Computación y Sistemas Inteligentes (MEI - UPC) Practical work 2016-2017

Lluís A. Belanche
belanche@cs.upc.edu

Abstract

This is the **guide** for the correct development of the Machine Learning (ML) practical work of the CSI course. The students must apply the different concepts and models lectured during the course to solve a real problem, and write a complete report describing the work carried out, the problems encountered and the solutions envisaged, as well as the final results and conclusions of the study.

1 General information

Students enrolled in CSI are required to complete a small practical project for the ML part. The goal is to develop a **classification** or **regression** model to solve one of the two supplied problems, that can be obtained from the supplied web addresses.

You are expected to write a complete **report** describing the work carried out, its motivation, the problems encountered and the solutions envisaged, and the final results and conclusions of the study. The main text is *strictly* limited to 12 pages (this includes graphics and tables; you can use an extra page for references); the *code* must be submitted *separately*).

The main programming language used for the modelling part must be `R`.[1] Remember that there are *many* packages for `R` which probably contain useful routines you can use; just be sure to mention them in your final document.[2] Other software can be used as long as it serves a specific or secondary purpose. Notice also that `R` can be interfaced with many other languages.[3] Any additional information on the methods or on the problems should be **acknowledged** and/or properly **cited**.

2 Evaluation

The grade will be partly based on the **clarity** of your report, so please make sure your final report is well organized and clearly written. There should be an introductory part explaining the basics of your work, and a conclusions section, basically stating what you know compared to what you knew before the work started; also any gaps, possible extensions or limitations in your development should be noted briefly discussed.

Your work will also be evaluated based on **technical quality**. This means that the techniques you use should be reasonable, the stated results should be accurate, and the technical results should be correct and complete.

In summary, these are the conditions for a high score (in this order):

[1] `http://cran.r-project.org/`
[2] `https://cran.r-project.org/web/packages/available_packages_by_name.html`
[3] `https://cran.r-project.org/manuals.html#R-admin`

1. The (good) use of techniques and methods presented in class
2. The care and rigor for obtaining the results (validation protocol, statistical significance)
3. The quality of the obtained results (generalization error, simplicity)
4. The quality of the written report (conciseness, completeness, clarity)

3 Detailed information

The **written report** that you must deliver should include the following **sections**:

1. A brief but self-contained description of the work and its goals, and of the available data, and any additional information that you have gathered and used
2. A brief description of related previous work and results
3. The data exploration process (pre-processing, feature extraction/selection, and visualization, if appropriate)
4. The resampling protocol (training/test, cross-validation, etc) that you have used
5. The results obtained using **at least three** methods (indicating the best set of parameters for each one):
   (a) If the task is **classification**, any of: logistic regression, LDA, QDA, RDA[4], Naive Bayes, nearest-neighbours, decision trees, Random Forests
   (b) If the task is **regression**, any of: linear regression, ridge regression, LASSO regression, nearest-neighbours, decision trees, Random Forests
6. A description and justification of the final model chosen and a honest estimation of its generalization error
7. A final part (one to two pages) containing:
   (a) A self-assessment of successes, failures and doubts (I suggest this to be a list of one-line items)
   (b) Scientific and personal conclusions
8. References to all your used sources: books, web pages, code, scientific papers, ...

On data pre-processing Each problem requires a different approach in what concerns data cleaning and preparation, and the selection of the particular information you are going to use can vary; this pre-process is very important because it can have a deep impact on future performance; it can easyly take you a significant part of the time. It is then strongly advised that you analyse well the data before doing anything, in order to gauge the best way to pre-process it. In particular, you shall pay attention to the following aspects (not necessarily in this order):

1. treatment of lost values (missing values)
2. treatment of anomalous values (outliers)
3. treatment of incoherent or incorrect values
4. coding of non-continuous or non-ordered variables (nominal or binary)
5. possible elimination of irrelevant or redundant variables (feature selection)
6. creation of new variables that can be useful (feature extraction)
7. normalization of the variables (*e.g.* standardization)
8. transformation of the variables (*e.g.* correction of serious skewness and/or kurtosis)

[4]Only if LDA or QDA cannot be applied.

On model selection and estimation of performance In accordance with the problem and the available data, you should design a set of experiments based on valid protocols to select models and to honestly estimate the generalization error (or any other measure of future performance) of the final proposed model or solution.

Some problems come with their own test data (data used for the estimation of true generalization error), some do not; in the latter case, you must obtain test data by splitting the full available data (once or several times, depending on the data size). For model selection, k-fold cross-validation will probably be necessary (the selection of the best value for k is your decision). It is methodologically prohibited to use as test data information that has already been used for the creation, adjustment or selection of the solution.

Mechanism for delivery: You have to deliver the **full code** (in separate files, and only electronically) and a brief text file with instructions on how to execute your code. The written report (a standard pdf file) should *not* include explanations of the methods seen in class. All deliveries are exclusively by e-mail in a **single compressed file**; you do not have to print anything.

1. Delivery date: **January 16, 2017** (this date is **strict**)
2. You are expected to form teams of 2 people