COMP 309 — *Machine Learning Tools and Techniques*

Assignment 2: Real-World Data Handling, Modelling and Visualisation

16% of Final Mark — Due: 11:59pm Monday 12th August 2019

1 Objectives

The goal of this assignment is to help you understand the Data manipulation and visualisation within tools for machine learning. The purpose is to implement common Data handling methods on real-world observations. Validation of the effectiveness of the implemented methods will be through using ML tools to perform analysis tasks to draw useful conclusions. In particular, the following topics should be reviewed:

- CRISP-DM,
- Feature manipulation, including feature selection and feature construction,
- Visualisation of results,
- Real-world uses of ML tools.

These topics are (to be) covered in lectures 01–09. Research into online resources for AI is encouraged, where the rabbit-hole will provide useful jumping off points for further exploration.

2 Question Description

"Whatever happens, over-fishing today will lead to collapses tomorrow, ... " Jun 14 2019 Stuff
Fishing in New Zealand waters has both economic and environmental consequences. Your task is to use real-world data in ML tools to investigate this issue.
Firstly, what evidence can be found to support (or dismiss) the newspaper headlines?
Secondly, what features can be considered as the underlying cause for any increase (or decrease) in fish numbers (stocks)?
Thirdly, what trends and predictions can be made from the data?
Finally, what findings would you communicate, based on your analysis [also consider the consequences of publicising the information found]? This assignment introduces the range of AI techniques. It demonstrates the ease of modern AI tools to produce knowledge from datasets. The task also asks you to consider how welcome, useful and trustworthy this knowledge is to the user.

2.1 Part 1: Core: Evidence related to fish stocks in New Zealand [40 marks]

The first part of this assignment is to explore the Data manipulation processes. You will need to obtain data from a public repository. The first place to start is Data NZ (https://www.data.govt.nz/). Search for data associated with 'fish' or other related key phrases.

The task is to use CRISP-DM and data manipulation in the construction of a dataset pipeline for future analysis. One dataset will be sufficient for this Core task. For the Completion and Challenge parts, it is necessary to use at least two datasets from various sources [these must be available to the public].

Requirements

Using pre-existing AI-tools, such as WEKA, a pipeline is to be constructed, such that proper data processes can be utilised.
This pipeline should be used to produce results, such as classification models, clustering results or regression models. These need to illustrate the most important aspects of NZ fisheries.

- (10 marks) Select a 'fish' dataset from https://www.data.govt.nz/ Import this data into a publicly available machine learning tool, such as WEKA. Analyse the data using various machine learning methods. Your judgement will be necessary in determining whether clustering, regression, classification or other techniques are most appropriate to analyse the data. Please choose two or more techniques to analyse the data. The analysis should be conducted using a pipeline for ease of data manipulation and selection of techniques.
- (10 marks) Describe the results of each technique used 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.
- (10 marks) Identify how these aspects of the techniques are different, e.g. how do the results from clustering differ from classification techniques. Please refer to the dataset when describing the differences in the techniques.
- (10 marks) Please revisit the business understanding based on your exploration of the data. It is noted that a simple question to ask is "is there any evidence of fish stocks collapsing in NZ waters?". Please create and describe two other questions that could be asked of the data as well.

These items needs to be summarised in a short description (half page per item plus any illustrations/figures) with examples of the methods employed being described.

2.2 Part 2: Completion: Feature importance to Fish stocks in New Zealand [40 marks]

Identify features that are important to determining Fish Stocks in New Zealand. For example, how important is the location, time of the year, weather conditions during the sample collection and so forth? Show how the machine learning tools can identify such features.

Requirements

- (10 marks) Select one of the three business understanding questions outlined above to address. Explain why it is an interesting question. Determined at least one additional publicly available datasets to complement the originally chosen dataset. 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 more obvious in some datasets than others. Credit will be given for interesting and appropriate choices of supplementary datasets. Describe why the dataset(s) chosen are appropriate to the business understanding case.
- (10 marks) Using suitable techniques, such as WEKA command-line, *merge* the datasets together. In the report, describe how this merging was achieved. Note that *appending* two datasets with identical features but different instances is different from *merging* distinct data sources that consider different features related to the same task/problem.
- (10 marks) Utilise dimensionality reduction technique(s) to identify which features are irrelevant and/or redundant to selected tools' performance, e.g. classification, regression and so fourth. Remove redundant/irrelevant data. Handle missing data. Discretise continuous ranges as necessary. Remove any unnecessary instances, e.g. redundant instances, outliers (take care whether this is necessary/appropriate) or non-effective instances.
- (10 marks) Now only a subset of the original features and instances should exist. Analyse the output of the ML tool on this processed dataset to identify which features are important to a selected tools' output, e.g. which features are near the start of the decision tree for classification, or which features have the highest weights in regression and so fourth.

Two pages of description plus results is required [half a page of text plus figures on each of the items above is appropriate].

2.3 Part 3: Challenge: Visualisation of results [20 marks]

ML Tools have visualisation methods as part of the pipeline. The task is to use such tools to highlight important findings from this work.
Results from using ML tools have to be communicated to the audience in a clear and meaningful manner.

Requirements

Design a one slide poster in PowerPoint (or equivalent package) of size no more than A3 and no font smaller than size 14 Times New Roman (we won't be checking, but if we can't read it we can't mark it!). It is the appropriate selection of tables, charts, graphs and insightful text that will be marked, not the design aesthetics.

You should submit the following poster electronically.

- (10 marks) Visualise the most interesting results from the dataset in an appropriate manner. For example, show important trends in fish numbers, show clusters of important natural resources and so forth.
- (10 marks) Consider the consequences and ethics of reporting your findings. Comment on any recommendations that your analysis suggests, e.g. why should businesses agree to a complete ban in fishing certain areas around NZ to prevent stock collapse?

  Please do not rely upon anecdotes and suppositions. Use constructed AI models to determine the effect of any new information on the analysis. Models are to be used (or constructed) in an attempt to provide insight into future hypothetical situations that might result from the release of above findings. For example, what would happen if the fishing doubled its catch per year? Please use your ML tool to investigate the effect on the data, e.g. investigate what catch size could make a difference to stock numbers (if this is an important factor).

Report

You should submit the following files electronically.

- `A report` in PDF or text format. The report should include the description/details as requested above.
- `Program code` for your system including both parameter set-up details and Command line input (e.g. code and scripts) can be included in an appendix of the report if not included as a screen capture is in the main body of the report.
- `A poster` in PowerPoint or equivalent format.

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/assignment2/`

1. https://www.data.govt.nz/

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 first week of the assignment to answer any questions regarding what is required and then in the week prior to the submission deadline.

5 Submission Guidelines

5.1 Submission Requirements

1. We reserve the right to individually assess your programmes running on ECS school machines. Thus, please keep a copy of all data, code, ML Tool setup (e.g. pipeline and parameters) and related material in the event that you are asked to demonstrate your results in person.

5.2 Submission Method

The Poster 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.