Conformal Prediction

Adithya K Anil, Rolla Siddharth Reddy, Pasupuleti Dhruv Shivkant, Nikhil Jamuda

March 28th, 2025

Summary

Why we need Conformal Prediction?

Machine Learning models are being extensively used in a lot of areas and that includes fields which involve a lot of risk. In such risky scenarios, it becomes a necessity to quantify the uncertainty involved in every prediction made by the predictor/prediction algorithm. This allows to make safer decisions rather than having to blindly trust the model. Conformal Prediction is a framework that provides a way to quantify the uncertainty involved in the predictions made by a model.

Conformal Prediction is distribution-free. What does that mean?

This means that conformal prediction doesn't rely on any properties of the underlying distribution of the data nor does it rely on any properties of the model. Also, it is not any approximation / asymptotic analysis of the model. It is backed by statistics and can be used with any model and any underlying distribution of the data.

Why we use Bernstein Quantile Estimator and not Naive Quantile Estimator?

Since we are using the score function to create an ordering among the residuals obtained from the Callibration data, we can use some properties of β distribution because order statistics are always distributed. From this we get that the the predictor $X_{\lceil n(1-\alpha)\rceil}$ underestimates the $1-\alpha$ quantile of the distribution of the residuals. So we take a slightly larger value of α to get a better estimate of the quantile. This is the reason we use the Bernstein Quantile Estimator.

Given a sample of size n, the empirical quantile estimator is chosen from the order statistics:

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

where $X_{(k)}$ represents the k-th smallest value in the sorted dataset.

Naive Quantile Estimation A simple way to estimate the $(1-\alpha)$-quantile is to select:

$$k = \lceil n(1-\alpha) \rceil$$

However, this underestimates the true quantile because the expected position of the order statistic is actually:

$$F^{-1}\left(\frac{k-0.5}{n}\right)$$

which is slightly lower than $F^{-1}(1-\alpha)$.

Bias Correction To correct this, we use:

$$k = \lceil (n+1)(1-\alpha) \rceil$$

This adjustment aligns the estimator with the expected quantile value and reduces bias.

Beta Distribution and Order Statistics The order statistic $X_{(k)}$ follows a Beta distribution with:

$$P(X_{(k)} \leq x) = I_x(k, n-k+1)$$

where I_x is the regularized incomplete Beta function. Using $k = (n+1)(1-\alpha)$ aligns the expectation correctly with the true quantile.

For further analysis, please read about the Beta Distribution and Order Statistics and Bernstein Quantile Estimator.

Examples of Conformal Prediction

Adaptive Prediction Sets

We now understand that conformal prediction aims at creating statistically rigorous uncertainity sets for black box prediction models and finds its greatest use in high risk settings such as the field of medicine. The advantage of conformal prediction is that it gives us valid (with respect to coverage) sets/conformal bands that are distribution free in the sense that the guarantees hold in a non-asymptotic manner without any underlying assumptions about the data or the model.

The following is the algorithm for split conformal prediction in a general setting;
1. Split the data into a train set and calibration set and train the blackbox model on the train set. Identify heuristic notion of uncertainity present in the output of the pre-trained (blackbox) model.
2. Define a score function. Although technically this can be any function of x and y, it is practically found that the more informative (say larger scores correlate with worser agreement between model prediction and the ground truth) score functions make the technique more powerful.
3. Compute $\hat{q}$ as the $\lceil (n+1)(1-\alpha) \rceil$ quantile of the calibration scores on the calibration set (X_i, Y_i), $i \in [N]$.
4. Form the prediction set corresponding to X_{test} as $\hat{C}(X_{test}) = \{y : s(X_{test}, y) \leq \hat{q}$.

There is a theoretical guarantee that following the above recipe to get the prediction band proviides the requisite coverage of $1-\alpha$. Further if it is known that ties occur among the scores of the calibration points with zero probability then we can upper bound $\mathbb{P}(Y_{test} \in \hat{C}(X_{test}))$ by $1 - \alpha + (1/(n+1))$. For proof of this refer paper1.

1 Easy example

Consider the scenario where we wish to apply conformal prediction to multi-category calssification as - Formally, suppose we have images as input and they each contain one of K classes. We begin with a classifier $\hat{f}$ that outputs estimated probabilities (softmax scores) for each class. This is our blackbox model that we want to quantify uncertainity on. Running the above algorithm with conformal scores as $s_i = 1 - \hat{f}(X_i)_{Y_i}$ gives us the marginal coverage property, i.e. $\mathbb{P}(Y_{test} \in \hat{C}(X_{test}) \geq 1 - \alpha$. Our heuristic notion of uncertainity in this case comes out to be the softmax scores.

2 An improvisation - APS

Although the method employed in the previous example gives us the desired marginal coverage and produces the smallest conformal band on average, it doesn't give us any guarantee on the conditional

coverage! (i.e tends to undercover harder examples and overcover easier examples). The conditional coverage guarantee refers to the probability $\mathbb{P}(Y_{test} \in \hat{C}(X_{test})|X_{test} = x) \geq 1 - \alpha$.

Why do we care about conditional coverage?
With $\alpha = 0.05$, we expect that the doctor's statement ("*...you can expect your blood pressure to go down by 10–15mmHg*") should hold with 95% probability. For marginal coverage, the probability is taken over both X_{n+1} and Y_{n+1}, while for conditional coverage, X_{n+1} is fixed and the probability is taken over Y_{n+1} only (and over all the training data in both situations).

This means that for marginal coverage, the doctor's statements have a 95% chance of being accurate on average over all possible patients that might arrive at the clinic (marginalizing over X_{n+1}), but might, for example, have 0% chance of being accurate for patients under the age of 25, as long as this is averaged out by a higher-than-95% chance of coverage for patients older than 25.

The stronger definition of conditional coverage, on the other hand, removes this possibility, and requires that whatever statement the doctor makes (different for each patient) has a 95% chance of being true for every individual patient, regardless of the patient's age, family history, etc.

Idea behind APS?
The APS tries to approximate the oracle classifier as described in the paper - 'Y. Romano, M. Sesia, and E. J. Cand‘es, "Classification with valid and adaptive coverage," arXiv:2006.02544, 2020' since to find the oracle classifier we need to know the probability distribution of Y given X.

To proceed, we define a score function inspired by the oracle algorithm as follows:

$$s(x, y) = \sum_{j=1}^{k} \hat{f}(x)\pi_j(x), \quad \text{where} \quad y = \pi_k(x),$$

where $\pi_j(x)$ represents the predicted probabilities for each class j, and the sum is taken over the classes until the true label is reached. This score utilizes the softmax outputs of all classes, not just the true class, and is a modification of the previous score from Section 1.

Next, as in all conformal procedures, we define the quantile $\hat{q}$ as follows:

$$\hat{q} = \text{Quantile}(s_1, \ldots, s_n; \lceil (n+1)(1-\alpha) \rceil / n),$$

where $s_1, \ldots, s_n$ are the scores computed for the calibration set, and α is the significance level (e.g., $\alpha = 0.05$).

Finally, the prediction set $C(x)$ is defined as:

$$C(x) = \{\pi_1(x), \ldots, \pi_k(x)\},$$

where k is the largest index k' such that

$$\sum_{j=1}^{k'} \hat{f}(x)\pi_j(x) < \hat{q}.$$

Thus, the prediction set $C(x)$ is the set of classes $\pi_1(x), \ldots, \pi_k(x)$, with k defined as

$$k = \sup \left\{ k' : \sum_{j=1}^{k'} \hat{f}(x)\pi_j(x) < \hat{q} \right\} + 1.$$

This procedure ensures that the prediction set includes the classes whose scores sum up to a value less than or equal to the quantile $\hat{q}$, and the class corresponding to the true label is included.

Quantile Regression

Just like how we accounted for a theoretical guarantee in classification problem, we can also incorporate a theoretical guarantee while doing regression predictions as well.

Outline of the procedure of prediction:

1. Firstly we aim to create 2 point predictor models $\hat{t}_{\frac{\alpha}{2}}$ and $\hat{t}_{1-\frac{\alpha}{2}}$. The intuition behind this is that let's say we know the underlying distribution of the dataset and we have the true value of the α quantile and $1-\alpha$ quantile. Denote this by $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$.

   Now we know from statistics that

   $$\mathbb{P}(x \in [t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]) = 1-\alpha$$

   This probabilistic guarantee motivates us to create the point predictors for these quantiles.

2. To create such a point predictor for the quantile, we can construct a normal regression problem with an alternative loss function called the **Quantile Loss/Pinball Loss** which is given by

   $$L_\gamma(\hat{t}_\gamma, y) = (y-\hat{t}_\gamma)\gamma 1_{y>\hat{t}_\gamma} + (\hat{t}_\gamma - y)(1-\gamma)1_{y\leq\hat{t}_\gamma}$$

   This can be alternatively written as

   $$L_\gamma(\hat{t}_\gamma, y) = max\{(\hat{t}_\gamma - y)(1-\gamma), \gamma(y-\hat{t}_\gamma)\}$$

   An interesting observation is that on setting $\gamma = 0.5$, we get the standard MSE Loss function

3. Once we have the point predictors, we need a "good" score function so as to get a tight fit on the interval in which the predicted value can lie with a theoretical guarantee. To do so, we define $s(x,y)$ as

   $$s(x,y) = max\{\hat{t}_{\frac{\alpha}{2}} - y, y - \hat{t}_{1-\frac{\alpha}{2}}\}$$

   where $x \in$ Calibration dataset and y corresponds to its value.

4. Now that we have the scores defined for each data point in the calibration data set, we do the same procedure as what we would do in case of split conformal prediction which is to order the scores and set

   $$\hat{q} = \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \text{ observation of the set} \{s_1, s_2, s_3, s_4, s_5 ... s_n\}$$

5. Finally for any x_{test}, we define the interval in which the prediction can lie as

   $$C(x) = [\hat{t}_{\frac{\alpha}{2}}(x) - \hat{q}, \hat{t}_{1-\frac{\alpha}{2}}(x) + \hat{q}]$$

This is just one way to arrive at a continuous interval and there are definitely many other ways however, there some advantages as to why Quantile Regression is widely used.

- The interval $[t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]$ by itself gives a good enough coverage by itself.
- This method gives an *asymptotically valid conditional coverage.*

  Conditional Coverage gives us guarantee for any x_{test} we pick and on mathematically formalizing it, we get that

  $$\mathbb{P}(y_{test} \in C(x_{test}) | x_{test}) \geq 1-\alpha$$

  whereas marginal coverage gives us that on an average among all x_{test} we have that

  $$\mathbb{P}(y_{test} \in C(x_{test})) \geq 1-\alpha$$

- The quantile loss function can be very easily modified to be incorporated in any regression model. In fact, the the standard MSE is just a specific case of the quantile loss function so we can easily interconvert between MSE Loss and Quantile Loss function

Conformalizing Scalar Uncertainty Estimates

This method rather than trying to create a confidence interval like quantile regression is trying to create an uncertainty interval around the point predictor so that the coverage guarantee still holds.

So in this type, we have a point predictor $\hat{f}(x)$ and an uncertainty predictor $\hat{u}(x)$. This function $\hat{u}(x)$ could represent standard deviation, variance, residuals or any type of function that is *negatively oriented*. This $\hat{u}(x)$ should be a good estimate of the function $u(x)$ it is trying to estimate so in reality, we would do some regression model on this parameter as well.

Now that we have these two predictor functions, we can define the score function for the calibration data as :

$$s(x, y) = \frac{|y - \hat{f}(x)|}{\hat{u}(x)}$$

Once we have this score function, we do the same procedure as sorting them in ascending order, getting the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile and setting it to $\hat{q}$.

Once we have found out $\hat{q}$, for any x_{test}, we can define the prediction interval to be

$$C(x) = [\hat{f}(x) - \hat{u}(x)\hat{q}, \hat{f}(x) + \hat{u}(x)\hat{q}]$$

This prediction interval gives us the same coverage guarantee.

Remark:
It is preferred to use quantile regression over uncertainty estimates because in quantile regression the predictors $\hat{t}_{\frac{\alpha}{2}}$ is directly related to the significance level α whereas in uncertainty estimates, both the predictors $\hat{f}(x), \hat{u}(x)$ are not related to α and hence we are not giving more importance to this parameter as much as we do in quantile regression.

In practice, quantile regression seems to perform slightly better than uncertainty estimation methods and this is has been cited in one of Angelopoulos' paper.

Nevertheless, uncertainty estimates is very easy to implement and very intuitive to understand, so it is still widely used.

Conformalizing Bayes

In this section we will cover conformal prediction using Bayesian Models. Bayesian predictors such as Bayesian Neural Network, which is nothing but extended version of NN with posterior inference inorder to control overfitting of data, usually they have a probability distribution on model parameters(weights and biases), usually studied in uncertainty quantification which rely on unverifiable and incorrect assumption which is not preferable to use in real word. Thus we use conformal prediction using the Bayesian model to create a prediction set that are Bayesian optimal(having the lowest risk) among all the prediction set that achieves $1 - \alpha$ coverage.

Similar to how the algorithm was performed with score function as the (1 - softmax score) of discrete classes, a simple change here is to replace the score function with continous probability distribution of classes given a test value, we'll show the procedure in the following,

- Given a Bayesian Model $\hat{f}(y \mid x)$ which is the posterior probability distribution of Y_{test} at label y given $X_{test} = x$. Under assumption of model being correctly specified (i.e., it perfectly represents the true data distribution) and we have a large enough dataset (where $n \to \infty$), then we can use this posterior to construct the best possible prediction set, given as:

$$S(x) = \left\{ y : \hat{f}(y \mid x) > t \right\}, \text{where } t \text{ is choosen s.t. } \int_{y \in S(x)} \hat{f}(y \mid x) dy = 1 - \alpha$$

holds. But our whole conformal prediction relies on no assumption of model or data. But we can take $\hat{f}$ as heuristic notion of uncertainity(described in the algorithm before).

- Thus now we can define our score function as:

$$s(x, y) = -\hat{f}(y \mid x)$$

This is a valid score function as it penalizes more when model is uncertain(gives low values).

- Using the same procedure for finding $\hat{q} = \lceil (1-\alpha)(n+1) \rceil / n$ quantile of score functions from calibration set. From here we get the prdiction set $\mathcal{C}(X_{test} = x)$ as:

$$\mathcal{C}(x) = \{y : s(x, y) \leq q\}, \quad \text{which is equivalent to saying}$$
$$\mathcal{C}(x) = \left\{y : \hat{f}(y \mid x) > -\hat{q}\right\}.$$

We can clearly see that this set gives the $1 - \alpha$ coverage and hence a valid set.

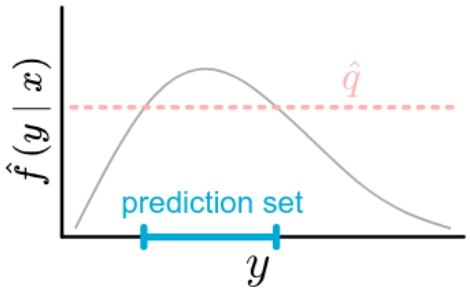

Figure 1: Prediciton set after Conformalizing Bayes

Above we can see how we narrowed the prediction set from some discrete values probabilities with Bayesian. What good about this is it gives the minimum average(over data and parameters) size of predicton set over any procedure which gives $1 - \alpha$ coverage with Bayes optimality, but under certain assumption that model gives correctly specified statistics and we have a large enough data set.

Proof Of Coverage

Here's we'll give the proof for the theorem of coverage upon which the whole Conformal Prediciton is relied on. The theorem can be split into two parts with, for lower and upper bound. Below we give the proof for the lower bound.

Theorem 1 Suppose we have $(X_i, Y_i)_{i=1,\ldots,n}$ and (X_{n+1}, Y_{n+1}) are *i.i.d.*, then we define,

$$\hat{q} = \inf\left\{q : \frac{|\{i : s(X_i, Y_i) \leq q\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\}.$$

and the prediciton set as,

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

Then we can say,

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

(Remarks: The above theorem is still valid for data being exchangeable, as we'll see)

Proof: First we define the score function $s_i = s(X_i, Y_i)$ for i = 1,. . . ,n and $s_{n+1} = s(X_{n+1}, Y_{n+1})$. For simplicity, we assume that there are no ties among the score function, i.e. they all are distinct with probability 1.

Now without loss of generality we assume that the score functions are sorted in order $s_1, s_2, \ldots, s_n$. Now we have two cases of alpha which decides the value of $\hat{q}$,

$$\hat{q} = \begin{cases} s_{\lceil (n+1)(1-\alpha) \rceil}, & \text{if } \alpha \geq \frac{1}{n+1}, \\ \infty, & \text{otherwise.} \end{cases}$$

The reasoning behind this choice is:

- When $\alpha \geq \frac{1}{n+1}$, the index $\lceil (n+1)(1-\alpha) \rceil$ corresponds to a valid position within the sorted calibration scores. Thus, we set $\hat{q}$ to the corresponding score to ensure the coverage probability requirement is met.
- When $\alpha < \frac{1}{n+1}$, we observe that:

  $$(n+1)(1-\alpha) > n \Rightarrow \lceil (n+1)(1-\alpha) \rceil = n+1.$$

  Since there are only n calibration scores available, the $(n+1)$th smallest score does not exist. In this case, we set $\hat{q} = \infty$, meaning the prediction set includes all possible labels $\mathcal{Y}$, ensuring 100% coverage.

Hence 2nd case is trivially satisfied, so we only have to deal with first case when $\alpha \geq \frac{1}{n+1}$. Now we can see that the following two sets are equivalent,

$$\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} = \{s_{n+1} \leq \hat{q}\}$$

Thus from the definition of $\hat{q}$ we get,

$$\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} = \left\{s_{n+1} \leq s_{\lceil (n+1)(1-\alpha) \rceil}\right\}$$

Now that we assumed *i.i.d.*(same holds for exchangeablity) se can say that for a given integer k, the probability,

$$\mathbb{P}(s_{n+1} \leq s_k) = \frac{k}{n+1}$$

which simply means that s_{n+1} has equal probability to fall anywhere between $s_1, s_2, \ldots, s_n$ with s_{n+1} where it can be added to the sorted indices in the position k or less than that.

Using the above result we can say that

$$\mathbb{P}(s_{n+1} \leq s_{\lceil (n+1)(1-\alpha) \rceil}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1 - \alpha.$$

Which was the desired result.

Now we will give the upper bound of the proof.

Theorem 2 When the score functions defined above have ties with probability zero(or continuous distribution) , we have,

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

Evaluation of conformal prediction

To confirm with the theoretical guarantees provided by **conformal prediction**, we take into account some formalizations to test and report how this behaves on real world data when implemented. This provides evaluation criteria for each of the models for conformal predictions described above.

- Evaluating Adaptivity - Adaptivity in general is not implied directly by conformal prediction. The conformal prediction only provides a coverage guarantee as described initially. But when implementing in a practical scenario, we need and require adaptivity of the prediction model.
- Evaluating correctness checks - Empirical checks to ensure practical coverage guaranteed by conformal prediction.

Note : Running evaluations is a very expensive task computationally since it requires empirical outcomes which need to be run over a large number of splits of the dataset to observe a proper value.

Adaptivity evaluation

Set sizes

When we analyze the predicted set sizes, if we obtain a large average set size, it in general indicates that the conformal procedure is not very precise.

The problem in general could be :

- Bad score function
- Bad underlying "point predictor" model

Normally a widespread distribution of these sets is generally acceptable and desirable. Although just this does not guarantee anything but just gives a rough notion of why it could be a proper adaptation.

Conditional Coverage

The initial coverage guarantee discussed above was - marginal coverage property. This intuitively means that over the entire set of datapoints, we achieve the bounds mentioned in the original conformal guarantee.

However in practice we intend to achieve a much stronger notion of prediction guarantee - **conditional coverage**. This explicitly means that for every value of the input X_{test}, we seek to return a prediction set with $1 - \alpha$ coverage guarantee.

Note : In the general case, conformal prediction is almost impossible to achieve, hence we need to find a heuristic to approximate it.

Intuitive example : Imagine there are two groups of people, group A and group B, with frequencies 90% and 10%. The prediction sets always cover Y among people in group A and never cover Y when the person comes from group B. Then the prediction sets have 90% coverage, but not conditional coverage. Conditional coverage would imply that the prediction sets cover Y at least 90% of the time in both groups. This is necessary, but not sufficient; conditional coverage is a very strong property that states the probability of the prediction set needs to be $\geq 90\%$ for a particular person. In other words, for any subset of the population, the coverage should be $\geq 90\%$.

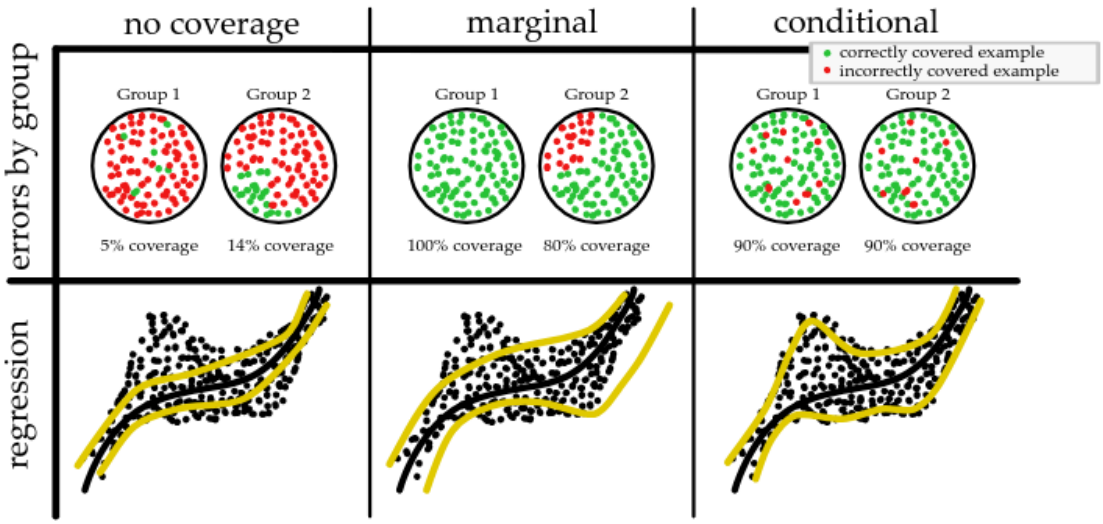

Figure 2: Example picture : credits : "gentle introduction paper"

Metrics

Mathematical formalization of what is described above.

- **Feature-stratified coverage metric** : Unequal coverage over different groups. groups are discrete features more or less.

Let $X_i^{(\text{val})}$ be features that take values in $\{1, \ldots, G\}$ for some G. Here, $i = 1, \ldots, n_{\text{val}}$ indexes the example in the validation set, and the first coordinate of each feature represents the group.

Let $I_g \subset \{1, \ldots, n_{\text{val}}\}$ be the set of observations such that $X_{i,1}^{(\text{val})} = g$ for $g = 1, \ldots, G$.

FSC metric:

$$\min_{g \in \{1,\ldots,G\}} \frac{1}{|I_g|} \sum_{i \in I_g} \mathbf{1}\left\{Y_i^{(\text{val})} \in \mathcal{C}\left(X_i^{(\text{val})}\right)\right\}$$

This is the observed coverage among all the instances when the discrete feature takes group g conditioned on the feature.

- **Size-stratified metric** : This is more general. It basically discretizes the possible cardinalities of the prediction set C(x) into different groups similarly as above. **SSC metric:**

$$\min_{g \in \{1,\ldots,G\}} \frac{1}{|I_g|} \sum_{i \in I_g} \mathbf{1}\left\{Y_i^{(\text{val})} \in \mathcal{C}\left(X_i^{(\text{val})}\right)\right\}$$

This is same as the previous one except that the I_g is defined to be a subset of the indices $\{1, 2, ...n_{val}\}$.

Effect of size of calibration set

The size of the calibration dataset is mainly the source of the finite-sample variability. Intuitively it feels that a larger isze would lead to a better conformation, however we need to quantize this notion.

Key idea : Coverage of conformal prediction conditionally on the calibration set is a random quantity.

Marginal coverage property guarantees $1 - \alpha$ coverage over the randomness in the calibration set. But if we fix one particular set, the coverage on a infinite validation set is not guaranteed to be $1 - \alpha$.

The distribution coverage in general has a form [Vladimir Vovk]:

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \{(X_i, Y_i)\}_{i=1}^n\right) \sim \text{Beta}\left(n + 1 - l, l\right),$$

$$l = \lfloor (n+1)\alpha \rfloor.$$

Correctness of coverage

Diagnostic to assess the procedure of conformal prediction when implemented. This is a very computationally heavy task. This involves R trials with different randomized splits of the calibration data and the calculating the mean of the empirical coverage for each of the runs.

$$C_j = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \mathbf{1}\left\{Y_{i,j}^{(\text{val})} \in C_j\left(X_{i,j}^{(\text{val})}\right)\right\}, \quad \text{for } j = 1, \ldots, R.$$

where n_{val} is the size of the validation set, $(X_{i,j}^{(\text{val})}, Y_{i,j}^{(\text{val})})$ is the ith validation example in trial j, and C_j is calibrated using the calibration data from the jth trial. A histogram of the C_j should be centered at roughly $1 - \alpha$. Likewise, the mean value,

$$\overline{C} = \frac{1}{R} \sum_{j=1}^{R} C_j,$$

This in general gives us a notion of how good our predictor has performed over the new test data and gives us the correlation between what the expected gurantee was and the empirical value.

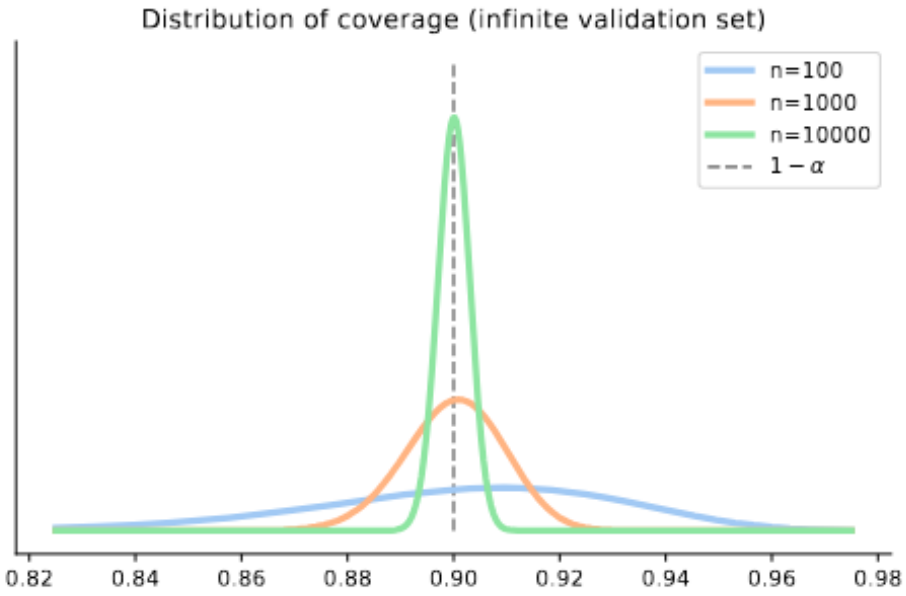

Figure 3: Beta distribution visualized for different values

What we implemented?

1. Adaptive Classification on the Fashion MNIST Dataset. We used multiple models and α to check the validity of the coverage guarantee.
2. Quantile Regression on the California Housing Dataset
3. Ran the codes present in the paper

Next target?

uncertainty prediction angelopoulos implemeniting