XVI LATIN AMERICAN CONGRESS OF PROBABILITY AND MATHEMATICAL STATISTICS

BOOK OF ABSTRACTS

IME - USP

JULY 10 - 14

IME.USP.BR/~16CLAPEM/

Institute of Mathematics and Statistics - University of São Paulo

In this talk we will present the main results in . we will discuss the problem of identifying homozygosity islands on the genome of individuals in a population. We present a method that directly tackles the issue of identification of the homozygosity islands at the population level, without the need of analysing single individuals and then combine the results. We propose regularized offline change-point methods to detect changes in the parameters of a multidimensional distribution when we have several aligned, independent samples of fixed resolution. We present a penalized maximum likelihood approach that can be efficiently computed by a dynamic programming algorithm or approximated by a fast binary segmentation algorithm. Both estimators are shown to converge almost surely to the set of change-points without the need of specifying a priori the number of change-points. In simulation, we observed similar performances from the exact and greedy estimators. Moreover, we provide a new methodology for the selection of the regularization constant which has the advantage of being automatic, consistent, and less prone to subjective analysis.

Acknowlegments: I thank Florencia Leonardi, Renan B. Lemes and Tábita Hünemeier for the collaboration in the paper.

Funding: This work was supported by the São Paulo Research Foundation, Brazil [2013/07699-0, 17/10555-0, 20/10136-0, 21/06860-8]; Coordination of Superior Level Staff Improvement, Brazil [Ph.D. fellowship to Lucas Prates]; and the National Council for Scientific and Technological Development, Brazil, [311763/2020-0 to Florencia Leonardi].

References

[1] Prates, Lucas and Lemes, Renan B and Hünemeier, Tábita and Leonardi, Florencia: *Population-based change-point detection for the identification of homozygosity islands*, Bioinformatics, 39, 4, pg.—(2023).

Comparative analysis of variables selection methods in classification problems

Luna Wagner Cunha, Cibele Maria Russo

Department of Applied Mathematics and Statistics, University of Sao Paulo,
Department of Statistics, Federal University of Sao Carlos.

Variable selection methods were proposed to address issues with large variable vectors by identifying the most relevant ones for the model. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017), serves as an aggregator of different interpretability techniques, detailing variable importance for each observation. This study compared variable selection methods using the SHAP package and Lasso. Practical applications using simulated and real databases, specifically the Lending Club dataset from Kaggle[1], yielded interesting results. The findings demonstrate the superiority, on both datasets, of the feature selection by the SHAP method over the commonly used Lasso method.

[1]Available in https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1?resource=download, Accessed in 12/06/2022

References

[1] Lundberg, S. M. e Lee, S.-I.: *A unified approach to interpreting model predictions.*, Em I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, e R. Garnett, editors, Advances in Neural Information Processing Systems 30 , pages 4765–4774. Curran Associates, Inc (2017)

A comparative study of cross validation techniques applied to imbalanced data model

Luiza Tuler Veloso

Universidade de São Paulo

Within the context of predictive modeling, choosing a model involves evaluating, through Expected Risk, the quality of predictions. Such risk, however, may be underestimated if obtained from the same sample utilized to adjust the model. To deal with such problem, Cross Validation strategies (Hold-out, K-Fold, Leave-one-out, Bootstrap) emerge, that seek to split the available data in Training Sample, in which the model will be adjusted, and Validation Sample, where the model will have its performance verified. When dealing with imbalanced data, in other words, data in which the event of interest ($Y = 1$) of the binary response variable occurs dozens to thousands of times less than the other category ($Y = 0$), might need some adaptations in the process of modeling and validation. In view of this, this paper seeks to evaluate the way in which model validation techniques behave, according to the degree of data imbalance and different sample sizes.

Estimation and model selection for mixing graphical models

Magno Tairone de Freitas Severino

University of São Paulo

In this work, which is part of my Ph.D. dissertation, we propose a global model selection criterion to estimate the graph of conditional dependencies of a random vector, whose distribution corresponds to the stationary distribution of a mixing stochastic process. By global criterion we mean the optimization of a function over the set of possible graphs, without the need of estimating the individual neighborhoods and then combining the results, as proposed in [1]. We prove the consistency of the approach and propose a simulated annealing efficient algorithm to estimate the graph of conditional dependencies as well as the parameters of the model.

Acknowlegments: Many thanks to Florencia Leonardi for her invaluable contribution and feedback during the development of this project.

Funding: This work was partially funded by CAPES, CNPq and FAPESP.

References

[1] Leonardi, F. & Carvalho, R.: Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions, arXiv, (2019), `https://arxiv.org/abs/1911.12198`