ulm university universität
uulm

University of Ulm | 89069 Ulm | Germany

Faculty of
Computer Science
Medical Systems
Biology

Conformal Prediction - A Basic Overview

Master Project Medical Systems Biology
Submitted for the Artificial Intelligence program.

Submitted by:
Michel Lutz
michel.lutz@uni-ulm.de
Matriculation number: 1048353

Project advisors:
Prof. Dr. Hans Kestler
Ulm University, Ulm

Supervisor:
Dr. Julian Schwab

Ulm, January 2024

Abstract

Contents

1. Introduction

Machine learning models have become more powerful and prominent in recent years and are increasingly being used in critical decision-making processes, for example in medical care. However, it is still unclear to what extent the predictions of these models can be trusted. Often these systems provide only point predictions with no indication about their quality, and even when they do, these are usually heuristic notations without any guarantee. Even if models make good predictions for most data points and demonstrating high accuracy, many systems tend to treat difficult and unusual data points with unjustified confidence. It is therefore impossible for users to distinguish bad from good predictions.

In order to make rational decisions on the basis of machine-generated predictions, a reliable uncertainty quantification is essential, which makes transparent in a valid way whether the system is certain of a prediction or whether it is wild speculation. Or in the words of Stanford University Prof. Emmanuel Candès: "Predictive models are used to make decisions that can have enormous consequences for people's lives. It's extremely important to understand the uncertainty about these predictions, so people don't make decisions based on false beliefs" (Candès, 2020).

1.1 Uncertainty Estimation

A meaningful uncertainty estimate should allow enable statements to be made about the credibility (how likely is the (point) prediction) and the confidence (how likely are the alternative predictions to the (point) prediction) of a system. This makes it possible to distinguish difficult data points with uncertain predictions from simpler ones with more certain predictions. In the interest of algorithmic fairness, it should also be possible to recognize whether the predictions of a system in a certain subset of data, e.g. health data of men and women, are subject to different degrees of uncertainty. This means that it must be made transparent whether, for example, there are greater uncertainties for female patients than for male patients.

Many machine learning models already have their own notation of their uncertainty by providing not only a point prediction but also, for example, a probability distribution over the entire label space. The softmax outputs of neural networks or the predictive posterior distributions of Bayesian models are typical examples. Other heuristic notations for the uncertainty of a system can be obtained by bootstrapping or additionally trained residual models. In models such as random forests, the variance between the trees can be interpret as a measure of their uncertainty. Nevertheless the main problem with all these approaches is that there is no reasonable guarantee that their predictive distributions are calibrated, i.e. that they actually deliver what they promise. So there is no guarantee that the prediction sets or intervals constructed on their basis contain the true labels with certainty. In other words they do not fulfill any coverage guarantee (Niculescu-Mizil & Caruana, 2005) (Lambrou, Papadopoulos, Nouretdinov, & Gammerman, 2012) (Manokhin, 2022b) (Dewolf, Baets, & Waegeman, 2023). For neural networks it is known that they are not well calibrated (Johansson & Gabrielsson, 2019) and for Bayesian approaches there is rarely a good reason to trust the proposed priors. In bootstrapping methods and residual models, it is known that they tend to underestimate the true variance (Hesterberg, 2015) (Manokhin, 2022b). To summarise in the words of Valery Manokhin a former stu-

dent of Vladimir Vovk: "many machine learning algorithms do not produce class membership probabilities and the ones that do often generate classification scores that do not correspond to class probabilities. In such cases the scores need to be transformed into well-calibrated probabilities that can be combined with utility scores for effective decision-making" (Manokhin, 2022b).

In the past, many approaches have been proposed to calibrate predictive distributions and thus to obtain a rigid notation for the uncertainty of a system. Conformal prediction (CP) is, to the best of my knowledge, the only framework that provides a model agnostic way to transform a heuristic notation for uncertainty into a rigid one in both classification and regression contexts (Figure 1. In other words, the output of a conformal method has a probabilistic guarantee that it covers the true outcome.

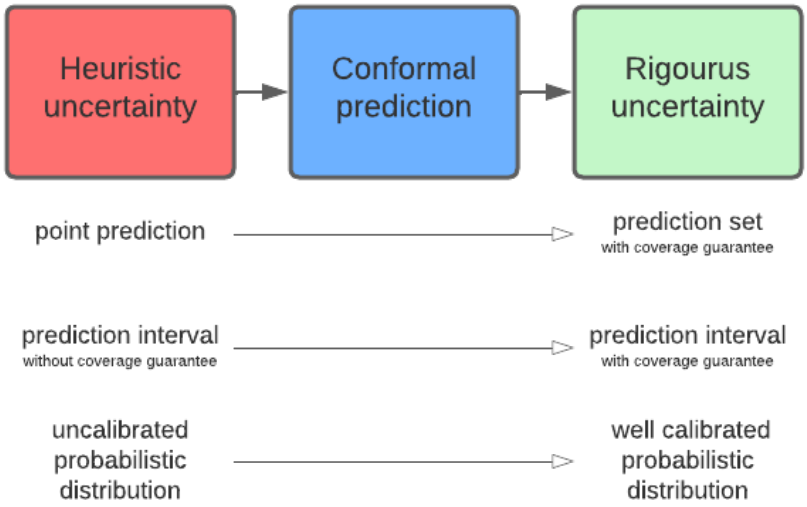


Figure 1: Conformal prediction uses a purely heuristic notation of uncertainty of any model and transforms it to a rigid uncertainty estimation with valid coverage guarantee.

1.2 Contribution

In this paper, the theoretical foundations of conformal prediction are first presented on an intuitive level to give the reader a first impression of the framework (Chapter 2). The difference between full and split conformal prediction (chapter 2.3) is then explained in more detail and split conformal prediction is presented as the more practically relevant special case of full conformal prediction. All CP methods use the same core steps and differ only in their specific score functions, this core algorithm in the sense of a conformal recipe is formulated in section 2.3.3. Split conformal methods use a hold-out dataset, often referred to as a calibration set. Theoretical considerations about its size are provided in chapter 2.4.3.

All conformal prediction methods guarantee marginal coverage, but conditional coverage guarantees are more crucial for many scenarios. An introduction to the difference between marginal coverage and conditional coverage is provided in section 2.4.

Even though conditional coverage cannot be theoretically guaranteed, methods to approach it exist with Mondrian conformal prediction 2.4.1 and 2.4.2. CP can be used for both classification and regression tasks, and the chapters 2.5 and 2.8 provide initial information on this.

Besides Split and full conformal prediction, CV+ and Jackknife+ are two methods that allow a trade-off between computational effort and data efficiency and are useful for many practically relevant scenarios with only a few available data points. Both methods are described in the 2.7 chapter.

In addition to this overview of the different varieties and methods of conformal prediction, chapter 3 provides an overview of the literature a historical account of the development from the initial work of Vladimir Vovk in the 1990s to the latest trends of recent years initiated by researchers such as Michael Jordan and Emanuel Candes.

In the experimental part of this work, the influence of different conformal methods, split, CV+ and jackknife+, on small datasets was investigated. For this purpose, both a real life dataset and an artificial dataset were artificially reduced in size and prediction sets with different conformal methods were created and compared with each other (Section 4).

2. Theory

Conformal Prediction (CP) is an innovative distribution-free, non-parametric, model-agnostic framework for conformance estimation with strict coverage guarantees at finite sample size. CP requires minimal assumptions, more specifically the framework only requires data exchangeability, a slightly weaker requirement than the iid requirement needed for many typical machine learning applications. CP comes in several flavors and requires only negligible additional computation time, at least in the inductive setting (split-conformal prediction). In this case, CP can be considered as a wrapper for any machine learning model that calibrates any heuristic notation of uncertainty provided by these models, thus guaranteeing mathematically valid coverage.

2.1 Coverage Guarantee

The conformal prediction guarantees that the output of each CP method contains the ground truth with a predefined probability $1 - \alpha$ on average. This property is usually referred to as coverage guarantee and is ensured by the fact that the method generates a prediction set $\mathbf{C}$ instead of a point predictor. In other words, the set contains the true label with a probability of exactly $1 - \alpha$. For labeled data points $z_i = (x_i, y_i)$, the coverage guarantee is expressed as follows, taking into account a small correction factor for finite sample size:

Definition 1. *Coverage guarantee*

$$1-\alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{x}_{test})\Big) \quad \leq \quad 1-\alpha+\frac{1}{n+1} \tag{1}$$

This statement holds for all sample sizes, models and distributions, without the need to make any further assumptions beyond the exchangeability of the data.

2.2 Intuition of Conformal prediction

In the following, the underlying principle of conformal prediction is explained on an intuitive level before a mathematically precise definition of the different conformal methods follows later. For this purpose, we consider the historically first developed case of full conformal prediction (transductive conformal prediction), since in this case the underlying rational becomes particularly clear and all other types of CP can be expressed as special cases of it (Fontana, Zeni, & Vantini, 2023).

2.2.1 Conformity Score

In a nutshell, Conformal Prediction uses past experience to generate precise levels of confidence in new predictions. Therefore, for one data point $z_{n+1} \in \mathbf{Z}$, e.g. the prediction, it is measured how "unusual" it looks compared to the bag of all exchangeable data points $\wr z_1, .., z_{n+1} \wr$. The inventors of the CP framework Gammerman, Vovk, and Vapnik (Gammerman, Vovk, & Vapnik, 1998a) refer to this as "a convenient measure of the evidence found to support this prediction". This concept of nonconformity or "strangeness" is quantified by a nonconformity measure (NCM) function, $S : Z^N \times Z \to \mathbb{R}$, which in principle can be defined arbitrarily without violating the coverage guarantees of CP. However, for the informativeness of a conformal method an appropriate choice of the NCM function is crucial. The strangeness of a data point $s_i = S(z_i)$ is now a days usually called the conformity score. Emphasis that the conformity score increases with increasing strangeness. The following chapters discusses how to chose a appropriate conformity score for regression as well as classification problems in more detail.

2.2.2 Constructing The Prediction Set

Thus, the question arises how to construct valid prediction sets from the conformity scores, which guarantee to deliver the true result with a definable certainty. Let us consider labeled exchangeable data $(X_1, Y_1), ..., (X_n, Y_n), (X_{n+1}, Y_{n+1})$ with the goal to predict for the point X_{n+1} the corresponding label Y_{n+1}. Exchangeability can be understood in a simplified way as the swap of any two data points. Meaning, the date obtained after swapping (X_{n+1}, Y_{n+1}) with (X_i, Y_i), cannot be distinguished from the non-exchanged bag of data.

We know that Y_{n+1} must live in the label space $\mathcal{Y}$. If we try every possible label $y \in \mathcal{Y}$ for the data point X_{n+1}, then due to the exchangability of the data points the pair (X_{n+1}, Y_{true}) must be interchangeable to the first n data points, i.e. the data point $(X_{n+1}, Y_{true} = y)$ is not "strange" with respect to all other data points, respectively has a low conformity score.

Full conformal prediction now directly exploits this principle by training for each possible label $y \in \mathcal{Y}$ a symmetric model f^y on the augmented dataset $(X_1, Y_1), ..., (X_N, Y_N), (X_{n+1}, y)$. In the next step, for each data point of the dataset, the corresponding conformity score is

computed as :

$$\begin{aligned} s_i^y &= S\Big((X_i, Y_i), f^y\Big) \quad \text{,for } i = 1, ..., n \\ s_{n+1}^y &= S\Big((X_{n+1}, y), f^y\Big) \end{aligned} \tag{2}$$

Now we define $\hat{q}$ as the $\frac{\lceil (1-\alpha)(n+1) \rceil}{n}$ quantile of the conformity scores $s_1^y, ..., s_n^y$, which is basically the $1-\alpha$ quantile, with a small correction for the finite sample size:

$$\hat{q}^y = \text{Quantile}\left(s_1^y, ..., s_n^y; \frac{\lceil (1-\alpha)(n+1) \rceil}{n}\right) \tag{3}$$

The prediction set $\mathbf{C}(X_{test})$ is now constructed by collecting all y that are sufficiently consistent with the previous data $(X_1, Y_1), ..., (X_N, Y_N)$:

$$\mathbf{C}(X_{test}) = \left\{ y : s_{n+1}^y \leq \hat{q}^y \right\} \tag{4}$$

The set constructed in this way satisfies the coverage guarantee (Equation 1). This can be explained by the fact, that if we order the conformal scores $s_1^y, ..., s_n^y$ of the individual data points by magnitude, the score s_{n+1}^y lies with uniform probability of $\frac{1}{n+1}$ between any two of these points (See Figure 2). Thus, the set $\mathbf{C}(X_{test})$ exactly contains $1-\alpha$ of the probabilistic density and therefore satisfies the coverage guarantee.

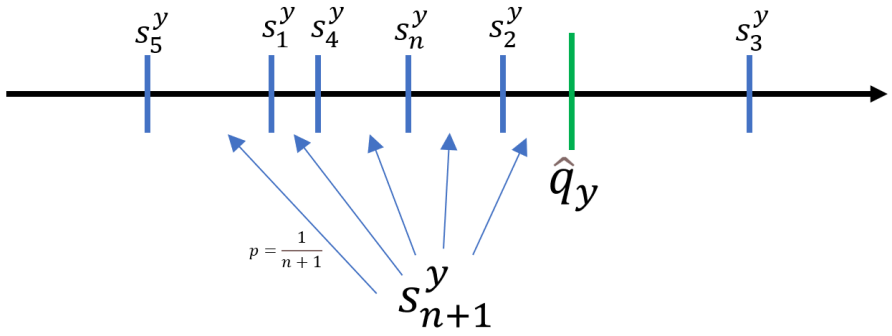


Figure 2: The conformity score $s_{n+1}^y = S\Big(X_{n+1}, y, f^y\Big)$ for a possible test data point $(X_{n+1}, y), y \in \mathbf{Y}$ lies with uniform probability of $p = \frac{1}{n+1}$ between each of the sorted conformity scores $s_1^y, ..., s_n^y$ of the previous data points $(X_1, Y_1), ..., (X_N, Y_N)$. Thus, all $y : s_{n+1}^y \leq \hat{q}^y$ lie within the $1-\alpha$ quantile.

With full conformal prediction, however, a model must be trained for each $y \in \mathcal{Y}$, which is not feasible for large or continuous label spaces or for computationally intensive models such as neural networks. For other approaches, such as split conformal prediction, a pre-trained model can be used so that this computational effort is not required. However, new data points that were not used during training would need to be utilized, making this approach less efficient in terms of using existing data. Split conformal prediction can be described as a special case of full conformal prediction and is presented in detail in section 2.3.

2.2.3 Score Function

At this point, the reader may continue to wonder how a statistically valid prediction set can be constructed, even if the underlying model and thus the heuristic notation of uncertainty may be arbitrarily bad. Intuitively, it must be stated that all information about the actual problem, the data and the underlying model is contained in the score function. To illustrate, if the scores s_i correctly classify the errors of the model for a given input, this leads to small prediction sets for simple inputs and large prediction sets for difficult inputs. However, if the scores do not correctly reflect this classification of difficult and easy inputs, e.g. because the underlying model only provides an inadequate notation of the uncertainty or because a non-informative score function was selected, the prediction sets become uninformative. In the extreme case, when the scores are just random noise, the resulting prediction sets are also a random sample of the labeling space, but they are large enough to still satisfy the coverage guarantee, i.e. they contain on average $1 - \alpha$ of the labeling space. Such a prediction set is therefore no more informative than guessing without prior information. This intuitive consideration leads to a fundamental property of all conformal methods. Conformal prediction methods fulfill the coverage guarantee in any case, but the informativeness and thus the usefulness of the prediction sets is determined by the scoring function. Depending on the quality of the scores obtained, which is determined by the choice of the score function and the quality of the underlying model, prediction sets with high informative value (approximate point predictions) or prediction sets with absolutely no informative value (the entire label space) are obtained.

2.3 Split Conformal Prediction

As we have seen, the fully conformal approach requires retraining the underlying model for every possible y in the label space $\mathcal{Y}$, which leads to a considerable computational effort. Therefore, recent work on this topic mostly employs the split-conformal approach, which can be seen as a special case of fully conformal prediction and originally goes back to work by Harris Papadopoulos (Papadopoulos, 2008) (Papadopoulos, Proedrou, Vovk, & Gammerman, 2002) and was introduced under the term inductive conformal prediction, in contrast to the fully conformal approach, which can be seen as a transductive approach.

2.3.1 Difference Of Inductive And Transductive Learning Systems

The distinction between transductive and inductive learning systems as used here goes back to the work of Wladimir Vapnik (Vapnik, 1998) and should be briefly explained here. Inductive systems generate in a first step from the available training data a general hypothesis which can be understood as a decision rule. With the help of this rule, a prediction for new examples can be deductively derived, without having to consider the training data any further. The advantage is that the derived rule is more compact than the original data, i.e. the information content is compressed, so that it can be stored more efficiently and a prediction based on it can usually be made more quickly. In other words, most of the computational effort is incurred during learning, also known as eager learning, with predictions being made quickly. Transductive systems, on the other hand, take a shortcut and do not generate a general decision rule but derive the prediction for a new example directly from the training data (Figure 3). Thus, such procedures are computationally inefficient, since they start

from scratch for each prediction and, in addition, all training data have to be kept. The actual computational effort is thus incurred in the prediction, while "learning" describes only the storage of the training data (Fontana et al., 2023). Therefore, transductive systems are often referred to as lazy learners.

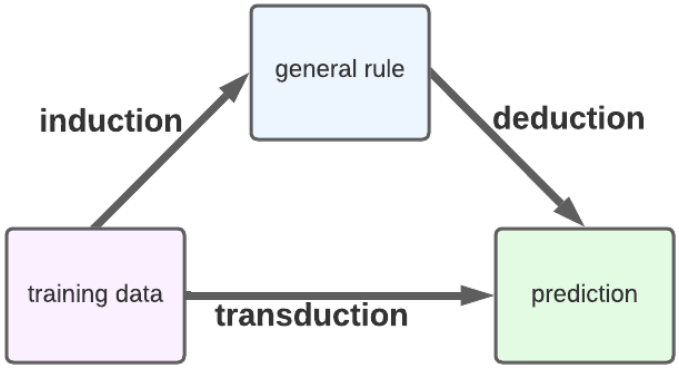


Figure 3: Inductive and transductive learning methods. Inductive learning generates general decision rules from the training data in the learning step and makes new predictions based on them. Transductive learning generates predictions based on the training data and the new data point only at the prediction time.

2.3.2 Principle Of Split Conformal Prediction

In the inductive approach for conformal prediction, the training dataset Z is split into two parts, a training dataset Z_{train} and a calibration set $Z_{calib} = (X_1, Y_1), ..., (X_n, Y_n)$. In a first step, an arbitrary underlying model $\hat{f}$ is trained on the training data Z_{train}. It is important to note that the model itself can also be a pre-trained model from any source. Split conformal prediction can thus be seen as an additional confidence layer, or wrapper, around any model, providing it with mathematically valid coverage guarantees.

For the application of a split conformal method it is only crucial that an additional calibration dataset exists, which is exchangeable to the used training dataset. A suitable calibration set should generally contain around $|Z_{calib}| = n = 500 - 1000$ data points. More detailed considerations about the calibration set size can be found in chapter 11.

In order to construct a prediction set $\mathbf{C}(X_{test})$ with coverage guarantee (Equation 1) from the pre-train model $\hat{f}$, the corresponding conformity scores $s_i = S\Big((X_i, Y_i), \hat{f}\Big)$ are now calculated for each data point $(X_1, Y_1), ..., (X_n, Y_n)$ from the calibration set Z_{calib}. In contrast to the full conformal approach, the model $\hat{f}$ is now fixed and thus no longer depends on a test data point X_{n+1}, and therefore all conformity scores s_i can be calculated in advance. Like in the transductive setting, $\hat{q}$ is then calculated as the $\frac{\lceil (1-\alpha(n+1)) \rceil}{n}$ quantile of the empirical conformity scores $s_1, ..., s_n$ and the prediction set for a new data point X_{n+1} is simply calculated as:

$$\mathbf{C}(X_{test}) = \Big\{ y : s(X_{test}, y) \leq \hat{q} \Big\}$$

.

The prediction sets obtained in this way meet the coverage guarantee, completely independent of the size of the calibration set, the underlying distribution of the data and the correctness of the underlying model (Fontana et al., 2023).

This approach can be derived as a special case from the full conformal method by using the algorithm for training the model $\hat{f}^y = \text{train}\Big((X_1, Y_1), ..., (X_n, Y_n), (X_{test}, y)\Big)$ in the full conformal method in such a way that it simply ignores the test data point (X_{test}, y). Thus, transductive scores $s_i^y = S((X_i, Y_i), \hat{f}^y)$ are identical to those from the inductive setting $s_i = S((X_i, Y_i), \hat{f})$.

2.3.3 The Conformal Recipe

As described in detail in the previous chapter, each split conformal method follows the same pattern, regardless of the actual underlying problem. Thus, a basic formulation of all split conformal methods is given here, which will serve as a recipe for all methods presented in the following chapters.

The following steps show the general recipe for all split-conformal methods with not necessarily discrete labeled calibration data $(X_1, Y_1), ...(X_n, Y_n)$ and an arbitrary pre-trained model $\hat{f}$ to create a prediction set $\mathbf{C}(X_{test})$ containing the ground truth with probability $1 - \alpha$ (Angelopoulos & Bates, 2021):

1. Identify a heuristic notion of uncertainty provided by $\hat{f}$
2. Define a score function $S\Big(X_i, Y_i; \hat{f}\Big)$ based on the heuristic notion of uncertainty.
3. Compute $\hat{q}$ as the $\frac{\lceil (1-\alpha(n+1) \rceil}{n}$ quantile of the calibration scores
$$s_1 = S\Big(X_1, Y_1; \hat{f}\Big), ..., s_n = S\Big((X_n, Y_n; \hat{f}\Big)$$
4. Calculate the prediction sets for a new data point X_{test} as:
$$\mathbf{C}(X_{test}) = \Big\{y : S(X_{test}, y; \hat{f}) \leq \hat{q}\Big\}$$

2.4 Conditional Coverage

As described in the previous chapters, all conformal methods satisfy the coverage guarantee. To be precise, this means that the prediction set for a new data point X_{test} contains the true label on average with a probability of $1 - \alpha$. This coverage guarantee is also called marginal coverage, since it is not adaptive to the difficulty of a single point X_{test} and can therefore easily be estimated as the percentage of the prediction sets that covers the ground truth for a new dataset (Molnar, 2023). For the sake of readability, the definition for marginal coverage will be repeated here (without considering the finite sample size correction for the upper bound).

Definition 2. ***Marginal coverage***

$$1-\alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test})\Big)$$

However, we have no guarantee that the coverage is also $1-\alpha$ for each individual data point, not even for specific groups such as individual classes. For example, consider a classification task with equal numbers of data points for men and women, where the prediction set for men contains the ground truth in all cases (easy inputs), but for women it contains it on average only in 90% of the cases (difficult inputs). The model still satisfies the coverage guarantee of 90%. Nevertheless, this property is not what is desired in many situations. Rather, we would like the coverage guarantee to apply not only on average, but also to specific subsets of the data. This property is commonly referred to as conditional coverage and can be formalized as follows (Angelopoulos & Bates, 2021) (See also Figure 4 for a intuitive description):

Definition 3. ***Conditional coverage***

$$1-\alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test}) \mid X_{test}\Big)$$

This means that for every single input X_{test} we provide prediction sets with $1-\alpha$ coverage. This property is much stronger than marginal coverage and cannot be guaranteed in general. However, there are some practically relevant special cases, such as class or group conditional coverage, for which a guarantee can be given. There are also conforming methods that approximate conditional coverage better than others. Such methods are generally referred to as adaptive conformal methods, as they take into account the higher uncertainty for difficult data points with larger prediction sets.

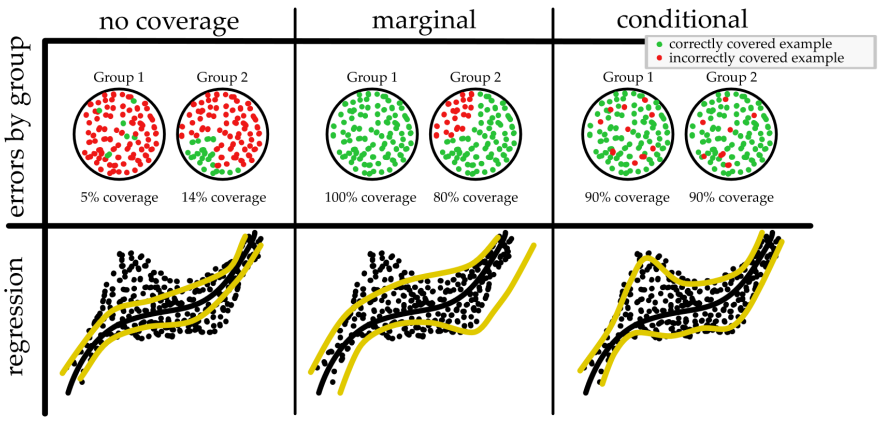


Figure 4: Overview of prediction set with no coverage, marginal and conditional coverage in the classification and regression case. In the marginal coverage case all errors happens in one group of data, or region of X. Conditional coverage enforces evenly distributed errors in all classes. Figure is taken from (Angelopoulos & Bates, 2021).

2.4.1 Mondrian Conformal Prediction

As seen in the previous chapter, conforming predictors within certain subsets of the data do not guarantee coverage. The proportion of errors in one group may be greater than the target significance level, which will result in fewer errors in other groups. The Mondrian-conformal predictors first proposed by Vovk (Vovk, Lindsay, Nouretdinov, & Gammerman, 2003) solve this problem by first dividing the data Z into certain categories or groups $g_1, ..., g_m \in G$. For this purpose, a measurable function $M : Z \Rightarrow G$ is formally introduced. A group $g_i = M(Z_i)$ can depend on the other data points, but ignores their order. Such a function is also called Mondrian taxonomy after the Dutch painter Piet Mondrian, because the partitioning of the data Z is reminiscent of his grid-like paintings (Figure 5).

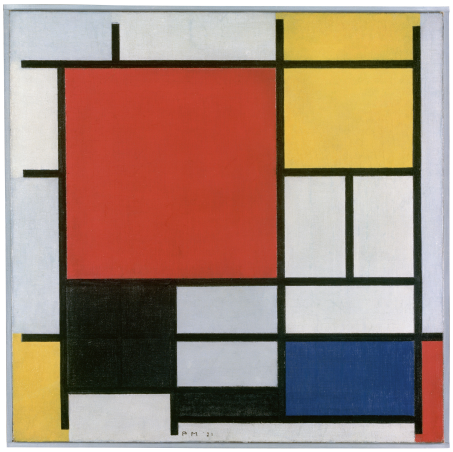


Figure 5: Piet Mondrian. Compositie met groot rood vlak, geel, zwart, grijs en blauw, 1921, Art Museum Den Haag

Using Mondrian taxonomy, we can define the group-balance coverage:

Definition 4. ***Group-balanced coverage***

$$1 - \alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test}) \mid X_{test} = g_i\Big) \quad : \quad \forall g \in G$$

This means we have $1 - \alpha$ coverage in all groups. This can be easily obtained by first dividing the data points into groups using the Mondrian taxonomy and then applying a standard conformal procedure for each group. That is we compute the scores $s_i^{(g)}$ for each point in the calibration set, where each data point belongs to a group g. We then calculate the $1 - \alpha$ quantile $\hat{q}^{(g)}$ for each group:

$$\hat{q}^{(g)} = \text{Quantile}\left(s_1(g), ..., s_n(g); \frac{\lceil (1-\alpha)(n(g)+1) \rceil}{n(g)}\right) \tag{5}$$

where $n(g)$ is the number of data points of the corresponding group. The prediction sets are now formed by using the relevant quantile for the new data point $X_{test}^{(g)}$:

$$\mathbf{C}(X_{test}^{(g)}) = \left\{y : s(X_{test}^{(g)}, y) \leq \hat{q}^{(g)}\right\}$$

This method has the not insignificant disadvantage that the calibration set has to be divided and therefore fewer calibration points are available for determining the quantise, making the estimate less reliable. Even if this does not violate the coverage guarantee (on average), it leads to a significantly higher variance (Angelopoulos & Bates, 2021). In practice, this means that the need for calibration points increases linearly with the number of different groups.

2.4.2 Class-Balanced Coverage

For classification tasks, coverage within each of the classes to be predicted should often be guaranteed. This leads to class-balanced coverage:

Definition 5. ***Class-balanced coverage***

$$1 - \alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test}) \mid Y_{test} = y\Big)$$

At first sight this problem seems to be similar to the group-balanced coverage and as we will see the solution is also quite similar. First, the corresponding quantile $\hat{q}^y$ is calculated separately for each class as in the group balanced case. The problem arises when constructing the prediction set for a new data point, because the corresponding class is not known. The solution is to use all class-wise predictors and the union of the resulting sets as the result.

$$\mathbf{C}(X_{test}) = \Big\{y : s(X_{test}, y) \leq \hat{q}^y\Big\}$$

This procedure guarantees class-balanced coverage, more precisely, the procedure exactly (Angelopoulos & Bates, 2021) satisfies the coverage for the most difficult class to classify, that is, the class y with the largest quantile $\hat{q}^y$. For all other classes, however, the coverage can be significantly higher than $1 - \alpha$ and thus the resulting prediction sets are usually significantly larger (Molnar, 2023).

2.4.3 Size Of The Calibration Set

As mentioned above, the coverage guarantee holds regardless of the size of the calibration set. However, it is intuitively clear that larger calibration sets lead to more stable conformal predictors. This intuition is correct, which can be attributed to the fact that the coverage guarantee (Equation 1) holds for a coverage of $1 - \alpha$ on average over the randomness in the calibration set. That is, with a fixed finite calibration set, the coverage will not be exactly $1 - \alpha$ even evaluated on infinite test points, but the deviation from this value will decrease as the calibration set size increases. This insight may seem sobering at first glance, but the fluctuation of coverage can be directly analyzed and controlled, since the coverage of conformal prediction conditionally on the calibration set is a random quantity following a Beta distribution (Vovk, 2012) (Figure 6):

$$\mathbb{P}\Big(Y_{test} \in C(X_{test}) \mid \{(X_i, Y_i)\}_{i=1}^{y}\Big) \sim Beta(n + 1 - l, l) \ , \ l = \lfloor (n + 1)\alpha \rfloor$$

With this statement we can now precisely determine the expected empirical coverage (evaluated with infinitely many test points). More precisely, it allows us to say how many

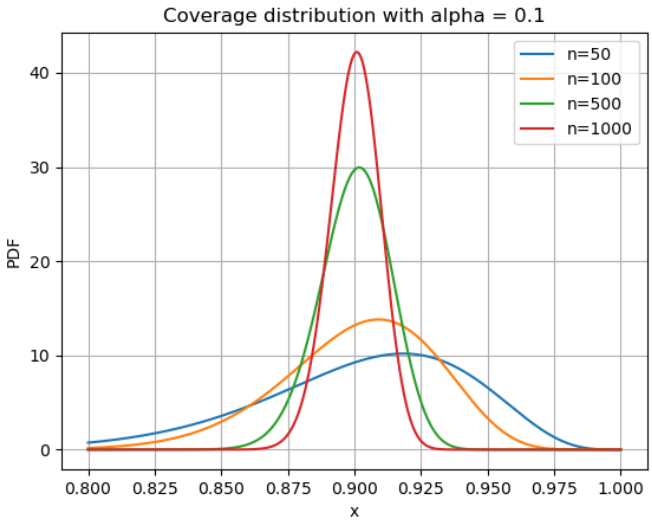


Figure 6: Distribution over the coverage considering a infinite validation set and different calibration set sizes n, following a beta distribution.

ϵ	0.1	0.05	0.01	0.001
$n(\epsilon)$	22	102	9812	244390

Table 1: Required calibrations set size $n(\epsilon)$ for a coverage of $1 - 0.9 \pm \epsilon$ with probability $\delta = 0.1$.

calibration points are needed for a coverage of $1 - \alpha \pm \epsilon$ with a probability of $1 - \delta$. Table 2.4.3 shows practically relevant example calibration set sizes (Angelopoulos & Bates, 2021):

From this consideration it can be deduced that for most applications calibration sets with about 1000 data points guarantee sufficient coverage.

2.5 Classification

Up to this point, the conformal prediction framework has been introduced in general. In the following chapter, several concrete conformal methods for classification tasks are presented. All approaches follow the conformal recipe (2.3.3) and differ in the choice of the score function.

2.5.1 Least Ambiguous set-valued Classifier

Many classifiers inherently provide a natural notation about the probability of the individual classes, such as the softmax layer of a neural network. These outputs are generally not calibrated, but can be transformed into valid prediction sets using a conformal method. Here we define the score function for the (softmax) outputs of a model $\hat{f}$ as follows:

$$s_i = 1 - \hat{f}(x_i)$$

Thus, the prediction set includes all classes for which the corresponding softmax output is greater than $1 - \hat{q}$ (Angelopoulos & Bates, 2021):

$$\mathbf{C}(X_{test}) = \left\{ y : \hat{f}(X_{test})_y \geq 1 - \hat{q} \right\}$$

2.6 Adaptive Prediction Sets (APS)

The Naive Score method produces prediction sets with the smallest average size (Sadinle, Lei, & Wasserman, 2019), but it is not adaptive. In practice he method tends to undercover hard subgroups and overcover easy subgroups. This is due to the fact that it uses only the outputs of the true class and ignores all others. Adaptive prediction sets, on the other hand, sum up all outputs, starting with the largest score up to the true class. Let $\pi(x)$ the permutation of the classes $1, ..., c$ that sorts $\hat{f}(X_{test})$ from most likely to least likely and y the true label, then the score function can be strongly compressed as follows:(Angelopoulos, Bates, Malik, & Jordan, 2020):

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

In an example, a classification algorithm outputs the (softmax) output of cat=0.3, lion=0.6 and dog=0.1 for a cat image (which tends to be hard to classify). The naive method calculates $s = 1 - 0.3 = 0.7$ and ignores the lion output. The adaptive method takes this into account and calculates $s = 1 - 0.6 - 0.3 = 0.9$. For easy to classify examples APS gives comparable scores as the naive method, but for difficult examples larger prediction sets are generated (Molnar, 2023).

2.6.1 Adaptive classification with split-conformal calibration

The description given up to this point was more for an intuitive understanding of how adaptive prediction sets can be constructed. However, some details like tie-breaking to guarantee marginal coverage are still missing for the actual formulation of the algorithm. In the following, a detailed formulation for adaptive classification with split-conformal calibration is given. This approach then leads directly to the CV+ and Jackknife+ methods which use the available data more efficiently (Section 2.7).

Considering an oracle classifier with perfect knowledge of the conditional distribution $\mathbb{P}_{Y|X}$, the construction of optimal prediction sets $C_\alpha^{\text{oracle}}(x_{test})$ would be trivial. Let $f_y(x) = \mathbb{P}[Y = y, X = y] \ , \ \forall y \in \mathbf{Y}$ with the order statistic $f_{(1)}(x) \geq f_{(2)}(x) \geq ... \geq f_{(C)}(x)$, then, without considering ties by now, we can define the generalized conditional quantile function for an arbitrary $\tau \in [0, 1]$:

$$L(x; f, \tau) = \min\{c \in 1, ..., C \ : \ F_{(1)}(x) + f_1(x) + ... + f_c(x) \geq \tau\} \tag{6}$$

and the resulting prediction set:

$$C_\alpha^{\text{oracle}}(x) = \{\text{corresponding } y \text{ for the } L(x; f, 1 - \alpha) \text{ largest } f_y(x)\} \tag{7}$$

With the already used example $f_{(cat)}(x)0.3$, $f_{(lion)}(x)0.3$ and $f_{(dog)}(x)0.3$ we get for $L(x; 0.9) = 2$ the prediction set $C_\alpha^{\text{oracle}}(x) = 1, 2$ and for $L(x; 0.5) = 1$ $C_\alpha^{\text{oracle}}(x) = 2$. The previous

formulation does not handle equal values for $f_{(c-1)}(x)$ and $f_{(c)}(x)$. The only way to theoretically guarantee coverage is to randomly add or discard the last label, which will simply break ties at random. For this we define the following function:

$$S(x, u; f, \tau) = \begin{cases} \text{corresponding } y \text{ for the } L(x; f, 1-\alpha) - 1 \text{ largest } f_y(x), & \text{if } u \geq V(x; f, \tau) \\ \text{corresponding } y \text{ for the } L(x; f, 1-\alpha) \text{ largest } f_y(x), & \text{otherwise} \end{cases} \tag{8}$$

where:

$$u \sim \text{Uniform}$$

$$V(x; f, \tau) = \frac{1}{f_{(L(x;f,\tau))}(x)} \left[\sum_{c=1}^{L(x;f,\tau)} f_{(c)}(x) - \tau \right] \tag{9}$$

With this formulation, we obtain tighter, tie breaking, and valid prediction sets (Romano, Sesia, & Candes, 2020):

$$C_{\alpha}^{\text{oracle}}(x_{test}) = S(x_{test}, U; f, 1-\alpha) \tag{10}$$

It is intuitively clear that any trained model$\hat{f}_y(x)$ can only approximate $f_y(x)$ and even be arbitrarily bad in theory, and thus the oracle approach cannot be used. However, the following method can construct valid prediction sets for any model$\hat{f}_y(x)$ by fitting the threshold τ based on a calibration set. The only restriction that applies to $\hat{f}_y(x)$ is that the algorithm treats the data points interchangeably, i.e. invariant to their order and that the output class probabilities are normalized, that is $\hat{f}_y(x) \in [0, 1],\ \sum_{y=1}^{C} \hat{f}_y(x) = 1\ ,\ \forall x, y$.

To calibrate τ on the basis of a hold-out set, we define the so-called generalized inverse quantile function $E(x, y, u; \hat{f}$ that computes the smallest possible value for τ such that $S(x, u; f, \tau)$ contains the true label y.

$$E(x, y, u; \hat{f}) = \min\{\tau \in [0, 1] : y \in S(x, u; f, \tau)\} \tag{11}$$

By this construction, the scores computed on the hold-out set (X_i, Y_i) are $E_i = E(X_i, Y_i, U_i; \hat{f})$ are uniformly distributed conditional on X for $f = \hat{f}$ and U_i independent uniform random variable. This property is not present in many other conformity score functions and makes the scores naturally comparable between different samples. In this way, we can now construct prediction sets with a provable marginal coverage guarantee of τ close to the $1 - \alpha$ quantile of $\{E_i\}_{i \in \mathbf{I}_2}$, where $\mathbf{I}_2$ is the hold-out or calibration set that was not used to train $\hat{f}$ (Romano et al., 2020).

The algorithm for calculating adaptive classification with split-conformal calibration is presented in 1.

The algorithms satisfies the marginal coverage guarantee as desired (for approximately distinct values for E_i the upper bound also holds):

$$1 - \alpha \leq \mathbb{P}\Big[Y_{test} \in C_{n,\alpha}^{\text{SC}}(x_{test})\Big] \leq 1 - \alpha + \frac{1}{|X_{calib}| + 1}$$

Algorithm 1 Adaptive classification with split-conformal calibration

Input: data $\{X_i, Y_i\}_{i=1}^n$, X_{test} , model $\hat{f}$, α
X_{train}, $X_{calib} \leftarrow$ train_test_split($\{X_i, Y_i\}_{i=1}^n$)
Train $\hat{f}$ on X_{train}
Compute $E_i = E(x_i, y_i, u_i; \hat{f})$ for each $x_i, y_i \in X_{calib}$ with function 11
Compute $\hat{Q}_{1-\alpha}(\{E_i\}_{i \in X_{calib}})$ as the $\lceil (1-\alpha)(1-|X_{calib}|) \rceil$th largest value in E_i
Output the prediction set:

$$C_{n,\alpha}^{\mathrm{SC}}(x_{test}) = S(x_{test}, u_{test}; \hat{f}, \hat{Q}_{1-\alpha}(\{E_i\}_{i \in X_{calib}}))$$

using the score function S defined in 8.

2.7 Adaptive classification with cross-validation+ and jackknife+ calibration

All split conformal methods have in common that they are not efficient, i.e. they do not use all data for the training of the underlying model, but require a hold-out dataset for calibration purposes. Even if for many applications about 1000 data points are sufficient (Angelopoulos & Bates, 2021) these are not always available. Full conformal prediction is a method to use all data efficiently, but it requires considerable computational effort and is therefore not suitable for computationally intensive models. However, CV+ and Jackknife+ are two methods that offer a compromise between computational complexity and data efficiency. We first consider CV+ and then introduce Jackknife+ as a special case of CV+.

CV+ uses a cross-validation approach and partitions the data into k distinct splits $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_k$. On each of the splits, the corresponding model $\hat{f}^{k(i)} = \hat{f}(\{X_i, Y_i\}_{i \in \{1,...,n\} \setminus \mathcal{I}_k})$ is trained. To form the prediction sets, it is now iterated over each possible label $y \in \mathcal{Y}$ and the y are unified to form the prediction set $C_{n,\alpha}^{\mathrm{CV+}}(x_{test})$ whose score $E(x_{test}, y, u_{test}; \hat{f}^{k(i)})$ is smaller than $\lceil (1-\alpha)(1-|X_{calib}|) \rceil$ hold-out scores $E(x_i, y_i, u_i; \hat{f}^{k(i)})$.

Algorithm 2 Adaptive classification with CV+ calibration

Input: data $\{X_i, Y_i\}_{i=1}^n$, X_{test} , model $\hat{f}$, number of splits $K \leq n$, α
Split data into k random distinct subsets $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_k$
for $k \in \{1, ..., k\}$:
    Train $\hat{f}^{k(i)}$ on $\{X_i, Y_i\}_{i \in \{1,...,n\} \setminus \mathcal{I}_k}$
Output the prediction set:

$$C_{n,\alpha}^{\mathrm{CV+}}(x_{n+1}) = \Bigg\{ y \in \mathcal{Y} : \sum_{n=1}^{n} \mathbf{1}\Big[E(x_i, y_i, u_i; \hat{f}^{k(i)}) \leq E(x_{n+1}, y_{n+1}, u_{n+1}; \hat{f}^{k(i)}) \Big] \leq \lceil (1-\alpha)(1-|n|) \rceil \Bigg\}$$

where $k(i) \in \{1, .., k\}$ denotes the fold containing the ith sample and using the function E defined in 11.

The algorithm theoretically fulfills a slightly weakened coverage guarantee:

$$\mathbb{P}\Big[Y_{test} \in C^{\text{CV+}}_{n,\alpha}(x_{test})\Big] \geq 1 - 2\alpha - \min\Big\{\frac{2(1-1\backslash K}{n\backslash K+1}, \frac{1-K\backslash n}{K+1}\Big\}$$

In the special case for $k = n$ we speak of the jackknife+ method with simplified bound of:

$$\mathbb{P}\Big[Y_{test} \in C^{\text{CV+}}_{n,\alpha}(x_{test})\Big] \geq 1 - 2\alpha$$

The authors of the method point out, however, that in practice for these methods mostly an empirical coverage of $1 - \alpha$ instead of $1 - 2\alpha$ is achieved. The following more conservative definition for the prediction set satisfies the coverage guarantee of $1 - \alpha$ also in theory but is resulting in larger prediction sets:

$$C^{\text{CV+mm}}_{n,\alpha}(x_{n+1}) = \Bigg\{ y \in \mathcal{Y} : \\ \sum_{n=1}^{n} \mathbf{1}\Big[E(x_i, y_i, u_i; \hat{f}^{k(i)}) \leq \min_{j\in\{1,\ldots,n\}} E(x_{n+1}, y_{n+1}, u_{n+1}; \hat{f}^{k(i)})\Big] \leq \lceil (1-\alpha)(1-|n|)\rceil \Bigg\}$$

2.7.1 Regularized Adaptive Prediction Sets (RAPS)

In classification tasks with many classes, APS tends to generate large prediction sets, especially when many classes are potentially possible. This is mainly due to the fact that there can be a long tail of (noise) classes with low probability. In these cases it can be helpful to use an additional regularization term. This is achieved by using a regularization term λ to penalize classes with a rank higher than k_{reg}. This has the effect that the prediction sets for classifications tasks with many labels become on average smaller than those produced by APS (Angelopoulos et al., 2020).

2.8 Regression

Although the focus of this thesis is on classification tasks, the basics of conformal regression methods are presented in the following chapter. These provide intervals that, like all conformal methods, adhere to the coverage guarantee 1.

2.8.1 Naive Conformal Regression

In the simplest case, the residuals of a regression model $\hat{f}$ are used directly as a score function:

$$s_i = |y_i - \hat{f}(x_i)|$$

As in all conformal methods, $\hat{q}$ is now calculated as the $\frac{\lceil (1-\alpha(n+1)\rceil}{n}$ quantile. The prediction interval is then calculated as follows:

$$\mathbf{C}(X_{test}) = \Big[\hat{f}(X_{test}) - \hat{q}\ ,\ \hat{f}(X_{test}) + \hat{q}\ \Big]$$

This method has the disadvantage that the resulting intervals all have exactly the same size and the method is therefore not adaptive in any way (Romano, Patterson, & Candes, 2019).

2.8.2 Conformal Regression For Scalar Uncertainty Estimates

To obtain more adaptive intervals, standardized residuals can be used. For this purpose, a scalar notation for the uncertainty of the residuals is employed.

A typical way construct those values is to train a second model $\hat{r}$ that estimates the residuals of the actual regression model $\hat{f}$. The intuition here is that if such a model $\hat{r}$ were a perfect oracle for the uncertainty of the residuals, the interval $\Big[\hat{f}(X_{test}) - \hat{r}(X_{test})\ ,\ \hat{f}(X_{test}) + \hat{r}(X_{test})\ \Big]$ would have perfect coverage. However, $\hat{r}$ is often a poor estimator in practice and we need to adjust the intervals using conformal prediction (Romano et al., 2019).

Generally speaking, we consider a second function $u(x_i)$, where larger values indicate greater uncertainty. This function can, for example, also describe the variance between an ensemble of models or the variance after slight input perturbations. With this heuristic notation of the uncertainty $u(x_i)$ the score function $s(x_i, y_i)$ is defined as follows:

$$s_i = \frac{|y_i - \hat{f}(x_i)|}{u(x_i)}$$

This uncertainty function can be taken as a correction factor for uncertainty $s_i * u(x_i) = |y_i - \hat{f}(x_i)|$. Subsequently, $\hat{q}$ is calculated as the $\frac{\lceil (1-\alpha(n+1) \rceil}{n}$ quantile of the scores, resulting in the following prediction sets:

$$\mathbf{C}(x_{test}) = \Big[\hat{f}(x_{test}) - u(x_{test})\hat{q}\ ,\ \hat{f}(x_{test}) + u(x_{test})\hat{q}\ \Big]$$

These thus fulfill the desired coverage guarantee:

$$\mathbb{P}\Big[s(X_{test}, y_{test}) \leq \hat{q}\Big] \geq 1 - \alpha \quad \Rightarrow \quad \mathbb{P}\Big[\Big|y_{test} - \hat{f}(X_{test})\Big| \leq u(X_{test})\hat{q}\Big] \geq 1 - \alpha$$

The prediction sets generated in this way are symmetric with respect to individual predictions $\hat{f}$, but it is not necessarily the case that the quantiles of uncertainty (e.g., the variance of the models) are directly related to the quantiles of label distribution. That is, they do not necessarily scale properly with α (Angelopoulos & Bates, 2021). For this reason, the literature points out that the conformalized quantile regression estimates the label distribution directly and thus has the better uncertainty heuristic, which can also be shown empirically (Angelopoulos, Kohli, Bates, Jordan, Malik, Alshaabi, Upadhyayula, & Romano, 2022).

2.8.3 Conformalized Quantile Regression

Many regression models can be easily transformed into quantile regressors by using a quantile loss , also known as pinnball loss. These quantile regression methods not only provide a point estimation but try to determine the γ quantile of a distribution $Y_{test}|X_{test} = x$ for each possible value of x. Conformal methods based on such models often yield better results in practice than the methods presented in the previous chapter, since their heuristic notation of uncertainty is directly related to the label space (Romano et al., 2019).

For the true quantiles $t_\gamma(x)$, the set $[t_\alpha, t_{1-alpha}]$ would have to have exactly a coverage of 2α, since by definition a fraction of α must be above t_α and below $t_{1-alpha}$, respectively.

However, there is no guarantee that the calculated quantiles are correct. Thus, conformal prediction can be used to calibrate them.

Let $\hat{t}_\alpha/2(x_i), \hat{t}_{1-alpha/2(x:i)}$ be the output of any quantile regression for the data point x_i, then the associated score function $s(x_i, y_i)$ can be defined as the distance from y_i to the nearest quantile:

$$s(x_i, y_i) = \max\Big\{\hat{t}_{\alpha/2}(x_i) - y_i \ , \ y_i - \hat{t}_{1-\alpha/2}(x_i)\Big\}$$

As in all conformal methods, $\hat{q}$ is now calculated as the $\frac{\lceil (1-\alpha(n+1)) \rceil}{n}$ quantile of the scores. The associated prediction intervals are then determined as follows:

$$\mathbf{C}(x_{test}) = \Big[\hat{t}_{\alpha/2}(x_i) - \hat{q} \ , \ \hat{t}_{1-\alpha/2}(x_i) + \hat{q}\Big]$$

Conformalized quantile regression can thus be understood as increasing or decreasing the interval provided by the underlying model for certain ranges of the label space to meet the coverage guarantee (Romano et al., 2019).

3. History And Recent Development

Conformal Prediction has seen a real explosion of interest particular in the last year, as evidenced by both numerous real-world applications and a veritable flood of publications. Nevertheless, the roots of the framework go back more than 50 years and are based on the work of Andrei Kolmogorov in Moscow. The real fathers of conformal prediction are Vladimir Vovk and Alexander Gammermann, who worked out the theoretical foundations together in London in the late 1990s. Especially V. Vovk has made numerous further fundamental applications and theoretical contributions in the following years until today. In the middle of the 2010s, the framework, which until then had lived in the shadows of academia, was given a new impetus by the work of the Professor of Statistics and Data Science at the Carnegie Mellon University Larry Wassermann in America and the group around Ulf Johansson and Henrik Boström in Sweden. Inspired by the work of Larry Wassermann, groups led by Emanuel Candes (Chair Mathematics and Statistics, Stanford), Ryan Tibshirani (Chair Statistics, UC Berkeley) and Michael Jordan (Chair Computer Science, Statistics, UC Berkeley) reformulated and popularized the framework and paved the way for a variety of new approaches and methods in recent years. Of course, this rough outline is highly simplistic and leaves out many researchers who also had significant influence on the development of Conformal Prediction, but should give the reader a sense of the most prominent dynamics in the field. In the following overviews, this rough description will be enriched by a slightly detailed timeline, which shows the most important developments and publications in the field, without claiming to be complete.

- **1960-1980** Andrei Kolmogorov starts at Moscow State University with work on notation of randomness, complexity and probability (Kolmogorov, 1968). Among other things, he studies algorithmically random sequences and finite Bernulli sequences (Kolmogorov, 1983). Vladimir Vovk becomes his student during this period.

- **1988** V. Vovk presents his PhD thesis "Predictability of algorithmically random sequences" under the supervision of A. Kolmogorov. Here he develops the Basis for a

first understanding of the role of finite-sample exchangeability in prediction problems for the study of Bernoulli sequences.

- **1996-1999** Vladimir Vovk, Alexander Gammerman and Vladimir Vapnik develop the framework now known as Conformal prediction together at the Royal Holloway University of London. They first used e-values (Gammerman, Vovk, & Vapnik, 1998b), later p-values (Vovk, Gammerman, & Saunders, 1999).

- **2002** Harris Papadopoulos, together with Vovk, develops what is now known as split-conformal prediction (Papadopoulos et al., 2002).

- **2003** Vladimir Vovk and Glen Shafer publish "Algorithmic Learning in a Random World" and coin the term "Conformal Prediction".

- **2003** Vovk and Gammermann lay the foundations for group balanced conformal prediction with the Mondrian Conformal predictors (Vovk et al., 2003).

- **2003** First Symposium on Conformal Prediction and its Applications (COPA) is organized in Greece by Harris Papadopoulos. This meeting has been held annually and is the most important event for the conformal prediction community.

- **2014** The group around Ulf Johansson, Henrik Boström and Henrik Linusson in Sweden take up the work of Vovk and develop conformal prediction methods especially for random forests (Johansson, Boström, Löfström, & Linusson, 2014). The group will also publish a number of papers and tutorials in the coming years.

- **2014** Larry Wassermann and Jing Lei begin their work in the field (Lei & Wasserman, 2014), making Vovk's work better known, especially in the United States. Among other things, they have done seminal work on class-balanced conformal prediction (Lei, 2014) and formulated a general framework for distribution-free predictive inference in regression (Lei, G'Sell, Rinaldo, Tibshirani, & Wasserman, 2018).

- **2019** The research group led by Emmanuel Candes at Stanford published the first of many papers on CP including. Conformalized Quantile Regression (Romano et al., 2019)

- **2020** Adaptive prediction sets (Romano et al., 2020)

- **2020** In the 2020 U.S. presidential election, The Washington Post estimated the yet-to-be-published election results for individual states using conformal prediction (Cherian & Bronner, 2021)

- **2021** Anastasios N. Angelopoulos and Stephen Bates, two PhD students of Michael Jordan show the use of conformal prediction on large-scale deep learning classifications tasks (Angelopoulos et al., 2020).

- **2021** General conformal risk control (Angelopoulos, Bates, Candès, Jordan, & Lei, 2021)

- **2021** Conformal Outlier detection (Bates, Candès, Lei, Romano, & Sesia, 2023)

- **2021** Detection of change points in time-series data (Vovk, 2021)
- **2021 and 2022** Dedicated tracks at ICML2021 and ICML2022. Keynote 'Conformal Prediction' at NeurlPS2022 by Emmanuel Candes
- **2022** Valery Manokhin, a PhD student of V. Vovk, created the Git repository "Awesome Conformal Prediction" [1]. which is today the most important overview platform for the rapidly developing field. Here you can find the most important papers, theses, books and many tutorials (Manokhin, 2022a).
- **2022** Release of the scikit-learn compatible package MAPIE with some basic conformal methods.[2]
- **2022** A. Angelopoulos and S. Bates publish "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification" (Angelopoulos & Bates, 2021), a highly readable introduction to conformal prediction including a good video tutorial.[3]
- **2022** CP beyond exchangeability (Barber, Candes, Ramdas, & Tibshirani, 2023)

4. Evaluation of different CP methods under small datasets

Conformal prediction comes in several flavors, with different properties in terms of required computation time and data efficiency, where higher efficiency is associated with increased computational effort. For large datasets, it is often possible to save a hold-out dataset, with about 1000 data points (Chapter 2.3), for calibration purposes and therefore to use a split conformal method. These have the advantage that the model only has to be trained once or an already pre-trained model can be used. In contrast, full conformal prediction methods require a model to be trained for each possible label in the sense of a lazy learner for each new test data point. This makes these methods impossible to use for many computationally intensive methods (2.2.2). With Jackknife+ and CV+ methods are available, which stand between full and split conformal prediction and thus represent an interesting alternative for more computationally intensive models and use cases with only few available data points (2.7). In a first evaluation the extent to which different splits of the data sets differ in terms of the conformal scores and the corresponding quantile was investigated. In a second experiment the performance of the split, CV+ and jackknife+ methods was evaluated on two different multi-class datasets both down sampled artificial into sets of different sizes. As data a artificially created and a real life dataset where used.

4.1 Methods

For the evaluation, two traditional classification models were used, namely a Gaussian Naive Bayes (**"GaussianNB"**) and a support vector machine with radial basis function kernel (**"SVM"**). Both models have the advantage that they are relatively robust to the

1. https://github.com/valeman/awesome-conformal-prediction
2. https://mapie.readthedocs.io/en/stable/
3. https://www.youtube.com/watch?v=nql000Lu_iE&t=1529s&ab_channel=AnastasiosNikolasAngelopoulos

required hyperparameters and thus their impact on training with different sized data sets is minimal. Another advantage is that the models require limited computational time, making the experiment feasible for jackknife+ methods. For both models, the implementation provided in the **scikit-learn** package[4] was used. For each model, a split conformal method and CV+ with $k = 5$ splits were evaluated. For the latter two variants were employed to aggregate the predictions obtained by cross-validation. The first simply calculates the average of the scores of the different models for a new test point (**"mean"**) and the second method (**"crossval"**)compares the individual conformity scores of the training points of the different models with the new test point as described in algorithm two 2. For the split method, 20% of the available data was used as a calibration set for each run. In addition, a jackknife+ procedure with **"crossval"** aggregation was evaluated for each model. All of these experiments were used with the least ambiguous set-valued classifier method 2.5.1 (**"score"**) as well as adaptive conformity scores (**"aps score"**) 2.6.1.
The artificial test data set was initialized with the method **make_classification** of the scikit-learn package[5] with 20 features, of which 15 were informative and 2 redundant, and 5 different classes. In total, the dataset consists of 10,000 data points. The other dataset used contains the data of 13,611 dry beans, which have to be classified into one of 7 variants based on 8 features such as length, roundness or firmness (Koklu & Ozkan, 2020).
In order to investigate the extent to which different splits of the data sets affect the conformal scores, the artificial dataset was first divided into 5 equally sized splits and the conformal scores and the calculated quantile were calculated separately for each of the splits.Both the least ambiguous set-valued classifier and the adaptive prediction set method were used as scoring methods.
For the actual experiment, 5000 points were used for both data sets to evaluate the different methods, i.e. they are not used for training or calibration. The remaining data points were dedicated to training, using a randomly distributed subset of these points for each experimental setup. Starting with 100 training points, the set was expanded by 100 points in each run. Thus, there were a total of 8 different strategies for 2 different models, with each resulting combination trained on datasets ranging in size from 100 to 5000 in steps of 100, and thus 800 different compliant runs. The experiment was repeated 5 times for each data set and the results averaged.

4.2 Results

In a first evaluation the extent to which different splits of the data sets differ in terms of the conformal scores and the corresponding quantile was investigated.

4.2.1 Influence of different splits on the conformal scores

The figures 7 and 8 show the results for the distribution of conformal scores and the resulting .95 quantile using the least ambiguous set-valued classifier method (lac) and the adaptive prediction set method (aps) for 5 different splits. The results for the Gaussian Naive Bayes model can be found in the appendix 7.1.

4. `https://scikit-learn.org/stable`
5. `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`

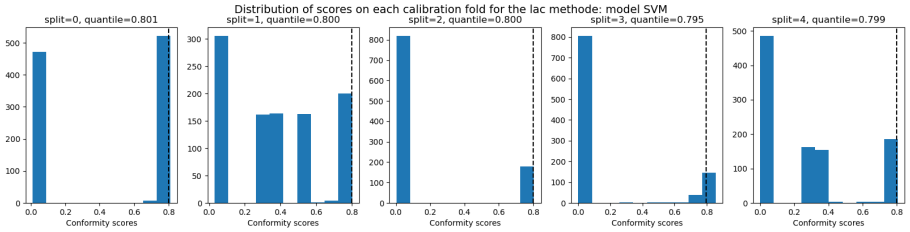


Figure 7: Distribution over the conformal scores using the least ambiguous set-valued classifier method (lac) and the resulting .95 quantile on 5 different splits based on a support vector machine with radial basis function kernel.

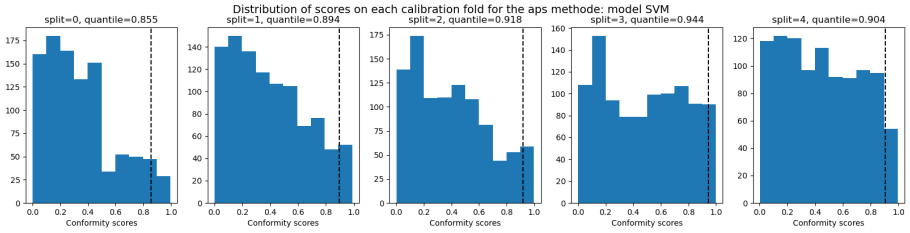


Figure 8: Distribution over the conformal scores using the adaptive prediction set method (aps) and the resulting .95 quantile on 5 different splits based on a support vector machine with radial basis function kernel.

4.2.2 Influence of small datasets on different conformal methods

In the following the results for the evaluation of different CP methods under datasets with increasing size for the least ambiguous set-valued classifier method (score) are presented. Since in the artificial dataset all classes are somewhat equally difficult to determine, an adaptive prediction score method had no real advantage and the resulting sets are only slightly larger and can be found in the appendix 7.2.

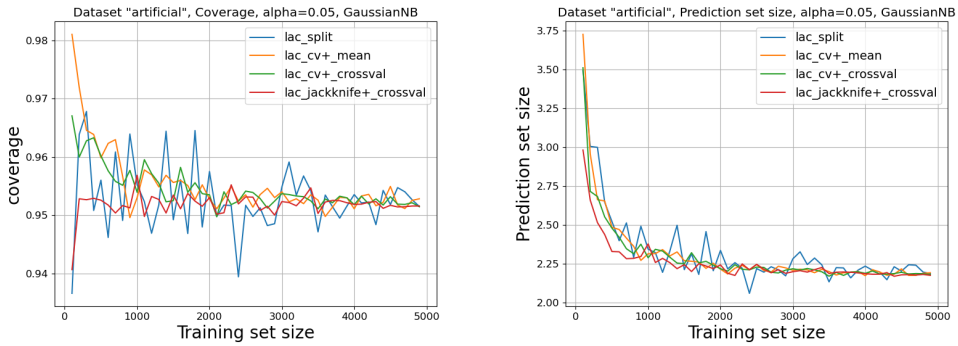
5. Discussion

5.1 Influence of small prediction sets on different CP methods

In a first experiment, the extent to which data sets with only a few data points influence different conformal methods, namely split, CV+ and jackknife+, was investigated. The split conformal method is the least computationally intensive, but most inefficient of these methods, as it uses a hold-out set for calibration in advance.

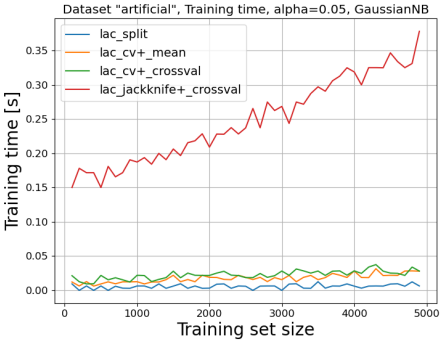
5.1.1 Influence of different split

It can be seen that even for the homoscedastic artificial data, different splits show a different distribution of scores across the calibration set. This means that a slightly different quantile is calculated for each split. This subsequently leads to fluctuations in the effective coverage. Furthermore, it is clear that the adaptive methods calculate the scores more evenly and thus, as intended, the difficulty of the single data points is taken more into account.

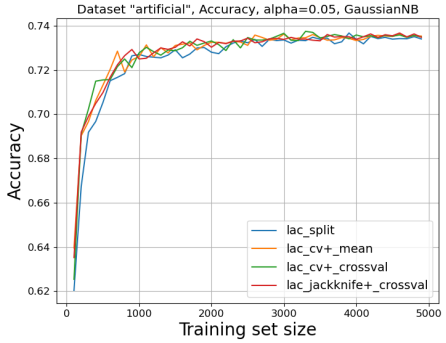


(a) Coverage

(b) Average prediction set size

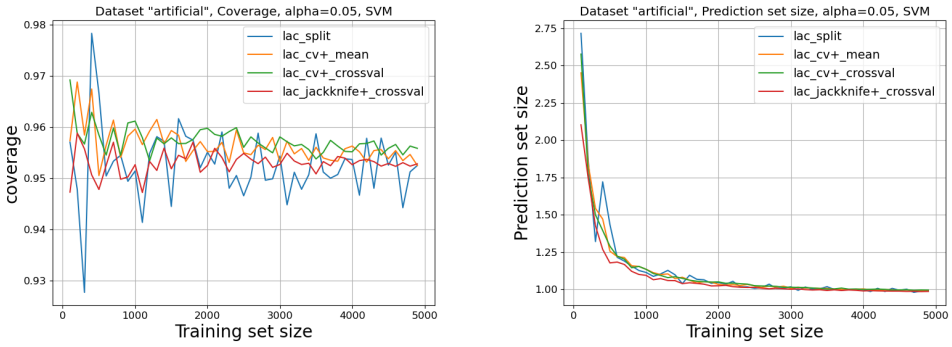


(c) Training time



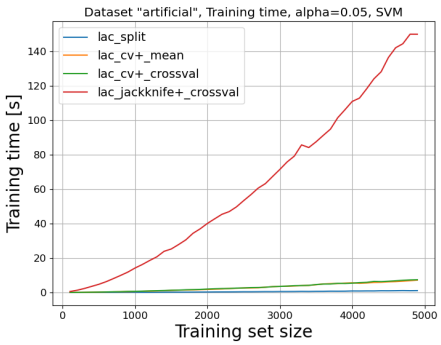
(d) Accuracy

Figure 9: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of an **artificial created dataset** with (20 features and 5 classes). Based on a Bayes Naive Gaussian model. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach according to (Romano et al., 2020). Conformal scores are computed using the least ambiguous set-valued classifier method (lac) for $\alpha = 0.05$.

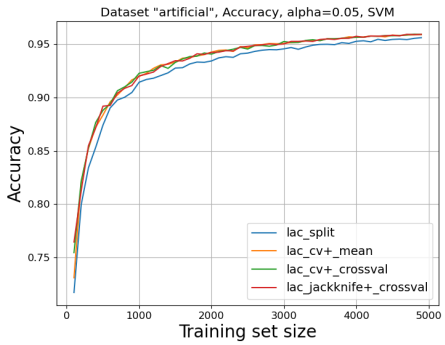


(a) Coverage

(b) Average prediction set size

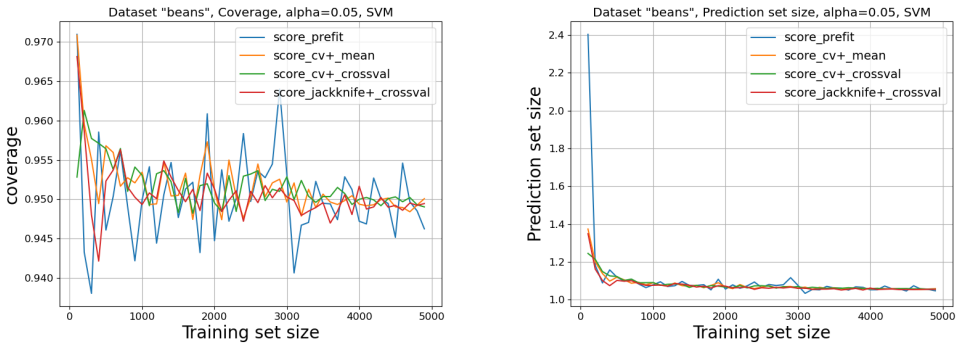


(c) Training time



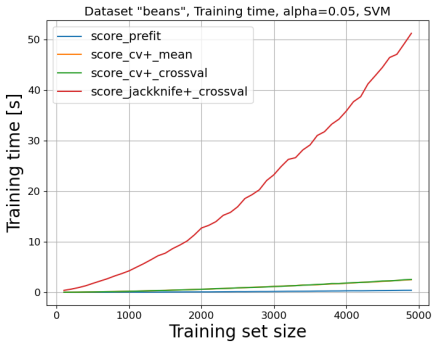
(d) Accuracy

Figure 10: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of an **artificial created dataset** with (20 features and 5 classes). Based on a support vector machine with radial basis function kernel. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach according to (Romano et al., 2020). Conformal scores are computed using the least ambiguous set-valued classifier method (lac) for $\alpha = 0.05$.



(a) Coverage

(b) Average prediction set size



(c) Training time

(d) Accuracy

Figure 11: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of the **dry beans dataset**. Based on a support vector machine with radial basis function kernel. CV+ scores are aggregated using an averaging (**mean**) method and the cross validation approach according to (Romano et al., 2020). Conformal scores are computed using the least ambiguous set-valued classifier method (lac) for $\alpha = 0.05$.

This observation also suggests that an unfortunate choice of calibration set can affect the robustness of any split conformal method.

5.1.2 Performance of the base model

It is intuitively clear that for the split methods, the underlying model can only be trained with less data and therefore, depending on the model and the complexity of the problem, the performance of the model is lower. As expected, the performance, measured as the accuracy of the base model, falls slightly for extremely small data sets for the split method compared to the other methods, regardless of the underlying model. The jackknife+ and CV+ methods use all data points and are therefore more performant for small datasets, whereby jackknife+ has no accuracy advantage over CV+, at least for the simple problems examined here.

5.1.3 Coverage of the conformal methods

For the split methods, theoretical assumptions can be made about the required size of the hold-out set, as described in chapter 2.4.3. This is due to the fact that the coverage of the randomness in the calibration set follows a beta distribution. This means that 1000 data points should be sufficient for most applications. In this work, 20% of the available data were used as the calibration set for all split conformal experiments. This means that for the largest investigated data set with 5000 data points, exactly 1000 points were used for calibration. As described, all conformal methods fulfill the coverage guarantee, however, smaller calibration sets result in larger fluctuations in the effective coverage. As for the split methods, this effect is clearly observable and amounts to differences of $\pm 3\%$ for small data sets. CV+ is significantly less affected by this as all data points are taken into account for the calibration, whereby the cross-validation aggregation function is significantly more robust than the mean function. As expected, jackknife+ shows the most stable results here, as every single available data point can actually be used for calibration.

5.1.4 Prediction set size

As with the coverage, the average size of the prediction sets fluctuates greatly under the split method, while it remains most stable for the jackknife+ method. The better, i.e. the more reliable the underlying model is, the smaller the prediction sets obtained, as the notation provided for uncertainty better reflects the true uncertainty in the data.

According to the theory, the least ambiguous set-valued classifier methods provide the smallest prediction sets on average, while they are significantly larger for the adaptive methods. The extent to which the prediction sets vary for the different classes was not investigated in this work and is planned for future work.

5.1.5 Computation time

The jackknife+ method, which calculates a model for each data point, is therefore extremely complex compared to the other methods and increases rapidly with the number of data points. The computational effort of the CV+ methods mainly depends on the number of splits used and is therefore adjustable. In the experiments presented here, $k = 5$ splits

were always used, whereby the performance using a cross-validation aggregation function only drops negligibly compared to the jackknife+ method. As described, jackknife+ can be regarded as a special case of CV+ with $k = n$. Thus, CV+ methods with a suitable number of splits represent a good trade-off between performance and computation time and should be preferred for all tasks with models that are reasonably complex to train.
In theory, the CV+ and Jackknife+ methods only fulfill the coverage guarantee of $1 - 2\alpha$. It turns out that for the SVM ($> 90\%$), which generally performs well on the artificial data set, a coverage of $1 - \alpha$ can be maintained empirically. The significantly worse performing GaussianNB models ($\backsim$ 70% accuracy) experiments on the other hand, are slightly below a .95 coverage for $\alpha = 0.95$. The effect is somewhat more pronounced for the adaptive prediction scores. It is therefore reasonable to assume that for well-performing base models, i.e. those with an accuracy comparable to α, the stricter coverage guarantee of $1 - \alpha$ is adhered to. On the other hand, for base models that are known to perform poorly and under alternative prediction set scores, the choice of α should be better adapted to the theoretically guaranteed $1 - 2\alpha$ coverage guarantee.

5.2 Real life experiment

6. Conclusion

References

Angelopoulos, A., Bates, S., Malik, J., & Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, -.

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, -.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., & Lei, L. (2021). Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, *51*.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., & Romano, Y. (2022). Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, *51*(2), 816–845.

Bates, S., Candès, E., Lei, L., Romano, Y., & Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, *51*(1), 149–178.

Candès, E. (2020). Stanford statisticians and washington post data scientists build more honest prediction models..

Cherian, J., & Bronner, L. (2021). How the washington post estimates outstanding votes for the 2020 presi- dential election..

Dewolf, N., Baets, B. D., & Waegeman, W. (2023). Valid prediction intervals for regression problems. *Artificial Intelligence Review*, *56*(1), 577–613.

Fontana, M., Zeni, G., & Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, *29*(1), 1–23.

Gammerman, A., Vovk, V., & Vapnik, V. (1998a). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, p. 148–155, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gammerman, A., Vovk, V., & Vapnik, V. (1998b). Learning by transduction, vol uai'98..

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The american statistician*, *69*(4), 371–386.

Johansson, U., Boström, H., Löfström, T., & Linusson, H. (2014). Regression conformal prediction with random forests. *Machine learning*, *97*, 155–176.

Johansson, U., & Gabrielsson, P. (2019). Are traditional neural networks well-calibrated?. In *2019 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Koklu, M., & Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, *174*, 105507.

Kolmogorov, A. (1968). Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, *14*(5), 662–664.

Kolmogorov, A. N. (1983). Combinatorial foundations of information theory and the calculus of probabilities. *Russian mathematical surveys*, *38*(4), 29.

Lambrou, A., Papadopoulos, H., Nouretdinov, I., & Gammerman, A. (2012). Reliable probability estimates based on support vector machines for large multiclass datasets. In *Artificial Intelligence Applications and Innovations: AIAI 2012 International Workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB, Halkidiki, Greece, September 27-30, 2012, Proceedings, Part II 8*, pp. 182–191. Springer.

Lei, J. (2014). Classification with confidence. *Biometrika*, *101*(4), 755–769.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111.

Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 71–96.

Manokhin, V. (2022a). Awesome conformal prediction..

Manokhin, V. (2022b). *Machine Learning for Probabilistic Prediction (PhD thesis, VALERY MANOKHIN)*. Ph.D. thesis, Royal Holloway University of London.

Molnar, C. (2023). *Introduction To Conformal Prediction With Python*. c/o MUCBOOK.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer.

Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, *32*.

Romano, Y., Sesia, M., & Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, *33*, 3581–3591.

Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, *114*(525), 223–234.

Vapnik, V. N. (1998). Adaptive and learning systems for signal processing communications, and control. *Statistical learning theory*, -.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, Vol. -, pp. 475–490. PMLR.

Vovk, V. (2021). Testing randomness online. *Statistical Science*, *36*(4), 595–611.

Vovk, V., Lindsay, D., Nouretdinov, I., & Gammerman, A. (2003). Mondrian confidence machine. *Technical Report*, -.

Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness. -, -.

7. Appendix

7.1 Results: Evaluation of different splits

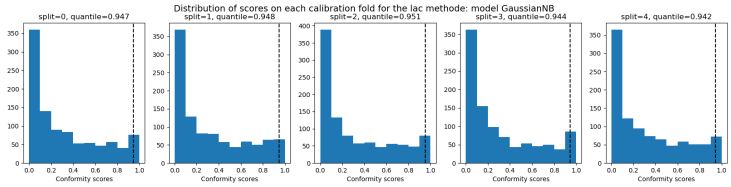


Figure 12: Distribution over the conformal scores using the least ambiguous set-valued classifier method (lac) and the resulting .95 quantile on 5 different splits based on a Gaussian Naive Bayes.

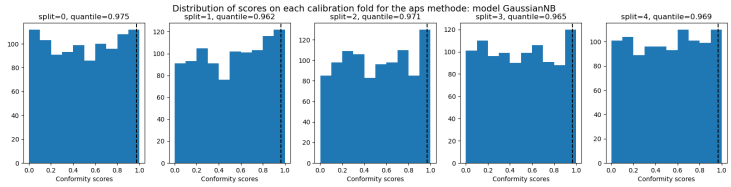
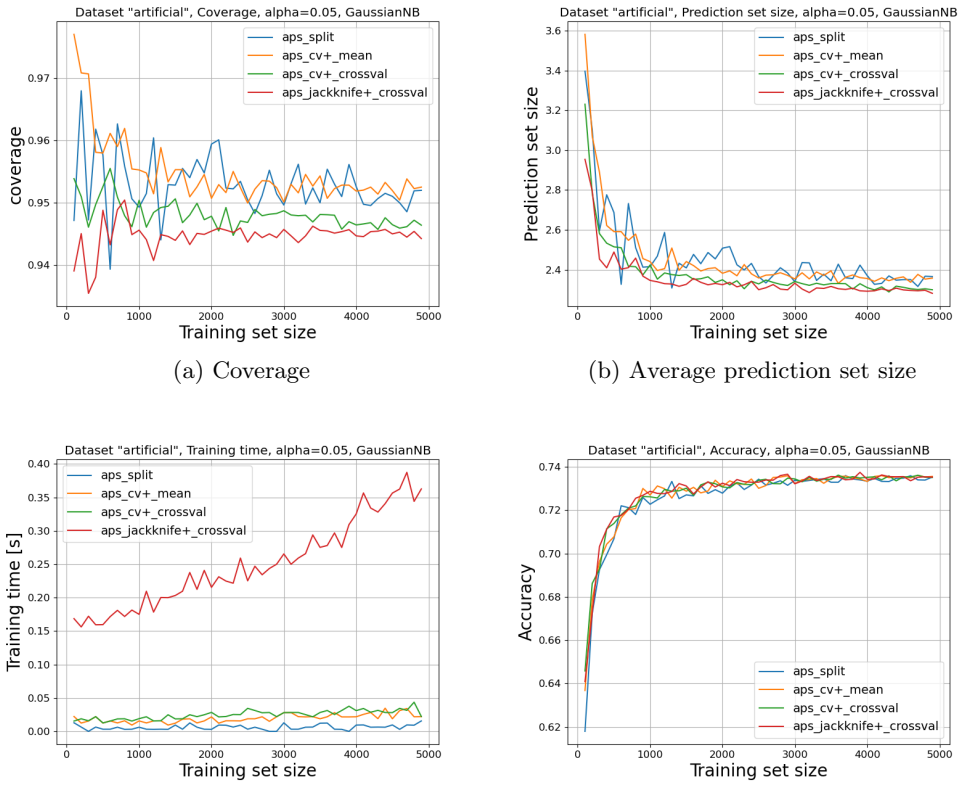


Figure 13: Distribution over the conformal scores using the adaptive prediction set method (aps) and the resulting .95 quantile on 5 different splits based on a Gaussian Naive Bayes.

7.2 Results: Evaluation of different CP methods under small datasets



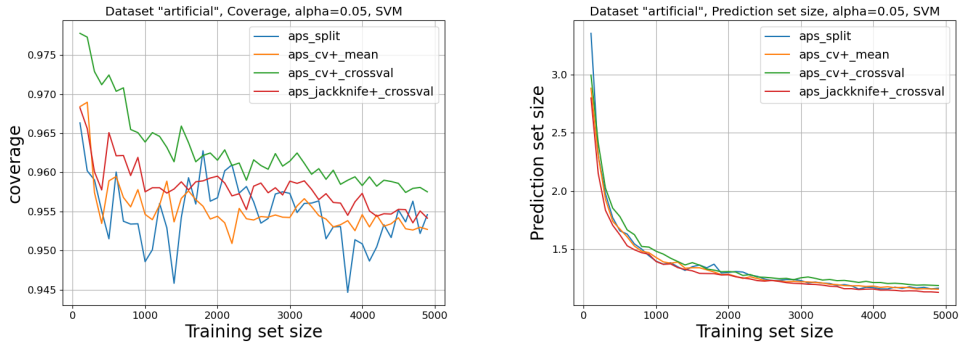
(a) Coverage

(b) Average prediction set size

(c) Training time

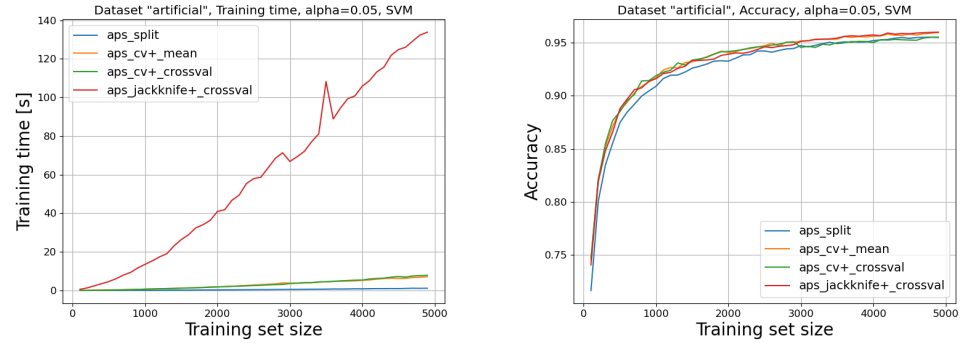
(d) Accuracy

Figure 14: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of an **artificial created dataset** with (20 features and 5 classes). Based on a Bayes Naive Gaussian model. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach according to (Romano et al., 2020). Conformal scores are computed using the adaptive prediction set method (aps) for $\alpha = 0.05$.



(a) Coverage

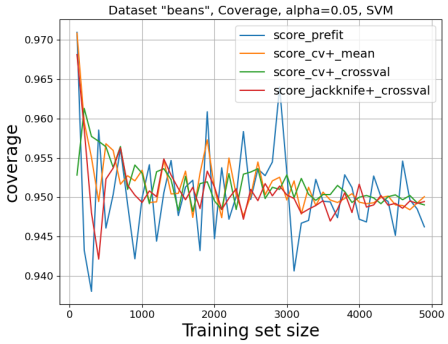
(b) Average prediction set size



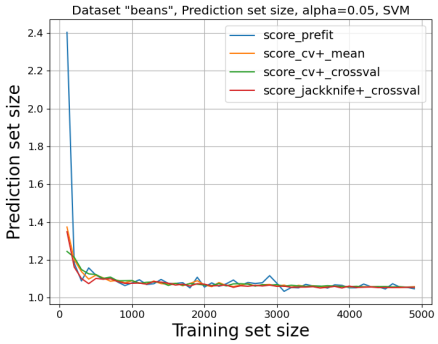
(c) Training time

(d) Accuracy

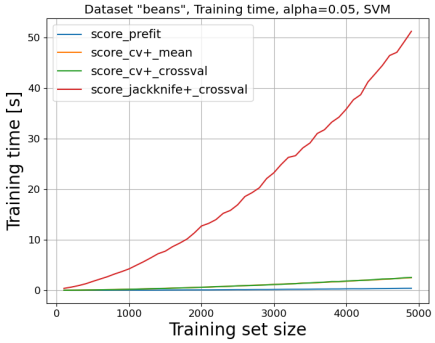
Figure 15: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of an **artificial created dataset** with (20 features and 5 classes). Based on a support vector machine with radial basis function kernel. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach following (Romano et al., 2020). Conformal scores are computed using the adaptive prediction set method (aps) for $\alpha = 0.05$.



(a) Coverage



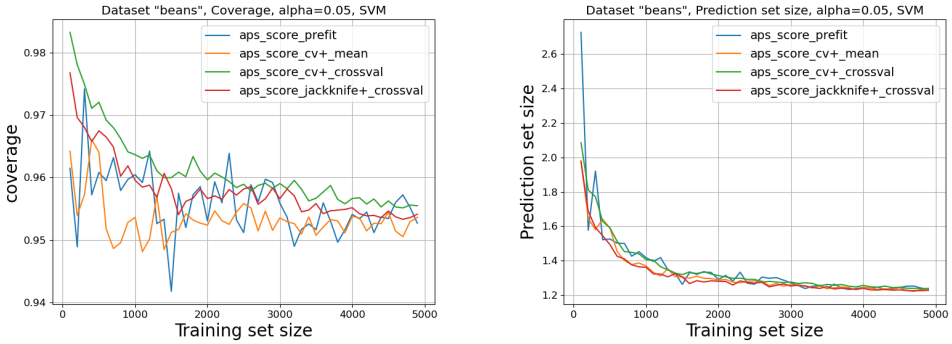
(b) Average prediction set size



(c) Training time

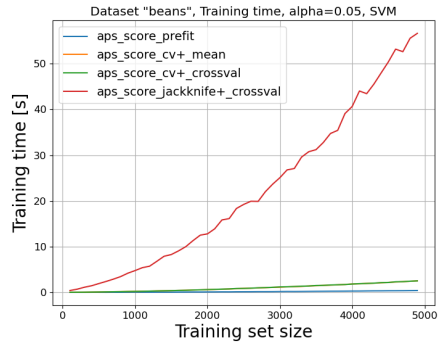
(d) Accuracy

Figure 16: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of of the **dry beans dataset**. Based on a support vector machine with radial basis function kernel. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach following (Romano et al., 2020).Conformal scores are computed using the least ambiguous set-valued classifier method (lac) for $\alpha = 0.05$.



(a) Coverage

(b) Average prediction set size



(c) Training time

(d) Accuracy

Figure 17: Effective coverage, average prediction set size, training time and accuracy for the different conformal methods split, CV+ with $k = 5$ and jackknife+ with increasing training set sizes of of the **dry beans dataset**. Based on a support vector machine with radial basis function kernel. CV+ scores are aggregated using an averaging (mean) method and the cross validation approach following (Romano et al., 2020). Conformal scores are computed using the adaptive prediction set method (aps) for $\alpha = 0.05$.