Statistical Learning: Conformal Prediction

Eszter Barabás, ELTE TTK

Supervisor: Balázs Csanád Csáji, ELTE TTK & HUN-REN SZTAKI

1. Introduction

In this semester my project is about conformal prediction, which is a statistical framework for quantifying uncertainty estimates for predictions. In essence it produces prediction sets that are guaranteed to contain the true data with a preset probability. This method can be useful to provide a range of likely outcomes instead of a single estimate. It can be used a myriad of real-world examples like predicting temperature ranges for weather forecast, equipment failure time ranges or future stock prices with 90% probability.

2. Conformal Predictors

To begin with we establish the precise theoretical framework. For the following part I cited sources [1] and [3] in my report. Consider our data in the form of ordered pairs (x_i, y_i) called examples. Each example consists of an object x_i coming from a measurable space $\mathcal{X}$ and its label y_i which is an element of a measurable space $\mathcal{Y}$.

We assume that $\mathcal{X}$ is non-empty and that $\mathcal{Y}$ contains at least two essentially different elements.[1] For more compact notation, we write $z_i = (x_i, y_i)$. We set

$$\mathcal{Z} := \mathcal{X} \times \mathcal{Y},$$

and call $\mathcal{Z}$ the *example space*.

Our standard assumption is that the examples are chosen independently from some probability distribution P on $\mathcal{Z}$.

Definition 1. *The exchangeability of P means that for every positive integer n, every permutation π of $\{1, \dots, n\}$, and every measurable set $E \subseteq \mathcal{Z}^n$,*

$$\mathbb{P}\big((z_1, \dots, z_n) \in E\big) = \mathbb{P}\big((z_{\pi(1)}, \dots, z_{\pi(n)}) \in E\big).$$

Definition 2. *A* simple predictor *is a measurable function*

$$D : \mathcal{Z}^* \times \mathcal{X} \to \mathcal{Y}.$$

that for any sequence of previous examples

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1}) \in \mathcal{Z}^*,$$

and any new object $x_n \in \mathcal{X}$, the prediction for the new label y_n is

$$D(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \in \mathcal{Y}.$$

We want to produce a set of possible predictions with a confidence level. In other words, our goal is to generate subsets of $\mathcal{Y}$ that are large enough to likely contain the real y_n value.

We require an additional input $\alpha \in (0, 1)$, called the *significance level*; the complementary quantity $1 - \alpha$ is the *confidence level*. Given these inputs, an adequate algorithm r outputs a subset of $\mathcal{Y}$:

$$\Gamma^\alpha(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \subseteq \mathcal{Y}.$$

Intuitively, smaller α corresponds to higher confidence in the prediction.

$$\text{If } \alpha_1 \geq \alpha_2, \quad \Gamma^{\alpha_1}(x_1, \dots, y_{n-1}, x_n) \supseteq \Gamma^{\alpha_2}(x_1, \dots, y_{n-1}, x_n). \tag{1}$$

Definition 3. *A confidence predictor is a measurable function*

$$\Gamma : (0, 1) \times \mathcal{Z}^* \times \mathcal{X} \to 2^{\mathcal{Y}}$$

that satisfies the monotonicity condition (1) for all significance levels $\alpha_1 \geq \alpha_2$, all $n \in \mathbb{Z}^+$, and all incomplete data sequences $(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$. Here $2^{\mathcal{Y}}$ denotes the set of all subsets of $\mathcal{Y}$. The function r to be measurable means that for each n, the set of sequences $(\alpha, x_1, y_1, \dots, x_n, y_n)$ satisfying

$$y_n \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

is a measurable subset of $(0, 1) \times (\mathcal{X} \times \mathcal{Y})^n$.

We now introduce a formal notation for the errors Γ makes when it processes the data sequence at significance level ϵ. The function whether Γ makes an error on the nth trial is the

[1] This ensures the prediction problem is non-trivial.

following:

$$\mathrm{err}_n^{(\epsilon)}(\Gamma,\omega) = \begin{cases} 1, & \text{error on the } n\text{th trial under } \omega, \\ 0, & \text{otherwise.} \end{cases}$$

The number of errors during the first n trials is then defined as

$$\mathrm{Err}_n^{(\epsilon)}(\Gamma,\omega) := \sum_{i=1}^{n} \mathrm{err}_i^{(\epsilon)}(\Gamma,\omega). \tag{2}$$

Definition 4. *A confidence predictor Γ is exactly valid if, for every exchangeable probability distribution P and for each significance level ϵ,*

$$\mathbb{P}[err_n(\Gamma,P)=1]=\epsilon, \quad \mathbb{P}[err_n(\Gamma,P)=0]=1-\epsilon,$$

and all $err_n(\Gamma,P)$ are independent.

Definition 5. *A confidence predictor Γ is asymptotically exact if, for any exchangeable distribution P and any $\epsilon \in (0,1)$,*

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma,P)}{n} = \epsilon \quad \textit{with probability one.}$$

Theorem 1. *No confidence predictor is exactly valid.*

Proof. The proof will show that even the property of being *weakly exact* is not satisfied, so that the independence of $\mathrm{err}_n(\Gamma)$ is left out and we assume randomness instead of exchangeability. Moreover, we will see that even for a fixed $n \in \mathbb{N}$ that $\mathbb{P}(\mathrm{err}(\Gamma)=1)=\epsilon$ for all $\epsilon \in (0,1)$ is not possible.

We may assume that the examples $z_1, z_2, \ldots$ are generated from a power distribution Q^∞ such that the probability distribution Q on $\mathcal{Z}$ is concentrated on the set $\{(x,y^{(1)}),(x,y^{(2)})\} \subseteq \mathcal{Z}$, for some arbitrarily fixed $x \in \mathcal{X}$ and $y^{(1)}, y^{(2)} \in \mathcal{Y}$ (we assumed $|\mathcal{X}| \geq 1$ and $|\mathcal{Y}| > 1$). Therefore, we assume, without loss of generality, that $\mathcal{Z} = \{0,1\}$. The argument is about impossibility, so if it fails even the smallest nontrivial case, it fails general. Fix $n \in \mathbb{N}$ and suppose that $\mathbb{P}(\mathrm{err}(\Gamma) = 1) = \epsilon$ for all $\epsilon \in (0,1)$.

For each k, let define:

$$f(k) = \mathbb{P}(\mathrm{err}(\Gamma)=1 \,|\, (z_1,\ldots,z_n) \text{ has } k \text{ ones})$$

(where we drop $z_{n+1}, z_{n+2}, \ldots$ from our notation since $\mathrm{err}(\Gamma)$ does not depend on them)

Since $\mathbb{E}[f(k)] = \epsilon$ with respect to any binomial distribution on $\{0,1,\ldots,n\}$, the standard completeness result (If the expected value of a function is constant for every binomial distribution, then the function must be constant) implies that $f(k) = \epsilon$ for all $k = 0,1,\ldots,n$. Therefore, $\epsilon\binom{n}{k}$ must be an integer for all k and ϵ, which cannot be true. [1] □

Proposition 1. *An exact confidence predictor is asymptotically exact.*

This proposition is an immediate consequence of the law of large numbers.

We want to measure the distinction between the new and the old examples, in order to achieve this we need to introduce the nonconformity measure. Before that we define the concept of a bag.

Definition 6. *A bag or multiset of size $n \in \mathbb{N}$ is a collection $\wr z_1,\ldots,z_n\wr$ of n elements from a measurable space $\mathcal{Z}$, where order is irrelevant and repetitions are allowed. We denote by $\mathcal{Z}^{(n)}$ the set of all bags of size n, and by $\mathcal{Z}^{(*)} = \bigcup_{n\geq 1} \mathcal{Z}^{(n)}$ the set of all finite bags.*

Definition 7. *A nonconformity measure is a measurable function*

$$A : \mathcal{Z}^{(*)} \times \mathcal{Z} \to \overline{\mathbb{R}}$$

that assigns to each bag of old examples and each new example $z \in \mathcal{Z}$ a score $A(\wr z_1,\ldots,z_n\wr, z)$ indicating how different z is from $\wr z_1,\ldots,z_n\wr$.

For regression problems where $z_i = (x_i, y_i) \in \mathbb{R}^2$ for a new example (x, y) we define $A(\wr z_1,\ldots,z_n\wr,(x,y)) = |y - \hat{y}(x)|$, where $\hat{y}(x)$ is the prediction from the bag.

We can also define functions that measure conformity, when we are comparing new examples to the old ones.

Definition 8. *A conformity measure is a function*

$$B : \mathcal{Z}^{(*)} \times \mathcal{Z} \to \overline{\mathbb{R}},$$

measuring how well z conforms with B. Nonconformity and conformity measures are related by strictly decreasing transformations, e.g. $A = -B$ or $A = 1/B$.

p-values. Given a nonconformity measure A and a bag $\wr z_1, \dots, z_n \int$, define

$$a_i = A_n(\wr z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n \int, z_i),$$
$$p_i = \frac{|\{j : a_j \geq a_i\}|}{n}.$$

Then p_i is the fraction called the *p-value* for z_i and it lies between 1/n and 1. If p_i is closer to the lower bound, then z_i is very nonconforming, and if it is closer to 1, then z_i is rather conforming.

Every nonconformity measure determines a confidence predictor. Given a new object x_{n+1} and a significance level ϵ, this predictor provides a prediction set $\Gamma_\epsilon(x_{n+1})$ that should contain the true label y_{n+1}.

Definition 9. *The conformal predictor determined by a nonconformity measure A is the confidence predictor Γ_ϵ defined as:*

$$\Gamma_\epsilon(x_{n+1}) = \{ y \in \mathcal{Y} \mid p_y > \epsilon \}$$

where the value p_y is the p-value *of (x_{n+1}, y) in the bag containing this example and the previous ones*

Definition 10. *A smoothed conformal predictor determined by a nonconformity measure A is a randomized confidence predictor Γ_ϵ defined as:*

$$\Gamma_\epsilon(x_{n+1}) = \{ y \in \mathcal{Y} \mid p_y > \epsilon \},$$

where the smoothed p-value p_y is given by

$$p_y = \frac{|\{ i : \alpha_i > \alpha_{n+1} \}| + \tau \cdot |\{ i : \alpha_i = \alpha_{n+1} \}|}{n+1},$$

and τ is a uniformly distributed random variable on $[0, 1]$.

The main difference from the standard conformal predictor is apparent for the cases $\alpha_i = \alpha_{n+1}$ for some i. Instead of adding a fixed $\frac{1}{n+1}$ for each equality, we add a random fraction $\frac{\tau}{n+1}$, introducing smoothness.

Proposition 2. *Any smoothed conformal predictor is exactly valid.*

3. Forming adjusted quantiles

We now return to the idea introduced at the beginning of the report and investigate the method using simulated data. For the following two sections, I used [2] article. Suppose that we have a sequence $Y_i \in \mathbb{R}$, $i = 1, \dots, n$ of real-valued response values and a significance level α.

Our goal is to find a one-sided prediction interval $C_n = (-\infty, q_n]$ such that

$$\mathbb{P}(Y_{n+1} \leq q_n) \geq 1 - \alpha.$$

As $\{(Y_i)\}_{i=1}^n$ is i.i.d., the rank of Y_{n+1} is uniformly distributed over the values $Y_1, \dots, Y_{n+1}$:

$$\mathbb{P}(Y_{n+1} \text{ is among the } (1-\alpha)(n+1) \text{ smallest of } Y_1, \dots, Y_{n+1}) \geq 1 - \alpha.$$

This is equivalent to

$$\mathbb{P}(Y_{n+1} \text{ is among the } (1-\alpha)(n+1) \text{ smallest of } Y_1, \dots, Y_n) \geq 1 - \alpha.$$

Accordingly, define

$$q_n = \lceil (1-\alpha)(n+1) \rceil\text{-th smallest of } Y_1, \dots, Y_n.$$

The comparison has changed from involving the entire sample $Y_1, \dots, Y_{n+1}$ to relying solely on the initial n observations, $Y_1, \dots, Y_n$.

Now we managed to make the right-hand side depend only on the first n data points, making it directly computable.

Remark. To see the equivalence, consider the complements of the events inside the probabilities. Let $k = (1-\alpha)(n+1)$. Then the statement

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

is clearly equivalent to

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

since Y_{n+1} cannot be strictly larger than itself.

This argument makes sense for $k \leq n$. For $k = n+1$, which happens when $\alpha < \frac{1}{n+1}$, we interpret the $(n+1)$-st smallest value of $Y_1, \dots, Y_n$ as $+\infty$.

4. Prediction to regression problems

Suppose that we observe (X_i, Y_i), $i = 1, \dots, n$, and want a prediction set for Y_{n+1} given X_{n+1}. Let f_n be any point predictor trained on the n samples. Train a nonconformity score, for example, we

define residuals:

$$R_i = |Y_i - f_n(X_i)|, \quad i = 1, \dots, n,$$

and the quantile

$$q_n = \lceil (1-\alpha)(n+1) \rceil\text{-th smallest of } R_1, \dots, R_n.$$

Then the naive prediction set is

$$C_n(x) = \left[f_n(x) - q_n,\ f_n(x) + q_n \right].$$

However, this generally undercovers, because the test residual

$$R_{n+1} = |Y_{n+1} - f_n(X_{n+1})|$$

is stochastically larger than the training residuals, since the point predictor is trained on the same data, so it tends to overfit, underestimating future residuals.

5. Split conformal prediction

Consider the regression setting above. Now, to fix the problem of undercovering, we partition the data indices into three parts: a training, a calibration, and a test set. The point predictor is trained on the proper training data, and the residuals and quantile is calculated on the calibration set. The split CP itself does not depend on the test data, it guaranties that the method is fair enough, so it is contained for theoretical convenience. For the following parts, I used ideas from sources [4] and [5].

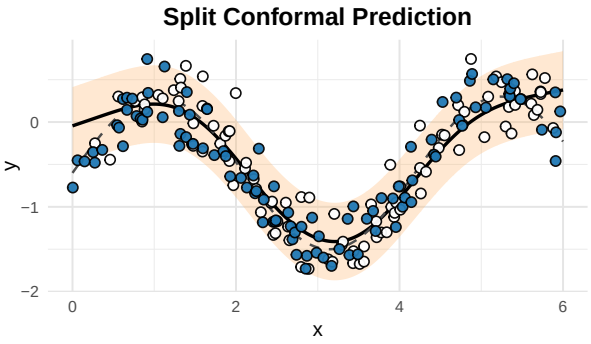

Figure 1. Split Conformal Prediction

Figure 1 shows the true underlying function (dashed line), the fitted smooth line and the shaded conformal band and the data points from the sets (training points are white, calibration points are blue).

Algorithm 1: Split Conformal Prediction Algorithm

Input: Training data $(X_i, Y_i)_{i=1}^n$, significance level α

Output: Prediction set $C_n(x)$ for a new input x

`// 1. Data splitting`

Split the training set into a proper training set D_1 of size n_1 and a calibration set D_2 of size n_2;

`// 2. Model fitting`

Fit predictor f_{n_1} using the data in D_1;

`// 3. Compute calibration residuals`

For each $i \in D_2$, compute

$$R_i = |Y_i - f_{n_1}(X_i)|$$

`// 4. Compute empirical` $(1-\alpha)$ `quantile`

Let q_{n_2} be the $\lceil (1-\alpha)(n_2+1) \rceil$-th smallest value among $\{R_i : i \in D_2\}$:

$$q_{n_2} = R_{(\lceil (1-\alpha)(n_2+1) \rceil)}$$

`// 5. Construct prediction set`

For any new input x, define

$$C_n(x) = \left[f_{n_1}(x) - q_{n_2},\ f_{n_1}(x) + q_{n_2} \right]$$

return $C_n(x)$;

I generated $n = 200$ data points[2] where the inputs are

$$x_i \sim \text{Uniform}(0, 6),$$

sorted in increasing order. The underlying regression function

$$f(x) = 0.9 \sin(1.5x) - 0.6.$$

I added independent noises, which satisfy

$$\varepsilon_i \sim \mathcal{N}(0,\ 0.25^2),$$

so the observed responses can be written in the form of

$$y_i = f(x_i) + \varepsilon_i.$$

The training and calibration sets are divided equally. A smoothing spline with 6 degrees of

[2] AI was used to generate parts of the code.

freedom is fitted on the training data and the residuals are calculated on the calibration set. The coverage level is 90% for the prediction band.

The naive prediction band typically under-covers, often when the fitted model overfits the training data. In contrast, the split conformal band is more resistant to overfitting because it compares test scores with those of a separate calibration set.

In the regression setting, with prediction sets

$$C_n(x) = [\, f_n(x) - q_n,\ f_n(x) + q_n \,],$$

and residuals $R_i = |Y_i - f_n(X_i)|$, the **empirical coverage** over a test set $\{(X_j, Y_j)\}_{j=1}^m$ is

$$\frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{\, |Y_j - f_n(X_j)| \le q_n \,\}.$$

If exchangeability holds and the sample sizes are large enough, the empirical coverage is close to the nominal level however, model overfitting can make it fluctuate. We can also define the **width** of prediction interval for the i-th observation:

$$W_i = \hat{y}_i^{\text{upper}} - \hat{y}_i^{\text{lower}}$$

where

$$\hat{y}_i^{\text{lower}} = \hat{y}_i - q_{1-\alpha}, \quad \hat{y}_i^{\text{upper}} = \hat{y}_i + q_{1-\alpha}.$$

For all test points i = 1, ... , n_{test}:

$$\mathcal{C}_i = [\hat{y}_i - q_{1-\alpha},\ \hat{y}_i + q_{1-\alpha}]$$

Thus, the (constant) interval width is:

$$W = 2q_{1-\alpha}$$

This lack of adaptivity is usually unfavorable, the band cannot adjust to the varying difficulty of prediction at different regions of the feature space.

In the final stage, we compare different machine learning methods for regression based on their empirical coverage and interval width to evaluate the performance of conformal prediction across models.

6. Synthetic Data Generation

We generate a synthetic linear regression dataset[3] with $n = 600$ observations and $p = 10$ predictors. Let

$$X \in \mathbb{R}^{n \times p},$$

where the entries are independent, and

$$X_{ij} \sim \mathcal{N}(0, 1).$$

The true regression coefficient vector is

$$\beta = (1.5,\ -1,\ 0.5,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)^\top.$$

Thus, the response variable $y \in \mathbb{R}^n$ is

$$y = X\beta + \varepsilon,$$

So,

$$y_i = 1.5\, X_{i1} - X_{i2} + 0.5\, X_{i3} + \varepsilon_i$$

where the noises are i.i.d.

$$\varepsilon_i \sim \mathcal{N}(0, 1).$$

Then, I split the data equally to train, calibration and test sets.

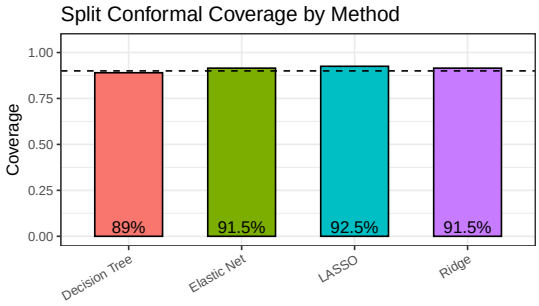

Figure 2. Coverage

Figure 3. Test Residuals

[3] I used AI to help generate parts of the code.

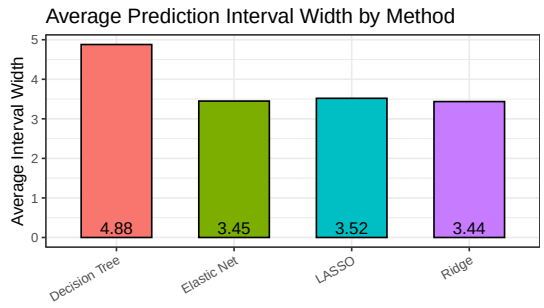

Figure 4. Width

For LASSO, Ridge, and Elastic Net point predictors, the regularization parameter λ is chosen by *cv.glmnet* for the best modell fit. The elastic Net here is a 50-50% combination of LASSO and Ridge. The Decision Tree, which is a Regression Tree here, is created using *rpart* package in R with default parameters (minsplit=20, maxdepth=30, cp=0.01).

Figure 2 shows similar coverage across methods, all close to 90 %. Coverage is controlled by the split conformal prediction.

The test residual distribution is visible in **Figure 3**, showing a boxplot. The horizontal line represents the median, the box captures the middle 50% of the data, the vertical line the middle 75%. The short box and median close to zero mean low residual variance.

Contrast to coverage, the interval width is different (**Figure 4**) - linear regularized methods give narrower band; Decision Tree, however, is producing a wider one compared to the other three. Width actually reflects model accuracy, this means that LASSO, Elastic Net, and Ridge are good model fit, while the performance of Decision Tree is poorer. This is due to the fact that Trees are piecewise constant approximators, not as smooth as the others. Ridge shrinks all coefficients, while LASSO can eliminate some, which produces smaller residuals. Elastic Net is the combination of the last two, so it is straightforward that its behaviour is between them.

References

[1] Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Boston, MA: Springer US, 2005.

[2] Tibshirani, Ryan. "Conformal prediction." UC Berkeley (2023).

[3] Shafer, Glenn, and Vladimir Vovk. "A tutorial on conformal prediction." Journal of Machine Learning Research 9.3 (2008).

[4] Fontana, Matteo, Gianluca Zeni, and Simone Vantini. "Conformal prediction: a unified review of theory and new challenges." Bernoulli 29.1 (2023): 1-23.

[5] Marques, F., and C. Paulo. "Universal distribution of the empirical coverage in split conformal prediction." Statistics & Probability Letters 219.C (2025).