference

Leave-one-out CV: simulation



Train: 90% coverage

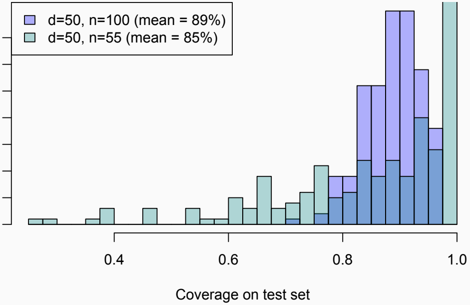


Test: 92% coverage

Using cross-validation for inference

However, no assumption-free theory for CV...

Example: least squares regression + leave-one-out CV



- Theoretical guarantees under asymptotic settings
- Unstable models may lead to undercoverage

Distribution-free prediction: aims

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

For the problem of predictive inference...

- Available training data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Given test points $X_{n+1}, X_{n+2}, \ldots$, predict $Y_{n+1}, Y_{n+2}, \ldots$

- Can we construct a prediction interval $\mathcal{C}(X_{n+i}) \subseteq \mathcal{Y}$ such that

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

Distribution-free prediction: aims

Goal: $\mathbb{P}\{Y_{n+i} \in \mathcal{C}(X_{n+i})\} \geq 1 - \alpha$ for test points (X_{n+i}, Y_{n+i})

- Avoid overly conservative or trivial solutions ($\mathcal{C}(X_{n+1}) = \mathcal{Y}$)
- Want to be able to use *any* regression method to construct $\mathcal{C}$ (classical or ML methods)

Conformal prediction theory & methodology

Intro to exchangeability

Introduction to exchangeability

For the rest of this talk: let $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

The i.i.d. data setting

Assume $\underbrace{Z_1, \dots, Z_n}_{\text{training}}, \underbrace{Z_{n+1}, Z_{n+2}, \dots}_{\text{test}}$ are i.i.d. from some distrib. P

Can we call this "distribution-free"?

- No assumptions on P (e.g., P does not need to be smooth)
- But, this does not allow for dependence across time / distribution shift / etc
- We will return to these settings later

Introduction to exchangeability

Exchangeability

Random variables $Z_1, \ldots, Z_n$ are exchangeable if

$$(Z_1, \ldots, Z_n) \stackrel{\mathrm{d}}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})$$

for every permutation $\sigma \in \mathcal{S}_n$.

- An important special case:
  if $Z_1, \ldots, Z_n \stackrel{\text{iid}}{\sim} P$ then $Z_1, \ldots, Z_n$ are exchangeable

Exchangeability for an infinite sequence

Random variables $Z_1, Z_2, \ldots$ are exchangeable
if $Z_1, \ldots, Z_n$ are exchangeable for each $n \geq 1$.

Introduction to exchangeability

Exchangeable but not i.i.d.—examples:

1. Sampling without replacement:

   $Z_1, \ldots, Z_n$ sampled uniformly without repl. from $\{z_1, \ldots, z_N\}$

2. A hierarchical model / latent variable model:

$$\begin{cases} \theta \sim P_0, \\ Z_1, \ldots, Z_n \mid \theta \overset{\text{iid}}{\sim} P(\cdot \mid \theta) \end{cases}$$

3. An equicorrelated model:

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & 1 \end{pmatrix} \right)$$

Useful alternative interpretations of exchangeability:

(a) If the n values are always unique—

- Condition on the unordered set $\{Z_1, \ldots, Z_n\}$
- Then $(Z_1, \ldots, Z_n)$ is a uniformly random perm. of this set

(b) If the data is real-valued (not necessarily unique)—

- Condition on the order statistics $Z_{(1)} \leq \cdots \leq Z_{(n)}$
- Then $(Z_1, \ldots, Z_n)$ is a uniformly random perm. of these values

Introduction to exchangeability

Useful alternative interpretations of exchangeability:

(c) More generally, we can condition on the empirical distribution

$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$$

δ_z = point mass on the value z

Then $(Z_1, \ldots, Z_n)$ is a uniformly random perm. of $\widehat{P}_n$

e.g., we observed 3 `A`'s and 1 `B` $\rightsquigarrow$ the data is equally likely to be
`A A A B` or `A A B A` or `A B A A` or `B A A A`

Introduction to exchangeability

The exchangeable data setting

Assume $\underbrace{Z_1, \ldots, Z_n}_{\text{training}}, \underbrace{Z_{n+1}, Z_{n+2}, \ldots}_{\text{test}}$ are exchangeable

- The i.i.d. data setting is a special case

Exchangeability & holdout data

Holdout set methods rely implicitly on exchangeability.

The data:

$$\underbrace{Z_1, \dots, Z_{n_0}}_{\substack{\text{pretraining} \\ \text{(fit model } \widehat{\mu}\text{)}}}, \underbrace{Z_{n_0+1}, \dots, Z_n}_{\substack{\text{calibration} \\ \text{(compute } \widehat{q}\text{)}}}, \underbrace{Z_{n+1}}_{\text{test}}$$

- Prediction intervals: for test point $n+1$,

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}$$

- Coverage?

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff \underbrace{|Y_{n+1} - \widehat{\mu}(X_{n+1})|}_{\text{residual on a test point}} \leq \underbrace{\widehat{q}}_{\substack{\text{residual quantile} \\ \text{from calibration data}}}$$

Exchangeability & holdout data

Suppose we have i.i.d. (or exchangeable) data

$$\underbrace{Z_1, \ldots, Z_{n_0}}_{\substack{\text{pretraining} \\ \text{(fit model } \widehat{\mu}\text{)}}}, \underbrace{Z_{n_0+1}, \ldots, Z_n}_{\substack{\text{calibration} \\ \text{(compute } \widehat{q}\text{)}}}, \underbrace{Z_{n+1}}_{\text{test}}$$

Exchangeability for holdout set methods

Let $\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_{n_0})$, for any regression algorithm $\mathcal{A}$.
Then the residuals

$$|Y_{n_0+1} - \widehat{\mu}(X_{n_0+1})|, \ldots, |Y_n - \widehat{\mu}(X_n)|, |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

are exchangeable.

Exchangeability & holdout data

Proof — the i.i.d. case:

Fact about exchangeability

If $Z_1, \ldots, Z_n$ are exchangeable and f is any function, then $f(Z_1), \ldots, f(Z_n)$ are exchangeable.

For the holdout set setting, since the data points are i.i.d., $|Y_{n_0+1} - \widehat{\mu}(X_{n_0+1})|, \ldots, |Y_{n+1} - \widehat{\mu}(X_{n+1})|$ are i.i.d. conditional on $\widehat{\mu}$.

And, conditionally-i.i.d. is a special case of exchangeability.

Proof — the general case:

Let σ be any permutation on $\{n_0+1, \ldots, n+1\}$.

We need to prove

$$R = (R_{n_0+1}, \ldots, R_{n+1}) \stackrel{\mathrm{d}}{=} (R_{\sigma(n_0+1)}, \ldots, R_{\sigma(n+1)}) = R_\sigma$$

Define a function $f : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}^{n_1+1}$ as

$$f\big(z_1, \ldots, z_{n+1}\big) = \Big(\big|y_j - \big[\mathcal{A}(z_1, \ldots, z_{n_0})\big](x_j)\big|\Big)_{n_0 < j \leq n+1}.$$

Then

$$R = f(Z_1, \ldots, Z_{n+1}) \stackrel{\mathrm{d}}{=} f(Z_1, \ldots, Z_{n_0}, Z_{\sigma(n_0+1)}, \ldots, Z_{\sigma(n+1)}) = R_\sigma$$

by exchangeability of the data $Z_1, \ldots, Z_{n+1}$.

Conformal prediction theory & methodology

Split conformal prediction

Split conformal prediction

The split conformal prediction method

1. Using pretraining data $Z_1, \ldots, Z_{n_0}$,
   construct fitted model $\widehat{\mu}$ using any regression algorithm:

$$\widehat{\mu} = \mathcal{A}\big(Z_1, \ldots, Z_{n_0}\big)$$

2. Compute quantile $\widehat{q}$ of calibration set residuals:

$$\widehat{q} = \mathrm{Quantile}_{(1-\alpha)(1+1/n_1)}\Big(\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \leq n}\Big)$$

3. For test point $n+1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}$$

Split conformal prediction

Theorem[3]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then split conformal satisfies:

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

Proof:

Define

$$R_i = |Y_i - \widehat{\mu}(X_i)|,$$

for each $i = n_0 + 1, \ldots, n, n+1$ (calibration & test data).

By construction of $\mathcal{C}(X_{n+1})$,

$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff R_{n+1} \leq \widehat{q} = \text{Quantile}_{(1-\alpha)(1+1/n_1)}\Big(\{R_i\}_{n_0 < i \leq n}\Big).$

[3] Vovk et al 2005, *Algorithmic Learning in a Random World*

Split conformal prediction

Fact about exchangeability

Let $\tau \in [0,1]$. If $Z_1, \ldots, Z_m \in \mathbb{R}$ are exchangeable,

$$\mathbb{P}\left\{Z_i \leq \text{Quantile}_\tau\Big(Z_1, \ldots, Z_m\Big)\right\} \geq \tau$$

From before, $R_{n_0+1}, \ldots, R_n, R_{n+1}$ are exchangeable. So,

$$\mathbb{P}\left\{\underbrace{R_{n+1} \leq \text{Quantile}_{1-\alpha}\Big(R_{n_0+1}, \ldots, R_n, R_{n+1}}_{\text{this event is the same as } R_{n+1} \leq \widehat{q}}\Big)\right\} \geq 1-\alpha.$$

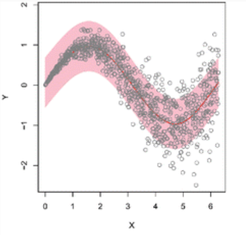
Fact about quantiles

Let $\tau \in [0,1]$. For any $z_1, \ldots, z_m, w \in \mathbb{R}$,

$$w \leq \text{Quantile}_\tau(z_1, \ldots, z_m, w) \;\Leftrightarrow\; w \leq \text{Quantile}_{\tau(1+1/m)}(z_1, \ldots, z_m)$$

The nonconformity score

Due to the construction of the split conformal method,
$\mathcal{C}(X_{n+1})$ has the same width regardless of the value of X_{n+1}



(figure from Lei et al 2018)

Why?

- $\mathcal{C}(X_{n+1}) = \left[\ \widehat{\mu}(X_{n+1}) \pm \widehat{q}\ \right] = \{y \in \mathbb{R} : |y - \widehat{\mu}(X_{n+1})| \leq \widehat{q}\}$
- Equivalently: we are using $|y - \widehat{\mu}(X_{n+1})|$ as a *score* to determine whether y is contained in $\mathcal{C}(X_{n+1})$ or not

The nonconformity score

The split conformal prediction method[4] (general score)

1. Using pretraining data $Z_1, \ldots, Z_{n_0}$,
   construct score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ using any algorithm
2. Compute quantile $\widehat{q}$ of calibration set scores:

$$\widehat{q} = \text{Quantile}_{(1-\alpha)(1+1/n_1)}\big(S_{n_0+1}, \ldots, S_n\big)$$

   where $S_i = s(X_i, Y_i)$
3. For test point $n+1$ return prediction interval

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

[4] Vovk et al 2005, *Algorithmic Learning in a Random World*

The nonconformity score

The residual score:

$$s(x, y) = |y - \widehat{\mu}(x)|, \text{ where } \widehat{\mu} \text{ fitted on pretraining data}$$

(Notation: $R_i \rightsquigarrow S_i$)

$$\implies \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}$$

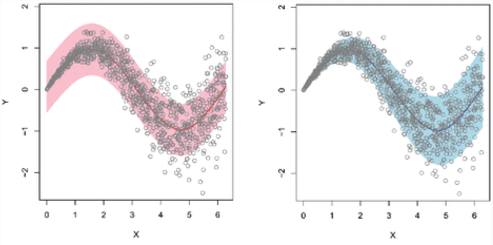
The nonconformity score

The rescaled residual score:[5]

$$s(x,y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)}, \text{ where } \widehat{\mu}, \widehat{\sigma} \text{ fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q} \cdot \widehat{\sigma}(X_{n+1})$$

Compare to residual score:



(figure from Lei et al 2018)

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

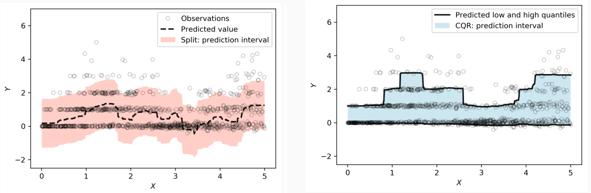
Conformalized quantile regression:[6]

$$s(x, y) = \max\{y - \widehat{\gamma}_{\text{hi}}(x), \widehat{\gamma}_{\text{lo}}(x) - y\}$$

where $\underbrace{\widehat{\gamma}_{\text{lo}},\ \widehat{\gamma}_{\text{hi}}}_{\text{estimated quantiles of } Y|X}$ fitted on pretraining data

$$\implies \mathcal{C}(X_{n+1}) = [\widehat{\gamma}_{\text{lo}}(X_{n+1}) - \widehat{q},\ \widehat{\gamma}_{\text{hi}}(X_{n+1}) + \widehat{q}]$$

Compare to residual score:



(figure from Romano et al 2019)

[6] Romano et al 2019, *Conformalized quantile regression*

The nonconformity score

Distributional conformal prediction:[7]

$$s(x,y) = \left|\widehat{F}(y|x) - 0.5\right| \text{ where } \underbrace{\widehat{F}(\cdot|x)}_{\substack{\text{estimated conditional CDF} \\ \text{of } Y \text{ given } X=x}} \text{ is fitted on pretraining data}$$

$$\Longrightarrow \mathcal{C}(X_{n+1}) = \left[\widehat{F}^{-1}(0.5 - \widehat{q} \mid X_{n+1}),\ \widehat{F}^{-1}(0.5 + \widehat{q} \mid X_{n+1})\right]$$

[7] Chernozhukov et al 2019, *Distributional conformal prediction*

The nonconformity score

Conditional density score:[8]

$$s(x,y) = -\widehat{f}(y|x) \text{ where } \underbrace{\widehat{f}(\cdot|x)}_{\substack{\text{estimated conditional density} \\ \text{of } Y \text{ given } X=x}} \text{ is fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \widehat{f}(y|X_{n+1}) \geq -\widehat{q} \right\}$$

[8] Izbicki et al 2020, *Flexible distribution-free conditional predictive bands using density estimators*

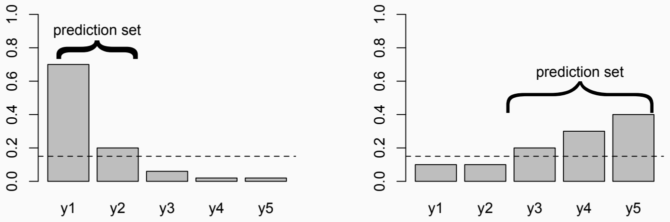
The nonconformity score — categorical data

If the response Y is categorical, with values $\mathcal{Y} = \{y_1, \ldots, y_K\}$—

The probability score:

$$s(x, y_k) = -\widehat{p}_k(x) \text{ where } \underbrace{\widehat{p}_k(x)}_{\text{estimates } \mathbb{P}\{Y = y_k \mid X = x\}} \text{ is fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \left\{y_k : \hat{p}_k(X_{n+1}) \geq -\widehat{q}\right\}$$

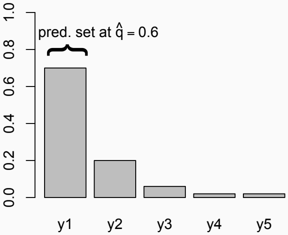


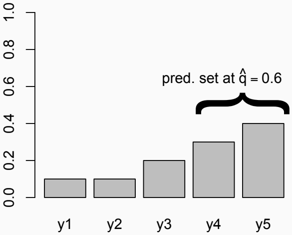
The nonconformity score — categorical data

A more efficient construction:[9] the cumulative probability score

$$s(x, y_k) = \sum_{k'} \hat{p}_{k'}(x) \cdot \mathbf{1}\{\hat{p}_{k'}(x) \geq \hat{p}_k(x)\}$$

Intuition: how far into the tail of the distribution is the label y_k?

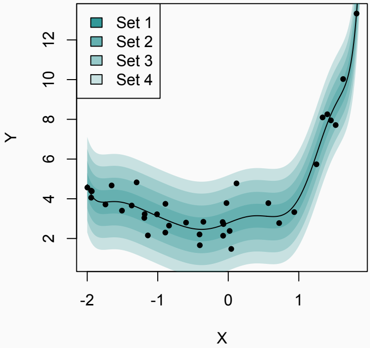




[9] Romano et al 2020, *Classification with Valid and Adaptive Coverage*; Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Split conformal prediction: summary

Another view of split conformal: nested sets[10]



- Use pretraining data to construct nested family of sets
- Use calibration data to choose one set within the family
- Flexibility to choose any score function

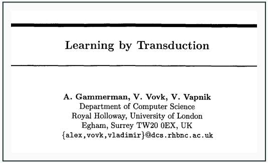
  $\rightsquigarrow$ shape of the set can adapt to the data distribution

[10] Gupta et al 2022, *Nested conformal prediction and quantile out-of-bag ensemble methods*

Conformal prediction: background

Background on the conformal prediction (CP) framework:

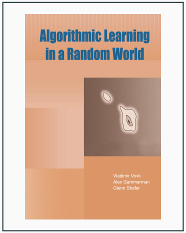
key idea = statistical inference via exchangeability of the data



Learning by Transduction

A. Gammerman, V. Vovk, V. Vapnik
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
{alex,vovk,vladimir}@dcs.rhbnc.ac.uk

Gammerman, Vovk, Vapnik
UAI 1998



Vovk, Gammerman, Shafer
2005 — see `alrw.net`

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1094–1111, Theory and Methods
https://doi.org/10.1080/01621459.2017.1307116

Taylor & Francis
Taylor & Francis Group

Check for updates

Distribution-Free Predictive Inference for Regression

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT
We develop a general framework for distribution-free predictive inference in regression, using conformal inference. The proposed methodology allows for the construction of a prediction band for the response variable using any estimator of the regression function. The resulting prediction band preserves the consistency properties of the original estimator under standard assumptions, while guaranteeing finite-sample marginal coverage even when these assumptions do not hold. We analyze and compare, both empirically and theoretically, the two major variants of our conformal framework: full conformal inference and split conformal inference, along with a related jackknife method. These methods offer different tradeoffs between statistical accuracy (length of resulting prediction intervals) and computational efficiency. As extensions, we develop a method for constructing valid in-sample prediction intervals called *rank-one-out* conformal inference, which has essentially the same computational efficiency as split conformal inference. We also describe an extension of our procedures for producing prediction bands with locally varying length, to adapt to heteroscedasticity in the data. Finally, we propose a model-free notion of variable importance, called *leave-one-covariate-out* or LOCO inference. Accompanying this article is an R package `conformalInference` that implements all of the proposals we have introduced. In the spirit of reproducibility, all of our empirical results can also be easily (re)generated using this package.

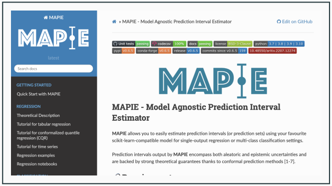
ARTICLE HISTORY
Received April 2016
Revised February 2017

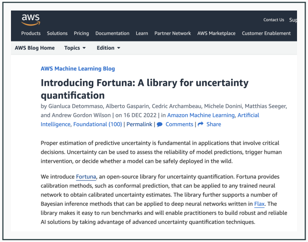
KEYWORDS
Distribution-free; Model misspecification; Prediction band; Regression; Variable importance

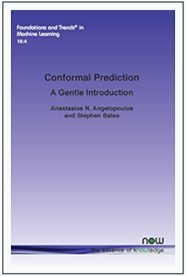
Lei, G'Sell, Rinaldo,
Tibshirani, Wasserman
JASA 2018

Conformal prediction: background

Recent developments — software packages & user-friendly tutorials

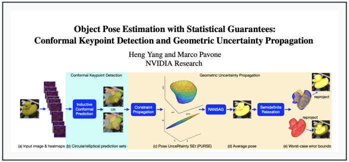


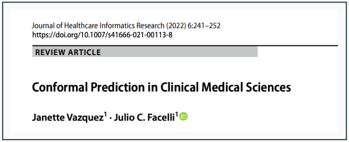


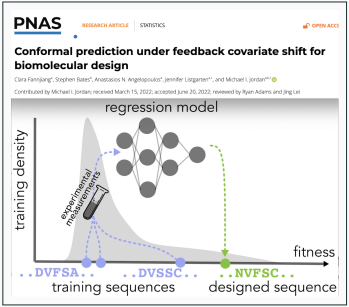


Conformal prediction: background

Recent developments — successful applications in biological sciences, machine learning, & many more domains







Summary & preview

Summary: lecture 1

Lecture 1 topics:

- Detailed intro to exchangeability
- Connection between exchangeability & holdout set methods:

  Pretrained model $\widehat{\mu} \rightsquigarrow$ exchangeable holdout & test residuals

Split CP allows us to start with any pretrained algorithm/model,
and then calibrate it to have valid predictive coverage
(as long as we can assume exchangeability!)

Preview: lecture 2

- The holdout set method (i.e., split CP) has DF guarantees,
  but model $\widehat{\mu}$/score s is less accurate due to data splitting
- And, CV type methods tend to work in practice
  but can fail in unstable settings / no distrib.-free theory

- Next we will cover *full conformal prediction*
  (and *cross-conformal prediction* in lecture 3)