universidade de aveiro
departamento de matemática

Fundação para a Ciência e a Tecnologia

Seminário

Grupo de Probabilidades e Estatística

29 de setembro de 2023 14:30 Sala Sousa Pinto

Linear Models for Distributional Data

Sónia Dias

ESTG, Instituto Politécnico de Viana do Castelo e LIAAD-INESC TEC

Abstract

In the classical data framework one numerical value or one category is associated with each individual. However, the interest of many studies lays in groups of records gathered according to characteristics of the individuals or classes of individuals. The classical solution for these situations is to associate with each individual or class of individuals a central measure, e.g., the mean or the mode of the corresponding records; however with this option the variability across the records is lost. For such situations, Symbolic Data Analysis proposes that a distribution or an interval of the individual records' values is associated with each unit, thereby considering new variable types, named symbolic variables (Brito, 2014). One such type of symbolic variable is the histogram-valued variable, where to each entity under analysis corresponds an empirical distribution. If for all observations each unit takes values on only one interval with weight equal to one, the histogram-valued variable is then reduced to the particular case of an interval-valued variable. So, it is necessary to adapt concepts and methods of classical statistics to new kinds of variables.

Currently, the development of models and methods for the representation, analysis, interpretation and organization of distributional data is growing (Dias and Brito, 2022). Linear models are the basis of several statistical methods, such as linear regression and linear discriminant analysis. The Distribution and Symmetric Distribution (DSD) linear regression model proposed in Dias and Brito (2015) allows predicting the distribution of the target variable from other histogram-valued variables, and is obtained optimizing a criterion based on the Mallows distance between the observed and the predicted distributions. As in classical analysis, it is possible to deduce a goodness-of-fit measure from the models whose values

This seminar is supported by CIDMA – Center for Research and Development in Mathematics and Applications through FCT – Fundação para a Ciência e a Tecnologia, within projects UIDB/04106/2020 and UIDP/04106/2020.

CIDMA CENTRO DE I&D EM MATEMÁTICA E APLICAÇÕES CENTER FOR R&D IN MATHEMATICS AND APPLICATIONS

REPÚBLICA PORTUGUESA

range between 0 and 1. A linear discriminant function is constructed using the linear combination definition that supports the DSD model (Dias, Brito and Amaral, 2021). The discriminant function allows defining a distribution score for each unit. Classification in two a priori groups is then based on the Mallows distance between the unit's score and the score obtained for the barycentric histogram of each a priori class. The observation is then assigned to the closest class. When considering more than two a priori classes (Santos, Dias, Brito and Amaral, 2023), one possible approach consists in dividing the multiclass classification problem into several binary classification subproblems. In this case, two well-known multiclass classification techniques may be applied: One-Versus-One (OVO) and One-Versus-All (OVA). The alternative approach consists in defining several linear discriminant functions, under the condition that each new discriminant function must be uncorrelated with all previous ones. Classification is then based on a suitable combination of the corresponding obtained scores, using the Mallows distance.

Based in the linear model proposed in Dias and Brito (2015), other works are being developed such as, for example, a Clusterwise Regression algorithm for interval-valued variables and an Outlier Detection method for distributional data.

References

Brito, P. (2014). Symbolic Data Analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4), 281–295.

Dias, S. and Brito, P. (2015). Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(2), 75-113.

Dias, S., Brito, P. and Amaral, P. (2021). Discriminant analysis of distributional data via fractional programming. *European Journal of Operational Research*, 294(1), 206-218.

Brito, P. and Dias S. (Eds) (2022). *Analysis of Distributional Data*. CRC Press LLC.

Santos, A., Dias S., Brito, P. and Amaral, P. (2023). Multiclass Classification of Distributional Data. In CLADAG 2023 - *Conference of the Classification and Data Analysis Group (ClaDAG) of the Italian Statistical Society*. https://https://www.statlab-unisa. it/cladag2023/.

Este seminário terá também transmissão via Zoom, através do link:
https://videoconf-colibri.zoom.us/j/98411941038?pwd=NUNqTWQvQ1JxT2Qwenp5MCtGVFFtQT09
ID da reunião: 984 1194 1038
Senha da reunião: 883345

This seminar is supported by CIDMA – Center for Research and Development in Mathematics and Applications through FCT – Fundação para a Ciência e a Tecnologia, within projects UIDB/04106/2020 and UIDP/04106/2020.

CIDMA CENTRO DE I&D EM MATEMÁTICA E APLICAÇÕES CENTER FOR R&D IN MATHEMATICS AND APPLICATIONS

REPÚBLICA PORTUGUESA