High-dimensional statistics

Academic Year 2022–2023

Project n°2 : (generalized) linear models and classification

Two QA sessions are scheduled: on 23/11 and on 7/12

1 Preliminary comment

This project may be done individually or in pairs (in the latter case, a unique project needs to be handed in, mentioning the two names). It is not compulsory to keep the same team as for project 1 and/or to keep working alone if that was the case for that project. When working in team, it is expected that all parts of the project have been developed in collaboration between the members of the team.

The project, written in English, is due on Wednesday 14 December 2022 (23h59) and needs to be submitted via eCampus. In the main body of the report (8 pages max), only the results, graphics and **interpretations** must be supplied and discussed (additional graphics or tables may be included in an annex). The `R` script used to compute the outputs of the analyses has to be submitted too as a complementary information.

2 Data

For this project, by default, the same data set as the one used for project 1 may be used (leaving out again the rows with missing values). However, if the data do not seem to fit with the objectives of this project (classification performed on the binary indicator), a new data set may be proposed, with the same constraints as those outlined in the statement of project 1 as far as the number and types of variables and individuals are concerned, but without the requirement of missing data. In the latter case, a text file with the new data has to be submitted together with a short description of the data.

3 Preliminaries for the supervised classification

The data are assumed to contain one binary indicator. In this project, classification rules based on a logistic GLM-regression model and on the LDA scores will be derived in order to allocate any new observation into one of the two groups defined by that binary indicator.

Before considering the two classification techniques, discuss, a priori (using the context of the collection of the data), the adequacy of the classification[1]. By means of some graphics or statistical summaries, determine whether some information about the classification might be available in the other variables (called explanatory variables from now on).

[1] In case there is no sense in trying to find a rule in order to classify new observations in one of the two possible categories of the binary indicator, another variable of the data set may be exploited, after dichotomizing its values in order to define a new binary indicator. If none of the available variables seem to fit with the objectives of a classification technique, find another data set.

4 Preparation of the classification rules

Only the quantitative variables of the data will be used for the definition of the classification rules. Moreover, a training data set will be used in order to derive the rules, while a test data set will be exploited in order to test the rules. To get these two sets, divide the data into these two sets by randomly splitting the data into two parts according to the following percentages: 80% of the data will be included in the training set and 20% of the data will consist of the test set.

The three analyses below have to be performed on the training data set.

1. Logistic regression

   Find a good logistic model explaining the probability of getting a success for the binary indicator. An objective strategy needs to be used in order to select the explanatory variables to include in the final model (paying attention to potential multicolinearity problems). Interpret the estimated model and look at the residuals and the fitted probabilities. Comment and interpret.

2. Linear Discriminant Analysis

   Derive the unique canonical variable, display the 1D-scores and determine the corresponding discriminant power. Comment. Try to simplify the expression of the canonical variable by suppressing some variables that did not look "discriminant" when doing the exploratory analysis of Section 3. Summarize the trials that seemed worthwhile to perform and discuss the potential loss of discriminant power when suppressing these variables. Select a final model.

3. ROC curves

   Still using the training data, represent, on the same plot, the ROC curves based on the "scores" derived by the two final models (one based on logistic regression, one based on LDA). Find an appropriate threshold for each classification technique.

5 Classification of the test data set

Classify the data of the test set and summarize the performance of the classifications by means of a confusion matrix based on these data. Compare the "errors" of classification made by the two methods and try to provide some explanation.