pyagree Documentation

Release 0.0.4

Alberto Casagrande

Aug 28, 2020

CONTENTS

pyagree is a simple Python module containing some of the main inter-rater agreement measures.

CHAPTER

ONE

INSTALLING *PYAGREE*

There are two main ways to install *pyagree*. The first one requires to clone pyagree's repository from GitHub by using the following command:

```
git clone https://github.com/albertocasagrande/pyagree.git
```

Once the command has fully been executed, the repository directory must be entered and the package can be installed by issuing as superuser (administrator or root) the following commands:

```
cd pyagree
python setup.py install
```

The second method to install *pyagree* is easier, but requires the tool *pip* (see pypi) and it may install a non-bleeding edge version of the package. As superuser, issue the command:

```
pip install pyagree
```

CHAPTER

TWO

USING *PYAGREE*

2.1 Importing *pyagree* Functions

In order to use *pyagree*, it is sufficient to import it, for instance, by using the statement:

```
import pyagree
```

After that, all the functions reported in the *API* can be invoked as follows:

```
pyagree.<function name>(<parameter1>, ...)
```

Alternatively, the desired functions can be individually imported and avoiding the package name prefix as in:

```
from pyagree import <function name>

<function name>(<parameter1>, ...)
```

2.2 Working Examples

For instance:

```
>>> from pyagree import bangdiwala_b, cohen_kappa

>>> A = [[10,  1],
...      [ 5, 10]]

>>> bangdiwala_b(A)

0.6060606060606061

>>> cohen_kappa(A)

0.5491329479768786
```

evaluates both Bangdiwala's B and Cohen's κ of the agreement matrix

$$A = \begin{pmatrix} 10 & 1 \\ 5 & 10 \end{pmatrix}$$

and print them in output.

2.2.1 NumPy Support

All the *pyagree* functions natively support both standard "list-of-list" representation of matrices and NumPy matrices.

```
>>> import numpy
>>> from pyagree import scott_pi

>>> A = [[0,1,2],
...      [3,4,5],
...      [6,7,8]]

>>> B = numpy.matrix(A)

>>> scott_pi(A)

-0.09090909090909094

>>> scott_pi(B)

-0.09090909090909094
```

2.2.2 Matrix Sizes and Exceptions

Whenever, the matrix size is not supported either by the agreement measure or by the corresponding *pyagree* function, an opportune *ValueError* is raised.

```
>>> from pyagree import cohen_kappa, yule_y

>>> A = [[0,1],
...      [2,3],
...      [4,5]]

>>> B = [[0,1,2],
...      [3,4,5],
...      [6,7,8]]

>>> C = [[0,1],
...      [2,3]]

>>> cohen_kappa(A)

Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/python3.8/site-packages/pyagree/standard.py", line 81, in cohen_kappa
    test_agreement_matrix(A)
  File "/usr/lib/python3.8/site-packages/pyagree/common.py", line 29, in test_
↪agreement_matrix
    raise ValueError("Non-squared matrix")
ValueError: Non-squared matrix

>>> cohen_kappa(B)

-0.0666666666666667

>>> yule_y(B)
```

(continues on next page)

(continued from previous page)

```
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/python3.8/site-packages/pyagree/standard.py", line 149, in yule_y
    raise ValueError("The agreement matrix must be a 2x2-matrix")
ValueError: The agreement matrix must be a 2x2-matrix

>>> yule_y(C)

-1.0
```

CHAPTER

THREE

AGREEMENT MEASURES

3.1 General Notions

Inter-rater agreement (also known as **inter-rater reliability**) is a measure of consensus among n raters in the classification of N objects in a k different categories.

In the general case, the rater evalutations can be represented by the **reliability data matrix**: a $n \times N$-matrix R such that $R[i,j]$ stores the category selected by the i-th rater for the j-th object.

A more succint representation is provided by a $N \times k$-matrix C whose elements $C[i,j]$ account how many raters evaluated the i-th object as belonging to the j-th category. This matrix is the **classification matrix**.

Whenever the numer of raters is 2, i.e., $n = 2$, the rater evaluations can be represented by the **agreement matrix**: a $k \times k$-matrix A such that $A[i,j]$ stores the number of objects that are classified at the same time as belonging to the i-th category by the first rater and to the j-th category by the second rater.

3.2 Bennett, Alpert and Goldstein's S

Bennett, Alpert and Goldstein's S is an inter-rater agreement measure on nominal scale (see [BAG54] and [War12]). It is defined as:

$$S \stackrel{\text{def}}{=} \frac{k * P_0 - 1}{k - 1}$$

where P_0 is the probability of agreement among the raters and k is the number of different categories in the classification.

3.3 Bangdiwala's B

Bangdiwala's B is an inter-rater agreement measure on nominal scale (see [MB97]). It is defined as:

$$B \stackrel{\text{def}}{=} \frac{\sum_i A[i,i]}{\sum_i A_{i\cdot} * A_{\cdot i}}$$

where $A_{i\cdot}$ and $A_{\cdot i}$ are the sums of the elements in the i-th row and i-th column of the matrix A, respectively.

3.4 Cohen's Kappa

Cohen's κ is an inter-rater agreement measure on nominal scale (see [Coh60]). It is defined as:

$$\kappa \stackrel{\text{def}}{=} \frac{P_0 - P_e}{1 - P_e}$$

where P_0 is the probability of agreement among the raters and P_e is the agreement probability by chance.

3.5 Scott's Pi

Scott's π is an inter-rater agreement measure on nominal scale (see [Sco55]). Similarly to Cohen's κ, it is defined as:

$$\pi \stackrel{\text{def}}{=} \frac{P_0 - P_e}{1 - P_e}$$

where P_0 is the probability of agreement among the raters (as in Cohen's κ) and P_e is the sum of the squared joint proportions (whereas it is the sum of the squared geometric means of marginal proportions in Cohen's κ). In particular, the *joint proportions* are the arithmetic means of the marginal proportions.

3.6 Yule's Y

Yule's Y (see [Yul12]), sometime called *coefficient of colligation*, measures the relation between two binary random variables (i.e., it can be computed exclusively on 2×2 agreement matrices). It is defined as:

$$Y \stackrel{\text{def}}{=} \frac{\sqrt{\text{OR}} - 1}{\sqrt{\text{OR}} + 1}$$

where OR is the *odds ratio* (e.g., see here):

$$\text{OR} \stackrel{\text{def}}{=} \frac{A[0,0] * A[1,1]}{A[1,0] * A[0,1]}.$$

3.7 Fleiss's Kappa

Fleiss's κ (see [Fle71]) is a multi-rater generalization of *Scott's Pi*.

If the classifications are represented in a classification matrix (see *General Notions*), the ratio of classifications assigned to class j is:

$$p_j \stackrel{\text{def}}{=} \frac{1}{N * n} \sum_{i=1}^{N} C[i,j]$$

and their square sum is:

$$\bar{P}_e \stackrel{\text{def}}{=} \sum_{j=1}^{k} p_j.$$

Instead, the ratio between the pairs of raters which agree on the i-th subject and the overall pairs of raters is:

$$P_i \stackrel{\text{def}}{=} \frac{1}{n * (n-1)} \left(\left(\sum_{j=1}^{k} C[i,j]^2 \right) - n \right)$$

and its mean is:

$$\bar{P} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} P_i.$$

Fleiss's κ is defined as:

$$\kappa \stackrel{\text{def}}{=} \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}.$$

3.8 Information Agreement

The *Information Agreement*, (IA), is an inter-rater agreement measure on nominal scale (see [CFG20a]) which gauges the dependence between the classifications of two raters.

The probability distributions for the evaluations of the rater $\mathfrak{X}$, those of the rater $\mathfrak{Y}$, and the joint evalutions $\mathfrak{X}\mathfrak{Y}$ on the agreement matrix A are:

$$p_{X_A}(j_0) \stackrel{\text{def}}{=} \frac{\sum_i A[i, j_0]}{\sum_i \sum_j A[i, j]}, \qquad p_{Y_A}(i_0) \stackrel{\text{def}}{=} \frac{\sum_j A[i_0, j]}{\sum_i \sum_j A[i, j]},$$

and

$$p_{X_A Y_A}(i_0, j_0) = \frac{A[i_0, j_0]}{\sum_i \sum_i A[i, j]},$$

respectively. The entropy functions for the random variables X_A, Y_A, and $X_A Y_A$ are:

$$H(X_A) \stackrel{\text{def}}{=} -\sum_i p_{X_A}(i) \log_2 p_{X_A}(i), \qquad H(Y_A) \stackrel{\text{def}}{=} -\sum_j p_{Y_A}(j) \log_2 p_{Y_A}(j),$$

and

$$H(X_A Y_A) \stackrel{\text{def}}{=} -\sum_i \sum_j p_{X_A Y_A}(i, j) \log_2 p_{X_A Y_A}(i, j).$$

The mutual information between the classification of $\mathfrak{X}$ and $\mathfrak{Y}$ is:

$$I(X_A, Y_A) \stackrel{\text{def}}{=} H(X_A) + H(Y_A) - H(X_A Y_A).$$

The *Information Agreement* of A is the ratio between $I(X_A, Y_A)$ and the minimum among $H(X_A)$ and $H(Y_A)$ as ϵ tends to 0 from the right, i.e.,

$$\text{IA} \stackrel{\text{def}}{=} \frac{I(X_A, Y_A)}{\min(H(X_A), H(Y_A))}.$$

3.8.1 Extension-by-Continuity of IA

IA was proven to be effetive in gauging agreement and solves some of the pitfalls of Cohen's κ. However, it is not defined over all the agreement matrices and, in particular, it cannot be directly computed on agreement matrices containing some zeros (see [CFG20b]).

The *extension-by-continuity of Information Agreement*, (IA_C), extends IA's domain so that it can deal with matrices containing some zeros (see [CFG20b]). In order to achieve this goal, the considered agreement matrix A is replaced by the symbolic matrix A_ϵ is defined as:

$$A_\epsilon[i, j] \stackrel{\text{def}}{=} \begin{cases} A[i, j] & \text{if } A[i, j] \neq 0 \\ \epsilon & \text{if } A[i, j] = 0 \end{cases}$$

where ϵ is a real variable with values in the open interval $(0, +\infty)$. On this matrix, mutual information of the variables X_{A_ϵ} and Y_{A_ϵ} and their entropy functions are defined. The *extension-by-continuity of Information Agreement* of A is the limit of the ratio between $I(X_{A_\epsilon}, Y_{A_\epsilon})$ and the minimum among $H(X_{A_\epsilon})$ and $H(Y_{A_\epsilon})$ as ϵ tends to 0 from the right, i.e.,

$$\mathrm{IA}_C(A) \stackrel{\text{def}}{=} \lim_{\epsilon \to 0^+} \frac{I(X_{A_\epsilon}, Y_{A_\epsilon})}{\min(H(X_{A_\epsilon}), H(Y_{A_\epsilon}))}.$$

$\mathrm{IA}_C(A)$ was proven to be defined over any non-null agreement matrix having more than one row/column and, if l and m are numbers of non-null columns and non-null rows in A, respectively, then:

$$\mathrm{IA}_C(A) = \begin{cases} 1 - \frac{m}{k} & \text{if } H(\overline{X_A}) = 0 \\ 1 - \frac{l}{k} & \text{if } H(\overline{Y_A}) = 0 \\ \frac{I(\overline{X_A}, \overline{Y_A})}{\min\left(H\left(\overline{X_A}\right), H\left(\overline{Y_A}\right)\right)} & \text{otherwise} \end{cases}$$

where $\overline{X_A}$, $\overline{Y_A}$, and $\overline{X_A Y_A}$ are three random variables having the same probability distributions of X_A, Y_A, and $X_A Y_A$ except for 0-probability events which are removed from their domains (see [CFG20b]).

3.9 References

CHAPTER

FOUR

API

4.1 *pyagree* API

`pyagree.`**`bennett_s`**(*agreement_matrix*)

Evaluate Bennett, Alpert and Goldstein's S

Compute the *Bennett, Alpert and Goldstein's S* of agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – An $n \times n$-agreement matrix

Returns The Bennett, Alpert and Goldstein's S of agreement_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`scott_pi`**(*agreement_matrix*)

Evaluate Scott's π

Compute the *Scott's Pi* of agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – An $n \times n$-agreement matrix

Returns The Scott's π of agreement_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`yule_y`**(*agreement_matrix*)

Evaluate Yule's Y

Compute the *Yule's Y* of a 2×2-agreement matrix agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – A 2×2-agreement matrix

Returns The Yule Y of agreement_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`bangdiwala_b`**(*agreement_matrix*)

Evaluate Bangdiwala's B

Compute the *Bangdiwala's B* of agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – An $n \times n$-agreement matrix

Returns The Bangdiwala's B of agreement_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`cohen_kappa`**(*agreement_matrix*)

Evaluate Cohen's κ

Compute *Cohen's Kappa* of agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – An $n \times n$-agreement matrix

Returns The Cohen's κ of agreement_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`fleiss_kappa`**(*classification_matrix*)

Evaluate Fleiss's κ

Compute the *Fleiss's Kappa* of a classification matrix classification_matrix.

Parameters **`classification_matrix`** (class:*numpy.ndarray*) – An $N \times k$-classification matrix

Returns The Fleiss's κ of classification_matrix

Return type `float`

Raises `ValueError`

`pyagree.`**`ia_c`**(*agreement_matrix*)

Evaluate *extension-by-continuity of Information Agreement*

Compute the *Extension-by-Continuity of IA* (IA_C) of agreement_matrix.

Parameters **`agreement_matrix`** (`numpy.ndarray`) – An agreement matrix

Returns The extension-by-continuity of Information Agreement of agreement_matrix

Return type `float`

Raises `ValueError`

CHAPTER

FIVE

INDICES AND TABLES

- genindex
- search

BIBLIOGRAPHY

[BAG54] E.M. Bennett, R. Alpert, and A.C. Goldsein. Communications Through Limited-Response Questioning. *Public Opinion Quarterly*, 18(3):303–308, 01 1954. doi:10.1086/266520.

[CFG20a] Alberto Casagrande, Francesco Fabris, and Rossano Girometti. Beyond Kappa: An Informational Index for Diagnostic Agreement in Dichotomous and Multivalue Ordered-Categorical Ratings. Submitted for the publication, 2020.

[CFG20b] Alberto Casagrande, Francesco Fabris, and Rossano Girometti. Extending Information Agreement by Continuity. 2020.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

[Fle71] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi:10.1037/h0031619.

[MB97] Sergio R. Munoz and Shrikant I. Bangdiwala. Interpretation of Kappa and B statistics measures of agreement. *Journal of Applied Statistics*, 24(1):105–112, 1997. doi:10.1080/02664769723918.

[Sco55] William A. Scott. Reliability of content analysis: the case of nominal scale coding. *The Public Opinion Quarterly*, 19(3):321–325, 1955. doi:10.1086/266577.

[War12] Matthijs J. Warrens. The effect of combining categories on bennett, alpert and goldstein's s. *Statistical Methodology*, 9(3):341–352, 2012. doi:10.1016/j.stamet.2011.09.001.

[Yul12] George Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912. doi:10.2307/2340126.

INDEX