COMP309 — *Machine Learning Tools and Techniques*

Assignment 1: Sprint on one dataset each

15% of Final Mark — Due: 11:59pm Friday 3 August 2018

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. In particular, the following topics should be reviewed:

- 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. Research into online resources for AI is encouraged.

2 Question Description

It is possible to pick an AI tool, apply it to a dataset and get apparently good results. However, it is difficult to know if these are the best achievable results[1].

This assignment introduces the range of AI techniques. It demonstrates the ease of modern AI tools to produce knowledge from datasets. This knowledge should be related to predictive classification of future instances, but can also include clustering, feature analysis, meta-data and so forth. The task to determine how optimum and trustworthy this knowledge is to the user.

Part 1: Investigate basic use of the five tribes of AI [40 Marks]

The first part of this assignment is to explore the five tribes of AI. You will be assigned a dataset from the UCI repository, please see course homepage for a list of student IDs to datasets. Note, Part 1 is designed to be a sprint, so try to get quick results (these will be refined in Parts 2 & 3).

Requirements

Using pre-existing AI-tools, such as Weka, this dataset is to be analysed using methods from the five tribes. One method must be selected from each tribe, giving a total of five experiments. The selected methods can be implemented from the same or different tools. This assignment concerns pre-existing tools, so please do not use technique code that you have written as results are time consuming to validate.

You should submit the following files and also a report electronically in week 3.

- (10 marks) Describe the results of each technique on the one dataset. Note the most appropriate form of results may differ between each technique. Exercise your skill and judgement to decide how the results should be communicated. Note, it is the description that is important here, not the experimental method, e.g. setting up training and test sets properly is explored in Part 2.

[1] Unless the dataset is small, such that enumerating all solutions is possible, or artificial so that the solutions are known beforehand (known as a 'toy' problem). In either case, apart from assessing the performance of a tool, there is little practical benefit in applying an AI tool here.

- (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.

- (10 marks) Identify how these aspects of the technique are different in relation to the dataset:

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

Part 2: Consider a pipeline for dataset processing [20 marks]

This part is to implement a data pipeline. Rather than naïvely applying an AI tool to the dataset a pipeline is to be created. Additional questions on k-fold cross validation need to be answered and discussed.

Problem Description

CRISP-DM and related knowledge discovery from dataset processes form a pipeline from raw data to deployable knowledge. Your tasks is to specify a pipeline suitable to handle the assigned dataset (Part 3 will use this pipeline).

Requirements

Your specification should detail the proposed pipeline in terms of platform, tool integration and important features.

You should submit the following report electronically.

- (20 marks) `A report` in any of the PDF, text or DOC formats. 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.

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.

Part 3: Use the pipeline to revaluate the five tribes of AI to classify the dataset [40 marks]

This part involves using the pipeline described in Part 2 with the techniques investigated in Part 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.

Any intermediate file(s) should be clearly described, e.g. imputed data, train/test sets and/or validation folds.

The final classifier together with testing results.

You should submit the following files electronically.

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

  (Optional) Compare and contrast the results between the different tools.

- (20 marks) A report in PDF or text format. 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.

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

4 Assessment

We will endeavour to mark your work and return it to you as soon as possible, hopefully in 2 weeks. The tutor(s) will run a number of helpdesks to provide assistance in the week prior to the submission deadline.

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 PDF, text or the DOC 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 should be completed on School machines, i.e. problems with personal PCs, internet connections 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.

Each student is allocated a dataset to investigate using their choice of AI tool. The dataset to use is allocated below (http://archive.ics.uci.edu/ml/index.php):

Surname (family name)	**Dataset**
A-B	Appendicitis
C-D	Breast cancer (Wisconsin)
E-F	Diabetes (Pima Indian)
G-H	Heart disease (Cleveland)
I-J	Hepatitis
K-L	Hypothyroid
M-N	Hepatobiliary disorders
O-P	Ionosphere
Q-R	Zoo
S-T	Sonar
U-V	Wine
W-X	Glass
Y-Z	Spect