Applied computational intelligence

Homework III: Models for classification

This exercise set relates to classification methods (logistic classification, linear and quadratic discriminant analysis, neural networks, k-nearest neighbors and support vector machines) applied to a set of data containing a number of observations for predictors and outcome.

Choose either Alternative 1 or Alternative 2, below. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Alternative 1

You are given a set of data[1] consisting of $N = 8708$ observations related to grant applications made between the years 2005 and 2008. For each observation of grant application, there are $D = 1882$ predictor variables. Given the extremely correlated predictors, the D variables have been filtered and only the `reducedSet` of predictors ($D_{\texttt{reducedSet}} = 252$) must be used for classification[2]. The class outcome, `successful` and `unsuccessful`, is contained in the column `Class`.

The observations for the predictors and the class are split between training and test sets and given in the `training` (N_{tr}=8190) and `testing` (N_{ts}=518) sets of data.

You must

1 Use the predictors in the training set to learn a linear classification model and test the model using the test set. For the task, you must select either a logistic classification or a linear discriminant analysis method. How many predictors would you use to fit the model? After training the model, compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of misclassifications made by the method.

2 Use the predictors in the training set to learn a nonlinear classification model and test the model using the test set. For the task, you must select at least one method among the following: quadratic discriminant analysis, neural networks, k-nearest neighbors and support vector machines. Depending on your selection of method, tune the model in a convenient way (if needed use cross-validation). After training the model, compute the confusion matrix and overall fraction of correct predictions.

[1] The data can be i) found enclosed to the homework assignment, or ii) downloaded from the Kaggle web site (`http://www.kaggle.com/c/unimelb`) and reproduced in R using the script (`'CreateGrantData.R'`) given in the package: `AppliedPredictiveModeling`.

[2] For a complete explanation of the predictor groups refer to the book M. Kuhn and K. Johnson, Applied predictive modeling, Springer (2014).

Explain what the confusion matrix is telling you about the types of misclassifications made by the method.

3 Compare the results obtained using the linear and nonlinear classification methods. Does the nonlinear structure improve the classification performance?

Alternative 2

Assuming you have at your disposal a set of data of your own interest and this dataset consists of a certain number of observations, each observation consists of a certain number of predictors and an outcome that you wish to classify, you might prefer to investigate the characteristics of your own data.

In this case, you must first describe your data and their features in terms of number of observations, number of predictor variables and outcome (if needed, you must pre-process the data). Then, you must split the set of data into training and test set and perform the steps defined in Alternative 1.

Guidelines

Regardless of your choice (Alternative 1 or 2), you must generate the following:

- Article: You must generate a report in the format of a conference paper following the template adapted from the ESANN (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning) conferences[3] that is attached to the homework assignment. The paper must be returned as PDF document and should not be longer than 6 pages and must include the following:
  - Title: Here, you summarize your paper in one sentence. [Spend time on it and try some alternatives[4]. As part of the preparation, this will help both you to write a clear abstract and the reader to grasp the content of the work.]
  - Abstract: Here, you introduce the main objective and overview of the work [Provide a short and informative view of the work, its scope and results].
  - Introduction: Here, you provide some context and background [Briefly, explore the literature in order to define classification and the models that can be used for it. Discuss some examples of application and provide the references.]
  - Methods: Here, you briefly describe your data set and the methods you use for classification. Provide a brief description of the methods, their pros and cons. [Also report and comment the main characteristics of the data. Each figure or table must be discussed in the text. Describe the features and the theoretical background of the methods you use for the classification.]
  - Results: Here, you compare the models. Are there any difference between the models? Is there statistical difference between them? [Report and comment the main results of the analysis.]
  - References: Here, you provide bibliographic references [Report the books and/or articles that you used for studying the methods and perform the analysis. Each reference reported in this section must be cited in the main text].
- Code listing: The code you used to perform the analysis. Regardless of your choice programming, your code must be executable/functioning. The code (and the relevant functions, if needed) can be either pasted at the end of the 6-page article (for instance as an appendix) or packaged together with the paper as a zip file.

The work can be done individually or in group of maximum 5 co-authors. You can chose to write your paper either in English or Portuguese[5] You can base your work on the book by M. Kuhn and K. Johnson, Applied predictive modeling, Springer (2014).

The work must be submitted by **November 26, 2018**. Note that **delays will be penalized** (<24h: 20% penalty; <48h: 40% penalty; etc.).

[3] The original template has been slightly modified to wider the page margins and include page numbers. If you wish extra space, you could use a use the two column format (to insert the two-column option in LaTeX: \documentclass[twocolumn]{esannV2})

[4] Avoid the obvious title "Homework 3: classification models".

[5] In LaTeX, specify \usepackage[portuguese]{babel} in the preamble to change the language.