Problem Set 5

Issued: Thursday, April 20th, 2023
Due Date: Monday, May 1st 2023, 11:59 PM ET

Problem 5.1: Time Series Quantile Regression with Uncertainty. [45 pts]

The quantile regression algorithm attempts to learn the α quantile of target $Y|X = x$ for each possible value of x, instead of the mean in standard least squared linear regression. Unlike regular linear regression which uses the method of least squares to calculate the conditional mean of the target across different values of the input, quantile regression estimates the conditional quantile of the target. For example, if we were to forecast the 25th quantile prediction, that would mean that there is a 25% chance the actual value is below the prediction, while there is a 75% chance that the value is above.

In this question, we will use quantile regression for timeseries with uncertainty intervals for predictions.

Data description: The data for this question has been kindly shared by Prof. Audun Botterud, a Principal Research Scientist in the MIT Laboratory for Decisions and Information Systems and a Senior Energy Systems Engineer at Argonne National Laboratory. The source of the data is the Electric Reliability Council of Texas (ERCOT). ERCOT is a non-profit organization that manages the electricity grid for most of Texas. ERCOT operates the grid and manages the electricity market for the state, ensuring that there is a reliable supply of electricity for a large part of the state of Texas. The organization is responsible for overseeing the transmission and distribution of electricity, as well as for managing the dispatch of power plants and the maintenance of the grid infrastructure. In the dataset we released, the spreadsheet contains the *hourly* wind throughput, as a percentage of the total installed capacity. The primary variable of interest in the file is the wind output percentage, expressed as a percentage of the total installed capacity. This is denoted by the column 'Wind Output, % of Installed' in the file. The column 'Time (Hour-Ending)' contains timestamps of data. More details about this and similar datasets will be provided in the guest lecture by Prof. Audun Botterud.

Train/validation/test data splits: Use the data from Jan-May for training, June-August for validation (or calibration, when appropriate), and September - December for testing for all questions that follow. Consider that the unit of time t is hour, unless specified otherwise. For the training set, consider data starting January 2022 (i.e., datapoints for which at least a 24-hour history is available).

Part (i)

(a) Compute the %change in wind output across hourly intervals on the train data split. That is, the percentage change in the time series from time t to $t + 1hr$. Plot the histogram of this value, and comment on whether you think autoregressive models are appropriate for modeling this data.

(b) If $F(t)$ denotes the wind output% at time t (i.e., the column 'Wind Output, % of Installed'), fit the data to a simple autoregressive model $F^{forecast}(t) = \sum_{i=24}^{1} \beta_i F(t-i) + \beta_0$ using quantile regression with $\alpha = 0.50$. Compute the mean squared error as well as the proportion of predictions that lie below the actual value on the test set for this value of α.

With Python, we recommend carefully choosing the quantile regression solver based on library versions. For example, the HiGHS linear programming solver (`highs` in Python) seems appropriate for `sklearn` versions $\geq 1.6.0$.

(c) Plot the 95% quantile region of predictions for the month of December 2022. Specifically, plot the predictions corresponding to quantile regression with $\alpha = 0.025$ as the lower bound and $\alpha = 0.975$ as the upper bound for hourly intervals in the month of December 2022. What do you note about the coverage and width of the intervals in each case?

(d) Next, we will use conformalization to obtain uncertainty bounds with guarantees, without making any assumptions on how the data is distributed. Read Section 5.3 of *Angelopoulous et al.* [1]. Link to paper: `https://arxiv.org/pdf/2107.07511.pdf`.

Consider the non-conformity score to be $\max\{F(t) - \tau_{0.975}, \tau_{0.025} - F(t)\}$ for each t, where τ_α denotes the predicted α-quantile corresponding to time t. Using the weighted conformalization procedure described in 5.3, and $K = 24$, produce uncertainty bounds. Note that in weighted conformalization, calibration is performed in a window of K for each time-point instead of on the whole set.

(e) Similar to part (c), plot the 95% conformalized bounds for each timepoint for the month of December 2022. Compare and contrast this plot to the one obtained from part (c), particularly in terms of interval width and coverage.

Part (ii)

(a) From the previous analyses, draw the conformalization score as a function of $F(t)$ for the month of August 2022 and briefly describe what this score and what thresholding it captures. (Hint: Consider how the position of the actual throughput – $F(t)$–with respect to the initial, non-conformalized prediction intervals from quantile regression influences this score).

Problem 5.2: Domain Adaptation. [15 pts]

Covariate shift is a specific type of domain shift in data: given input x, and label y, the input distribution x shifts between domains while the conditional probability of $y|x$ remains unchanged. Formally, we consider a "source" domain – the data domain where our machine learning model was trained on – and a "target" domain, where the model may be tested. For the scope of this question, assume that there are *only two domains of interest*: a source S and a target T.

In this question, we are interested in evaluating the loss or empirical risk of any *class-label classifier* h.

Data assumptions: We have access to a large number of labeled examples from source domain examples. Thus, we have both $P_S(x)$ and $P_S(y|x)$. We also have a large number of unlabeled examples from the target domain so we can construct $P_T(x)$. Assume also that the support of $P_T(x)$ is contained within the support of $P_S(x)$.

Auxilary information: To identify "risky domains" or test domains with high loss, a friend of ours was kind enough to train a probabilistic *domain classifier*. This classifier outputs the probability that a given input x came from domain S or T. Note that this is separate from our classifier h.

Unfortunately for us, in training the domain classifier, the friend assumed that target examples were *three times* as likely to occur overall than the source examples. The domain classifier was trained to maximize the log-probability of the correct domain for each x when x was sampled from $P_T(x)$ thrice as often as from $P_S(x)$.

(a) Write down an expression for the friend's probabilistic domain classifier $Q(d = T|x)$ as a function of $P_S(x)$ and $P_T(x)$.

(b) Now, we use the domain classifier to evaluate the target risk of any class-label classifier $h(x)$. You can assume that empirical $\text{Loss}(y, h(x))$ is given. Provide an expression for the target risk of h as a function of $P_S(x)$, $P_S(y|x)$, h, $\text{Loss}(y, h(x))$, and the friend's domain classifier $Q(d = T|x)$ (Hint: Use the covariate shift definition.).

References

[1] Angelopoulos, A. N., and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).