Optimal Scaling: New Insights Into an Old Problem

Gilbert Saporta

Optimal Scaling: New Insights Into an Old Problem

Gilbert Saporta

Abstract Processing qualitative variables with a very large number of categories in Machine Learning is an opportunity to revisit the theory of optimal scaling and its applications.

Key words: optimal scaling, encoding, correspondence analysis.

1 A Brief History

Coding (or scoring) a qualitative variable consists in assigning numerical values to its modalities, thus transforming it into a discrete numerical variable allowing the use of methods designed for numerical data. Scoring qualitative variables has a long history going back to Hirschfeld [8], Fisher [4], Williams [17], Guttman [6], Hayashi [7,16], among others. It was the origin of correspondence analysis [11,13].

It is interesting to note, as did Kendall & Stuart [9], that Lancaster's theorem [10] implies that the search for separate scoring systems for two categorised variables such as to maximise their correlation, comes down to trying to produce a bivariate normal distribution by operations upon the marginal distributions.

The 1970s and early 1980s were a particularly fertile period for the development of optimal scaling (ie scoring) in supervised and unsupervised contexts, performed with alternating least squares between model parameters and data parameters (the codings) [5,18].

Then, for almost 40 years, the topic did not generate much research; applications became routine, such as risk scores in banking and insurance [2,15].

[1] Gilbert Saporta, Cedric-Cnam, Paris, France; gilbert.saporta@cnam.fr

2 Machine Learning and Variable Encoding

With the availability of massive data, machine-learning researchers and practitioners were confronted with categorical data, ill-suited to neural networks with moreover a large number of categories (eg zip codes).

Generally ignoring the old works of statisticians, dozens of encoding methods [14], some quite arbitrary or rediscovered, have flourished in ML literature, like Hash encoding, methods where the encoding only depends on the response variable (conditional average), as well as the One-Hot Encoding (OHE). OHE is nothing else than the well-known disjunctive form of categorical variables with as many indicators as categories. It should be noted that OHE is more a representation of categories than an encoding, since it is not a transformation into a single numerical variable. A. Di Ciaccio [3] proposes a review of encoding methods from the point of view of a statistician.

Machine Learning also offers methods adapted to massive data. Linear methods can be modelled by neural networks, which in this case are not very efficient computationally, but the use of networks such as auto-encoders with nonlinear links and minimising cross-entropy may give better results than the ALS methods [1].

One of the essential contributions of Machine Learning lies in the learning-validation approach to avoid overfitting. A large number of categories for some variables raises problems of stability and overfitting. These issues were neglected in usual statistical applications where the number of modalities is small. [12] shows how regularization may be applied in the context of regression with optimal scaling features.

3 Conclusions and Perspectives

The confrontation of Statistics and Machine Learning worlds allows us to consider a renewal of the coding methods, from both a theoretical and a practical point of view. It must be emphasised that there is no optimal coding per se: it depends on the problem and on the criterion to be optimised.

References

1. Abdi, H., Di Ciaccio, A., Saporta, G.: Old and New Perspectives on Optimal Scaling. In: Beh, E.J., Lombardo, R., Clavel, J.G. (eds.) Analysis of Categorical Data from Historical Perspectives. Behaviormetrics: Quantitative Approaches to Human Behavior, vol 17, pp. 131-154. Springer, Singapore. (2023)
2. Bouroche, J. M., Saporta, G.: Les méthodes et les applications du credit-scoring. In: 34° Riunione Scientifica della Sociéta Italiana di Statistica, Siena, pp. 19-26. (1988)
3. Di Ciaccio, A.: Optimal Coding of categorical data in machine learning. In: L. Grilli, M. Lupparelli, E. Rocco, C. Rampichini, M. Vichi (eds.), Statistical Models and Methods for Data Science. Studies in Classification, Data Analysis and Knowledge Organization, pp. 39-51. Springer, Cham (2023)

4. Fisher, R.A.: The precision of discriminant functions. Ann. Eugenics **10**, 422–429 (1940)
5. Gifi, A.: Nonlinear Multivariate Analysis. Wiley, Chichester (1990)
6. Guttman, L.: The quantification of a class of attributes: a theory and method of a scale construction. In: Horst, P., Wallin, P., Guttman, L. (eds.) The Prediction of Personal Adjustment, 321–348. Social Science Research Council, New York (1941)
7. Hayashi, C.: On the quantification of qualitative data from the mathematico-statistical point of view. Ann. Inst. Stat. Math. **2**, 35–47 (1950)
8. Hirschfeld, H.O.: A connection between correlation and contingency. Math. Proc. Camb. Philos. Soc. **31**, 520–524 (1935)
9. Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics, vol. II. Charles Griffin, London (1961)
10. Lancaster, H.O.: Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika **44**, 289–292 (1957)
11. Lebart, L., Saporta, G.: Historical elements of correspondence analysis and multiple correspondence analysis. In: Blasius, J., Greenacre, M. (eds.) Visualization and Verbalization of Data, 73–86. Chapman & Hall/CRC, Boca Raton (2014)
12. Meulman, J.J., van der Kooij, A.J., Duisters, K.L.: ROS regression: integrating regularization with optimal scaling regression. Stat. Sci. **34**, 361–390 (2019)
13. Nishisato, S.: Analysis of Categorical Data: Dual Scaling and Its Applications. University of Toronto Press, Toronto (1980)
14. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. **175**(4), 7–9 (2017)
15. Saporta, G., Niang-Keita, N.: Correspondence analysis and classification. In: Greenacre, M., Blasius, J. (eds.) Multiple Correspondence Analysis and Related Methods, pp. 371–392. Chapman and Hall/CRC, Boca Raton (2006)
16. Tanaka, Y.: Review of the methods of quantification. Environ. Health Perspect. **32**, 113–123 (1979)
17. Williams, E.J.: Use of scores for the analysis of association in contingency tables. Biometrika **39**, 274–289 (1952)
18. Young, F.W.: Quantitative analysis of qualitative data. Psychometrika **46**, 357–388 (1981)