University of Dayton

eCommons

Undergraduate Mathematics Day: Proceedings and Other Materials

Department of Mathematics

2022

Efficient Conformal Binary Classification under Nearest Neighbor

Maxwell Lovig
University of Louisiana at Lafayette

Follow this and additional works at: https://ecommons.udayton.edu/mth_epumd

 Part of the Mathematics Commons

eCommons Citation

Lovig, Maxwell, "Efficient Conformal Binary Classification under Nearest Neighbor" (2022). *Undergraduate Mathematics Day: Proceedings and Other Materials*. 43.
https://ecommons.udayton.edu/mth_epumd/43

This Article is brought to you for free and open access by the Department of Mathematics at eCommons. It has been accepted for inclusion in Undergraduate Mathematics Day: Proceedings and Other Materials by an authorized administrator of eCommons. For more information, please contact mschlangen1@udayton.edu, ecommons@udayton.edu.

Elec. Proc. Undergrad. Math. Day
7 (2022), 1–12.

EFFICIENT CONFORMAL BINARY CLASSIFICATION UNDER NEAREST NEIGHBOR

MAXWELL LOVIG

Communicated by Paul Eloe

Abstract. There are many types of statistical inferences that can be used today: Frequentist, Bayesian, Fiducial, and others. However, Vovk introduced a new version of statistical inference known as Conformal Predictions. Conformal Predictions were designed to reduce the assumptions of standard prediction methods. Instead of assuming all observations are drawn independently and identically distributed, we instead assume exchangeability. Meaning, all N! possible orderings of our N observations are equally likely. This is more applicable to fields such as machine learning where assumptions may not be easily satisfied. In the case of binary classification, Vovk provided the nearest neighbors (NN) measure which is a ratio of in-class versus out-of-class distance. Later on, Papodopolous introduced normalizing constants for NN for the regression case, we extend this work to the classification case. We provide an asymptotic guarantee which shows what is known empirically. The normalization of NN produces smaller confidence sets on average compared to standard NN. A small synthetic simulation is also presented to shown the viability in a non-asymptotic case.

Keywords: *Efficiency, conformal predictions, set prediction, asymptotic, confidence sets*

MSC (2010): Primary 62G20

1. Introduction

As we begin to analyze more complex structures, we find ourselves faced with new issues to address. First, we must find methods which can relax statistical assumptions that might not be valid. Second, we must create methods which are easily applicable to complex non-linear models. Shafer and Vovk introduced the conformal prediction framework [5] which addresses these two questions. This framework is designed to reduce the classical assumption that $Z_1, \cdots, Z_N \overset{iid}{\sim} f_Z(z)$ (independently identically distributed). Instead of assuming all observations are drawn iid, Conformal Predictions assumes exchangeability. Meaning, the $N!$ possible orderings of our observations are equally likely. Written formally, with Ω as a set of the permutation of our observations,

$$\forall \omega_1, \omega_2 \in \Omega \quad f_{Z_{\omega_1(1)}, \cdots, Z_{\omega_1(N)}}(z_{\omega_1(1)}, \cdots, z_{\omega_1(N)}) = f_{Z_{\omega_2(1)}, \cdots, Z_{\omega_2(N)}}(z_{\omega_2(1)}, \cdots, z_{\omega_2(N)}).$$

Under the assumption our observations have this property, we can implement Conformal Classification Predictions. For classification, this requires a set of labelled observations $Z = z_1, \cdots, z_n = (x_1, y_1), \cdots, (x_n, y_n)$, where $x_i \in \mathbb{R}^n$ is our observation and $y_i \in Y$ is its label which is drawn from the set of possible labels Y. We also require a measurable function which takes in a set $\tilde{Z}$ and single labeled observation $\tilde{z}$ and returns a score which denotes the "non-conformity" or "oddity" of observation $\tilde{z}$. Written formally, when $\#\tilde{Z} = u$ and $\#\tilde{z} = v$

$$A : \mathbb{R}^{u \times v} \times \mathbb{R}^v \mapsto \mathbb{R}, \quad A(\tilde{Z}, \tilde{z}) \uparrow \Longrightarrow \text{ a more non-conformal occurrence of } \tilde{z}.$$

It is common to compare the non-conformity of a single z_i to the other observations in Z, in this case we write $A(Z \setminus z_i, z_i)$. When this is done for each z_i we create a distribution of non-conformity scores

which we can compare the score of an observation-label pairing. This comparison can then produce a p-value as a measure of likelihood. With a set of labelled observations Z, conformal measure A, possible label set Y, desired level of error ε and unlabelled observation x_{n+1}, we present the Conformal Prediction algorithm to construct prediction set Γ^A_ε:

Algorithm 1: Conformal Prediction Algorithm

Data: $Z = \{z_1, \cdots, z_n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$
Result: Γ^A_ε
for $z_i \in Z$ **do**
  $\alpha_i \leftarrow A(Z \setminus z_i, z_i)$
end
for $y_i \in Y$ **do**
  $z_{n+1} \leftarrow (x_{n+1}, y_i)$
  $\alpha_{n+1} \leftarrow A(Z, z_{n+1})$
  $p_y \leftarrow \frac{\#\{i=1,\cdots,n \text{ s.t.} \alpha_i \geq \alpha_{n+1}\}}{n+1}$
  if $p_y > \varepsilon$ **then**
    $y_i \in \Gamma^A_\varepsilon$
  end
end

Once the experimenter is given a particular data-set Z they have only two choices to make for the algorithm. The first is the error rate ε, which can be easily decided as $1 - \varepsilon$ is the desired accuracy of the confidence intervals. The choice for A, however, is not as clear. It is hard to decided a-priori what a good measure of non-conformity is, this is why we rely on the use of simple functions. One such simple function is the nearest-neighbor (NN) measure proposed by Vovk [5]. With $x \in Z_{y_i}$ denoting the set of observations from Z with label y_i, and norm $||\cdot||$, we have

$$A^{NN}(Z \setminus z, z) = A^{NN}(\tilde{Z}, (x^*, y_i)) = \frac{\min_{x \in \tilde{Z}_{y_i}} ||x - x^*||}{\min_{x \in \tilde{Z}_{\neg y_i}} ||x - x^*||}. \tag{1.1}$$

One issue with this measure is that it does not utilize any extra information about if our labels are difficult to predict. In the regression case, Papadopolous discussed the advantages of allotting differing constants σ which regulate how hard a label y is to predict. We extend this work to the classification case by assigning the difficulty to predict the label y as σ_y. As σ_y increases in size the label y is considered easier to predict. This leads to the generic normalized non-conformity function introduced by Papopdopolous [4],

$$A^*(Z, z) = A^*(Z, (x^*, y)) = \frac{A(Z, (x^*, y))}{\sigma_y}. \tag{1.2}$$

More complexity can be added, such as normalization terms and different criterion to try to minimize our prediction set size with using intuition on our labels. We can even begin to compare the difficultly of predicting given observation x as well, extending our terms to $\sigma_{x,y}$. Lim and Belotti gave many empirical measurements for diverse normalization in the regression case. They showed that there is influence on the efficiency of the prediction sets from the choice of normalization on the Ames housing data, but there is no theoretical connection as of yet [3].

One can then ask the question, why not add as much normalization as possible? Can't we just set aside a certain proportion of Z, say $\hat{Z}$, which we can compare the size of our prediction sets and then simply choose the measure which gives us the smallest average sets. This is a fair idea until one considers the normalization function

$$\sigma^*_{(x,y)} = \begin{cases} \infty, & \text{if } (x, y) \in \hat{Z}, \\ 0, & \text{otherwise.} \end{cases} \tag{1.3}$$

This would lead all of our prediction sets from $\hat{Z}$ being with $\#\Gamma^{A^*}_\varepsilon \leq 1$ but all other future predictions will have $\#\Gamma^{A^*}_\varepsilon = \#Y$. There must be some guidance on how we recommend a normalization method from Conformal Predictions. To do this, we must define what it means for a choice of A_1 to be better than the choice A_0. This will allow us to give theoretically sound guidance on ways to improve

non-conformity which is true for a large class of functions. This will prevent intentional over-fitting as irregular choices for normalization will be recognizable and thoroughly investigated at the time of review. As the popularity field of conformal predictions increases is it necessary that we establish criterion for these guarantees.

2. Efficient Conformal Predictions For Binary Classification Under Nearest Neighbors

2.1. **Definitions.**

Before these definitions are given one thing should be noted. A non-conformity measure A is not just a measurable function. We can also have A be a transformation of a set of random variables, making A itself a random variable.

DEF: Let us have $A(Z, z)$, where $Z = Z_1, \cdots, Z_n$ random variables, we define $\overline{A}(z)$ when $n \to \infty$

$$A_n(Z, z) \xrightarrow{p} \overline{A}(z).$$

For example: Another nonconformity measure proposed by Vovk [5] is the mean distance non-conformity, defined as

$$A^M(Z, z) = A^M(Z, (x^*, y_i)) = \left|\left| x^* - \sum_{x \in Z_{y_i}} x \Big/ \#Z_{y_i} \right|\right| \tag{2.1}$$

which then has $\overline{A^M}(z)$,[a]

$$\overline{A^M}(z) = \overline{A^M}(x^*, y_i) = ||x^* - E[x \in Z_{y_i}]||. \tag{2.2}$$

[a]This is shown in lemma 1 in the appendix

DEF: Let us have two non-conformity functions A_0 and A_1, A_1 is *asymptotically more efficient* (AME) than A_0 if for all ε, $E[\#\Gamma_\varepsilon^{\overline{A_1}}] \leq E[\#\Gamma_\varepsilon^{\overline{A_0}}]$ and $\#\Gamma_\varepsilon^{\overline{A_1}} < \#\Gamma_\varepsilon^{\overline{A_0}}$ for some ε.

For example: If A_1 has $E[\Gamma_\varepsilon^{\overline{A}_1}] = 1 - \varepsilon^2$ and A_2 has has $E[\#\Gamma_\varepsilon^{\overline{A}_2}] = 1 - \varepsilon$, then A_2 is AME than A_1.

The choice of the mean set size is a commonly used metric to evaluate the utility of our sets. There are other measures such as the proportion of singleton sets or considering the variance of set sizes. These complexities were discussed by Johanessen et al. [1]. They have found that the choice of non-conformity measure has a large effect on the utility of prediction sets. Of particular note, different measures of set utility, were optimized by different non-conformity measures. For the average set size to have meaning, we must have a situation where our labels are exclusive (an observation cannot have both labels 0 & 1). Krstajic discussed the issues using the mean prediction set size when labels are not exclusive [2].

DEF: If $A_*(Z, z) = \frac{A(Z,z)}{\sigma_*}$ is more efficient than all other $A = \frac{f(x,y)}{\sigma}$ with $\overline{A} \neq \overline{A_*}$, then A_* is the *asymptotically most efficient non-conformity under A* (AMEUA).

For our results, in order to avoid approximating (1.3) we restrict our σ to a function of y, making our normalization σ_y.

2.2. **Results.**

Let us consider $Z = (X, Y)$ drawn from a bounded space $S \subset \mathbb{R}^n$, these vectors can have the corresponding label 0 when $x \in S_0 \subset S$ and can have label 1 when $x \in S_1 \subset S$, with $S_0, S_1, S_0 \cap S_0$ having a non-empty interior. We define the probability distribution of X, Y as

$$f_Z(z) = f_{X,Y}(x, y) = f_Y(0)f_0(x)I_{x\in S_0} + f_Y(1)f_1(x)I_{x\in S_1}, \tag{2.3}$$

where $f_0(x)$ and $f_1(x)$ can be any bounded probability distribution and $f_Y(x)$ is defined as

$$f_Y(y) = \begin{cases} y_0, & \text{if } y = 0, \\ y_1, & \text{if } y = 1. \end{cases}$$

where $y_0 > 0$ and $y_1 > 0$ with $y_0 + y_1 = 1$.

We take A from (1.1) and combined with (1.2), letting $\sigma_y = \frac{1}{\varsigma_y}$, $\forall\ y, \varsigma > 0$, we have

$$A_*^{NN}(Z, z) = A_*^{NN}(Z, (x^*, y_i)) = \varsigma_y A(Z, (x^*, y_i)) \quad \varsigma_y = \begin{cases} \varsigma_0, & \text{if } y = (1, 0), \\ \varsigma_1, & \text{if } y = (0, 1). \end{cases} \tag{2.4}$$

We introduce the following results,

Theorem 1: *Under (2.3), the normalized non-conformity function A_*^{NN} is AME than A^{NN} when $S_0 \cap S_1 \neq \varnothing$. If S_0 and S_1 are disjoint then neither functions are AME.*

Proof. (For notational simplicity, A^{NN} will be suppressed as A for this proof.)

We see that as the number of draws is infinite for A we have the piecewise function $\overline{A} : S_0 \cup S_1 \mapsto \{0, 1, \infty\}$,[1]

$$\overline{A} : S_0 \cup S_1 \mapsto \{0, 1, \infty\}, \quad \overline{A}(z) = \overline{A}(x, y) = \begin{cases} 0, & \text{if } x \in S_0,\ x \notin S_1, y = 0, \\ 1, & \text{if } x \in S_0,\ x \in S_1, \\ 0, & \text{if } x \notin S_0,\ x \in S_1, y = 1, \\ \infty, & \text{if } x \in S_0,\ x \notin S_1, y = 1, \\ \infty, & \text{if } x \notin S_0,\ x \in S_1, y = 0. \end{cases} \tag{2.5}$$

We can now calculate the p_y for each of our possible combinations.

$$p_y = \begin{cases} 1, & \text{if } x \in S_0,\ x \notin S_1, y = 0, \\ P(x \in S_1 \cap S_0), & \text{if } x \in S_0,\ x \in S_1, \\ 1, & \text{if } x \notin S_0,\ x \in S_1, y = 1, \\ 0, & \text{if } x \in S_0,\ x \notin S_1, y = 1, \\ 0, & \text{if } x \notin S_0,\ x \in S_1, y = 0. \end{cases}$$

As the function of P_y is not surjective the interval $[0, 1]$, then there are intervals (or singletons) of ε where the expected interval size is unchanged. These intervals are:

$$\{[0, P(x \in S_0 \cap S_1)), [P(x \in S_0 \cap S_1), 1), \{1\}\}.$$

[1]See Lemma 2 in the appendix to show $A^{NN} \xrightarrow{p} 1$ for $x \in S_0 \cap S_1$:

Making our expected efficiency our prediction sets, $\#\Gamma_\varepsilon^{\overline{A}}$ given an error rate ε as

$$E[\#\Gamma_\varepsilon^{\overline{A}}|\varepsilon] = \sum_{x=0}^{2} x \cdot P(\#\Gamma_\varepsilon^{\overline{A}} = x|\varepsilon) = P(\#\Gamma_\varepsilon^{\overline{A}} = 1|\varepsilon) + 2 \cdot P(\#\Gamma_\varepsilon^{\overline{A}} = 2|\varepsilon)$$

$$= \begin{cases} 0, & \text{if } \varepsilon = 1, \\ P(x \in S_0) + P(x \in S_1) - 2P(x \in S_0 \cap S_1), & \text{if } P(x \in S_0 \cap S_1) \leq \varepsilon < 1, \\ P(x \in S_0) + P(x \in S_1), & \text{if } \varepsilon < P(x \in S_0 \cap S_1). \end{cases} \tag{2.6}$$

We can now consider A_* as seen in (2.4). $\overline{A}_*$ has form,

$$\overline{A_*} : S_0 \cup S_1 \mapsto \{0, \varsigma_0, \varsigma_1, \infty\}, \quad \overline{A}(z) = \overline{A}(x, y) = \begin{cases} 0, & \text{if } x \in S_0,\ x \notin S_1, y = 0, \\ \varsigma_0, & \text{if } x \in S_0,\ x \in S_1, y = 0, \\ \varsigma_1, & \text{if } x \in S_0,\ x \in S_1, y = 1, \\ 0, & \text{if } x \notin S_0,\ x \in S_1, y = 1, \\ \infty, & \text{if } x \in S_0,\ x \notin S_1, y = 1, \\ \infty, & \text{if } x \notin S_0,\ x \in S_1, y = 0. \end{cases} \tag{2.7}$$

This leads to 2 different possibilities in the distribution of non-conformity scores, one where $\varsigma_0 < \varsigma_1$ and another where $\varsigma_1 < \varsigma_0$. This makes $E[\#\Gamma_\varepsilon^{\overline{A_*}}|\varepsilon]$ have two separate possibilities, one where $\varsigma_0 < \varsigma_1$ and another where $\varsigma_0 > \varsigma_1$.

When $\varsigma_0 < \varsigma_1$, we have

$$P_y = \begin{cases} 1, & \text{if } x \in S_0,\ x \notin S_1, y = 0-, \\ P(x \in S_1 \cap S_0), & \text{if } x \in S_0,\ x \in S_1, y = 0, \\ P(x \in S_1 \cap S_0, y = 1), & \text{if } x \in S_0,\ x \in S_1, y = 1, \\ 1, & \text{if } x \notin S_0,\ x \in S_1, y = 1, \\ 0, & \text{if } x \in S_0,\ x \notin S_1, y = 1, \\ 0, & \text{if } x \notin S_0,\ x \in S_1, y = 0. \end{cases}$$

$$\implies E[\#\Gamma_\varepsilon^{\overline{A_*}}|\varepsilon]$$

$$= \begin{cases} 0, & \text{if } \varepsilon = 1, \\ P(x \in S_0) + P(x \in S_1) - 2P(x \in S_0 \cap S_1), & \text{if } P(x \in S_0 \cap S_1), \leq \varepsilon < 1 \\ 1, & \text{if } P(x \in S_0 \cap S_1, y = 0) < \varepsilon < P(x \in S_0 \cap S_1), \\ P(x \in S_0) + P(x \in S_1), & \text{if } \varepsilon < P(x \in S_0 \cap S_1, y = 0). \end{cases} \tag{2.8}$$

When $\varsigma_0 > \varsigma_1$, we have

$$P_y = \begin{cases} 1, & \text{if } x \in S_0,\ x \notin S_1, y = 0, \\ P(x \in S_1 \cap S_0), & \text{if } x \in S_0,\ x \in S_1, y = 1, \\ P(x \in S_1 \cap S_0, y = 0), & \text{if } x \in S_0,\ x \in S_1, y = 0, \\ 1, & \text{if } x \notin S_0,\ x \in S_1, y = 1, \\ 0, & \text{if } x \in S_0,\ x \notin S_1, y = 1, \\ 0, & \text{if } x \notin S_0,\ x \in S_1, y = 0. \end{cases}$$

$$\implies E[\#\Gamma_\varepsilon^{\overline{A_*}}|\varepsilon]$$

$$= \begin{cases} 0, & \text{if } \varepsilon = 1, \\ P(x \in S_0) + P(x \in S_1) - 2P(x \in S_0 \cap S_1), & \text{if } P(x \in S_0 \cap S_1) \leq \varepsilon < 1, \\ 1, & \text{if } P(y = 1, x \in S_0 \cap S_1) \leq \varepsilon < P(x \in S_0 \cap S_1), \\ P(x \in S_0) + P(x \in S_1), & \text{if } \varepsilon < P(y = 1, x \in S_0 \cap S_1). \end{cases} \tag{2.9}$$

∎

If we are willing to estimate how $P(y=0, x \in S_0 \cap S_1)$ compares to $P(y=1, x \in S_0 \cap S_1)$, then we can get an even stronger statement. With this estimate we can construct the best normalization constants when we desire to use the nearest neighbor measure.

Theorem 2: *For bounded binary classification, if we restrict σ to only a function of y,*

$$A_\varsigma^{NN}(Z,z) = \varsigma_y A^{NN}(Z,(x^*,y_i)), \quad \varsigma(y_i) = \frac{\#Z_{y_i}}{\#\{Z_{y_i} \ s.t. \ A(Z \setminus z, z) \geq \eta\}} \tag{2.10}$$

where $\eta > 0$, A_ς^{NN} is AMEF under A^{NN}.

Proof.

$$A_\varsigma^{NN}(Z,z) = \frac{\#Z_{y_i}}{\#\{Z_{y_i} \text{ s.t. } A(Z \setminus z, z) \geq \eta\}} A^{NN}(Z,z)$$

$$= \frac{\#Z_{y_i}\#Z}{\#\{Z_{y_i} \text{ s.t. } A(Z \setminus z, z) \geq \eta\}\#Z} A^{NN}(Z,z)$$

$$= \left(\frac{\#Z_{y_i}}{\#Z}\right)\left(\frac{\#\{Z_{y_i} \text{ s.t. } A(Z,z) \geq \eta\}}{\#Z}\right)^{-1} A^{NN}(Z,z).$$

$$\text{As } n \to \infty, \ \overline{A_\varsigma^{NN}}(Z,z) = P(y=y_i)(P(y=y_i, x \in S_0 \cap S_0))^{-1}\overline{A^{NN}}(Z,z)$$

$$= \frac{1}{P(y=y_i | x \in S_0 \cap S_1)}\overline{A^{NN}}(Z,z)$$

$$\text{as such } \varsigma(y_i) < \varsigma(y_j) \implies P(y=y_i | x \in S_0 \cap S_1) > P(y=y_j | x \in S_0 \cap S_1).$$

$$\text{Meaning } P(\#\Gamma_\varepsilon = 1 | \varepsilon) \text{ is at a maximum } \forall \varepsilon \text{ when}$$

$$\min\{\varsigma(y_0), \varsigma(y_1)\} < \varepsilon < P(x \in S_0) + P(x \in S_1) - 2 \cdot P(x \in S_0 \cap S_1)$$

$$\implies A_\varsigma^{NN} \text{ is more efficient then } A_*^{NN}(2.4), \text{ when } A_\varsigma^{NN} \text{ has the reversed inequality}$$

$$\text{(i.e. when } A_\varsigma^{NN} \neq A_*^{NN})$$

$$\implies A_\varsigma^{M} \text{ is the most efficient under (1.1).}$$

∎

The choice of η in (2.10) is still left to the experimenters discretion. This is due to any $\eta > 0$, as the sample size approaches infinity, will discern the observations which belong in $S_0 \cap S_1$. However a good choice of η is not exactly obvious outside of the asymptotic case. A canonical choice is 1 but for smaller sample sizes the choice of η is unclear. As η gets smaller, we are adding more observations which we classify as belonging to $S_0 \cap S_1$. As η gets larger, we are removing observations which we classify as belonging to $S_0 \cap S_1$. In future work it would be helpful to see how the choice of η changes the efficiency of prediction sets, this can be done in numerical simulations of some varied probability distributions satisfying (2.3).

3. Practical Obstacles and Synthetic Simulations

3.1. **Choice of Distance Function.**

Asymptotically, we defined a generic $||\cdot||$ for our results. In computation, however, we do have to make this choice and it will effect the sample size for which we can invoke asymptotic results. Previous work has been done to bridge the nearest neighbors to a more computationally stable method, an example of this is k-nearest neighbors [4]. We propose a slightly different approach, by discretizing $||\bullet - \circ||$ in (1.1) to semi-metric $d(\bullet, \circ)$, we can create a more stable non-asymptotic nearest neighbor type function. With $\kappa > 0$ and $\delta \gtrsim 0$ we have

$$d(\bullet, \circ) = \begin{cases} 0, & \text{if } ||\bullet - \circ|| = 0, \\ \delta, & \text{if } 0 < ||\bullet - \circ|| \leq \kappa, \\ 1, & \text{if } ||\bullet - \circ|| > \kappa. \end{cases}$$

The choice of κ is a hyper-parameter and should be experimented with to see which values lead to optimal results. It is used as a cutoff for which points within a κ distance are considered close together and which points outside a κ distance are considered far apart. As the number of data-points increases, we can be more sure if two points are close given the context of the data and we can adjust κ accordingly. This is a similar struggle to the choice of η seen in section 2. Note that asymptotically d will result in the same nonconformity values as those produced in $||\cdot||$ when using the nearest neighbor method if we allow $\kappa \to 0$ as $n \to \infty$.

3.2. **Synthetic Simulation.**

To illustrate how these methods would be implemented in practice we present the following simulation. With U referring to a uniform distribution, we generate observations from a distribution (2.3) according to the following parameters

$$S_0 = [0,\ 125],\ S_1 = [75,\ 200]\ , f_Y(0) = .7,\ f_Y(1) = .3,\ f_X(0) = U(S_0),\ f_X(1) = U(S_1).$$

Unbeknownst to the experimenter, there is a higher likelihood an observation which falls in the interval [75,125] will have true response 0. This imbalance implies there should be a benefit to normalizing based on our conditioned response. We will now describe a simulation performed in the programming language Julia. We initialize our parameters and generate our synthetic data. The code's prefix and functions used below can be found in the appendix.

```
a0, b0 = 0, 125
a1, b1 = 75, 200
n, p = 1000, .7
ς0, ς1 = 2, 4
κ = 1

Data = SyntheticData(a0, b0, a1, b1, p, n)
```

We then calculate the values of A and A^* with $||\cdot||_A = ||\cdot||_1$ and generating $||\cdot||_B$ as described in section 3.1

```
A = NNorm(Data[1], Data[2], 1, 1, κ)
AN = NNorm(Data[1], Data[2], ς0, ς1, κ)
histogram(A)
histogram(AN)
```

This generates two plots. On the left, we have the histogram of the standard non-conformity scores (blue). On the right, we have the histogram for the normalized non-conformity scores (red). These fit our asymptotic expectations quite well.

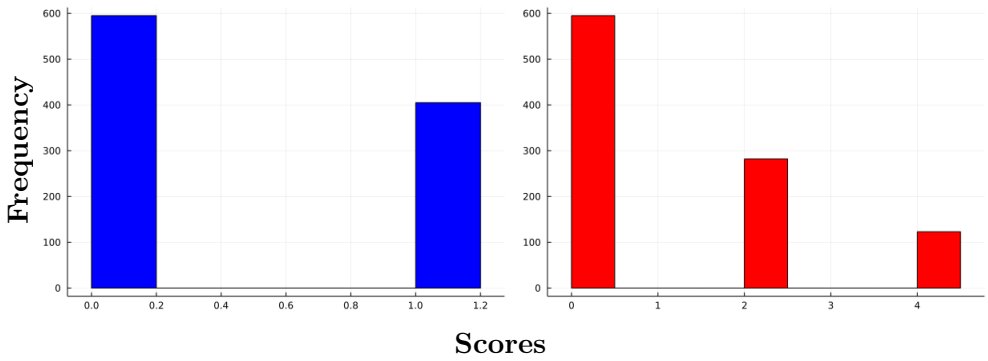

FIGURE 1. The histograms of our simulated non-conformity scores for un-normalized (left, blue) and normalized(right, red)

We can now begin to test the efficiency of these non-conformity scores by using a Monte-Carlo simulation to estimate the expected size of our prediction sets. We generate another 10000 test observations from our distribution. Then, calculate the p-value of these testing observations in both the un-normalized and normalized setting for response 0 versus response 1.

```
Test = vcat(SyntheticData(a0, b0, a1, b1, p, 10000)...)

P0, P1, PN0, PN1 = [], [], [], []

for x in ProgressBar(Test)
    M = FindPVal(x, Data[1], Data[2], A, 1, 1, κ)
    MN = FindPVal(x, Data[1], Data[2], AN, ς0, ς1, κ)
    push!(P1, M[2])
    push!(P0, M[1])
    push!(PN1, MN[2])
    push!(PN0, MN[1])
end
```

By comparing these p-values to a range of ε values, we can then begin to see if our simulation confirms our asymptotic results. Does normalization aid in improving the efficiency of prediction set sizes? We choose ε in steps of size .001 from the range $[0, .5]$.

```
Eff, EffN = [], []

for ϵ in range(-.0001, .5, step = .001)
    Predict1 = A1 .> ϵ
    Predict0 = A0 .> ϵ
    push!(Eff, mean(Predict0 .+ Predict1))
    PredictN1 = AN1 .> ϵ
    PredictN0 = AN0 .> ϵ
    push!(EffN, mean(PredictN0 .+ PredictN1))
```

```
end

plot(Eff)
plot!(EffN)
```

This will produce a plot with functions of ε vs average set size for the un-normalized case (solid blue) and the normalized case (dotted red). Whichever line is lower will be the one with better set efficiency. Our theory dictates that the normalized case will be AME, meaning there should exist some area where the red line is below the blue line. Our simulations show this holds true.

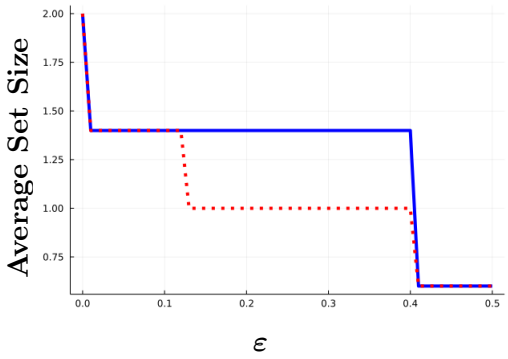

Figure 2. Average set sizes given a ε for un-normalized (solid blue) versus normalized (dotted red)

We can also reverse the normalization constants (swapping ς_0 and ς_1) and repeat this experiment to get the normalization in the opposite manner. Although it is less efficient than the correct orientation of the normalization constants, it is still superior to the un-normalized non-conformity. The histogram of scores for the reversed constants and the comparison to previous efficiencies (with reversed constants as dashed green) are seen below. This confirms the theoretical result stating that any amount of normalization will improve efficiency. This is independent of the assignment of the normalization constants.

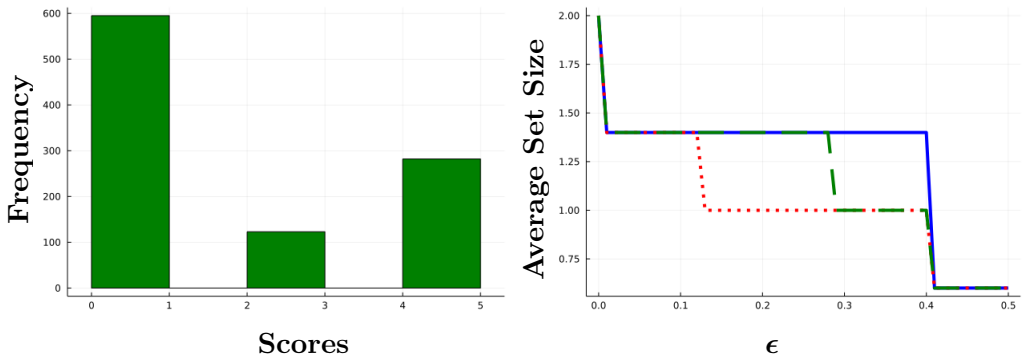

Figure 3. Histogram of reversed normalized non-conformity scores (left), Average set sizes given a ε for un-normalized (solid blue) vs normalized (dotted red) vs reversed normalized (dashed green)(right)

4. Conclusion

Conformal predictions are a double-edged sword. On one hand they offer a reduced level of assumptions and can provide very convenient prediction sets for classification problems. On the other hand they are ripe for unneeded complications and over-fitting which can create falsified results in research. As such, there is a need for guarantees for better performing non-conformity measures under a broad class of probability functions. This paper created such a guarantee. Under the well known nearest neighbor non-conformity measure [5], we showed asymptotically, normalization proposed by Papodopolus [4] produces better prediction sets. Further research needs to explore comparing the nearest neighbor measure (2.5) to the mean measure (2.1). It is postulated that this is due to the connected-ness and variance of the distribution. As well as showing if the relaxation or constriction of η in (2.10) has an effect on the efficiency of prediction sets with small n. Initial simulations show with a good choice in $||\cdot||$ that the asymptotic theory is viable to be applied in non-asymptotic cases but more work needs to be done.

Acknowledgements

I would like to thank Claire Vincent for helping the proof writing process and providing some subtle changes to improve the quality of this paper.

References

[1] Ulf Johansson et al. "Model-agnostic nonconformity functions for conformal classification". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017, pp. 2072–2079. DOI: 10.1109/IJCNN.2017.7966105.

[2] Damjan Krstajic. "Critical Assessment of Conformal Prediction Methods Applied in Binary Classification Settings". In: *Journal of Chemical Information and Modeling* 61.10 (Sept. 2021), pp. 4823–4826. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.1c00549.

[3] Zhe Lim and Anthony Bellotti. "Normalized nonconformity measures for automated valuation models". In: *Expert Systems with Applications* 180 (2021), p. 115165. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115165.

[4] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. "Regression Conformal Prediction with Nearest Neighbours". In: *J. Artif. Int. Res.* 40.1 (Jan. 2011), pp. 815–840. ISSN: 1076-9757.

[5] Glenn Shafer and Vladimir Vovk. *A tutorial on conformal prediction*. 2007. arXiv: 0706.3188 [cs.LG].

5. Appendix

Lemma 1: $A^M(Z,z) \xrightarrow{p} \overline{A^M}(z)$

Proof. A sufficient condition is that $\sum_{x \in Z_{y_i}} x \Big/ \#Z_{y_i} \xrightarrow{p} E[x \in Z_{y_i}]$, the weak law of large numbers affirms this.

∎

Lemma 2: For $x \in S_0 \cap S_1$, $A^{NN}(Z,(x,y)) \xrightarrow{p} 1$

Proof. Let $\varepsilon > 0$

$S_0 \cap S_1$ has no isolation points, thus we can define an ε-ball around any point in this set and have non-empty intersection with $S_0 \cap S_1$

$$\text{As } f(x,y) > 0 \text{ for } x \in S_0 \cap S_1,\ y = \{0,1\}$$

$$\implies \mathbb{F}_i = \int_{x_0 \times y} f(x_0, y) > 0 \text{ for } x_0 \in \varepsilon\text{-ball} \cap S_0 \cap S_1,\ y = i$$
$$\implies 1 - \mathbb{F}_i < 1 \text{ for } x_0 \in \varepsilon\text{-ball} \cap S_0 \cap S_1,\ i = 0 \text{ or } 1$$

We can then define $1 - \mathbb{F}_0 < \phi_0 < 1$ & $1 - \mathbb{F}_1 < \phi_1 < 1$

$$\implies 0 \leq P(|A^{NN}(Z, (x, y)) - 1| > \varepsilon) < \max\{\phi_0^{n-1}, \phi_1^{n-1}\}$$

This is because $|A^{NN}(Z, (x, y)) - 1| > \varepsilon$ when there is a single observation within an epsilon and all others outside it. As $\lim_{n\to\infty} \phi_0^n = 0$ and $\lim_{n\to\infty} \phi_1^n = 0$ then $\lim_{n\to\infty} \max\{\phi_0^n, \phi_1^n\} = 0$, we invoke the squeeze theorem to have

$$\lim_{n\to\infty} P(|A^{NN}(Z, (x, y)) - 1| > \varepsilon) = 0 \implies A^{NN} \xrightarrow{p} 1$$

∎

Code Prefix and Functions

```
using Plots
using Distributions
using Random
using ProgressBars
using Base.Sort

global δ= .0001

function SyntheticData(a0, b0, a1, b1, p, n)
    D = Binomial(n, p)
    D0 = Uniform(a0, b0)
    D1 = Uniform(a1, b1)
    n0 = rand(D)
    n1 = n - n0
    X0 = rand(D0, n0)
    X1 = rand(D1, n1)
    return X0, X1
end

function NNorm(X0, X1, ς0, ς1, κ)
    A = []
    for x in ProgressBar(X0)
        push!(A, ς0 * ((partialsort(abs.(X0 .- x), 2) > κ) + δ)/((minimum(abs.(X1 .- x)) > κ
            ) + δ))
    end
    for x in ProgressBar(X1)
        push!(A, ς1 * ((partialsort(abs.(X1 .- x), 2) > κ) + δ)/((minimum(abs.(X0 .- x)) > κ
            ) + δ))
    end
    return A
end

function FindPVal(x, X0, X1, A, ς0, ς1, κ)
    α1 = ς1 * ((minimum(abs.(X1 .- x)) > κ) + δ) / ((minimum(abs.(X0 .- x)) > κ) + δ)
    α0 = ς0 * ((minimum(abs.(X0 .- x)) > κ) + δ) / ((minimum(abs.(X1 .- x)) > κ) + δ)
```

```
    p0 = mean(A .>= α0)
    p1 = mean(A .>= α1)
    return p0, p1
end
```

Department of Mathematics, University of Louisiana, Lafayette, LA, USA
E-mail address: maxwelllovig@gmail.com

Received November 8, 2021; revised January 11, 2022.