Introduction to Conformal Prediction

Presenter: Aarshvi Gajjar

Motivation

- Most ML models are point predictors.
- These predictions can trigger important decisions so it becomes necessary to also report an uncertainty along with the predictions.
- For instance, after observing some data, we want to report a range of possible values such that for an unseen sample, the label lies in the range 90% of the time.
- If the interval is wide, then at least we know what we don't know about the prediction.

Goal

- Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be i.i.d. pairs of (features,labels) sampled from a distribution P on $\mathcal{X} \times \mathcal{Y}$.
- Let (X_{n+1}, Y_{n+1}) be a new independent observation from P.
- Fix an error level $\alpha \in (0, 1)$.
- The goal is to find a prediction band, $\hat{C}_n$ without any assumptions on P.

Definition (Prediction Band)

For a given $\alpha \in (0, 1)$, $\hat{C}_n$ is a map from $\mathcal{X}$ to subsets of $\mathcal{Y}$ such that for a new observation (X_{n+1}, Y_{n+1}),

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

- For example, if $\alpha = 0.1$, we want Y_{n+1} to belong to $\hat{C}_n(X_{n+1})$ w.p. 0.9.

Dumb method

- Set $\hat{C}_n = \mathcal{Y}$ always. Then, $\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} = 1$.
- But this does not give us any useful information.
- We want the prediction band to be as small as possible, while still giving us valid coverage results.

Conformal Prediction

- **Conformal Prediction**: This is a relatively new framework for converting point predictions into prediction sets with finite sample coverage.
- Classical linear regression prediction intervals are based on well specified model assumptions.
- First introduced by Vovk, Gammerman and Vapnik in 2005, *Algorithmic Learning in a Random World*.
- Later, in mid 2010s, it was repopularized and translated in less esoteric language by Lei, Wasserman, et. al. where they give a general framework for distribution free predictive inference.
- Can be applied to black box prediction methods.

First Key Idea: Rank based statistics

- Consider the simple setting of $Y_1, Y_2, \ldots Y_n$ sampled i.i.d. from some distribution on $\mathbb{R}$.
- Suppose we want to find a one sided interval, $(-\infty, \hat{q}_n]$, where $\hat{q}_n$ is a function of the data $\{Y_1, Y_2, \ldots, Y_n\}$, such that

$$\mathbb{P}\{Y_{n+1} \leq \hat{q}_n\} \geq 1 - \alpha.$$

- Since $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ is i.i.d, the *rank* of Y_{n+1} among $Y_1, Y_2, \ldots, Y_{n+1}$ is uniformly distributed.
- Suppose R_{n+1} is the rank of Y_{n+1}, or more precisely, $R_{n+1}(Y_{n+1}) = \left|j \in [n+1] : Y_j \leq Y_{n+1}\right|$
- Then $\mathbb{P}\{R_{n+1} = k\} = \frac{1}{n+1}$ for any $k \in [n+1]$ and probability that Y_{n+1} is among the smallest k elements is $\frac{k}{n+1}$.
- Therefore, the probability that Y_{n+1} is among the $\lceil (1-\alpha)(n+1) \rceil$ smallest values is $\geq 1 - \alpha$.

Claim

$$\mathbb{P}\{Y_{n+1} \textit{ is among the } k \textit{ smallest of } Y_1, \dots Y_n, Y_{n+1}\} \geq 1 - \alpha$$

$$\Longleftrightarrow$$

$$\mathbb{P}\{Y_{n+1} \textit{ is among the } k \textit{ smallest of } Y_1, \dots Y_n\} \geq 1 - \alpha$$

Proof.

Consider the complement event:

$$\{Y_{n+1} > k \text{ smallest elements of } Y_1, \dots Y_{n+1}\}$$

$$\Longleftrightarrow$$

$$\{Y_{n+1} > k \text{ smallest elements of } Y_1, \dots Y_n\}$$

□

- Again, we are considering order statistics for $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$. But since we do not know Y_{n+1}, we can only work with $Y_1, Y_2, \ldots, Y_n$.
- The previous claim allows us to do that.
- Define

$$\hat{q}_n = \begin{cases} Y_{(\lceil (1-\alpha)(n+1) \rceil)} & \text{if } \lceil (1-\alpha)(n+1) \rceil \leq n \\ \infty \quad \text{Otherwise} \end{cases}$$

- $Y_{(k)}$ is the kth order statistic.
- As just mentioned, the computation of $\hat{q_n}$ can be done using just $Y_1, Y_2, \ldots, Y_n$.

Exchangeability

- If you carefully look at our analysis, we did not use the full power of i.i.d. We only needed exchangeability everywhere.
- Exchangeability is defined as

$$(Y_1, Y_2, \ldots Y_{n+1}) \stackrel{d}{=} (Y_{\sigma(1)}, Y_{\sigma(2)}, \ldots Y_{\sigma(n+1)})$$

  for every permutation $\sigma : [n+1] \to [n+1]$.
- Under the exchangeability assumption, the indexing of the random variables is immaterial.

Application to Regression

- Suppose that $\hat{f}_n$ is a point predictor trained on $(X_i, Y_i)_{i=1}^n$.
- We want to give a prediction set for Y_{n+1}.
- We could look at the residuals, $R_i = \left|Y_i - \hat{f}_n(X_i)\right|$ for $i \in [n]$, and construct $[\hat{f}_n(X_{n+1}) - \hat{q}_n, \hat{f}_n(X_{n+1}) + \hat{q}_n]$, where just like previously $\hat{q}_n = R_{(\lceil (1-\alpha)(n+1) \rceil)}$.
- But because our model $\hat{f}_n$ has already been trained on $(X_i, Y_i)_{i=1}^n$, the residuals will be unnaturally small (intuitively, residual for a new data point will be generally bigger). And so, the interval just constructed undercovers.
- More precisely, $Y_{n+1} \in \hat{C}_n(X_{n+1}) \iff R_{n+1} \leq k$th smallest of $(R_i)_{i=1}^n$ will not hold with probability $\geq 1 - \alpha$ because R_{n+1} is not exchangeable with $R_1, \ldots, R_n$.

Second Key Idea: Maintain Exchangeability

- Split the indices $\mathcal{I} = [n]$ into two disjoint sets: $\mathcal{I}_1$ and $\mathcal{I}_2$, training and calibration sets.
- $|\mathcal{I}_1| = n_1$ and $|\mathcal{I}_2| = n_2$.
- Fit $\hat{f}_{n_1}$ on data with indices $\mathcal{I}_1$.
- Obtain the residuals $R_1, R_2, \ldots, R_{n_2}$ on data with indices $\mathcal{I}_2$.
- $\hat{q}_{n_2} = R_{(\lceil (1-\alpha)(n_2+1) \rceil)}$.
- Conformal set: $[\hat{f}_{n_1}(x) - \hat{q}_{n_2}, \hat{f}_{n_1}(x) + \hat{q}_{n_2}]$.
- $$\mathbb{P}\left(Y_{n+1} \in \hat{C}_{n+1}(X_{n+1}) \mid (X_i, Y_i), i \in \mathcal{I}_1\right) \geq 1 - \alpha$$

  .
- This holds because conditioned on $\mathcal{I}_1$, the calibration residuals of data with indices $\mathcal{I}_2$, $R_1, R_2, \ldots, R_{n_2}$ and R_{n+1} are i.i.d.

Split Conformal Prediction Algorithm

Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$
$\widehat{\mu} = \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\big)$
$R_i = |Y_i - \widehat{\mu}(X_i)|$, $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$
Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

Remarks

- Instead of residual, we can take any *conformity score*, $R_i = V(X_i, Y_i)$.
- However, the length of the prediction band is constant and does not adapt to the local hardness of the problem.
- Split conformal prediction sacrifices statistical effeciency by splitting the data.

Effect of the quality of $\hat{f}$

- Note that we did not comment on the the prediction accuracy of $\hat{f}$.
- Better the point predictor, $\hat{f}$, tighter the prediction band.
- Average length of a prediction set: $\mathbb{E}_{(X_i,Y_i)\sim P, i\in\mathcal{I}_2}\left[\int \int_{\hat{C}_n(x)} d\mu(y) dP_X(x)\right]$, μ =Lebesgue measure.
- Coverage: $\mathbb{E}_{(X_i,Y_i)\sim P, i\in\mathcal{I}_2}\left[\int \int_{\hat{C}_n(x)} dP_{Y|X}(y) dP_X(x)\right]$
- An inefficient algorithm must somehow put mass at low density regions, which does not hurt it's coverage but inflates the length.

Full Conformal Prediction

- For efficiency reasons we don't want to split the data.
- Fix any $x \in \mathcal{X}$.
- Suppose we want to find out whether an arbitrary $y \in \mathbb{R}$ should be in the prediction set $\hat{C}_n(x)$.
- Suppose we train our prediction algorithm on an augmented training set : $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (x, y)$ and obtain a point predictor $\hat{f}_{n,(x,y)}$.
- Define

$$R_i^{(x,y)} = \begin{cases} \left| Y_i - \hat{f}_{n,(x,y)}(X_i) \right|, & i \in [n] \\ \left| y - \hat{f}_{n,(x,y)}(x) \right|, & i = n+1 \end{cases}$$

- The full conformal set is defined as

$$\hat{C}_n = \{y : R_{n+1}^{(x,y)} \leq \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } R_i^{(x,y)} \text{ for } i \in [n]\}$$

.

- The subtle point is that we can get the guarantee of $\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1}\} \geq 1 - \alpha$ once we plug in $(x, y) = (X_{n+1}, Y_{n+1})$, in which case all the residuals are exchangeable.
- This is true only if $\hat{f}_{n,(x,y)}$ does not use the knowledge of the order of the training data.

Full Conformal Prediction

Algorithm 1 Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$, points $\mathcal{X}_{\text{new}} = \{X_{n+1}, X_{n+2}, \ldots\}$ at which to construct prediction intervals, and values $\mathcal{Y}_{\text{trial}} = \{y_1, y_2, \ldots\}$ to act as trial values

Output: Predictions intervals, at each element of $\mathcal{X}_{\text{new}}$

for $x \in \mathcal{X}_{\text{new}}$ **do**

  for $y \in \mathcal{Y}_{\text{trial}}$ **do**

    $\widehat{\mu}_y = \mathcal{A}\big(\{(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)\}\big)$

    $R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \ldots, n$, and $R_{y,n+1} = |y - \widehat{\mu}_y(x)|$

    $\pi(y) = (1 + \sum_{i=1}^{n} \mathbb{1}\{R_{y,i} \leq R_{y,n+1}\})/(n+1)$

  end for

  $C_{\text{conf}}(x) = \{y \in \mathcal{Y}_{\text{trial}} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\}$

end for

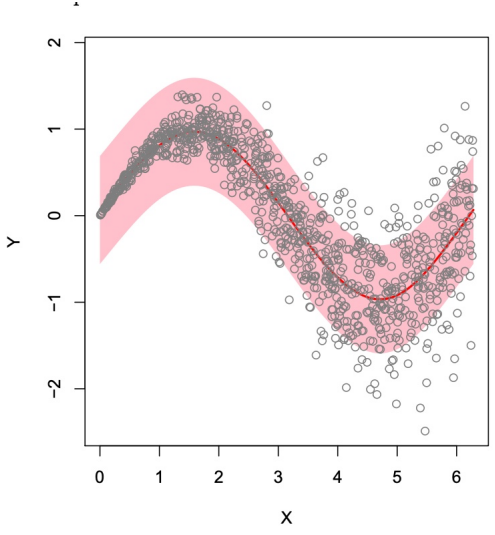
Return $C_{\text{conf}}(x)$, for each $x \in \mathcal{X}_{\text{new}}$

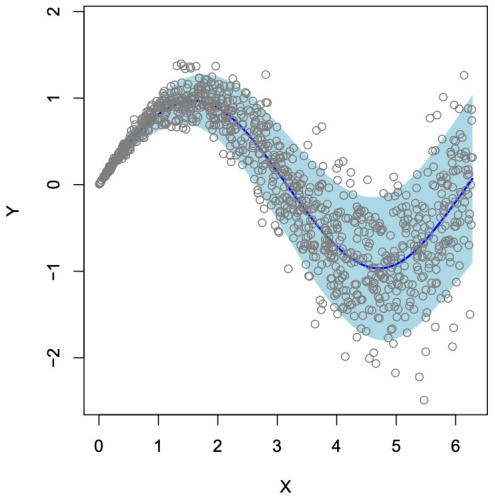
Remarks

- This method is extremely computationally intensive – for every x, we need to refit $\hat{f}_{n,(x,y)}$ for all $y \in \mathbb{R}$. This is infinitely expensive.
- Can work practically for prediction algorithms which have a fast way to refit the point predictor.
- Some methods are proposed that lie between split and full conformal prediction. (Barber et.al, 2021).

Adaptive size of prediction set: Studentized Residuals

- Consider split prediction. On $\mathcal{I}_1$, we fit both, a point predictor $\hat{f}_{n_1}$ and a variance predictor, $\hat{\sigma}_{n_1}$ which fits the standard deviation of the residual, $\left|Y - \hat{f}_{n_1}(X)\right|$.
- We compute normalized residuals on $\mathcal{I}_2$. $R_i = \frac{\left|Y_i - \hat{f}_{n_1}(X_i)\right|}{\hat{\sigma}_{n_1}(X_i)}$.


Y
X
2
1
0
−1
−2
0
1
2
3
4
5
6


Y
X
2
1
0
−1
−2
0
1
2
3
4
5
6

Extensions and trends

- Designing conformal methods which have good practical performance small set sizes or balanced coverage across regions in feature space.
- Distribution shift: test point has a different distribution from the calibration.
- Beyond the exchangeability assumption.
- Full conformal prediction to asymmetric algorithms.
- Prediction sets that preserve the privacy of the data.
- Conformal Predictive distribution – which outputs a probability distribution over the space $\mathcal{Y}$.
- And many more ...

Resources

- Distribution-Free Predictive Inference For Regression, by Lei, Wasserman et. al. 2017.
- A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification by Angelopoulos & Bates, 2022.
- Distribution Free Prediction Bands, by Lei and Wasserman, 2014.
- Conformal prediction beyond exchangeability, by Candes, 2022.