LSMF2013 project
« Quantitative Data Analysis »

Professor:	Marco Saerens marco.saerens@uclouvain.be
Address :	Université catholique de Louvain
	LSM/ICTEAM
	Place des Doyens 1
	B-1348 Louvain-la-Neuve
	Belgique
Phone :	010 47.92.46.
Fax :	010 47.83.24.
Assistants :	Bertrand Lebichot betrand.lebichot@uclouvain.be
	Benoit Sluysmans benoit.sluysmans@student.uclouvain.be
	Sylvain Courtain sylvain.courtain@student.uclouvain.be

Objectives

The objective is to put into practice quantitative data analysis techniques through a case study requiring a data analysis software.

Data

In this project, you will have to use several methods introduced in the lectures by applying these on a data set of your choice. Data sets can come from websites of well-known data providers (links available on Moodle) or from any other source you might find. Restrictions are:

- Data sets must contain at least one categorical variable, to allow the usage of classification methods. One of the categorical variables will then be the dependent variable (variable to predict)
- Avoid data sets related to time series or network data, as they require different preprocessing and data analysis methods than those introduced in your lectures
- Data should contain less than 50.000 observations but more than 300
- Data should contain less than 50 explanatory variables but more than 5

Groups should avoid to use the same data set. Feel free to use the forum on Moodle to publish the name of your dataset and communicate with other groups or students.

For training and playing with the statistical software, you could use a data set called “Spambase”, which is a good example of what we are expecting. This dataset will be used for the TP presentation.

The work and the report

Project evaluation will be based on a written report. Different steps are required in the report; for each item, the minimum content is indicated but to a certain extend, more theoretical and experimental work means a higher grade. Keep in mind that, since your datasets are all different, some of the following steps can be more or less important, depending on your problem. The different steps you will have to perform on your data set, and that have to stand in your report are:

- **Theory reminders**: Every classification method used will be introduced. The different hypothesis will be clearly stated and theoretical principles summarized, together with the most important equations. All the concepts that you use must be clearly explained and referenced.
- **Data exploration**: This is the first step in your data modeling process (particularly important). To familiarize yourself with the data and possibly detect errors, every variable (also called feature)

will be visualized via one or several graphics. Some correlations between the dependent variable and explanatory variables will be computed, as well as correlations between explanatory variables. If you need a reminder about data exploration, some references will be available on Moodle.

- Univariate plots : histograms for continuous features or bar plots for discrete features.
- Multivariate plots : mainly histograms/bar plots of explanatory variables for different values of the dependent variable.
- Univariate statistics : mean, max, min, std, skewness and kurtosis.
- Multivariate statistics : impact of each explanatory variable on the categorical dependent variable. Typically, a t-test for continuous features or chi-square test for discrete features. The correlation (redundancy) between explanatory variables can also be computed.

❖ **Variable selection, transformation and recoding**: Sometimes, it is necessary to transform variables to extract indicators or to satisfy a model hypothesis. Also, variables providing redundant information and outliers could be deleted. Finally, variables can be recoded or replaced to create new variables that are combinations of others but keep the same quantity of information. In this step, you are asked to test *at least one feature extraction/selection technique* (PCA, correspondence analysis, discriminant analysis, stepwise logistic regression, …) and analyze its impact on the classification rate.

- Imputation is required if you have missing values.
- PCA must be used for continuous features (try to compare the model performances with and without PCA).
- Multiple correspondence analysis should be performed on categorical features.

❖ **Modelling**: You must select *at least* 3 classification methods (choose, e.g., the ones you have seen during the lectures or in the reference books mentioned in the slides). The performances of each model will have to be compared using external validation data (cross-validation or bootstrap) and the best performer will be chosen according to this criterion. The creativity in the selection of the classification methods will be rewarded in term of "bonus". Don't hesitate to compare more than 3 classification models and to use more recent/advanced methods, not shown during the lectures.

- You must compare the models with *at least* one performance measure (the miss-classification rate on test set in most of the cases).

❖ **Scenario with cost***: Once the best model has been selected, create a two-classes scenario in which the relative cost of each classification error is different. In this context, you must define yourself a cost matrix by setting a scenario "with costs" and then deduce *theoretically* the optimal classification rule taking into account the defined costs (the a-posteriori classification threshold above which you categorize the observation as positive). Then, the optimal classification rule with costs must be verified empirically on the data.

- You must verify that the total cost obtained on the entire population is lower than if you rely on a standard "Bayesian" classifier, without integrating the costs.
- You must modify the decision threshold and observe empirically that the total cost is minimal around the optimal theoretical threshold.

❖ **Study of the "reject option"***: You are asked to study the possibility to rely on a "reject option". You will have to provide a chart showing the evolution of either the total cost or the classification rate depending on the "reject option" threshold.

*For the scenario with cost and the reject option, you must work on a binary classification problem. Just deduce a special binary classification problem from the original data. In other words, you can simply turn the problem into a classification problem with two classes.

Software tools

For building the classification models and evaluating the results by cross-validation or bootstrap, if you are an INGE student with no programming background, you can use the Tanagra open source

software. For INGE students with programming background, please use R. For computer science students, I recommend R, Matlab/Octave or Python. You can find documentation about these programs on Moodle.

Notice that you do not have to restrict yourself to a single software – *the use of several one is highly recommended*. For instance, you could use Excel, SAS JMP or SPSS for data transformation and exploration and then use a more specialized software for classification like Tanagra, R, Matlab or Python. That's what every data scientist does.

Several books introducing classification techniques with these softwares will be mentioned in the lectures and the references are available on Moodle.

Evaluation

This project must be done by groups of exactly 3 students. It is highly recommended that you share the work between members of the group and exchange information with other groups (by, e.g., using the forum on Moodle). The final printed report must be given to your assistant before the 19th of May at 18:00 pm.

For the S6 (13th of March), you will meet an assistant and show the results of data exploration and at least one classification method (on the training set only). You will have to present shortly and informally these two parts to your assistant during 5-10 minutes. This intermediate report will not be graded but it allows us to have an idea of your advancement of your work and to verify if your data set is OK.

This final report (maximum 10 pages in total) counts for 40% of your final grade. This report should be provided without annex. Thus, the report will include only the most relevant graphics and tables of your work. The group score will be shared between all the members of the group. The oral exam counts for the remaining 60% of the grade.

It is recommended to write the report in the form of a high-quality, professional, scientific paper (use Latex, Lyx or Word, but the result should really look as a professional research paper). Notice that the grade scoring grid is available on Moodle for information. Don't hesitate to check on the grid if you answer all the required criteria.