Efficient Conformal Classification Under Nearest Neighbor

Maxwell Lovig
Univ of Louisiana, Dept of Mathematics

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 which might not be valid.
- Second, we must create methods which are easily applicable to complex non-linear models.

1 Introduction

Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization

Overzealous Normalization
Malicious Non-conformity

Definitions

Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof

Conclusion

References

Conformal Predictions

- Shafer and Vovk introduced the conformal statistical framework [3].
- Instead of assuming all observations are drawn $Z_1, \cdots, Z_N \overset{iid}{\sim} f_Z(z)$, Conformal Predictions assumes exchangeability.

Meaning, the $N!$ possible orderings of our observations are equally likely. Written formally, with Ω as a set of all possible permutation of our observations

$\forall \omega_1, \omega_2 \in \Omega,$

$$f_{Z_{\omega_1(1)},\cdots,Z_{\omega_1(N)}}\left(z_{\omega_1(1)}, \cdots, z_{\omega_1(N)}\right) = f_{Z_{\omega_2(1)},\cdots,Z_{\omega_2(N)}}\left(z_{\omega_2(1)}, \cdots, z_{\omega_2(N)}\right)$$

Maxwell Lovig

Introduction
2 Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

- With this we can implement Conformal Classification Prediction.
- 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 label $y_i \in Y$.
- 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" of observation $\tilde{z}$.

Written formally, when $\#\tilde{Z} = u$ and $\#\tilde{z} = v$

$$A : \mathbb{R}^{u \times v} \times \mathbb{R}^v \mapsto \mathbb{R}$$

$$A(\tilde{Z}, \tilde{z}) \uparrow \implies \text{ a more non-conformal occurrence of } \tilde{z}$$

Maxwell Lovig

Introduction

3 Conformal Predictions

The Algorithm

Nearest Neighbor Measure

Normalization

Overzealous Normalization

Malicious Non-conformity

Definitions

Results

Assumptions

Theorem 1

Proof

Theorem 2

Proof

Conclusion

References

The Algorithm

- 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 in the future.

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

Algorithm 1: Conformal Prediction Alogorithm

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

Maxwell Lovig

Introduction
Conformal Predictions
5 The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

Nearest Neighbor

- 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 measure (NN) proposed by Vovk [3].
- With $x \in Z_{y_i}$ denoting the set of observations from Z with label y_i, with norm $||\cdot||$, we have

$$A^{NN}(Z \setminus z, z) = A^{NN}(\tilde{Z}, (x^*, y_i)) = \frac{\min_{x \in \tilde{Z}_{y_i}} ||x - x^*||}{\min_{x \in \tilde{Z}_{\neg y_i}} ||x - x^*||} \qquad (1)$$

Normalization

- Papadopolous discussed the advantages of allotting differing constants σ which regulate how hard a label y is to predict.
- We assign the difficulty to predict the label y as σ_y. Where, as $\sigma_y \uparrow$ the value is considered easier to predict. This leads to the generic normalized non-conformity function introduced by Papopdopolous [2],

$$A^*(Z, z) = A^*(Z, (x^*, y)) = \frac{A(Z, (x^*, y))}{\sigma_y} \tag{2}$$

$$A^{NN}_*(Z, z) = A^{NN}_*(Z, (x^*, y_i)) = \varsigma_y A(Z, (x^*, y_i)) \tag{3}$$

$$\varsigma_y = \begin{cases} \varsigma_0 & \text{if } y = (1, 0) \\ \varsigma_1 & \text{if } y = (0, 1) \end{cases}$$

Overzealous Normalization

- We can add extra normalization terms and different criterion to try to minimize our prediction set size with intuition on what our intervals prefer.
- 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 showed that there is influence on the efficiency of the prediction sets empirically from the choice of normalization on the Ames housing data, but there is no theoretical connection as of yet [1].

Malicious Non-conformity

One can then ask the question, why not add as much normalization as possible? This is a fair idea until one considers the normalization function

$$\sigma^*_{(x,y)} = \begin{cases} \infty & \text{if } (x,y) \in \hat{Z} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This would lead all of our prediction sets from $\hat{Z}$ being with $\Gamma^{A^*}_{\epsilon} \leq 1$ but all other future predictions will have $\#\Gamma^{A^*}_{\epsilon} = \#Y$.

Maxwell Lovig

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
9 Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

Definitions

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 [3] is the mean distance non-conformity, defined as

$$A^M(Z, z) = A^M(Z, (x^*, y_i)) = \left|\left| x^* - \sum_{x \in Z_{y_i}} x \Big/ \#Z_{y_i} \right|\right| \quad (5)$$

$$\implies \overline{A}(z) = \overline{A}(x^*, y) = ||x^* - E[x \in Z_{y_i}]||$$

Maxwell Lovig

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
10 Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

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 $E[\#\Gamma_\varepsilon^{\overline{A_1}}] = \varepsilon$ and A_2 has $E[\Gamma_\varepsilon^{\overline{A_2}}] = \varepsilon^2$, then A_2 is AME than A_1.

DEF: If $A_*(Z, z) = \frac{A(Z,z)}{\sigma_*}$ is more efficient then 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 (4) we restrict our σ to a function of y, making our normalization σ_y.

Maxwell Lovig

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
11 Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

Let us consider $Z = (X, Y)$ drawn from a bounded space $S \subset \mathbb{R}^n$, these vectors can have the corresponding label y_0 when $X \in S_0 \subset S$ and can have label y_1 when $X \in S_1 \subset S$. We define the probability distribution of X, Y

$$f_Z(z) = f_{X,Y}(x, y) = f_Y(0) f_0(x) I_{x \in S_0} + f_Y(1) f_1(x) I_{x \in S_1} \quad (6)$$

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 $y_0 + y_1 = 1$.

Theorem 1: *Under (6), the normalized non-conformity function A_*^{NN} is AME than A^{NN} then $S_0 \cap S_1 \neq \varnothing$. If S_0 and S_1 are disjoint then neither measure 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 peisewise 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

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

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
14 Proof
Theorem 2
Proof
Conclusion
References

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:

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

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

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

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

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

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

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

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

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

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
Conclusion
References

When $\varsigma_0 > \varsigma_1$ we have

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

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

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

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

$$A_{\varsigma}^{NN}(Z, z) = \varsigma_y A^{NN}(Z, (x^*, y_i))$$
$$\varsigma(y_i) = \frac{\#Z_{y_i}}{\#\{Z_{y_i} \text{ s.t. } A(Z \setminus z, z) \geq \eta\}} \tag{7}$$

where $\eta > 0$, A_{ς}^{NN} is AMEF under A^{NN}

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
19 Theorem 2
Proof
Conclusion
References

Proof.

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

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

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

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

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

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

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

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

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

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

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

Conclusion

- Conformal predictions are a double-edged sword, they offer a reduced level of assumptions but they are ripe for unneeded complication and over-fitting which can create falsified results in research.
- We need better guidance on how to choose a NC measure. Under the well known nearest neighbor non-conformity measure [3], we showed asymptotically, normalization proposed by Papodopolus [2] produces better prediction sets.
- Further research needs to be explores into comparing the nearest neighbor measure (1) to the mean measure (5). As well as showing if the relaxation or constriction of η in (7) has an effect on the efficiency of prediction sets with small n.

Introduction
Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization
Overzealous Normalization
Malicious Non-conformity
Definitions
Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof
21 Conclusion
References

References

[1] 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.

[2] 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.

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

Thank You!

Introduction

Conformal Predictions
The Algorithm
Nearest Neighbor Measure
Normalization

Overzealous Normalization
Malicious Non-conformity

Definitions

Results
Assumptions
Theorem 1
Proof
Theorem 2
Proof

Conclusion

23 References