STATS 300C: Theory of Statistics **Spring 2022**

Lecture 19 — June 2, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Joachim Sasson*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by Zhaomeng Chen and Disha Ghandwani.*

Outline

Agenda: Conformal Prediction

1. Enhanced nonconformity scores
2. A taste of counterfactual inference

19.1 Enhanced nonconformity scores

19.1.1 Review of Conformal Quantile Regression (CQR)

In Lecture 18, we described the method of Conformal Quantile Regression. There are two main motivations for this method. First, split conformal and CV+ methods return intervals of uniform width. Second, the intervals they return are based on an estimation of $y(x)$ rather than an estimation of the quantiles $q_{1-\alpha}(x)$, $q_\alpha(x)$. The CQR method enables to overcome these two problems by constructing prediction intervals whose length is adaptive (i.e the length depends on x) and that are based on quantile estimations. The CQR procedure to produce prediction intervals with coverage $1-\alpha$ is the following:

1. Split the data into a proper training set and a calibration set.
2. Fit two conditional quantile functions, $\text{lower}(x)$ and $\text{upper}(x)$, on the training set by minimizing a pinball loss (see Lecture 18).
3. For the ith point in the calibration set, compute

$$S_i = \max\{\text{lower}(X_i) - Y_i, Y_i - \text{upper}(X_i)\}$$

   Note that S_i is a signed distance to the boundary, where S_i is negative if $\text{lower}(X_i) \leq Y_i \leq \text{upper}(X_i)$ and positive otherwise.
4. Compute Q as the $(1-\alpha)$th quantile of the S_i's. Intuitively, Q is positive if the initial intervals are too narrow.

5. Define the prediction interval as

$$C(x) = [\text{lower}(x) - Q, \text{upper}(x) + Q]$$

As seen in the previous lecture, this method enjoys theoretical guarantees for coverage when the data are assumed to be exchangeable.

19.1.2 Calibration via adaptive coverage

What we have done so far is to leverage quantile regression to get estimates of the quantiles for each new observation x and move them up and down by the same amount Q. Though this approach is adaptive in the sense that it returns intervals whose length depends on x, we would like to expand our prediction intervals in a non-uniform way. One possibility is to calibrate the quantiles via adaptive coverage, as in [3].

1. Fit a model $\hat{F}_{Y|X}$ of $F_{Y|X}$ (for example via quantile regression) on the training set
2. For different values of $\tau \in [0, 1]$ with an arbitrary step size and for each point x in the calibration set, construct the naive intervals

$$C^{\text{naive}}(x, 1-\tau) = \left[\hat{F}^{-1}_{Y|X}\left(\frac{\tau}{2}\right), \hat{F}^{-1}_{Y|X}\left(1-\frac{\tau}{2}\right)\right]$$

   and measure the coverage $\text{cov}(\tau) := \frac{\#\left\{(X_i,Y_i)\in\text{calibration set s.t } Y_i\in C^{\text{naive}}(X_i,1-\tau)\right\}}{\#\{(X_i,Y_i)\in\text{calibration set}\}}$
3. Choose $\hat{\tau} := \inf\{\tau \in [0, 1] \text{ s.t } \text{cov}(\tau) \geq 1 - \alpha\}$
4. For a new observation x, the prediction interval is

$$C(x) = C^{\text{naive}}(x, 1 - \hat{\tau})$$

Remark: Similar to split conformal, we fit $\hat{F}_{Y|X}$ on the training set and then subsequently select $\hat{\tau}$ (i.e. we calibrate our model) based on the calibration data in order to avoid overfitting.

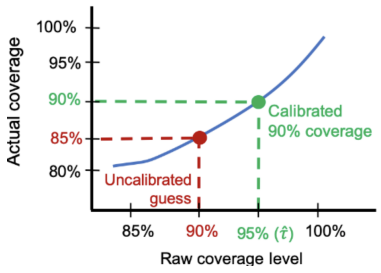

Figure 19.1. Calibration of $\hat{\tau}$

This method can easily be adapted to the classification task to build prediction sets, as in [4]. Let us denote by m the number of classes that we have. The previous algorithm can be adapted as follows to the classification framework.

1. Estimate the conditional distribution $\hat{\pi}(y|x)$ on the training set (for example as the output of a Neural Net with a softmax in its last layer)

2. As previously, for $\tau \in [0,1]$ with arbitrary step size and for all x in the calibration set, construct the naive sets as $C^{\text{naive}}(x, 1-\tau) = \left\{\text{classes for } \hat{\pi}_{(m)}(x), ..., \hat{\pi}_{(k(x))}(x)\right\}$ where $\hat{\pi}_{(1)}, \hat{\pi}_{(2)} \ldots, \hat{\pi}_{(m)}$ are the ranked $\hat{\pi}_i(x)$ and $k(x) := \inf\left\{1 \leq \ell \leq m \mid \sum_{j=\ell}^{m} \hat{\pi}_{(j)}(x) \geq 1-\tau\right\}$

3. The calibration of $\hat{\tau}$ is then exactly the same as in the regression case.

19.2 Counterfactual inference

19.2.1 Motivation

In a lot of randomized experiments, say for example in healthcare applications, patients are assigned either a control group (does not receive the treatment) or a treatment group depending of the value of their covariates (age, past diseases, lifestyle etc...). More precisely, we assign the treatment by a coin toss for each subject based on their **propensity score** $e(x)$.

$$\mathbb{P}(\text{treated}|x) = e(x)$$
$$\mathbb{P}(\text{control}|x) = 1 - e(x)$$

We place ourselves in the Neyman-Rubin causal framework where each subject has potential outcomes $(Y(1), Y(0))$ and the observed outcome Y^{obs}. More precisely, for each individual from the control or treatment group, we observe a response Y^{obs}. However, we do not know the **counterfactual responses** that concretely is: How would individuals from the control group respond to the treatment? Or more formally: How to infer $Y(1)$ for a control individual?

One important problem is that here the data for which we want to make inference (the control group) does not have the same distribution as the data for which we observe a response to the treatment since we recall that the treatment was assigned based on propensity scores $e(x)$. In other words, we are in a case of **covariate shift**. Figure 19.2 illustrates this situation. In the next section we are going to describe a way to adapt conformal inference to the above situation of distribution mismatch using **weighted conformal inference**.

19.2.2 Weighted conformal inference

The goal that we want to reach can be formalized as follows. We want to use i.i.d. (training) samples $(X_i, Y_i) \sim P_X \times P_{Y|X}$ to construct $\hat{C}(X)$ such that:

$$\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha \tag{19.1}$$

with $(X, Y) \sim Q_X \times P_{Y|X}$. In other words, we have that the distribution of the covariates X is different between the training and test sets. In this setting, we consider the likelihood

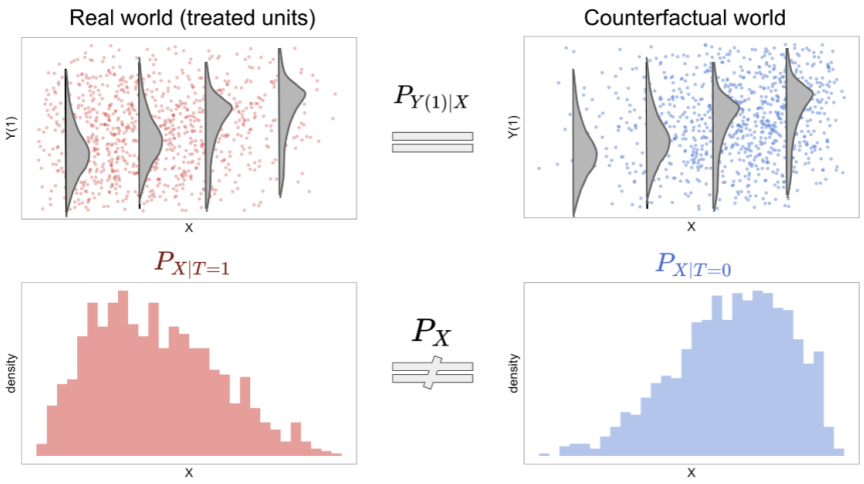

Figure 19.2. The counterfactual inference problem illustrated

ratio between measures Q and P that we define as $w(x) := \frac{dQ_X}{dP_X}(x)$. In the counterfactual inference framework we have

$$w(x) = \frac{dP_{X|T=0}}{dP_{X|T=1}}(x) \propto \frac{1-e(x)}{e(x)}$$

In this context, it can be shown that re-weighting the distribution of our conformity scores (the S_i's in 19.1.1) by the weights $w(X_i)$ enables building prediction intervals (or sets) that satisfy (19.1). The necessary framework is given by the following theorem whose proof is in [1].

Theorem 1. Assuming that Q is uniformly continuous with respect to P. For any conformity score function S and $\alpha \in [0,1]$. The set

$$\hat{C}_n(x) := \left\{ y \in \mathbb{R} \text{ s.t. } V_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha, \sum_{i=1}^{n} p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_\infty\right)\right\}$$

where $p_i^w(x) := \frac{w(X_i)}{\sum_{j=1}^n w(X_j)+w(x)}$ and $p_{n+1}^w(x) := \frac{w(x)}{\sum_{j=1}^n w(X_j)+w(x)}$, satisfies

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

To prove this theorem, we will first need to introduce the notion of weighted exchangeability which generalized the notion of exchangeability.

Definition 1. We call random variables $V_1, ..., V_n$ weighted exchangeable, with weight functions $w_1, ..., w_n$, if the density f of their joint distribution can be factorized as

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

where g is a symmetric function i.e. for any $\sigma \in \mathcal{S}_n$ we have $g(v_{\sigma(1)}, ..., v_{\sigma(1)}) = g(v_1, ...v_n)$

The concept of weighted exchangeability encompasses among others the case of covariate shift. This is the object of the following lemma.

Lemma 2. Let $Z_i \sim P_i$ for $1 \leq i \leq n$ be independent draws, where each P_i is absolutely continuous with respect to P_1, for $i \geq 2$. Then $Z_1, ..., Z_n$ are weighted exchangeable, with weight functions $w_1 := 1$, and $w_i = \frac{dP_i}{dP_1}$ for $i \geq 2$.

The proof of this lemma follows directly from definition 1. Now, let us give the key lemma needed to prove Theorem 1, this result and its proof can be found in [5].

Lemma 3. Let Z_i for $1 \leq i \leq n+1$ be weighted exchangeable with weight functions $w_1, ..., w_{n+1}$. Let $V_i := \mathcal{S}(Z_i, Z_{1:n+1}))$ for $1 \leq i \leq n$ and $\mathcal{S}$ an arbitrary nonconformity score function. Define for $1 \leq i \leq n+1$:

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

where the sums are taken over the permutations of $\{1, ..., n+1\}$. Then, for any $\alpha \in [0, 1]$

$$\mathbb{P}\left(V_{n+1} \leq \text{Quantile}\left(1-\alpha, \sum_{i=1}^{n} p_i^w(Z_1, ..., Z_{n+1})\delta_{V_i} + p_{n+1}^w(Z_1, ..., Z_{n+1})\delta_\infty\right)\right) \geq 1-\alpha$$

Proof. Denote by E_z the event that $\{Z_1, ..., Z_{n+1}\} = \{z_1, ..., z_{n+1}\}$, and denote $v_i = \mathcal{S}(z_i, z_{1:(n+1)})$ for $1 \leq i \leq n+1$. Let f denote the joint density of $Z_1, ..., Z_{n+1}$. Assuming that the V_i's are almost surely distinct for $1 \leq i \leq n+1$ we have:

$$\mathbb{P}(V_{n+1} = v_i \mid E_z) = \mathbb{P}(Z_{n+1} = z_i \mid E_z) = \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, ..., z_{\sigma(n+1)})}{\sum_\sigma f(z_{\sigma(1)}, ..., z_{\sigma(n+1)})}$$

Using the fact that $Z_1, ..., Z_{n+1}$ are weighted exchangeable we have

$$\begin{aligned}
\frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, ..., z_{\sigma(n+1)})}{\sum_\sigma f(z_{\sigma(1)}, ..., z_{\sigma(n+1)})} &= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_{\sigma(1)}, ..., z_{\sigma(n+1)})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_{\sigma(1)}, ..., z_{\sigma(n+1)})} && (19.2)\\
&= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_1, ..., z_{n+1})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_1, ..., z_{n+1})} && (19.3)\\
&= p_i^w(z_1, ..., z_{n+1}) && (19.4)
\end{aligned}$$

Therefore, the conditional distribution $V_{n+1} \mid E_z$ is characterized as follows:

$$V_{n+1} \mid E_z \sim \sum_{i=1}^{n+1} p_i^w(z_1, ..., z_{n+1})\delta_{v_i}$$

Therefore

$$\mathbb{P}\left(V_{n+1} \leq \text{Quantile}\left(1-\alpha, \sum_{i=1}^{n+1} p_i^w(z_1, ..., z_{n+1})\delta_{v_i}\right) \mid E_z\right) \geq 1-\alpha$$

Which is equivalent to

$$\mathbb{P}\left(V_{n+1} \leq \text{Quantile}\left(1-\alpha, \sum_{i=1}^{n+1} p_i^w(z_1, ..., z_{n+1})\delta_{V_i}\right) \mid E_z\right) \geq 1-\alpha$$

Which after marginalizing gives

$$\mathbb{P}\left(V_{n+1} \leq \text{Quantile}\left(1-\alpha, \sum_{i=1}^{n+1} p_i^w(z_1, ..., z_{n+1})\delta_{V_i}\right)\right) \geq 1-\alpha$$

□

From the above result, it is easy to prove 1. The following theorem that was proven in [2] is a specific version of 1 that applies to the problem of counterfactual inference i.e. with $w(x) = \frac{1-e(x)}{e(x)}$.

Theorem 4. Let $w(x) = \frac{1-e(x)}{e(x)}$, where $e(x)$ is known. Let $\hat{C}(x)$ denote the prediction sets obtained by weighted conformal inference. Then, the following inequality holds:

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

Further, if we assume our nonconformity scores are almost surely distinct and $w(X)$ is almost surely bounded (overlap condition) then:

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

where n denotes the size of the training set and C is a positive constant.

It is important to notice that for this result to hold we need to know the propensity score $e(x)$, which is not the case in real situation though we are able to have good approximations of it. In [2], the authors prove that this theorem approximately holds if we have a good estimate of either $e(x)$ or of the quantiles of $Y(1) \mid X$.

19.2.3 Simulation

In this section, we are going to present an example of the concepts presented above on simulated data. This example is a variant of [6]. The setting is as follows:

1. n samples of i.i.d Gaussian vectors in $\mathbb{R}^{100}$ (the distribution of these vectors has a general covariance matrix Σ)
2. $Y(1) \mid X \sim \mathcal{N}(\mu(X), \sigma(X)^2)$ where
   - μ is a smooth function of X_1, X_2
   - $\sigma(X) = 1$ (homoscedastic) or $\sigma(X) = -\log(1 - \phi(X_1))$ (heteroscedastic)
3. $0.25 \leq e(X) \leq 0.5$ and e is a smooth function of X_1

On figure 19.3, we can see that conformal prediction methods (top 3 rows) reach 95% coverage (red line) in both homoscedastic and heteroscedastic settings and perform far better than other methods widely discussed in the literature for this kind of problem.

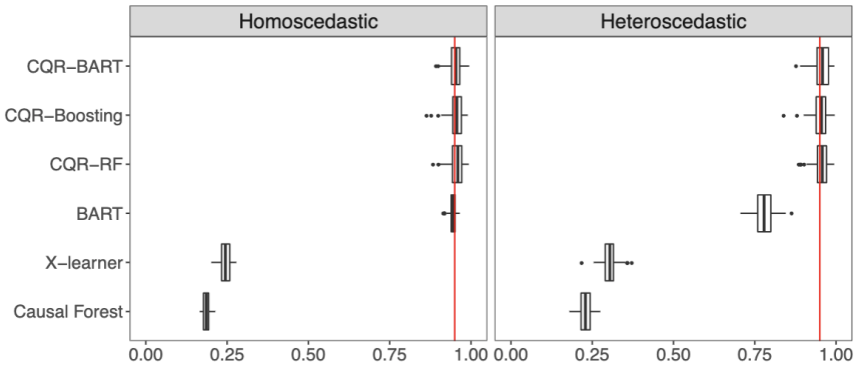

Figure 19.3. Coverage of prediction intervals associated to various models

Besides, as we can see on figure 19.4, Random Forest and Boosting seem to be the methods returning the tightest prediction intervals. We also notice that of course conformal prediction methods do not yield intervals with the length we would obtain if we knew perfectly the conditional distribution $Y(1) \mid X$ (blue line) but we see that we are not very far from it.

Finally, figure 19.5 shows the conditional coverage of $Y(1)$ as a function of x for each of the 6 methods considered. We can see that accross a range of values of x we have approximate 95% conditional coverage for conformalized methods (three right plots) while other classical methods are far from reaching this coverage.

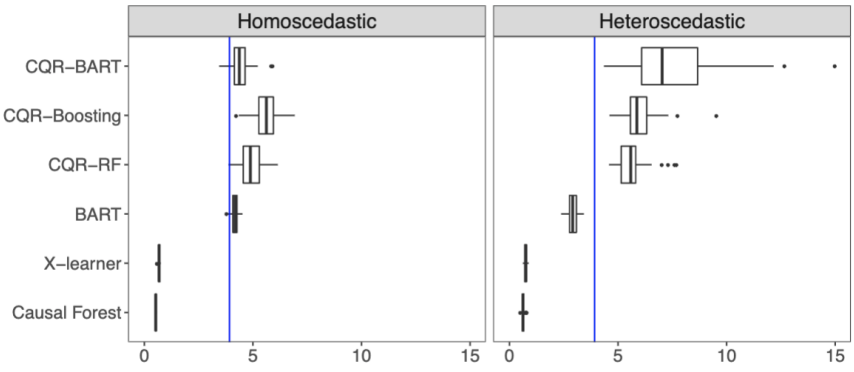

Figure 19.4. Length of prediction intervals associated to various models

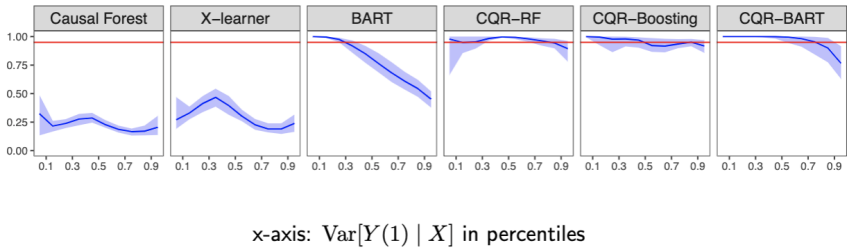

Figure 19.5. Conditional coverage of $Y(1)$

Bibliography

[1] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. 2021.

[2] Lihua Lei and Emmanuel Candes. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society*, 2020.

[3] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[4] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[5] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. pages 2526–2536, 2019.

[6] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.