6.S951 Modern Mathematical Statistics Fall 2024

Lecture 15 — November 7, 2024

Prof. Stephen Bates

Scribe: Ashutosh Tripathi

1 Outline

Recap: Last time: Predictive Inference

1. Goals: CIs for test point
2. Challenge: training residuals too small
3. Linear model calculation
4. Started discussing Conformal Prediction

Agenda: Conformal Prediction

1. Residual Case
2. General case, conditional coverage

2 Conformal Prediction: Residual Case

Setting: Consider $(X_1, Y_1), \cdots (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P$, where $X_i \in \mathcal{X} = \mathbb{R}^n$ and $Y_i \in \mathcal{Y} = \mathbb{R}$.

Define the training data set $\mathcal{D} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$.

We also have a test data point $(X_{\text{test}}, Y_{\text{test}}) \sim P$, where Y_{test} is not observed.

High-Level Idea of Conformal Prediction: Consider $y \in \mathbb{R}$. We then test if

$$(X_{\text{test}}, y) \overset{d}{=} (X_i, Y_i)$$

by forming a test $\phi_y(X_{\text{test}}; \mathcal{D})$.

Then the confidence interval for Y_{test} is

$$\text{CI}(X_{\text{test}}) = \{y \mid \phi_y(X_{\text{test}}, \mathcal{D}) = 0\}.$$

Let $\hat{f}\left(x; \mathcal{D} \cup (X_{\text{test}}, y)\right)$ be a prediction of Y given $X = x$ based on training data $\mathcal{D} \cup (X_{\text{test}}, y)$.

Here, $\hat{f}$ can be anything, but must be symmetric in $\mathcal{D} \cup (X_{\text{test}}, y)$ (i.e. invariant to ordering) (†).

Given this, the algorithm for full conformal prediction is:

Algorithm 1 Full Conformal Prediction: Residual case

for $y \in \mathbb{R}$ **do**
  Calculate $r_i^y = Y_i - \hat{f}\left(X_i; \mathcal{D} \cup (X_{\text{test}}, y)\right)$
  Calculate $r_{\text{test}}^y = y - \hat{f}\left(X_{\text{test}}; \mathcal{D} \cup (X_{\text{test}}, y)\right)$
  Calculate

$$p^y = \frac{1 + \sum_{i=1}^n \mathbb{1}\{|r_{\text{test}}^y| \leq |r_i^y|\}}{n+1}$$

end for
Return $\text{CI}(X_{\text{test}}) = \{y \mid p^y > \alpha\}$

Theorem 1. *Consider $\hat{f}$, as defined earlier, which is symmetric (as in (†)). Also assume*

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{test}, Y_{test}) \overset{i.i.d.}{\sim} P.$$

Then Algorithm 1 works as intended, i.e.

$$P(Y_{test} \in CI(X_{test})) \geq 1 - \alpha.$$

Proof. By duality of CI and hypothesis tests, we note that $Y_{\text{test}} \in \text{CI}(X_{\text{test}}) \iff p^{Y_{\text{test}}} > \alpha$.

Since we are interested only in checking coverage guaranteed, it suffices to only check for Y_{test}, even though running the full conformal prediction requires us to train the model using values $y \neq Y_{\text{test}}$ as well. As a result, it suffices to show that

$$P(p^{Y_{\text{test}}} > \alpha) \geq 1 - \alpha,$$

i.e. $p^{Y_{\text{test}}}$ is super-uniform.

Now, condition on the **unordered data** $\mathcal{D} \bigcup \{(X_{\text{test}}, Y_{\text{test}})\}$, i.e. we condition on knowing the $n+1$ data points, but we don't know the ordering.

Critical Step: Notice that $|r_{\text{test}}^{\text{test}}|$ is a random draw from

$$\left\{|r_1^{\text{test}}|, \cdots, |r_n^{\text{test}}|, |r_{\text{test}}^{\text{test}}|\right\},$$

by symmetry. This step is rather subtle, and works as we are assuming all our $n+1$ points are i.i.d. As a result, the test residual can be thought of being randomly sampled from the set of all the residuals.

From definition, we note that

$$p^{Y_{\text{test}}} = \frac{1 + \sum_{i=1}^n \mathbb{1}\{|r_{\text{test}}^{Y_{\text{test}}}| \leq |r_i^{Y_{\text{test}}}|\}}{n+1}$$

We claim that this is super-uniform due to exchangeability. Indeed, this follows as, if we assume no ties, then $\sum_{i=1}^n \mathbb{1}\{|r_{\text{test}}^{Y_{\text{test}}}| \leq |r_i^{Y_{\text{test}}}|\} \sim \text{Unif}(\{0, \cdots, n\})$, and the p-value follows a discrete uniform distribution,

$$p^{Y_{\text{test}}} \sim \text{Unif}\left(\{\frac{1}{n+1}, \cdots, \frac{n}{n+1}, 1\}\right).$$

The case with ties follows easily as well. □

Remarks

- We don't have $(n+1)!$ due to the the inclusion of symmetry.
- We can simplify the algorithm by doing various forms of data-split
- The argument is rich: it can be extended beyond the i.i.d setup.
- This can be extened to online learning, outlier detection, image segmentation, etc.

3 Score Functions

We are going to replace the residual $|y - \hat{f}(X_{\text{test}}; \mathcal{D} \cup (X_{\text{test}}, y))|$ with a generic measurement of agreement.

Definition A **symmetric score function** $S((X_{\text{test}}, y); \mathcal{D})\colon (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}$ is a function that is invariant to ordering of $\mathcal{D}$, i.e

$$S((X_{\text{test}}, y); \mathcal{D}) = S((X_{\text{test}}, y); \mathcal{D}_\sigma)$$

Idea: Large value encodes worse agreement.

Example (Residual score) An example of residual score was something we just saw - the (training) residuals. In particular,

$$S^{\text{resid}}((X_{\text{test}}, y); \mathcal{D}) = |y - \hat{f}(X_{\text{test}}; \mathcal{D} \cup (X_{\text{test}}, y))|,$$

where $\hat{f}$ is symmetric as earlier.

Example (Leave one out score) Another example of a symmetric score function is leave one out,

$$S^{\text{loo}}((X_{\text{test}}, y); \mathcal{D}) = |y - \hat{f}(X_{\text{test}}; \mathcal{D})|,$$

which is a variation of residual score.

Example (Heteroskedastic score) Define the Heteroskedastic score function as

$$S^{\text{het}}((X_{\text{test}}, y); \mathcal{D}) = \frac{S^{\text{resid}}((X_{\text{test}}, y); \mathcal{D})}{\hat{\sigma}(X_{\text{test}})}.$$

Example (Quantile score) Say, $\hat{\tau}_{\alpha/2}(x;\mathcal{D}), \hat{\tau}_{1-\alpha/2}(x;\mathcal{D})$ are quantile estimates of $Y_{\text{test}} \mid X_{\text{test}} = x$. Then, the quantile regression score function is defined as

$$S^{\text{QR}}((X_{\text{test}}, y); \mathcal{D}) = \max\left(\hat{\tau}_{\alpha/2}(X_{\text{test}};\mathcal{D}) - y, y - \hat{\tau}_{1-\alpha/2}(X_{\text{test}};\mathcal{D})\right).$$

Note that the algorithm for the full conformal prediction is similar to the one earlier.

Algorithm 2 Full Conformal Prediction: Generic case

for $y \in \mathbb{R}$ **do**
  Calculate $r_i^y = S\left(X_i; (\mathcal{D} \setminus (X_i, Y_i)) \cup (X_{\text{test}}, y)\right)$
  Calculate $r_{\text{test}}^y = S\left(X_{\text{test}}; \mathcal{D}\right)$
  Calculate

$$p^y = \frac{1 + \sum_{i=1}^n \mathbb{1}\{r_{\text{test}}^y \leq r_i^y\}}{n+1}$$

end for
Return $\text{CI}(X_{\text{test}}) = \{y \mid p^y > \alpha\}$

Theorem 2. *If S is a symmetric score function. Also assume*

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{test}, Y_{test}) \overset{i.i.d.}{\sim} P.$$

Then Algorithm 2 works as intended, i.e.

$$P(Y_{test} \in CI(X_{test})) \geq 1 - \alpha.$$

Proof. The proof is analogous to that of Theorem 1. □

Key Takeaway: In general, the score function changes the shape.

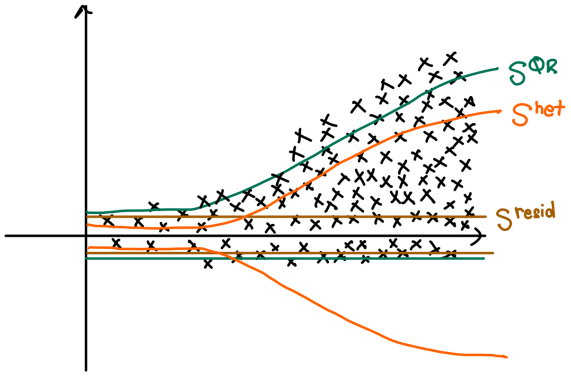

Figure 1: For the given dataset, we plot the CI bands for the different score functions $S^{\text{resid}}, S^{\text{het}}, S^{\text{QR}}$ as labeled in the plot. We note that compared to the residual and heteroskedastic score functions, the quantile score function seems to more closely fit the data, and thus give a more reasonable confidence interval.