High-dimensional data analysis

Academic Year 2019–2020

Project n°3 : Supervised classification

1 Preliminary comment

This project may be done individually or in groups of 2/3 students (in the latter case, a unique project needs to be handed in, mentioning all the names). It is not compulsory to keep the same team as for projects 1 or 2 and/or to keep working alone if that was the case for these projects. When working in team, it is again 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 11 December 2019 and a **paper version** must be handed in (max 8 pages). In the main body of the report, 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 sent via email (to G.Haesbroeck@uliege.be) as a complementary information.

2 Data

For this project, by default, the same data set as the one used for the two other projects may be used. However, if the data do not seem to fit with the objectives of this project (supervised 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 (it is no longer compulsory to have missing values). In the latter case, a text file with the new data has to be provided by email and additional information needs to be supplied when some of the results derived for the two first projects seem to be relevant for the current project.

3 Preliminaries for the supervised classification

The data are assumed to contain at least one binary indicator. In this project, classification rules based either on a logistic regression model or on the LDA scores will be derived in order to classify any data point into one of the two possible categories of that binary indicator.

Before considering the two possible 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 (probably already discussed in Project 1), 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 Classification using the logistic regression model

1. Using all the data and all the variables (if not, justify why some variables are left out by default), 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. Interpret the estimated model, define precisely the classification rule that may be derived from it, look at the residuals and at the fitted values. Comment and interpret.

2. Resort to a leave-one out cross-validation technique[2] in order to characterize the classification performance of the procedure. At each step of the CV, a similar model as the "optimal" model derived in question 1 needs to be fitted in order to derive the classification rule. Construct a confusion matrix and measure its corresponding error rate. Comment.

5 Classification based on LDA scores

Here, the explanatory variables are the quantitative variables of the data.

1. As there are only two groups, a single canonical variable is available. Give the expression of that canonical variable, display the 1D-scores and determine the corresponding discriminant power. Comment.

2. 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.

3. Using the model you consider the most appropriate (either the first one or the simplified one), derive a classification rule using the LDA scores and test it with the same data. Comment.

4. Discuss whether the homoscedasticity assumption seems to be valid for the data.

[2] A brief description of the K-fold cross validation technique (the particular case of the n-fold CV corresponds to the "leave-one-out" version) can be found in Section 7.10.1 of the reference *The elements of Statistical learning* of Hastie, Tibshirani and Friedman.