Conformal Prediction Example

Catherine Chen cyc2152

November 2023

1 Linear Regression

1.1 Assumptions

1. Weak Exogeneity
2. Linearity
3. Constant variance (Homoscedasticity)
4. Independence of Errors
5. Lack of Perfect Multicollinearity

2 Quantile Regression

Estimate a given quantile, i.e. the median, of Y conditional on X. The conditional distribution function of Y given $X = x$ is

$$F(y \mid X = x) := \mathbb{P}\{Y \leq y \mid X = x\}$$

and the α th conditional quantile function is

$$q_\alpha(x) := \inf\{y \in \mathbb{R} : F(y \mid X = x) \geq \alpha\}.$$

Suppose fixed lower and upper quantiles: $\alpha_{\text{lo}} = \alpha/2$ and $\alpha_{\text{hi}} = 1 - \alpha/2$. Given $q_{\alpha_{\text{lo}}}(x)$ and $q_{\alpha_{\text{hi}}}(x)$, the corresponding lower and upper conditional quantile functions, the conditional prediction interval for Y given $X = x$, with miscoverage rate α is

$$C(x) = [q_{\alpha_{\text{lo}}}(x), q_{\alpha_{\text{hi}}}(x)]$$

By construction, this interval satisfies

$$\mathbb{P}\{Y \in C(X) \mid X = x\} \geq 1 - \alpha.$$

The length of the interval $C(X)$ can vary greatly depending on the value of X, which reflects the uncertainty in the prediction of Y.

Classical regression analysis estimates the conditional mean of the test response Y_{n+1} given the features $X_{n+1} = x$ by minimizing the sum of squared residuals on the n training points:

$$\hat{\mu}(x) = \mu(x; \hat{\theta}), \quad \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mu\left(X_i; \theta\right)\right)^2 + \mathcal{R}(\theta).$$

where θ are the parameters of the regression model, $\mu(x; \theta)$ is the regression function, and $\mathcal{R}$ is a regularizer.

Analogously, quantile regression estimates a conditional quantile function q_α of Y_{n+1} given $X_{n+1} = x$. This can be cast as the optimization problem

$$\hat{q}_\alpha(x) = f(x; \hat{\theta}), \quad \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha \left(Y_i, f\left(X_i; \theta\right)\right) + \mathcal{R}(\theta),$$

where $f(x;\theta)$ is the quantile regression function and the loss function ρ_α is the "pinball loss"

$$\rho_\alpha(y, \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0, \\ (1-\alpha)(\hat{y} - y) & \text{otherwise} \end{cases}$$

Quantile Regression Procedure: construct a prediction band with nominal miscoverage rate α, estimate $\hat{q}_{\alpha_{\text{lo}}}(x)$ and $\hat{q}_{\alpha_{\text{hi}}}(x)$ using quantile regression, then output $\hat{C}\left(X_{n+1}\right) = \left[\hat{q}_{\alpha_{\text{lo}}}\left(X_{n+1}\right), \hat{q}_{\alpha_{\text{hi}}}\left(X_{n+1}\right)\right]$ as an estimate of the ideal interval $C\left(X_{n+1}\right)$. This yields intervals that are adaptive to heteroscedasticity. However, it is not guaranteed to satisfy the coverage statement

$$\mathbb{P}\{Y \in C(X) \mid X = x\} \geq 1 - \alpha.$$

when $C(X)$ is replaced by the estimated interval $\hat{C}\left(X_{n+1}\right)$ due to the absence of any finite sample guarantee, which can result in undercoverage.

3 Conformal Prediction

Given n training samples $\{(X_i, Y_i)\}_{i=1}^{n}$, we want to predict the unknown value of Y_{n+1} at a test point X_{n+1}. Assume that all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are drawn exchangeably-for instance, drawn i.i.d.-from an arbitrary joint distribution P_{XY} over the feature vectors $X \in \mathbb{R}^p$ and response variables $Y \in \mathbb{R}$. We aim to construct a marginal distribution-free prediction interval $C\left(X_{n+1}\right) \subseteq \mathbb{R}$ that is likely to contain the unknown response Y_{n+1}. That is, given a desired miscoverage rate α, we ask that

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

for any joint distribution P_{XY} and any sample size n. The probability in this statement is marginal, being taken over all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$.

To accomplish this, training data is split into two disjoint subsets, a proper training set and a calibration set. Fit two quantile regressors on the proper training set to obtain initial estimates of the lower and upper bounds of the prediction interval. Then, using the calibration set, conformal prediction is applied and, if necessary, the prediction interval is corrected. Unlike the original interval, the conformalized prediction interval is guaranteed to satisfy the coverage requirement regardless of the choice or accuracy of the quantile regression estimator

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

The split conformal method begins by splitting the training data into two disjoint subsets: a proper training set $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ and calibration set $\{(X_i, Y_i) : i \in \mathcal{I}_2\}$. Then, given any regression algorithm $\mathcal{A}$, a regression model is fit to the proper training set:

$$\hat{\mu}(x) \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\right).$$

Then, the absolute residuals are computed on the calibration set:

$$R_i = \left|Y_i - \hat{\mu}\left(X_i\right)\right|, \quad i \in \mathcal{I}_2.$$

For a given level α, compute a quantile of the empirical distribution of the absolute residuals,

$$Q_{1-\alpha}\left(R, \mathcal{I}_2\right) := (1-\alpha)\left(1 + 1/\left|\mathcal{I}_2\right|\right)\text{-th empirical quantile of } \left\{R_i : i \in \mathcal{I}_2\right\}.$$

Then, the prediction interval at a new point X_{n+1} is given by

$$C\left(X_{n+1}\right) = \left[\hat{\mu}\left(X_{n+1}\right) - Q_{1-\alpha}\left(R, \mathcal{I}_2\right), \hat{\mu}\left(X_{n+1}\right) + Q_{1-\alpha}\left(R, \mathcal{I}_2\right)\right].$$

4 Conformalized Quantile Regression Regression

As in split conformal prediction, split the data into a proper training set, indexed by $\mathcal{I}_1$, and a calibration set, indexed by $\mathcal{I}_2$. Given any quantile regression algorithm $\mathcal{A}$, fit two conditional quantile functions $\hat{q}_{\alpha_{\mathrm{lo}}}$ and $\hat{q}_{\alpha_{\mathrm{hi}}}$ on the proper training set:

$$\{\hat{q}_{\alpha_{\mathrm{lo}}}, \hat{q}_{\alpha_{\mathrm{hi}}}\} \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\right).$$

Next, compute conformity scores that quantify the error made by the plug-in prediction interval $\hat{C}(x) = [\hat{q}_{\alpha_{\mathrm{lo}}}(x), \hat{q}_{\alpha_{\mathrm{hi}}}(x)]$ on the calibration set:

$$E_i := \max\left\{\hat{q}_{\alpha_{\mathrm{lo}}}\left(X_i\right) - Y_i, Y_i - \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_i\right)\right\}$$

for each $i \in \mathcal{I}_2$.

1. If Y_i is below the lower endpoint of the interval, $Y_i < \hat{q}_{\alpha_{\mathrm{lo}}}\left(X_i\right)$, then $E_i = |Y_i - \hat{q}_{\alpha_{\mathrm{lo}}}\left(X_i\right)|$ is the magnitude of the error incurred by this mistake. (Undercoverage)
2. If Y_i is above the upper endpoint of the interval, $Y_i > \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_i\right)$, then $E_i = |Y_i - \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_i\right)|$. (Undercoverage)
3. If Y_i correctly belongs to the interval, $\hat{q}_{\alpha_{\mathrm{lo}}}\left(X_i\right) \leq Y_i \leq \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_i\right)$, then E_i is the larger of the two non-positive numbers $\hat{q}_{\alpha_{lo}}\left(X_i\right) - Y_i$ and $Y_i - \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_i\right)$ and so is itself non-positive. (Overcoverage)

The conformity score thus accounts for both undercoverage and overcoverage.

Finally, given new input data X_{n+1}, we construct the prediction interval for Y_{n+1} as

$$C\left(X_{n+1}\right) = \left[\hat{q}_{\alpha_{\mathrm{lo}}}\left(X_{n+1}\right) - Q_{1-\alpha}\left(E, \mathcal{I}_2\right), \hat{q}_{\alpha_{\mathrm{hi}}}\left(X_{n+1}\right) + Q_{1-\alpha}\left(E, \mathcal{I}_2\right)\right]$$

where

$$Q_{1-\alpha}\left(E, \mathcal{I}_2\right) := (1-\alpha)\left(1 + 1/\left|\mathcal{I}_2\right|\right)\text{-th empirical quantile of } \left\{E_i : i \in \mathcal{I}_2\right\}$$

conformalizes the plug-in prediction interval.

Theorem 1. *If $\left(X_i, Y_i\right), i = 1, \ldots, n+1$ are exchangeable, then the prediction interval $C\left(X_{n+1}\right)$ constructed by the split CQR algorithm satisfies*

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

Moreover, if the conformity scores E_i are almost surely distinct, then the prediction interval is nearly perfectly calibrated:

$$\mathbb{P}\left\{Y_{n+1} \in C\left(X_{n+1}\right)\right\} \leq 1 - \alpha + \frac{1}{\left|\mathcal{I}_2\right| + 1}$$