High-dimensional data analysis

Academic Year 2018–2019

Project n°2 : Supervised classification using logistic regression and multivariate ranks

1 Preliminary comment

This project may be done individually or together with another student of the course (in the latter case, a unique project needs to be handed in, mentioning both 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 project 1. When working in pairs, it is expected that all parts of the project have been developed in collaboration between the two members of the team. A specific question on the project will be included in the exam questionnaire in order to further develop or explain some aspects of it.

The project, written in English, is due on Wednesday 9 January 2019 and a **paper version** must be handed in (max 10 pages). In the main body of the report, only the results, graphics and **interpretations** must be supplied and discussed. It is not compulsory to use the software `R`. However, if available, the `R` script used to compute the outputs of the analyses may be sent via email (to G.Haesbroeck@uliege.be) as a complementary information (e.g. valuable details not reported in the text might be obtained by looking at the script).

2 Data

For this project, by default, the same data set as the one used for the first project may be used[1]. However, if the data do not seem to fit with the objectives of the two parts of this project, a new data set may be proposed, with the same constraints as those outlined in the statement of project 1. In the latter case, a text file with the new data has to be provided by email.

3 Supervised classification with logistic regression

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

1. Discuss, a priori (using the context of the collection of the data), the adequacy of the classification[2]. By means of some graphics or statistical summaries, determine whether

[1] In case of a change in the composition of one group, both members of that group may keep working on the same data set.

[2] 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.

some information about the classification might be available in the other variables (called explanatory variables from now on) of the data.

2. Using all the data, 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 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 and comment.

3. Constructing, at each step of the procedure, models similar to the "optimal" model derived in the preceding question, use a cross-validation technique in order to measure the classification performance of the logistic model. Comment.

4 Multivariate ranks

1. Using only the quantitative variables of the data set, compute the ranks of the observations using the Mahalanobis depth function, either classically estimated by means of the sample mean vector and covariance matrix or robustly estimated with the MCD estimator (with a justification for the choice of one of the two strategies). Represent the data on the first principal plane, together with varying symbols/colors allowing to interpret the different depths (or ranks) of the observations. Discuss the specificities (if there are some) of the extreme cases on one hand and of the deepest cases on the other hand.

2. Focusing on the two most interesting quantitative variables (chosen according to a criterium to define), compute, for each of these variables, a vector of 1D-ranks for the observations when defining the ranks as a center-outward procedure starting from the median.

3. Still using only these two variables, represent the bag plot of the data in 2D and comment. Measure, in some way, the discrepancy between this 2D-ranking and the multivariate ranking derived in question 1, as well as with the 1D-rankings computed in question 2. Comment.