Conformal Prediction - Extensions and practical applications

April 3rd, 2025

Extensions of Conformal Prediction

We try extending the notion of conformal prediction to more general and practical scenarios and see what additional constraints or problems would have to deal with in it's implementation.

Group-Balanced Conformal Prediction

As an extension to conditional coverage for this method that we highly desire, in the more general setting, we require certain groups in specific to satisfy the coverage guarantee at least. Eg. medical reports predicted from different ethinicities or groups of people.

Formally, Let the inputs $X_{i,1},\ i = 1, \dots, n$, take values in a discrete set $\{1, \dots, G\}$, corresponding to categorical groups. We then require group-balanced coverage:

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}) \,|\, X_{\text{test},1} = g\right) \geq 1 - \alpha,$$

for all groups $g \in \{1, \dots, G\}$.

Note: These groups may be partitioned after the calibration. Since we cannot guarantee the conditional coverage in each of the groups, we make a slight change to our approach : We run the conformal prediction over each of the group separately.

We do this by computing the conformal quantile for each of the groups within themselves and then running the conformal algorithm. Let $n(g)$ denote the number of calibration examples belonging to group g. We define the group-specific quantile $q(g)$ as:

$$q(g) = \text{Quantile}\left(s_1, \dots, s_{n(g)}; \lceil (n(g) + 1)(1 - \alpha) \rceil / n(g)\right),$$

where s_i are the conformity scores of the calibration points in group g.

Finally, we form the prediction set $\mathcal{C}(x)$ by selecting the appropriate quantile based on the group membership of x:

$$\mathcal{C}(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}(x_1)\},$$

where x_1 denotes the group attribute of x, and $\hat{q}(x_1)$ is the threshold corresponding to that group.

This $\mathcal{C}$ satisfies the required coverage property across the various groups that we desire.

Class-Conditional Conformal Prediction

Since conditional coverage is not achievable, we can try to achieve marginal coverage over every class as a subset. This process of achieving class-conditional coverage requires slight modifications compared to regular conformal prediction methods.

As seen in evaluation methods, we can split the calibration data set into groups where each group corresponds to a particular class.

For a given class k, we can define the score of the ith occurrence in this class as

$$s_i^k = s(X_j, y_j)$$

This means that after splitting into groups, the jth data point is the ith data point in the kth class.

Now for each class we can define the quantile to be

$$\hat{q}^k = \frac{\lceil (n^k+1)(1-\alpha) \rceil}{n^k} \text{ Quantile of the set} \{s_1^k, s_2^k, ...s_{n^k}^k\}$$

where n^k is the number of data points in the kth class.Now we can define the prediction set to be

$$C(x) = \{y : s(x, y) \leq \hat{q}^y\}$$

The proof of why this choice of prediction set satisfies marginal coverage can be found in [3, 5]

Conformal Risk control

' For most machine learning problems, we don't necessarily bound our generalization loss just by the expression :

$$E[1_{Y_{test} \notin C(X_{test})}]$$

But we may use a general loss function on this value. Thus we would want a guarantee of the form :

$$\mathbb{E}\left[\mathbb{I}\left\{Y_{\text{test}} \notin C(X_{\text{test}})\right\}\right] \leq \alpha$$

We show that conformal prediction in general can even provide this guarantee for some general and arbitrary bounded loss function l if we tweak our algorithm a bit.

In practice, we often apply a post-processing step to a base model f in order to construct a prediction set $C_\lambda(\cdot)$. The parameter λ influences the expressiveness of the prediction model - typically a larger value of λ gives a bigger prediction set.

To evaluate the performance of such prediction sets, we consider a loss function $\ell(C_\lambda(x), y)$, which is bounded above by a constant $B < \infty$ and is non-increasing in λ. That is, increasing λ (making the prediction set more conservative) does not worsen the loss.

We define the empirical risk over a calibration set of size n as:

$$\hat{R}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \ell(C_\lambda(X_i), Y_i).$$

To ensure valid risk control, we select the smallest value of λ, denoted $\hat{\lambda}$, such that the empirical risk is at most a slightly adjusted threshold:

$$\hat{\lambda} = \inf \left\{ \lambda : \hat{R}(\lambda) \leq \alpha - \frac{B - \alpha}{n} \right\}.$$

This adjustment makes the procedure slightly more conservative than simply using the threshold α, which accounts for finite-sample effects. For example, when $B = 1$, $\alpha = 0.1$, and $n = 1000$, the threshold becomes 0.0991 instead of 0.1, thus promoting a more robust selection of $\hat{\lambda}$.

Outlier Detection

There are several distance-based methods, clustering-based methods, density based methods, etc. to detect outliers in unsupervised learning. Outliers are points that do not belong to the same dataset and enter the dataset due to some error in measurement or some other reason.

These methods associated in detecting outliers have some uncertainty involved and we can use conformal prediction to quantify this uncertainty.

Using conformal prediction, the main goal is to reduce the False positives because we wouldn't want to loose training data by considering it to be an outlier.

The procedure is not different at all from the standard methods:

1. Assume we have a model for outlier detection with the criteria that a **Large Score from model $\implies$ Outlier**. This basically means that we have a negatively oriented score function.
2. Use this score function $s_i = s(X_i)$ on the calibration data set to get the sorted score values.
3. Now define $\hat{q}$ to be the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$
4. For every test point, we define the prediction (*Note that here we are making predictions and not prediction set*)
$$C(x) = sign(s(x) - \hat{q})$$
Where $C(x) \leq 0 \implies$ inlier and $C(x) > 0 \implies$ outlier.

Proof of marginal coverage

Refer [6, 1, 2]

Conformal Prediction Under Covariate Shift

In all the methods, the score values are computed on the calibration data and they are simply extended to the test set with the assumption that the probability distribution of the calibration data set is same as the probability distribution of the test data set.

However this might not always be true. **Covariate Shift** is a special case where the distribution of (X_{test}, y_{test}) is different from (X_i, y_i) however the distribution of $X_{test}|y_{test}$ is **EQUAL** to the distribution of $X_i|y_i$.

To get better results, we can use weighted conformal prediction as in [4]. The basic intuition is that we now have weights corresponding to each data point and we would increase the weights of those points which are more likely to be in the test set distribution and reduce the weights of those which are less likely.

A simple choice of weight function for this would be related to the likelihood ratio:

$$w(x) = \frac{f_{test}(x)}{f(x)}$$

Where $f_{test}(x)$ is the probability density of the test data set and $f(x)$ is the probability density function of the train data set. Now we can define the updated weights as:

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(x)}$$

Now we do the usual adaptive conformal prediction based on cumulative sums of scores where the scores are going to be re-calculated as follows:

$$\hat{q}(x) = min(s_j : \sum_{i=1}^{j} p_i^w(x) s_i(x) \geq 1 - \alpha)$$

Note that $\hat{q}(x)$ is now dependent on the test data point we pick.

Finally define the prediction set as

$$C(x) = \{y : s(x, y) \leq \hat{q}(x)\}$$

Proof of marginal coverage in [4]

Full Conformal Prediction

In split conformal prediction, we train the model only once over a *Proper Training Data Set* with a tradeoff on the size of this dataset because we are splitting the given train data set into *Proper Training Data Set* and *Calibration Data Set.* Since data is expensive, this is one drawback of split conformal prediction methods.

Full Conformal prediction uses the entire training data set with the tradeoff on execution time. This method involves training the model n times where n is the size of the training data set.

Full conformal prediction like split conformal prediction also requires only exchangeable data points as a requirement. A standard form of full conformal prediction algorithm is:

1. For each $y \in Y$ we fit a new model $\hat{f}^y$ on the data set.
2. For each y we compute a score function of the form
$$s_i^y = s(X_i, Y_i, \hat{f}^y)$$
and let $S^y = \{s_i^y\}$
3. For every y, we compute a quantile point prediction $\hat{q}^y$ as the quantile of its corresponding score array S^y.
4. Now for any new X_{test} we can define the prediction set as
$$C(X_{test}) = \{y : s^y(X_{test}, y, \hat{f}^y) \leq \hat{q}^y\}$$

Apart from this general method, there are many specific methods as described below.

0.1 Jackknife

0.2 Jackknife+

0.3 Jackknife Minimax

0.4 CV

0.5 CV+

0.6 CV Minimax

Will be filled with reference to the jackknife paper by Tibshirani

References

[1] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia. Testing for outliers with conformal p-values. *arXiv:2104.08279*, 2021.

[2] L. Guan and R. Tibshirani. Prediction and outlier detection in classification problems. *arXiv:1905.04396*, 2019.

[3] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114:223–234, 2019.

[4] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32*, pages 2530–2540, 2019.

[5] V. Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25, pages 475–490, 2012.

[6] V. Vovk, I. Nouretdinov, and A. Gammerman. Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 768–775, 2003.