COMP 309 — *Machine Learning Tools and Techniques*

Assignment 1: Sprint on One Dataset Each

16% of Final Mark — Due: 11:59pm Monday 29 July 2019

1 Objectives

The goal of this assignment is to help you understand the basic concepts and algorithms within tools for machine learning. The purpose is to implement common Artificial Intelligence (AI) algorithms, use these algorithms to perform classification tasks, and analyse the results to draw useful conclusions. The Core and Completion components will introduce the WEKA data mining tool, whereas the Challenge component will introduce the KEEL software tool. The assignment will introduce you to:

- Machine learning concepts,
- Machine learning common tasks, paradigms and methods/algorithms,
- The concept of data-mining pipelines,
- The five tribes of AI.

These topics are (to be) covered in lectures 01–05.

2 Question Description

2.1 Core: Investigate Basic Use of The Different Tribes of AI [40 marks]

The first part of this assignment is to explore the four out of the five tribes of AI (i.e. Connectionist, Symbolic, Bayesian and Kernel-based approaches). You will be assigned a dataset from the UCI repository, please see course homepage (or below) for a list of students to datasets (please use the initial of your family/surname). Note, Part 2.1 is designed to be a sprint, so try to get quick results (these will be refined in Parts 2.2 & 2.3).

Requirements

Using the preexisting AI-tool, WEKA, this dataset is to be analysed using methods from the different tribes. Note that WEKA does not contain an evolutionary classification algorithm, so this will be tested in the Challenge component. One method must be selected in WEKA from each of the non-EC tribes, giving a total of four experiments. This assignment concerns utilising preexisting tools, so do not use technique code that you have written.

You should submit the following files and also a report electronically by the due date/time.

- (10 marks) Choose one classification method from WEKA for each of the (non-EC) AI-tribes.

  Perform classification using your chosen method on the assigned dataset (initially, use the whole dataset without splitting, which will be implemented in the later parts of this assignment).

  Present and describe in the report what you consider to be the important aspects of the results of each technique on the one dataset (approximately two paragraphs of text plus figures).

  Note the most appropriate form of presentation of results (e.g. decision trees, classification rules, decision boundaries and so forth) may differ between each technique. Exercise your skill and judgment to decide how the results should be communicated. Note, it is the techniques/results description that is important here, not the experimental method, e.g. setting up training and test sets properly is explored in subsequent parts.

- (20 marks) The report should detail why each selected technique from the stated tool belongs to a given tribe, i.e. identify the important aspects of the technique in terms of

  1. General description, especially the
  2. representation,
  3. evaluation method,
  4. optimization driver.

  (approximately four paragraphs of text)

- (10 marks) These important aspects may be different between each technique. Illustrate any important differences using the dataset. For example, use an instance in the dataset to show the differences in representation of each technique. (approximately three paragraphs of text)

Insight into 'why one or more aspect of a technique is suited to a given dataset' will be needed to achieve high grades.

2.2 Completion 1: Consider a Pipeline for Dataset Processing [20 marks]

This part is to implement a data pipeline.

Problem Description

Dataset pre-processing using a pipeline from raw data to deployable knowledge is very common in industry (e.g. CRISP-DM, which is covered later in the course). Your tasks is to specify a pipeline suitable to handle the assigned dataset.

Requirements

Instead of naively applying an AI tool to the whole dataset, a pipeline is to be created.
You will need to specify the pipeline in terms of the important components (features and parameter settings) used.

This will include k-fold cross validation.
You should include the following in the report:

- (20 marks) The report should include:

  1. Business understanding - consider the business aspects of the dataset, e.g. why was the data gathered? what did the acquisition hope to achieve? Note, that this may be mere obvious in some datasets than others.
  2. Data understanding - not only should the metadata be described (which is readily available in the UCI repository), but any interesting factors should be noted, e.g. mixed attribute type, high epistasis, outlier/noisy/missing data instances.
  3. Data preparation - state how the pipeline could assist in the preparation of the data prior to the technique being applied.
  4. Modelling - state whether this pipeline suits one or more of the five tribes of AI.
  5. Evaluation - similarly, state whether this pipeline supports one or more methods to evaluate a solution.
  6. Deployment - explain whether the model produced can easily be deployed or whether additional effort is required.

(approximately six paragraphs of text)
Note: "state" requires a direct answer, with one or two lines of additional insight only. "Discuss", "describe" and "consider" can be longer as different viewpoints on the arguments can be presented.

2.3 Completion 2: Use the pipeline to reevaluate the selected techniques in Part 2.1 used to classify the dataset [20 marks]

This part involves using the pipeline described in Part 2.2 with the techniques investigated in Part 2.1.

Problem Description

The main dataset is to be passed through the pipeline to generate deployable knowledge.

Requirements

Your pipeline should take the dataset file as input.
Please use intermediate file(s) that need to be clearly described in the report, e.g. imputed data, train/test sets and/or validation folds.
A final classifier together with testing results needs to be produced.

Compare the results of the naive approach (part 2.1) with the results of this pipeline approach (part 2.3). You should consider any changes in apparent classification accuracy, confidence in the results and likelihood of replicating the results.
You should submit the following files electronically.

- (10 marks) `Program code` for your classifiers (both set-up details (e.g. code and scripts) and executable program capable of running on the ECS School machines). The program should print out the classifiers in as human readable form (text form is fine) as possible.

  Compare and contrast the results between the different tools.

- (10 marks) The report should include:

  1. Accuracy in terms of the fraction of the test instances that it classified correctly.
  2. Report a snapshot of the learned classifiers discovered by your program.
  3. Compare the accuracy of your techniques before and after using a data pipeline approach. Please comment on any differences, suggesting reasons.

  (approximately three paragraphs of text plus figures)

2.4 Challenge: Use the KEEL to evaluate the Evolutionary Computation tribe on the dataset in Part 2.1 to classify the dataset [20 marks]

Problem Description

Evolutionary Computation is one of the five main tribes in Artificial Intelligence. KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source (GPLv3) Java software tool that can be used for a large number of different knowledge data discovery tasks.

Requirements

KEEL has an associated dataset repository, Including 'Standard Classification data sets', where it is worth noting that the cross-fold validation sets are premade for many common datasets.
There are over 500 algorithms included in KEEL, where you are to select one of the Classification Algorithms.
Choose one of the 'Rule Learning for Classification' → 'EVOLUTIONARY CRISP RULE LEARNING FOR CLASSIFICATION' algorithms so that the understandable nature and accuracy of the results can be compared with the other tribes.
Classify the dataset using the Evolutionary Computation technique selected.

- (10 marks) `Program code` for your classifiers (both set-up details (e.g. code and scripts) and executable program (capable of running on the ECS School machines). The program should print out the classifiers in as human readable form (text form is fine) as possible.
- (10 marks) The report should include:

  1. Accuracy in terms of the fraction of the test instances that it classified correctly.
  2. Report a snapshot of the learned classifiers discovered by your program.
  3. Compare the Evolutionary Computation approach with the other AI tribes from earlier parts.
  4. Compare the WEKA tool with the KEEL tool commenting on ease-of-use, performance and any other aspects you consider important (e.g. data format, documentation, online tutorials and so forth).

  (as this is the Challenge component, please decide how much you should write noting that too little could lose marks and too much will have diminishing returns. The art of clear and concise communication is an important skill for a data miner).

3 Relevant Data Files and Program Files

The relevant data files, information files about the data sets, and some utility program files can be found on-line. A soft copy of this assignment is available in the following directory:

`/vol/comp309/assignment1/`

1. The UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/index.php`),
2. WEKA can downloaded from the following website, which also contains tutorials:
   `https://www.cs.waikato.ac.nz/ml/index.html`
3. KEEL can downloaded from the following website, which also contains useful information:
   `http://www.keel.es/`

4 Assessment

We will endeavour to mark your work and return it to you as soon as possible, ideally in 2 weeks. The tutor(s) will run a number of helpdesks to provide assistance for parts 1 and 2 prior to the submission deadline. Anyone who submits the assignment a day or more late will not have it returned promptly as it will miss the moderation meetings.

5 Submission Guidelines

5.1 Submission Requirements

1. Programs for all individual parts. To avoid confusion, all the individual parts should use directories `part1/`, `part2/`, `...` and all programs should be stored in their corresponding directories. Within each directory, please provide a `readme file` that specifies how to compile and run your programs on the ECS School machines. A script file called `sampleoutput.txt` should also be provided to show how your program run properly. If you programs cannot run properly, you should provide a `buglist` file.
2. A document that consisting of the report of all the individual parts. The document should mark each part clearly. The document can be written in text or the DOC, but should be submitted in PDF format.

5.2 Submission Method

The programs and the PDF version of the document should be submitted through the web submission system from the COMP309 course web site **by the due time**.

There is NO required hard copy of the documents.

KEEP a backup and receipt of submission.

Submission need to be completed on School machines, i.e. problems with personal PCs, internet connections, file dependencies and lost files, which although eliciting sympathies, will not result in extensions for missed deadlines.

5.3 Late Penalties

The assignment must be handed in on time unless you have made a prior arrangement with the lecturer or have a valid medical excuse (for minor illnesses it is sufficient to discuss this with the lecturer). The penalty for assignments that are handed in late without prior arrangement is one grade reduction per day. Assignments that are more than one week late will not be marked.

6 Data File allocation

Each student is allocated a dataset to investigate using their choice of AI tool. Please see the data set information for the task, e.g. 'Prediction task is to determine whether a person makes over 50K a year.' The dataset to use is allocated below (`http://archive.ics.uci.edu/ml/index.php`):

Surname (Family name)	**Dataset**
A–B	Adult
C–D	Heart disease (Cleveland)
E–F	Abalone
G–H	Breast cancer (Wisconsin)
I–J	Hepatitis
K–L	Mushroom
M–N	Soybean (Small)
O–P	Ionosphere
Q–R	Zoo
S–T	Sonar
U–V	Wine
W–X	Glass
Y–Z	Spect