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

Combined coursework reassessment for CE802

Set by: Dr Luca Citi (lciti@essex.ac.uk)
Dr Vito De Feo (vito.defeo@essex.ac.uk)

Submission mode: as instructed by the School Office

Assignment objectives

This document specifies the reassessment coursework assignment for CE802. Main aim of this assignment is to learn to identify machine learning techniques appropriate for a particular practical problem and then to undertake a comparative evaluation of several machine learning procedures when applied to the specific problem. It also assesses the general knowledge of the topics covered during the module.

Assignment description

1. General Questions on Machine Learning

Question 1

Define "Machine Learning" and in particular describe:

- some examples of problems for which it is used successfully;
- some examples of problems for which it is unfit;
- the importance of making assumptions about the target function (inductive bias).

Question 2

Explain the difference between supervised and unsupervised learning. Then briefly describe two examples of supervised learning algorithms and two examples of unsupervised learning algorithms.

Question 3

Explain "overfitting" in the context of machine learning and in particular discuss the factors contributing to it. Also explain the relationship between the bias-variance tradeoff and overfitting.

2. 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 an online computer store contacts you to investigate the feasibility of using machine learning to make their email advertisement campaigns more effective. Throughout the year, the company runs a series of email campaigns to advertise their new products and uses unique promotion codes to monitor whether each email sent led to a purchase (this is called "conversion"). They are currently using the so called batch-and-blast approach (which consists in sending all emails to their entire database), but are afraid that this may lead to high unsubscribe rates and lower the overall effectiveness of the campaigns. They want your help to pursue targeted email marketing: for each campaign, they only want to send emails to those contacts that most likely will buy the advertised product. The manager has access to historical data of each previously sent email and whether it led to a sale.

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 successfully to 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 success of your system.

You can assume that the manager has some knowledge of machine learning and you only need to briefly explain how the recommended learning method works. Also 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.

3. 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 1000 examples with 8 features for each email sent in the past and one label representing whether the receiver bought a product using the corresponding unique promotion code. These data are organized in the file `CE802_Ass_Resit.csv` available here[1]. In this part of the assignment, you are asked to investigate the performance of a number of machine learning procedures on this dataset using Python/scikit-learn.

You need to perform a comparative study of a number of machine learning procedures:

- a Naïve Bayes classifier;
- decision trees;
- at least two more ML technique to predict whether an email led to a sale.

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 the Kappa statistics for each prediction algorithm. 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 scikit-learn's output). Your report should have an appendix section with the code that you implemented (in the form of a Jupyter notebook).

The rest of the report study 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.

[1] `https://essexuniversity.box.com/s/ec1u89bs35xhtqaat0yvdl3h132gqqs1`

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.

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), etc.: see CE802 moodle page.

Marking criteria

This assignment is worth 50% of the module mark and will be assessed based on:

- General Questions on Machine Learning
  - Question 1 .... 15%
  - Question 2 .... 15%
  - Question 3 .... 15%
- Pilot Study Report
  - Correctness of identified type of predictive task 3%
  - Validity of examples of possibly informative features 4%
  - Appropriateness of learning procedure(s) suggested 5%
  - Correctness of evaluation methods suggested 5%
  - Overall clarity of presentation 5%
- Comparative Study Report
  - Correctness and completeness of investigation 12%
  - Quality of presentation and discussion of results 7%
  - Justification of conclusions drawn 6%
  - Quality of the code submitted (appendix) 8%

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
15/06/2023
Luca Citi & Vito De Feo