STATS305A - Lecture 19

John Duchi
Scribed by Michael Howes

12/02/21

Contents

1 Announcements

Final homework and etude are avialable on the website.

2 Predictive inference

2.1 Setting and motivation

Given data $(X_i, Y_i)_{i=1}^n$ we would like to find a confidence set mapping $\widehat{C} : \mathcal{X} \to \ell$ where $\mathcal{X}$ is the space of X and $\ell = \{[a, b] : a \leq b\}$ is the set of all intervals in $\mathbb{R}$. We would want the confidence set mapping $\widehat{C}$ to have the property that if we have a new sample (X_{n+1}, Y_{n+1}), then $Y_{n+1} \in \widehat{C}(X_{n+1})$ some prescribed high probability i.e. above some threshold $1 - \alpha$.

There are two ways we can formalize this property:

(a) We could ask for *conditional coverage* (CC). That is we want a procedure $\widehat{C}$ that satisfies

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(x) | X_{n+1} = x) \geq 1 - \alpha,$$

for (almost) all $x \in \mathcal{X}$. Unfortunately this is not possible as shown by Vovk/Lei & Wasserman: If the procedure $\widehat{C}$ satisfies (CC) for all $\mathbb{P}$, then $\mathbb{E}[\text{length}(\widehat{C}(x))] = +\infty$ for almost all $x \in \mathcal{X}$ that are not atoms of $\mathbb{P}$.

(b) We can achieve *marginal coverage*. That is we can achieve CC on average over x. We will see a procedure $\widehat{C}$ which is a function of $(X_i, Y_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}$ such that

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

The procedure we will describe is distribution and model independent and very general. The procedure can be viewed as a protective wrapper around any black box model that will still guarantee coverage and validty. The basic idea is that we will construct confidence sets to evaluate the "weirdness" of new data Y_{n+1}.

3 Permutations, qunatiles and p-values

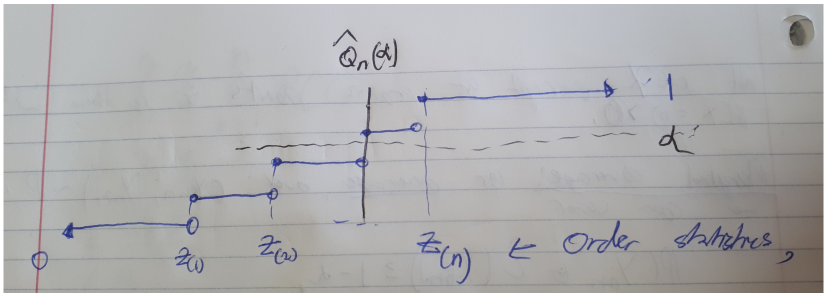
Suppose that $(Z_i)_{i=1}^n$ are exchangeable. That is

$$(Z_i)_{i=1}^n \overset{\text{dist}}{=} (Z_{\pi(i)})_{i=1}^n,$$

for all permuations π. Let $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$ be the order statistics of Z_i. Recall that we previously define the emperical CDF $F_n(t)$ to be the function

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Z_i \leq t).$$

The emperical CDF looks something like the below blue curve:



We also defined the quantile function $\widehat{Q}_n(\alpha)$ given by

$$\widehat{Q}_n(\alpha) = \inf\{t \in \mathbb{R} : F_n(t) \geq \alpha\} = Z_{(\lceil \alpha n \rceil)}.$$

The quantile function $\widehat{Q}_n$ is also illustrated in the above picture. When studying permutation tests we proved the below lemma.

Lemma 1. *If $(Z_i)_{i=1}^n$ are exchangeable, then*

$$\mathbb{P}(Z_n \leq \widehat{Q}_n(\alpha)) \geq \alpha.$$

Likewise, if Z_i are distinct with probablity 1, then

$$\mathbb{P}(Z_n \leq \widehat{Q}_n(\alpha)) \leq \alpha - \frac{1}{n}.$$

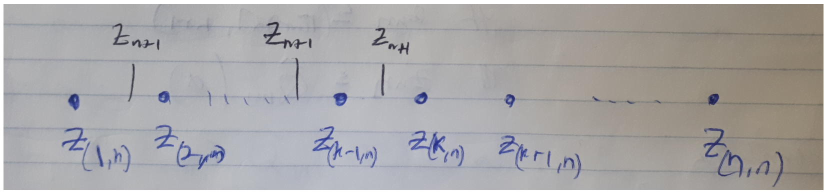
What is we want to add a new point Z_{n+1}. Suppose now that $(Z_i)_{i=1}^{n+1}$ are exchangeable. Can we use $(Z_i)_{i=1}^n$ to measure how "weird" Z_{n+1} is?

Notation 1. Let $Z_{(i,n)}$ be the order statistics of $(Z_i)_{i=1}^n$ and let $Z_{(i,n+1)}$ be the order statistics of $(Z_i)_{i=1}^{n+1}$.

Lemma 2. *Let $k \in \{1, \ldots, n\}$, then*

$$Z_{n+1} \leq Z_{(k,n)} \iff Z_{n+1} \leq Z_{(k,n+1)}.$$

Proof. First suppose that $Z_{n+1} \leq Z_{(k,n)}$. Our order statistics thus look something like this:



We will now consider two cases. If $Z_{n+1} \leq Z_{(k-1,n)}$, then $Z_{(k,n+1)} = Z_{(k-1,n)}$. On the other hand if $Z_{n+1} > Z_{(k-1,n)}$, then $Z_{(k,n+1)} = Z_{n+1}$. Thus we have

$$Z_{(k,n+1)} = \max\{Z_{n+1}, Z_{(k-1,n)}\}.$$

Therefore $Z_{n+1} \leq Z_{(k,n+1)}$.

Now conversely suppose that $Z_{n+1} \leq Z_{(k,n+1)}$. We always have $Z_{(k,n+1)} \leq Z_{(k,n)}$. We can thus conclude that $Z_{n+1} \leq Z_{(k,n)}$. □

Corollary 1. *Suppose $(Z_i)_{i=1}^{n+1}$ are exchangeable. For any $\alpha \in [0,1]$,*

$$\mathbb{P}\left(Z_{n+1} \leq \widehat{Q}_n\left(\frac{n+1}{n}\alpha\right)\right) \geq \alpha.$$

Furthermore if $(Z_i)_{i=1}^{n+1}$ are distinct with probability 1, then

$$\mathbb{P}\left(Z_{n+1} \leq \widehat{Q}_n\left(\frac{n+1}{n}\alpha\right)\right) \leq \alpha + \frac{1}{n+1}.$$

Proof. Note that

$$\widehat{Q}_n\left(\frac{n+1}{n}\alpha\right) = Z_{(\lceil n\frac{n+1}{n}\alpha\rceil, n)} = Z_{(\lceil (n+1)\alpha\rceil, n)}.$$

Thus

$$\begin{aligned} Z_{n+1} \leq \widehat{Q}_n\left(\frac{n+1}{n}\alpha\right) &\iff Z_{n+1} \leq Z_{(\lceil (n+1)\alpha\rceil, n)} \\ &\iff Z_{n+1} \leq Z_{(\lceil (n+1)\alpha\rceil, n+1)} \\ &\iff Z_{n+1} \leq \widehat{Q}_{n+1}(\alpha). \end{aligned}$$

Thus the result follows from Lemma 1. □

This corollary implies that

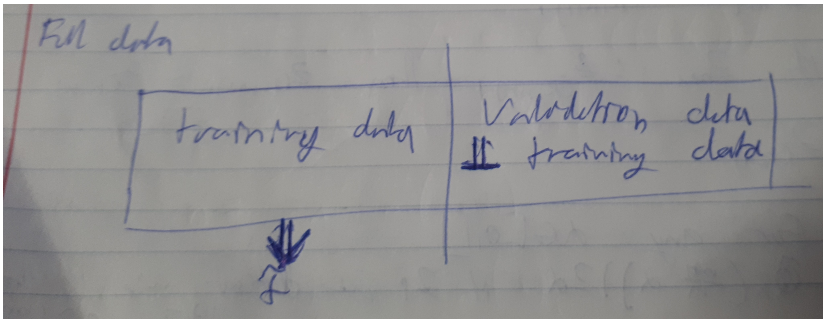
$$\mathbb{P}\left(Z_{n+1} > \widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right)\right) \leq \alpha,$$

and if $(Z_i)_{i=1}^{n+1}$ are distinct with probability one, then

$$\mathbb{P}\left(Z_{n+1} > \widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right)\right) \geq \alpha - \frac{1}{n+1}.$$

4 Split conformal inference

How can we use these quantile ideas to make confidence intervals? Suppose that we have a model $\widehat{f}:\mathcal{X}\to\mathbb{R}$ fitted on training data. Suppose also that we have validation data in the form of an i.i.d. sample $(X_i,Y_i)_{i=1}^n$ that is independent of $\widehat{f}$. In practice this means splitting our data like so:



We will use the validation set $(X_i,Y_i)_{i=1}^n$ to find a valid confidence set. For each i define $Z_i = \left|Y_i-\widehat{f}(X_i)\right|$ to be the *non-conformity score* of (X_i,Y_i). We can let $\widehat{Q}_n$ be the quantile function for $(Z_i)_{i=1}^n$. Then on a new data point (X_{n+1},Y_{n+1}) we will have

$$\mathbb{P}\left(Z_{n+1}>\widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right)\right)\le\alpha.$$

How do we transform this into a confidence set? For $x\in\mathcal{X}$ and $\tau\ge 0$ define

$$C_\tau(x):=[\widehat{f}(x)-\tau,\widehat{f}(x)+\tau].$$

Then

$$y\in C_\tau(x)\iff y\in[\widehat{f}(x)\pm\tau]\iff\left|\widehat{f}(x)-y\right|\le\tau.$$

For a given $\alpha\in[0,1]$, take

$$\widehat{\tau}_n=\widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right),$$

and define $\widehat{C}(x)=C_{\widehat{\tau}_n}(x)$.

Proposition 1. *With $\widehat{C}$ as above we have*

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

Furthermore if $Z_i=\left|Y_i-\widehat{f}(X_i)\right|$ are distinct with probability 1, then

$$\mathbb{P}(Y_{n+1}\in\widehat{C}(X_{n+1}))\le 1-\alpha+\frac{1}{n+1}.$$

Proof. We have already proven all of the main ingredients. Note that

$$\begin{aligned}Y_{n+1}\in\widehat{C}(X_{n+1})&\iff\left|\widehat{f}(X_{n+1})-Y_{n+1}\right|\le\widehat{\tau}_n\\&\iff Z_{n+1}\le\widehat{\tau}_n\\&\iff Z_{n+1}\le\widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right).\end{aligned}$$

Thus the result follows from Lemma 2. □

Remark 1. We are getting confience sets for Y_{n+1} not for $\mathbb{E}[Y_{n+1}|X_{n+1}]$ or anything like that. Note that the size of $\widehat{C}(X)$ depends on how well the model $\widehat{f}(X)$ performs on the independent validation set $(X_i,Y_i)_{i=1}^n$.

4.1 A general recipe

We could have chosen lots of things other that $\left|\widehat{f}(X_i) - Y_i\right|$ for our conformal score Z_i. Here is one very general approach for creating conformal scores.

Suppose that our Y now live in an arbitrary space $\mathcal{Y}$ and that we have a nested collection of confidence sets indexed by $\tau \in \mathbb{R}$. That is for each $\tau \in \mathbb{R}$ and $x \in \mathcal{X}$ we have a subset $C_\tau(x) \subseteq \mathcal{Y}$ such that for all $\tau_0 \leq \tau_1$ and all $x \in \mathcal{X}$ we have

$$C_{\tau_0}(x) \subseteq C_{\tau_1}(x).$$

Example 1. Our previous choice of $C_\tau(x)$ was $C_\tau(x) = [\widehat{f}(x) - \tau, \widehat{f}(x) + \tau]$ which are indeed nested.

We can use the collection $C_\tau(x)$ to define confidence scores Z_i by

$$Z_i = \inf\{t \in \mathbb{R} : Y_i \in C_t(X_i)\}.$$

Example 2. If $C_\tau(x) = [\widehat{f}(x) - \tau, \widehat{f}(x) + \tau]$, then

$$\inf\{t \in \mathbb{R} : Y_i \in C_t(X_i)\} = \left|Y_i - \widehat{f}(X_i)\right|,$$

which matches the previous section.

Define $\widehat{Q}_n$ on $(Z_i)_{i=1}^n$ and let $\widehat{\tau}_n = \widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right)$. Then, as before, define

$$\widehat{C}(x) := C_{\widehat{\tau}_n}(x).$$

Theorem 1. *With $\widehat{C}$ as above*

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

Proof. Note that

$$\begin{aligned} Y_{n+1} \in \widehat{C}(X_{n+1}) &\iff Y_{n+1} \in C_{\widehat{\tau}_n}(X_{n+1}) \\ &\iff Z_{n+1} \leq \widehat{\tau}_n \\ &\iff Z_{n+1} \leq \widehat{Q}_n\left(\frac{n+1}{n}(1-\alpha)\right) \end{aligned}$$

□

Example 3. If $\mathcal{Y}$ is a metric space with metric d, then we can define

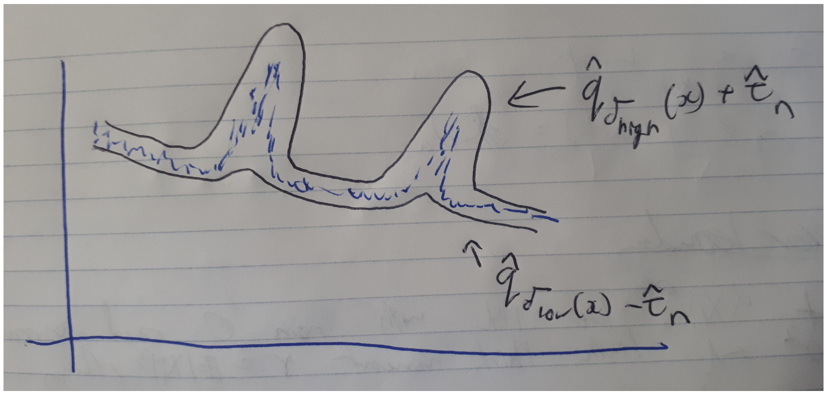
$$C_\tau(x) = \{y \in \mathcal{Y} : d(y, \widehat{f}(x)) \leq \tau\},$$

where $\widehat{f}$ is some predictor trained on data that is independent of $(X_i, Y_i)_{i=1}^n$. This lets us make conformal confidence sets for lots prediction problems (not just problems where the response is real valued).

Example 4 (Etude 4)**.** Suppose that we have a real response $Y \in \mathbb{R}$ and we use M-estimation to fit $\widehat{q}_{\delta_{\text{low}}}(x)$ and $\widehat{q}_{\delta_{\text{high}}}(x)$ which are meant to predict the δ_{low} and δ_{high} quantiles of Y. We can then do split conformal inference by using the confidence sets

$$C_\tau(x) = \left[\widehat{q}_{\delta_{\text{low}}}(x) - \tau, \widehat{q}_{\delta_{\text{high}}}(x) + \tau\right].$$

This gives us conformal confidence intervals which can be asymmetric and can vary in size with x. For example we could produce confidence intervals that look like the below image


$\leftarrow \hat{q}_{\sigma_{high}}(x) + \hat{\tau}_n$
$\uparrow \hat{q}_{\sigma_{low}}(x) - \hat{\tau}_n$

5 Class summary

People use models $Y = f(X) + \varepsilon$. If we make assumptions on f and ε, then we can prove analytic results. If we do not want to make assumptions, we have to do some sort of validation.