CSM6420/SEM6420 – Machine Learning for Intelligent Systems

Second Assignment, April 2016

This is the second of two assignments for CSM6420/SEM6420, and comprises 60% *of the total marks for the module. It will be assessed according to the Department's assessment criteria for essays (see online Appendix AC of Student Handbook). In particular, marks will take account of understanding of the problem, challenge of the chosen task, completion of the task, accuracy of the prediction on the test set and quality of the presentation. Other marks will cover knowledge of the literature, justification of the approaches taken, quality of analysis and amount of work involved.*

Please submit your work through Blackboard before 5pm, Monday, 9th May 2016

---

1 Dataset

This assignment is based on the Thoracic Surgery Dataset[1]. Some modifications have been applied to the original dataset for this assignment.

The training dataset (train_risk.arff or train_risk.csv) is composed of various preoperative clinical information for 300 patients who underwent major lung resections for primary lung cancer, and the target attribute "Risk1Yr" indicating whether or not the patients survived a year after the surgery (0 for survival and 1 for death). The main aim of the data task is to predict the risk of postoperative death within a year given the patient clinical information prior to thoracic surgery.

The test dataset (test_risk.arff or test_risk.csv) contains 100 patient records of the same attributes as the training dataset, except the class label of "Risk1Yr".

The training dataset can be used directly to perform experiments in WEKA[2] or in python with python libraries Scikit-learn [3] and Pandas [4].

The publication associated with this data is: Zieba, M., Tomczak, J. M., Lubicz, M., and Swiatek, J. (2013). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing, 14(A), pp. 99-108.

2 Assignment specification

For this assignment, you are to investigate the performance of at least three classifiers (of your own choosing) when applied to this data, using WEKA or python libraries Scikit-learn and Pandas. You should then write a report on this, taking into account the guidelines given below. The datasets described above are available on Blackboard for you to use.

- The .arff or .csv files that you will be using are available from the module website.
- The following criteria will be used for marking the report:
  (1) Discuss any observations you have about the data and also describe what data preprocessing and/or dimensionality reduction you have performed and why. (10%)
  (2) Describe the classifiers you have selected, explaining how they work and discussing the reasons behind why you chose them. (10%)
  (3) Discuss and investigate the options you chose for configuration, i.e. the hyper-parameter settings. The default settings may not be the optimal ones for many classifiers. (20%)
  (4) Describe the experiments performed and discuss the results. Issues you might like to consider: the impact of data preprocessing or dimensionality reduction, how the classifier performances compare, etc. (30%)

[1] http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] http://scikit-learn.org/stable/
[4] http://pandas.pydata.org/

(5) Build one classifier with the learning methods and possible hyper-parameters that you select based on your experiments and analysis, then use this classifier to predict the class labels with associated output probability/score for the given 100 test cases, and submit the prediction results in a separate csv file for automatic evaluation. The submitted csv file should consist of three columns with the header of: "test_id", "predicted_label"(0/1) and "predicted_output" (probability or score). You receive marks according to the resulting area under the ROC curve (AUC) for the test set, e.g. marks of 10 for an AUC of 1.0; and 5.5 for an AUC of 0.55). Note that the AUC will be calculated based on your predicted output, not labels. (10%)

(6) To make your work reproducible, please provide in appendix the WEKA configurations or relevant python code, whichever appropriate for building the final prediction model. This should cover important steps such as data preprocessing, cross-validation for associated hyper-parameter tuning and model fitting. Also please indicate the version of WEKA, or python and relevant libraries that you've used for this work. (10%)

(7) The quality of the report will be marked as 10%, assessing the structure and presentation, including writing style, citation and formatting issues.

- You should aim to keep your answers concise, while conveying the important information. Graphs and/or tables should be included where appropriate to present the results of your comparisons and experiments. Between 3000-4000 words (excluding references and appendices) might be appropriate for this report. A report which is not in .pdf format or whose length is larger than 4500 words may be penalised.

Plagiarism: One of the dangers of this assignment is the temptation to use paragraphs from web documents or papers that you have read. Please resist this temptation and do not do it. Otherwise, you will be heavily penalised. The report should be completely in your own words. If it is appropriate and absolutely necessary to include sentences and materials from elsewhere, then they should be clearly indicated as quotes, and references should be cited.

Please do not show your report to any other students.

Chuan Lu (cul), 13th April 2015