The Journal of Open Source Software

PyCM: Multiclass confusion matrix library in Python

Sepand Haghighi[1], **Masoomeh Jasemi**[1], **Shaahin Hessabi**[1], and **Alireza Zolanvari**[2]

1 Sharif University of Technology **2** Amirkabir University of Technology

DOI: 10.21105/joss.00729

Software

- Review
- Repository
- Archive

Submitted: 28 April 2018
Published: 28 May 2018

Summary

In the field of machine learning and specifically for statistical classification a confusion matrix - also known as error matrix - is a specific table layout that allows visualization of the algorithm performance, and is mostly used in supervised learning. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predict class (or vice versa)(Powers 2011; Sammut and Webb 2010; Shepperd, Bowes, and Hall 2014; X. Deng et al. 2016).

PyCM is a multi-class confusion matrix library written in Python that supports both input data vectors and direct matrix, and a proper tool for post-classification model evaluation that supports most classes and overall statistics parameters (Landis and Koch 1977; Fleiss 1971; Altman 1990; Gwet 2008; Scott 1955; Bennett, Alpert, and Goldstein 1954; Cicchetti 1994; Davies 1980; Kullback and Leibler 1951; Goodman and Kruskal 1972, 1963; Byrt, Bishop, and Carlin 1993).

We can categorize these statistics in 3 sections:

1. Basic
2. Class Statistics
3. Overall Statistics

PyCM is also capable of generating report in HTML, CSV and .pycm formats.

To sum it up, PyCM is the swiss-army knife of confusion matrices, targeted mainly at data scientists that need a broad array of metrics for predictive models and an accurate evaluation of large variety of classifiers (Haghighi, Jasemi, and Hessabi 2018).

Website : http://pycm.shaghighi.ir

References

Altman, Douglas G. 1990. *Practical Statistics for Medical Research.* CRC press.

Bennett, E. M., R. Alpert, and A. C. Goldstein. 1954. "Communications Through Limited Response Questioning." *Public Opinion Quarterly* 18 (3). Oxford University Press (OUP):303. https://doi.org/10.1086/266520.

Byrt, Ted, Janet Bishop, and John B. Carlin. 1993. "Bias, Prevalence and Kappa." *Journal of Clinical Epidemiology* 46 (5). Elsevier BV:423–29. https://doi.org/10.1016/0895-4356(93)90018-v.

Cicchetti, Domenic V. 1994. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment* 6 (4). American Psychological Association (APA):284–90. https://doi.org/10.1037/1040-3590.6.4.284.

Davies, Robert B. 1980. "Algorithm AS 155: The Distribution of a Linear Combination of Chi-Squared Random Variables." *Applied Statistics* 29 (3). JSTOR:323. https://doi.org/10.2307/2346911.

Deng, Xinyang, Qi Liu, Yong Deng, and Sankaran Mahadevan. 2016. "An Improved Method to Construct Basic Probability Assignment Based on the Confusion Matrix for Classification Problem." *Information Sciences* 340-341 (May). Elsevier BV:250–61. https://doi.org/10.1016/j.ins.2016.01.033.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin* 76 (5). American Psychological Association (APA):378–82. https://doi.org/10.1037/h0031619.

Goodman, Leo A., and William H. Kruskal. 1963. "Measures of Association for Cross Classifications III: Approximate Sampling Theory." *Journal of the American Statistical Association* 58 (302). Informa UK Limited:310–64. https://doi.org/10.1080/01621459.1963.10500850.

———. 1972. "Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances." *Journal of the American Statistical Association* 67 (338). Informa UK Limited:415–21. https://doi.org/10.1080/01621459.1972.10482401.

Gwet, Kilem Li. 2008. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." *British Journal of Mathematical and Statistical Psychology* 61 (1). Wiley-Blackwell:29–48. https://doi.org/10.1348/000711006x126600.

Haghighi, Sepand, Masoomeh Jasemi, and Shaahin Hessabi. 2018. "Pycm : Multi Class Confusion Matrix Library in Python." Zenodo. https://doi.org/10.5281/zenodo.1157173.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1). Institute of Mathematical Statistics:79–86. https://doi.org/10.1214/aoms/1177729694.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1). JSTOR:159. https://doi.org/10.2307/2529310.

Powers, David Martin. 2011. "Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness and Correlation." Bioinfo Publications. https://doi.org/10.9735/2229-3981.

Sammut, Claude, and Geoffrey I. Webb, eds. 2010. *Encyclopedia of Machine Learning*. Springer US. https://doi.org/10.1007/978-0-387-30164-8.

Scott, William A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding." *Public Opinion Quarterly* 19 (3). Oxford University Press (OUP):321. https://doi.org/10.1086/266577.

Shepperd, Martin, David Bowes, and Tracy Hall. 2014. "Researcher Bias: The Use of Machine Learning in Software Defect Prediction." *IEEE Transactions on Software Engineering* 40 (6). Institute of Electrical; Electronics Engineers (IEEE):603–16. https://doi.org/10.1109/tse.2014.2322358.