DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Project – Multivariate Analysis

MECD, MMAC, OTHERS, 1^{st} SEMESTER, 2022/2023

Handed out on October 10, 2022.

To be handed back by January 7, 2023.

Presentations and discussion on 10/11 January 2023.

The groups must work independently.

1. Make a preliminary analysis of the data and discuss what you have learned from this analysis.
2. The purpose of the project is to apply the multivariate analysis techniques covered in our course to an actual data collection. You are allowed to conduct any analysis you deem pertinent and instructive. Any software can be used to perform calculations, however R is recommended. You should add key tables or attach graphics to your report, but avoid submitting pages of raw computer data. Remember that there is no single best response. You will be scored on how well your analyses highlight intriguing elements of the data set, how well you explain your findings, how well you defend your analytical methods, and the general calibre of your writing.
3. Include in your discussion all options that you have made, advantages and disadvantages of each alternative. Discuss limitations of the analysis you have done and provide suggestions for future work.

- **About the dataset:**
  - The data sets were from a study that Rosenwald and colleagues carried out in 2002. Retrospectively, 240 patients were chosen based on the availability of tumor-biopsy samples. All of the patients had had chemotherapy using anthracyclines. The 240 diffuse large-B-cell lymphomas' biopsy samples were used to extract the microarray gene expression. On the project link you have the "microarray" data to get the gene expression data collection. Another data set called "patients" has the gene expression data as well as various clinical characteristics, gene expression signatures, follow-up time, and status during the follow-up. At `https://web.tecnico.ulisboa.pt/~ist13493/AM2023/Project`, both data sets are available.
  - You can link these two data sets by comparing the `DLBCL sample (LYM number)` in the "patients" with the `LYM number` in column names of the "microarray" data.
  - The reference: *Rosenwald, A., Wright, G., Wing, C., Connors, J., Campo, E. et al. (2002). The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma. The New England Journal of Medicine, 346(25), 1937-1947* offers further information regarding these two data sets.
- **About the presentation:**
  - Duration of 15 minutes plus 10 minutes for discussion.
  - Slides must be hand back with the report, on the due time.
- **About the report:**
  - The report should not exceed 10 pages, in the form of a scientific paper.
- Do not forget topics such as:
  1. Description of the problem under study;
  2. Objectives;
  3. Estimation and validation methods;
  4. Discussion of the results and interpretation of the findings;
  5. Conclusions;
  6. References.
- The R code (or other code), the data cleaned or transformed (where applicable), the slides presentation as well as the report must be send by email to:
  `conceicao.amado@tecnico.ulisboa.pt`