Conformal Prediction Regression Cheat Sheet

By Gianfrancesco Angelini; v0.2

Main Components of Conformal Prediction

- A *predictive model* $M = M(x)$
- *Nonconformity function* $s(y, \hat{y})$ (e.g., $s(y, \hat{y}) = |y - \hat{y}|$).
- A *significance* level $\alpha \in [0, 1]$.
- A set of candidates for the *prediction interval* $y_j : s_j < s_\alpha$, where s_α is the $\lceil (1-\alpha)(n+1) \rceil$-th quantile of the set comprising all the nonconformity scores (calibration scores) computed according to the method used. In the regression case, candidates can come from a grid of y in the prediction's $\hat{y}$ surroundings, which excludes all y outside the calibration quantile.

Validity: Under exchangeability, C.P. methods guarantee a coverage probability of at least $1 - \alpha$ for the prediction set.

References

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.

[2] Vladimir Vovk. "Cross-conformal predictors". In: *Annals of Mathematics and Artificial Intelligence* 74 (2015), pp. 9–28.

[3] Yaniv Romano, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression". In: *Advances in neural information processing systems* 32 (2019).

[4] Vladimir Vovk et al. "Mondrian confidence machine". In: *Technical Report* (2003).

[5] Rina Foygel Barber and Emmanuel J. Candès. "Predictive Inference with the Jackknife+". In: *Annals of Statistics* 49.1 (2021), pp. 486–507.

Full Conformal Prediction [1]

Calibration:

1. Choose a nonconformity function $s(y, \hat{y})$ (e.g., $s(y, \hat{y}) = |y - \hat{y}|$).
2. Train your model M on the full dataset $D = \{(x_i, y_i)\}_{i=1}^n$.
3. Compute calibration scores: $s_i = s(y_i, M(x_i))$ for all $(x_i, y_i) \in D$.

Evaluation: For a new instance x_{new}, and for each candidate label y (e.g., all classes or a grid for regression):

1. Form the augmented dataset:
$$D^{(y)} = D \cup \{(x_{\text{new}}, y)\}.$$
2. (If needed) Refit the model $M^{(y)}$ on $D^{(y)}$ and compute updated scores $s_i^{(y)} = s(y_i, M^{(y)}(x_i))$.
3. Compute the new instance's score:
$$s_{\text{new}}^{(y)} = s\big(y, M^{(y)}(x_{\text{new}})\big).$$
4. Calculate the p-value:
$$p(y) = \frac{|\{i : s_i^{(y)} \geq s_{\text{new}}^{(y)}\}| + 1}{|D^{(y)}| + 1}.$$

Prediction Set: For significance level α,
$$C(x_{\text{new}}) = \{y : p(y) > \alpha\}.$$

Split Conformal Prediction

Core Idea: The dataset is split into a training set to train the model, and a calibration set to determine a quantile threshold for the nonconformity scores.

1. Split the available dataset $D = \{X, Y\}$ into a training set D_T and a calibration set D_C.
2. Train your predictive model M on the training set D_T.
3. Compute non-conformity scores $s_i(y_i, \hat{y}_i)$, on the calibration set D_C, for each x_i. E.g. for each example (x_i, y_i) where $i \in D_C$, calculate
$$s_i = |y_i - \hat{y}_i|,$$
where $\hat{y}_i = M(x_i)$.
4. Determine the quantile $\hat{q}$ of the calibration scores such that
$$\hat{q} = \frac{\lceil (|D_C| + 1)(1 - \alpha) \rceil}{|D_C|},$$
where $|D_C|$ is the number of examples in the calibration set.
5. Finally, for each x_{new} compute the *Confidence Interval* for x_{new},
$$C(x_{\text{new}}) = \{y_i : s_i(y_i, \hat{y}_i) < \hat{q}\}.$$
following the steps below:
   - 5.1 Choose a grid of y_j candidates centered on $\hat{y}_{\text{new}}$ and such that they recede left and right from $\hat{y}_{\text{new}}$.
   - 5.2 Compute $s_j(y_j, \hat{y}_{\text{new}})$.
   - 5.3 If $s_j < \hat{q}$, accept y_j in C_j; otherwise, reject it.

Cross Conformal Prediction [2]

Core Idea: The dataset is partitioned into k folds. For each fold, a model is trained on the remaining folds, and nonconformity scores are computed for the held-out fold. The aggregated scores from all folds produce a robust prediction set that utilizes the full dataset.

1. Split the dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ into k disjoint subsets (folds) $D_1, ..., D_k$.
2. For each fold $k = 1, ..., K$:
    1. Train a model M_k on the training set $D \setminus D_k$
    2. Compute the nonconformity scores for each example $(x_i, y_i) \in D_k$:
    $$s_i = |y_i - M_k(x_i)|$$
3. Aggregate all nonconformity scores into a single set $S = \{s_1, ..., s_n\}$.
4. Compute the empirical quantile $\hat{q}$ as:
    $\hat{q} =$ the $\lceil (n+1)(1-\alpha) \rceil$ -th smallest value in S
    where $n = |D|$ and α is the significance level.
5. For a new observation x_{new}:
    5.1 Generate a set of candidates $\{y_1, ..., y_m\}$ around $\hat{y}_{\text{new}}$
    5.2 For each candidate y_j:
    - Compute $s_j = |y_j - \hat{y}_{\text{new}}|$
    - Include y_j in the confidence interval if $s_j \leq \hat{q}$

Conformalized Quantile Regression [3]

Core Idea: Quantile regression is used to estimate initial prediction intervals, and a correction factor is derived from the calibration residuals to adjust these intervals, ensuring valid coverage.

1. Split the dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ into:
    - Training set D_{train}
    - Calibration set D_{cal}
2. Train two quantile regression models on D_{train}:
    $$\hat{q}_{\alpha/2}(x),\ \hat{q}_{1-\alpha/2}(x)$$
    where α is the desired error level.
3. Compute the nonconformity scores on the calibration set:
    $$E_i = \max\left\{\hat{q}_{\alpha/2}(x_i) - y_i,\ y_i - \hat{q}_{1-\alpha/2}(x_i)\right\}$$
    Set $E_i = 0$ if both terms are negative.
4. Determine the correction factor δ as:
    $$\delta = \text{Quantile}\left(\{E_i\}_{i \in D_{\text{cal}}},\ 1-\alpha\right)$$
    using the correctly adjusted empirical formula:
    $$\delta = \text{the } \lceil (1-\alpha)(1+|D_{\text{cal}}|) \rceil \text{-th value in } \{E_i\}$$
5. Construct the final prediction interval for a new x_{new}:
    $$C(x_{\text{new}}) = \left[\hat{q}_{\alpha/2}(x_{\text{new}}) - \delta,\ \hat{q}_{1-\alpha/2}(x_{\text{new}}) + \delta\right]$$

Mondrian Conformal Prediction [4], [1]

Core Idea: Partition the data into homogeneous groups (or categories) and perform conformal prediction within each group.

Procedure:

- **Define Groups:** Use known labels or cluster the feature space to form Mondrian categories.
- **Compute Scores:** For each group g, compute nonconformity scores
    $$s_i^{(g)} = s(y_i, \hat{y}_i), \quad \forall (x_i, y_i) \text{ in group } g,$$
    using a chosen nonconformity function (e.g., $s(y, \hat{y}) = |y - \hat{y}|$).
- **Group-Specific Quantile:** For a significance level α, determine the quantile $s_\alpha^{(g)}$ as the $\lceil (|D^{(g)}| + 1)(1-\alpha) \rceil$-th smallest score within group g.
- **Prediction Set:** For a new instance x_{new}, identify its group g and construct:
    $$C(x_{\text{new}}) = \{y : s(y, \hat{y}) \leq s_\alpha^{(g)}\}.$$

Jackknife+ [5]

Core Idea: Use leave-one-out predictions to construct valid prediction intervals with finite-sample guarantees.

1. **Leave-One-Out Training:** Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
   for each $i = 1, \ldots, n$:

   1. Train the model M_{-i} on $D \setminus \{(x_i, y_i)\}$.
   2. Compute the leave-one-out prediction for training:

   $$\hat{y}_{-i}(x_i) = M_{-i}(x_i).$$

   3. Compute the residuals:

   $$r_i = |y_i - \hat{y}_{-i}(x_i)|\,.$$

2. **Constructing the Prediction Interval for a New Observation $x_{\mathbf{new}}$:**

   2.1 Compute the leave-one-out predictions using the models trained before:

   $$\hat{y}_{-i}(x_{\text{new}}) = M_{-i}(x_{\text{new}}) \quad \text{for } i = 1, \ldots, n.$$

   2.2 Compute L and U:

   $$L = \{\hat{y}_{-i}(x_{new}) - r_i\}\big|_{i=1}^{n}$$

   $$U = \{\hat{y}_{-i}(x_{new}) + r_i\}\big|_{i=1}^{n}$$

   2.3 Compute the quantiles:
   $q^{-}_{n,\alpha}(L) = \lfloor \alpha(n+1) \rfloor$-th smallest value in L
   $q^{+}_{n,\alpha}(U) = \lceil (1-\alpha)(n+1) \rceil$-th smallest value in U

   2.4 Construct the interval bounds using quantiles:

   $$L_\alpha = q^{-}_{n,\alpha}(L),$$

   $$U_\alpha = q^{+}_{n,\alpha}(U).$$

   2.5 The prediction interval is given by:

   $$C(x_{\text{new}}) = [L_\alpha,\ U_\alpha]\,.$$