Practical work on Multivariate Analysis

MIRI - Data Science

Any data matrix contains information about the generating phenomenon. The practice will consist in choosing a real problem and applying multivariate techniques to reveal the hidden information contained in the data set, as a prior step to modeling. The problem can be chosen from the machine learning repository http://archive.ics.uci.edu/ml/, or from any other repository. Moreover, the students can provide their own dataset, previous approval of the professor.

The student must perform a multivariate approach of the data matrix (visualization, clustering and interpretation) plus a prediction model suitable for the undertaken problem (with preprocessing of the raw data of course). The student must write a complete report upon the solution envisaged.

Steps for conducting the practice

1. The student will choose a problem and read the corresponding documentation trying to understand what is the objective of the problem and the available data.
2. Pre-process of data. The student will perform a first summary of the data, and, eventually, detect errors, outliers and missing values and take the appropriate measures of correction. According to the problem and data, it may be necessary to perform a selection of variables (feature selection) and /or a derivation of new explanatory variables (feature extraction) if the problem needs it.
3. The student will choose the type of protocol for the validation (i.e. holdout or test sample to assess the quality of the final model). (Depending on the data size, it won't make sense to have a separate

test data file). Test data should approximate as possible future unseen cases of the phenomenon, and not a random sample of the available data, in order that the predicted test accuracy would be a good estimation of what we can expect in the future.

4. The student will perform a multivariate exploratory analysis of the training data set, taking the test data (if it exists) as supplementary. The Multivariate exploration will consist on the visualisation of the information, detection of the hidden latent factors. The synthesis of the complexity by clustering and interpretation of the results.
5. Then, according the undertaken problem, the student will choose a model for prediction of the response variable within the ones explained in the MVA course, will find its optimal parameters and will evaluate its performance (generalization error) according the established validation protocol.

The report should include:

1. A description of the problem and available data
2. The pre-process of data
3. The protocol of validation
4. The visualisation performed
5. The interpretation of the latent concepts.
6. The clustering performed
7. The interpretation of the found clusters.
8. Discussion about the differences of the test sample respect to the training one.
9. The prediction model with its best parameterization and its generalization error
10. Scientific and personal conclusions