University of Essex
School of Computer Science and Electronic Engineering
CE802 Machine Learning and Data Mining

Assignment: Design and Application of a Machine Learning System for a Practical Problem

Set by:	Dr Luca Citi (lciti@essex.ac.uk)
Distributed to students:	week 9
Submission deadline:	week 16 (see FASer for exact official date and time)
Feedback:	three weeks from submission deadline
Submission mode:	electronic only via FASer

Assignment objectives

This document specifies the coursework assignment to be submitted by students taking CE802. Aims of this assignment are: a) to learn to identify machine learning techniques appropriate for a particular practical problem; and b) to undertake a comparative evaluation of several machine learning procedures when applied to the specific problem.

Assignment description

1. Pilot-Study Proposal

Imagine that you work as Data Mining and Machine Learning independent consultant, providing scientific advisory and consulting services to companies seeking to apply data analytics to their business activities.

The manager of a large restaurant chain contacts you to investigate the feasibility of using machine learning to predict whether a new restaurant opened in a given location will be profitable or not. The manager has access to historical data of successful and unsuccessful restaurants opened under the chain's brand and she offers to provide you with geographical and socio-economical data about the locations and neighbourhoods.

In the first part of your assignment, you are asked to write a detailed proposal for a pilot study to investigate whether machine learning procedures could be used to successfully solve this problem. Your report should discuss several aspects of the problem, including the following main points:

- the type of predictive task that must be performed (e.g., classification, regression, clustering, rules mining, ...);
- examples of possibly informative features that you would like to be provided with;
- the learning procedure or procedures (e.g., DTs, k-NN, k-means, linear regression, Apriori, SVMs, ...) you would choose and the reason for your choice;
- how you would evaluate the performance of your system before deploying it.

You can assume that the manager has some knowledge of machine learning and you do not need to explain how the recommended learning method works. Simply discuss your recommendation and back it with sound arguments.

This document should consist of approximately 500–750 words of narrative (i.e. excluding references, pictures, and diagrams). Please report your word count on the title page.

2. Comparative Study

Thanks to the convincing arguments in your pilot-study proposal, the company decides to collect the data that you suggested and to hire you to perform the proposed study. They provide you with a training set of historical data made of 1500 examples with 14 features and one label representing whether the restaurant was profitable or not. These data are organized in the file `CE802_Ass_2018_Data.csv` inside the `CE802_Ass_2018.zip` archive available from the CE802 moodle page. In this part of the assignment, you are asked to investigate the performance of a number of machine learning procedures on this dataset. In particular, you are required to perform a comparative study of the following machine learning procedures:

- a pruned Decision Tree classifier;
- at least two more ML technique to predict the success of a given restaurant.

The company uses Python internally and therefore Python with scikit-learn is the required language and machine learning library for the problem.

After conducting this study, you are asked to write a report containing an account of your investigation. There should be a brief summary of the experiments performed followed by one or more tables summarizing the performance of the different solutions. In particular you should at least report the accuracy and/or the Kappa statistics for each approach attempted. Any numerical data that you include should be in a suitable graphical or tabular form. You should not include any numerical data that is not relevant to your discussion of the relative performances (do not trivially copy/paste raw output produced by the library). Your report should have an appendix section with the code that you implemented (for example in the form of a Jupyter notebook).

The rest of the report should concentrate on your interpretation of the results that the software produced for you and what they tell you about the relative strengths and weaknesses of the alternative methods when applied to the given data.

This second document should consist of approximately 750–1500 words of narrative (i.e. excluding references, code, pictures, and diagrams). Please report your word count on the title page.

3. Prediction on Test Set

An additional dataset, `CE802_Ass_2018_Test.csv`, is provided inside the `CE802_Ass_2018.zip` archive. Binary outcomes are withheld for this test set (i.e. the “Class” column is empty). In this last part of the assignment you are required to produce class predictions of the records in the test set using one approach of your choice among those tested in the comparative study (for example the one achieving the best performance). These data must not be used other than to test the algorithm trained on the training data.

As part of your submission you should submit a new version of the file `CE802_Ass_2018_Test.csv` in CSV format with the missing class replaced with the output predictions obtained using the approach chosen. This last part of the assignment will be marked based on the prediction accuracy on the test.

Suggested material

`Scikit-learn` online documentation and tutorials: `http://scikit-learn.org`.
Lecture notes on machine learning and Lab notes on `scikit-learn`, `pandas` (to read/write CSV files), and `tensorflow`: see CE802 moodle page.

Submission

Your work must be submitted to the university's online coursework submission system at the address `http://faser.essex.ac.uk/` by the deadline given on the system. No other mode of submission is acceptable. You are strongly advised to submit a draft submission well before the deadline, and then update it up to the deadline.
You are required to submit a **ZIP archive** (not a RAR or 7z) containing the following files:

1. a single document with two sections (the pilot study report and the comparative study report) exported to **PDF format** (no doc or docx, etc.);
2. the file `CE802_Ass_2018_Test.csv` with your predictions replacing all missing values.

DO NOT WAIT UNTIL CLOSE TO THE DEADLINE TO MAKE YOUR FIRST SUBMISSION. Difficulties with the submission system will not be accepted as an excuse for a missing submission.

Marking criteria

This assignment is worth 20% of the module mark and will be assessed based on:

- Pilot Study Report
  - Correctness of identified type of predictive task 4%
  - Validity of examples of possibly informative features 6%
  - Appropriateness of learning procedure(s) suggested 7%
  - Correctness of evaluation methods suggested 7%
  - Overall clarity of presentation 8%
- Comparative Study Report
  - Correctness and completeness of investigation 18%
  - Quality of presentation and discussion of results 10%
  - Justification of conclusions drawn 8%
  - Quality of the code submitted (appendix) 10%
- Prediction on Test Set
  - Accuracy of predictions 20%
- Others
  - Compliance with submission instructions (e.g., archive and file formats) 2%

Problems

If any problems arise in the assignment, please start a thread on the moodle forum.

Late Submission and Plagiarism

Please refer to the Postgraduate Students' Handbook for details of the Departmental policy regarding submission and University regulations regarding plagiarism.

Revision 1.0
28/11/2018
Luca Citi