Aprendizagem Automática 2016/2017
Course Project – Assessing Classifiers

28/11/2016, version 1.3

Overview

For the academic year of 2016/17, the project of curricular unit *Aprendizagem Automática* targets the prediction of credit card default. Given a set of features related with credit card usage, the objective is to predict whether a given client will default payment.

Objectives

The project's objective is to implement and assess three classifiers for the credit card default problem. The implementation is expected to use the R programming language.

The target data set is available from the UCI repository [1]: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [1]. The credit card default data set is detailed in reference [5].

You must assess the use of support vector machines (SVMs) and multilayer perceptrons (MLPs). In addition, you are asked to select another classifier. You may use one of the classifiers studied in the course, or you may use some other classifier. Examples references to consider include the course's main textbooks [3, 6] or other recommended references [4, 2].

A project to receive full marks is expected to develop a dedicated implementation of MLPs. Alternatively, a penalization of 1~2 (out of 20) marks will be given to projects that only use R packages.

Report

The report should briefly describe the classifiers used, the training and testing procedure, the organization of the code, and the overall conclusions. A possible organization is:

1. Introduction: summarize objectives of the project, and provide overview of your solution.

[1] A copy of the Excel file containing the dataset is available from: http://www.di.fc.ul.pt/~jpms/Courses/aa-201617/datasets/. A summary of the columns is included in this project description.

2. Problem statement: describe the problem being solved.
3. Classifiers used: briefly detail the classifiers used; add references if necessary.
4. Experimental results: include a brief evaluation of the three classifiers on training and test data.
5. Source code organization: briefly summarize any main design decisions of your solution.
6. Conclusions: overview the project, what was achieved and what was not, given the project's objectives.

Note: The page limit for the report is 8 pages.

Submission

The project is to be submitted by email to the course lecturer (email: `jpms@ciencias.ulisboa.pt`) by the due date. The submission should be an archive with the R source code and a report, in PDF. The report should be organized as indicated above.

Groups

The project is to be implemented in groups of 2 students. Exceptions require auhorization from the course lecturer. Groups of more than 2 students are expected to implement a more ambitious solution.

Dates

- Project published: 10/11/2016.
- Project due: 22/12/2016, 23:59, Lisbon time.

Omissions & Errors

Any detected omissions or errors will be added to future versions of this document.
Any required clarifications will be made available through the course's official website.

Versions

10/11/2016,	version 1.0:	Original version.
14/11/2016,	version 1.1:	Updated with clarifications on the format.
15/11/2016,	version 1.2:	Fixed project due date.
28/11/2016,	version 1.3:	Local distribution of dataset.

Clarifications

2016/11/11: Dataset Format

The format description on the UCI website is not consistent with the dataset. The authors provided the following clarification:

```
This research employed a binary variable, default payment (Yes=1, No=0),
as the response variable. This study reviewed the literature and used
the following 23 variables as explanatory variables:
X1: Amount of the given credit (NT dollar): it includes both the
individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1=male; 2=female).
X3: Education (1=graduate school; 2=university; 3=high school; 0,4,5,6=others).
X4: Marital status (1=married; 2=single; 3=divorce; 0=others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records
(from April to September, 2005) as follows:
X6= the repayment status in September, 2005;
X7= the repayment status in August, 2005;
...;
X11= the repayment status in April, 2005.
The measurement scale for the repayment status is:
-2: No consumption;
-1: Paid in full;
0: The use of revolving credit;
1= payment delay for one month;
2= payment delay for two months;
...;
8= payment delay for eight months;
9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar).
X12= amount of bill statement in September, 2005;
X13= amount of bill statement in August, 2005;
...;
X17= amount of bill statement in April, 2005.
```

```
X18-X23: Amount of previous payment (NT dollar).
X18= amount paid in September, 2005;
X19= amount paid in August, 2005;
...;
X23= amount paid in April, 2005.
Y: client's behavior; Y=0 then not default, Y=1 then default
```

2016/11/28: ALternative Link to Dataset

The dataset to be used in the project has been made locally available at FCUL. The link is: http://www.di.fc.ul.pt/~jpms/Courses/aa-201617/datasets/.

References

[1] UCI Machine Learning repository. https://archive.ics.uci.edu/ml/datasets.html.

[2] E. Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[3] P. Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[4] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach, 3ed*. Prentice Hall, 2010.

[5] I. Yeh and C. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.*, 36(2):2473–2480, 2009.

[6] M. J. Zaki and W. M. Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.