Discussion of “Conformal Prediction in 2022” by Emmanuel Candès

Davide Risso - University of Padova
Aldo Solari - University of Milano-Bicocca

Statistical methods and models for complex data
800 years of research to understand a complex world
September 21, 2022

Conformal Prediction

General framework for constructing prediction sets $\hat{C}_n$ with

1. Finite-sample coverage guarantee (exact)
2. For any data distribution (distribution-free)
3. For any predictive model (model-free)

$$\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} = 1 - \alpha$$

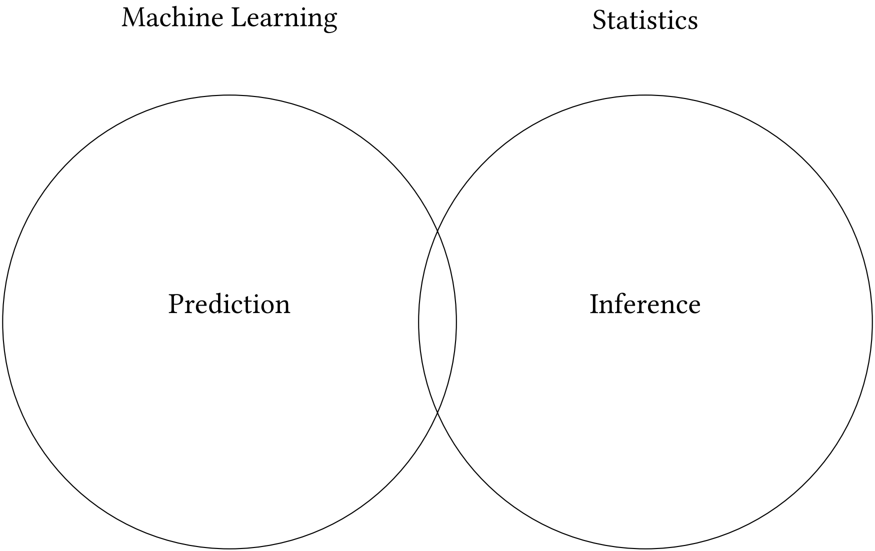
In his talk, Prof. Candès addresses the two main limitations:

1. Marginal coverage $\leftarrow$ Conformalized Quantile Regression
2. Exchangeability assumption $\leftarrow$ Adaptive Conformal Inference

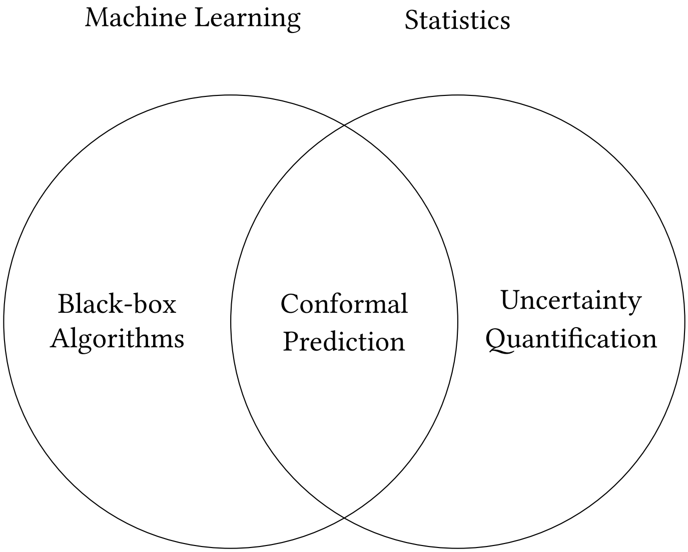
- Conformal inference for online prediction with arbitrary distribution shifts. 2022
- Conformal prediction beyond exchangeability. 2022
- Sensitivity analysis of individual treatment effects: A robust conformal inference approach. 2021
- Learn then test: Calibrating predictive algorithms to achieve risk control. 2021
- Adaptive conformal inference under distribution shift. 2021.
- Testing for outliers with conformal p-values. 2021.
- Conformalized survival analysis. 2021.
- Distribution-free conditional median inference. 2021
- Conformal inference of counterfactuals and individual treatment effects. 2020
- Achieving equalized odds by resampling sensitive attributes. 2020
- Classification with valid and adaptive coverage. 2020
- A comparison of some conformal quantile regression methods. 2019
- With malice toward none: Assessing uncertainty via equalized coverage. 2019
- Predictive inference with the jackknife+. 2019
- Conformalized quantile regression. 2019
- Conformal prediction under covariate shift. 2019
- The limits of distribution-free conditional predictive inference. 2019

Table of Contents

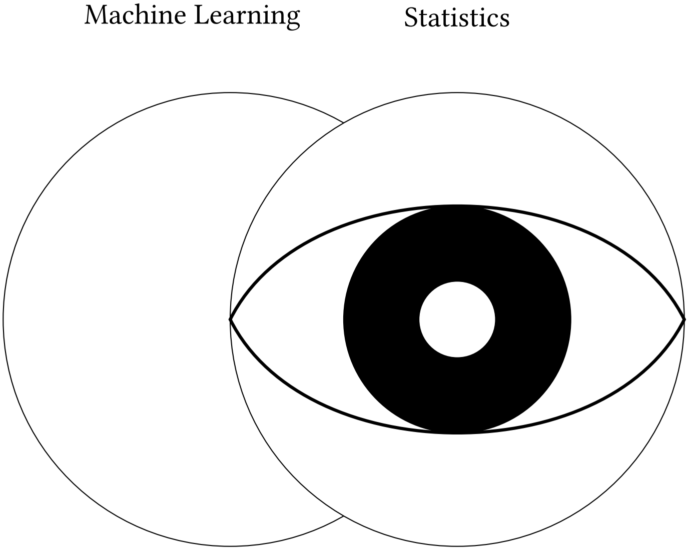
2001



2022


Machine Learning
Statistics
Black-box Algorithms
Conformal Prediction
Uncertainty Quantification

Statistical Point of View



Nonparametric Statistics

- Machine Learning has strong historical roots in Nonparametric Statistics
- K-Nearest Neighbors was introduced by two statisticians (students of Jerzy Neyman), Evelyn Fix and Joseph Hodges (Fix and Hodges, 1951)
- Conformal Prediction turns out to have roots in Permutation Testing (Fisher, 1925; Pesarin, 2001; Efron, 2021)

Prediction interval for Y_{n+1} (Vovk et al., 2005)	Confidence interval for Δ (Lehmann, 1963)
Supervised learning Training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ Test point (X_{n+1}, Y_{n+1})	Two-sample location shift model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F(x)$ $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} F(y - \Delta)$
$H_y : Y_{n+1} = y$	$H_d : \Delta = d$
$(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y)$	$x_1, \ldots, x_n, y_1 - d, \ldots, y_m - d$
$\hat{C} = \{y : p_y^* > \alpha\}$	$\hat{C} = \{d : p_d^* > \alpha\}$

Statistical Methods for Research Workers.
Fisher R.A.
Oliver & Boyd, 1925.

Nonparametric Confidence Intervals for a Shift Parameter.
Lehmann E.L.
The Annals of Mathematical Statistics, 1963, 34:1507–1512.

Multivariate Permutation Tests: with Applications in Biostatistics.
Pesarin F.
Wiley, 2001.

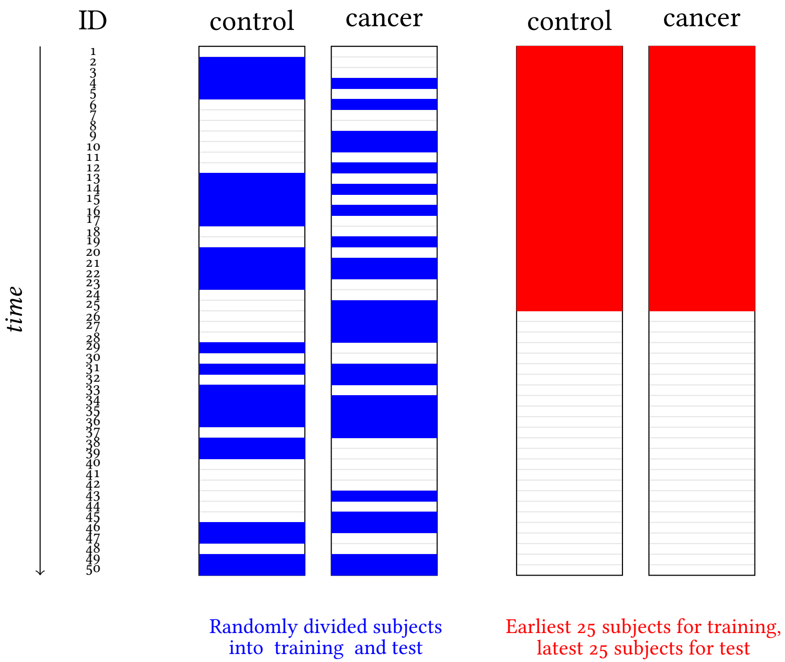
Resampling Plans and the Estimation of Prediction Error.
Efron B.
Stats, 2021, 4:1091–1115.

Table of Contents

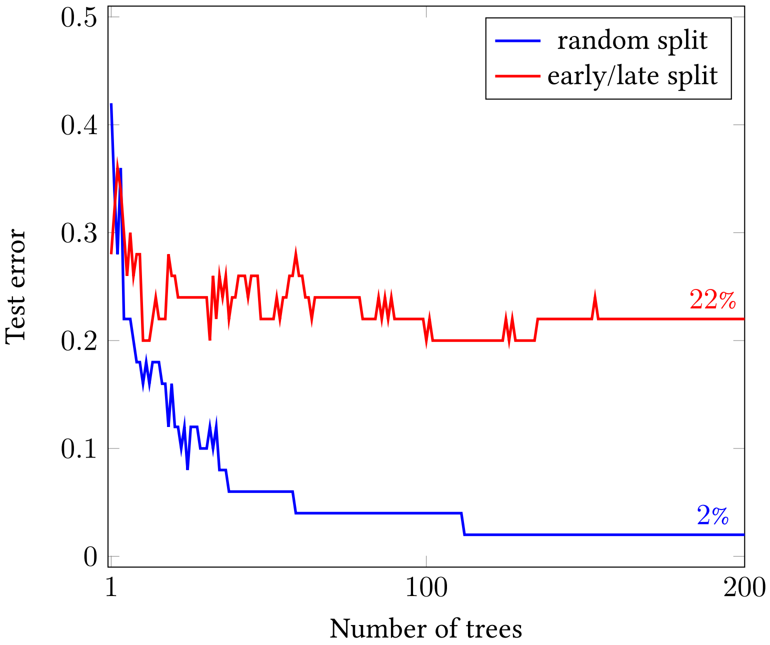
The Prostate Cancer Microarray Study

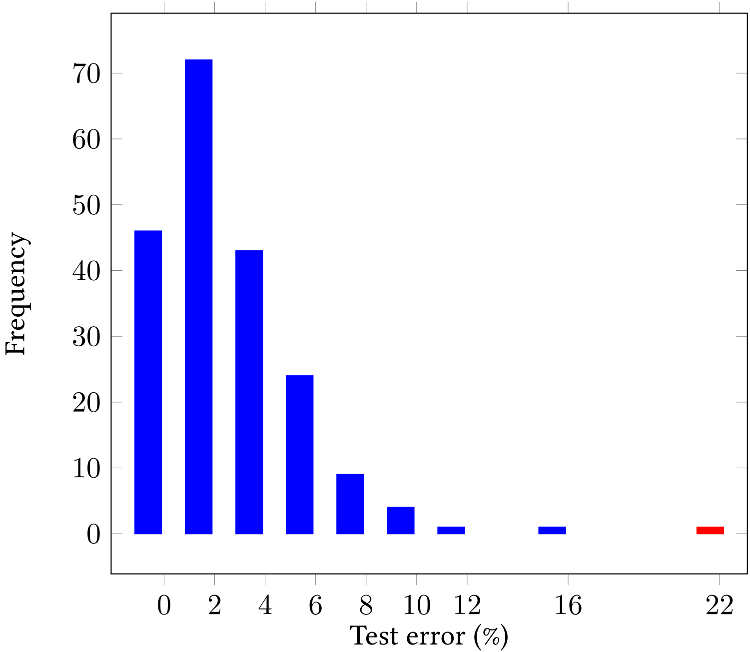
- $n = 100$ men: 50 prostate cancer, 50 normal controls
- For each man measure activity of $p = 6033$ genes

- Split into Training / Test (Efron, 2020; Candès and Sabatti; 2020)
  - At random
  - According to the time of entry in the study



ID
control
cancer
control
cancer
time
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
Randomly divided subjects into training and test
Earliest 25 subjects for training, latest 25 subjects for test


random split
early/late split
22%
2%
Test error
Number of trees
0
0.1
0.2
0.3
0.4
0.5
1
100
200


70
60
50
40
30
20
10
0
Frequency
0
2
4
6
8
10
12
16
22
Test error (%)

There Are Known Unknowns ...

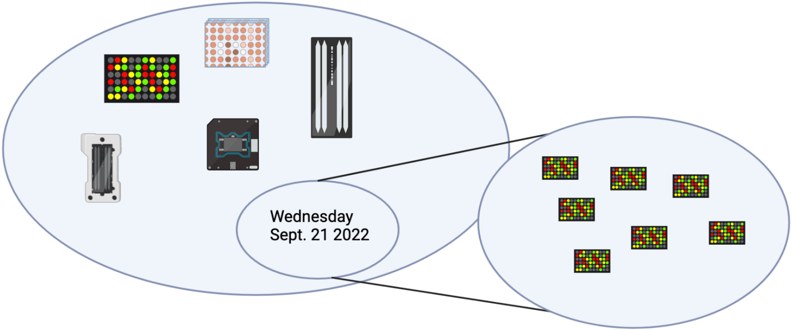
Is time a proxy for hidden technical confounders?

- (X_i, Y_i) assumed exchangeable.
- observe $X_i + \Delta_i$ rather than X_i.

Many approaches to estimate and account for latent confounders in this context (Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Lopez et al., 2018).

…And Unknown Unknowns

What if all the samples are processed in the same “batch”?



See also “cross-study validation” (Bernau et al., 2014).

Table of Contents

Questions

1. Stochastic vs. Epistemic Uncertainty
   - By focusing on the former do we create an illusion of confidence and risk to neglect the latter?

2. Measuring vs. Communicating Uncertainty
   - What is the impact of conveying uncertainty effectively, like in the Washington Post's example?

3. Learners vs. Learned
   - "In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists." - Eric Hoffer (Hand, 2006).

Prediction, Estimation, and Attribution.
Efron B.
Journal of the American Statistical Association, 2020, 115:636–655.

Classifier Technology and the Illusion of Progress.
Hand D.
Statistical Science, 2006, 21:1–15.

Capturing heterogeneity in gene expression studies by surrogate variable analysis.
Leek J. T. and Storey J. D.
PLOS Genetics, 2007, 3:e161.

Using control genes to correct for unwanted variation in microarray data.
Gagnon-Bartsch J. A. and Speed T. P.
Biostatistics, 2012, 13:539–552.

Deep generative modeling for single-cell transcriptomics.
Lopez R., Regier J., Cole M. B., Jordan M. I., Yosef N.
Nature Methods, 2018, 15:1053–1058.

Cross-study validation for the assessment of prediction algorithms.
Bernau C., Riester M., Boulesteix A., Parmigiani G., Huttenhower C., Waldron L., Trippa L.
Bioinformatics, 2014, 30:i105–i112.