Brandon Salter
596 Final Project
December 8, 2023

Literature Review: Conformal Prediction and Global Models

Seeing as our course is sectioned into regression and time series analysis units, for my final project I present a literature review of topics in both areas. Specifically, I review research on the topics of conformal prediction, conformalized quantile regression, and global models for time series forecasting, having three sections corresponding to the three topics, with individual papers for each.

The paper, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," by Anastasios N. Angelopoulos and Stephen Bates, is intended as a self-contained, hands-on introduction to the topic of conformal prediction (a.k.a. conformal inference). It describes the benefits over other methods of uncertainty quantification, the general procedure, numerous worked examples, extensions, and proves the coverage guarantees of the conformal prediction framework.

As we know, machine learning models are integrated ubiquitously into processes that drive real world decisions. According to the authors, conformal prediction is 'a user-friendly paradigm for creating statistically rigorous uncertainty sets/intervals for the predictions of such models.' Said differently, conformal prediction takes an output from a model, containing a heuristic measure of uncertainty, and converts this value to a rigorous one. It achieves this without making any strong assumptions. The one assumption is that of exchangeability, which occurs if the joint distribution of a sequence of observations is a symmetric function of its n arguments (e.g., a special case of exchangeability is when all datapoints are random instances of the same data generating process). The advantages discussed in the paper of conformal prediction over other methods of uncertainty quantification are the following:

1. Conformal prediction provides a theoretical coverage guarantee, with minimal assumptions, that, on average, the prediction region will cover the true outcome with probability set by the user.
2. It is very easy to implement, given an existing model.
3. It is generally applicable to any supervised machine learning model (i.e., it is model agnostic).
4. It works for any sample size (non-asymptotic).

The general procedure for conformal prediction is, as outlined by the authors, the following:

Step 1: Training

1. Split data into training, calibration, and test sets
2. Train a model for a supervised classification or regression task

Step 2: Calibration

1. Using some in-built heuristic notion of uncertainty regarding the conditional prediction of the label (e.g., softmax outputs), compute non-conformity scores for calibration data such that more certainty is ranked lower (e.g., 1 – the softmax output of the true class), then sort these scores in ascending order.
2. Find the quantile $q = \left\lceil \frac{(n+1)(1-\alpha)}{n} \right\rceil$ (the $(1-\alpha)^{th}$ empirical quantile corrected for finite sample) of the above scores. Scores below this quantile value are said to 'conform' to the rest.

Step 3: Prediction

1. Predict the test set.
2. For each prediction, choose all values that produce a score below q to form a prediction set or interval with guaranteed marginal coverage.

Given that we have followed the above steps, the notion of guaranteed marginal coverage is stated formally as follows:

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

The proof of this statement can be found in Appendix D. This statement is true for split conformal prediction (the most common kind), which is when part of the training set is reserved for conformalization, or calibration. This is the same method outlined in the above algorithm.

Another important technical aspect of conformal prediction is the choice of score function. This largely determines the size of the prediction set or interval and thus, the quality of the overall inference process. Therefore, constructing an adequate score function is a key engineering detail that must not be overlooked. Since interval/set size is highly correlated with the quality of score, the former could serve as an approximate means to assess and compare between choices of score functions.

On the topic of evaluation, the authors discuss two main ways to assess the relative goodness of conformal prediction regions, the first being evaluating adaptivity. Adaptivity is the changing in size of the prediction region depending on the difficulty of prediction imposed by different model inputs. This is a valuable feature for practical deployment, as it indicates higher certainty for tighter regions and vice versa. Methods suggested to evaluate adaptivity are plotting histograms of set sizes and computing conditional coverage (which is stronger than the guaranteed marginal coverage). Feature-stratified and size-stratified coverage metrics are two means by which the authors propose conditional coverage can be assessed. The second method for evaluation is checking to make sure that coverage guarantees are met (i.e., that your implementation is correct). This can be done computing the realized marginal coverage of test prediction regions and ensuring the true label is contained in at least $(1 - \alpha)\%$ of the regions.

For brevity, the remainder of the paper discusses both practical and theoretical extensions to split conformal prediction, historical notes, current trends, and numerous worked examples including the primary topic of the next paper I will discuss, conformalized quantile regression.

The paper, "Conformalized Quantile Regression," by Yaniv Romano, Evan Patterson, and Emmanuel J. Candès, discusses general conformal prediction as was discussed in the previous paper, briefly overviews quantile regression, conformalizes quantile regression, and shows empirically how conformalized quantile regression outperforms other methods in terms of both interval coverage and width while providing theoretically valid intervals.

In discussing standard conformal prediction, the authors note how in the case of regression, the conformalization procedure is typically done by estimating the conditional mean of Y given X. The downside to this approach is that the resulting prediction intervals are not adaptive, meaning that their length is fixed rather than changing depending on the certainty of the model of its prediction. However, conformal prediction as outlined by Angelopoulos and Bates, serves as the basis to which the conformalized quantile regression procedure improves upon. Regarding the formation of prediction intervals, the authors state two properties for which any uncertainty quantification should be measured by, those being non-asymptotic validity for finite samples without strong distributional assumptions and the length of prediction interval.

The paper then gives a high-level overview of traditional quantile regression. Quantile regression can be used to form prediction intervals by fitting the conditional quantile function to the $\left(\frac{\alpha}{2}\right)\%$ and $\left(1 - \frac{\alpha}{2}\right)\%$ levels. This procedure, unlike conditional mean regression, is adaptive to local variability (heteroskedasticity of the label in the feature space). This finding is intuitive when considering the nature of the α^{th} conditional quantile function and the data that lie below it. The main point made in this section was that while under certain regularity conditions, estimates of conditional quantiles can be shown to be asymptotically consistent, these adaptive intervals are without finite sample guarantees and are not generally valid.

Moving to conformalized quantile regression (CQR), the general procedure outlined in the paper is as follows:

1. Split the data into training and calibration sets.
2. Fit two quantile functions (based on the desired $(1-\alpha)\%$ marginal coverage) to the training data using some quantile regression algorithm.
3. Compute the conformity score (by equation 9). Importantly, the equation accounts for both undercoverage and overcoverage. Sort (ascending) and take finite sample corrected empirical quantile.
4. Construct prediction intervals for new data (by 10 and 11) using the quantile value from the conformity scores in the previous step to 'correct' (or calibrate) the original quantile intervals.

The prediction intervals for new data are then shown to be both valid and, if conformity scores are almost surely distinct, nearly perfectly calibrated. The proof of this theorem can be found at the end of the section.

The final sections describe a thorough experiment made to compare the CQR method to the standard and locally adaptive versions of split conformal prediction. For typical conformal prediction, ridge regression, random forest, and neural net models were fit to eleven benchmark datasets. As mentioned, after conformalization these models produce fixed size prediction intervals and hence are not adaptive. Locally adaptive versions of these same models were then fit to the same datasets. Following these, CQR and traditional quantile versions of the random forest and neural net models were fit. Across all models, hyperparameters were controlled for and an 80-20 train-test split was used. For the conformal models, half of the training data was reserved for calibration. Out of 2,200 experiments, the authors found that on average, CQR produces significantly shorter prediction intervals, even over the non-conformalized quantile regression methods which had access to more training data. The marginal coverage of the CQR models was also matching their theoretical guarantee. On individual datasets, the overall trend is also generally reflected, as can be seen in the figures following the references section.

Shifting now to the topic of time series, the paper, “Global models for time series forecasting: A Simulation Study,” by Hansika Hewanalage, Christoph Bergmeir, and Kasun Bandara empirically investigates factors effecting the performance of global forecasting models (GFMs), which are models trained across multiple time series, exploiting series relatedness.

Through the 20th century and into the current decades, time series datasets have been treated independently for building forecasting models, leading to classical time series models such as ETS and ARIMA rising to prominence. These models typically perform on par with or better than more complex machine learning and deep learning models when trained only on a single dataset of modest size. The paradigm has been shifting in recent times with the collection of large amounts of data from related series (e.g., sales across departments) and the building of models simultaneously across all of it. Currently, the prestigious M4 and M5 forecasting competitions have been dominated exclusively by GFMs, which could be a harbinger for their widespread adoption in industry and further exploration in academia.

Posed by the authors as perhaps the ‘most open question in time series forecasting,’ explaining the ‘unreasonable effectiveness of global models’ is what the paper seeks to accomplish. It does so by controlling factors relating to the homogeneity of global datasets across series, their complexity patterns, model complexity, and dataset size through multiple extensive empirical evaluations across simulated time series datasets. The simulation techniques range from simple data generating processes (DGPs) such as auto regressive (AR), seasonal AR, and Fourier terms to more complex DGPs such as chaotic logistic map, self-exciting threshold AR, and Mackey-Glass equations. Like the M-competitions, modern machine learning techniques (PR, FFNN, LGBM, RNN) are pitted against classical methods (AR, SAR, SETAR, ARIMA, DHR-ARIMA) under the aforementioned controlled conditions and the interplay of these conditions is assessed in order to assess their effect on GFM performance.

Permutations of simulated dataset conditions can be seen in Figure 1 on page 3. After simulating the data and creating both homogenous and heterogenous scenarios (in building a global dataset), the authors then trained a suite of both local and global models, which can also be seen in Figure 1. Complexity of global models is varied by the two model setups, GFM.All, which takes the global dataset input as is, and GFM.Cluster, which uses clustering analysis to further segment the data. Their main findings were the following:

- The performance of both local and global models improves with longer series.
- The performance of local models does not improve when trained across multiple different series (globally).
- The appropriateness of the model, whether global or local, related to the DGP matters substantially in the model’s generalization performance.

- For heterogenous data, local models are competitive over simpler linear global models, more so as series length increases. This makes sense intuitively, since a local model trained globally is complex when considering the number of parameters trained across all datasets.
- Nonlinear global models are very competitive for all lengths of series for heterogenous data.
- Generally, as global datasets increase in heterogeneity, complex global models dominate. Further, complex global model can perform reasonably well irrespective of the DGP, while linear models work well only when there are known linear patterns present in the data.
- With limited data, switching model family can be a better alternative to increasing model complexity rather than increasing the complexity of an existing linear model.

The authors verify the conclusions made from the simulated datasets by using real-world datasets matching the homogeneity of various simulated ones measured in terms of tsfeatures, which is a topic of much previous work.

In conclusion, this literature review delves into the realms of conformal prediction and global models, addressing key aspects in both regression and time series analysis. The exploration of conformal prediction, as elucidated by Angelopoulos and Bates, reveals its user-friendly paradigm for uncertainty quantification, grounded in minimal assumptions and offering theoretical coverage guarantees. The subsequent examination of conformalized quantile regression by Romano, Patterson, and Candès extends the conformal prediction framework to enhance adaptivity in prediction intervals while reducing interval length. It also showcases how empirically, CQR outperforms both traditional quantile and non-quantile-conformal methods of regression uncertainty quantification. Shifting focus to time series forecasting, Hewanalage, Bergmeir, and Bandara's study on global models unveils a paradigm shift from independent models to those trained across multiple series, particularly evident in the domination of global models in recent M4 and M5 forecasting competitions. The authors scrutinize factors influencing the performance of global models, providing insights into the effectiveness of such models across various conditions. This comprehensive review not only synthesizes current knowledge but also paves the way for further exploration and application of these methodologies in diverse analytical domains.