COMP 309 — *Machine Learning Tools and Techniques*

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

15% of Final Mark — Due: 11:59pm Monday 20th August 2018

1 Objectives

The goal of this assignment is to help you understand the data manipulation and visualisation within the context of machine learning (ML) tools. The purpose is to implement common data handling methods on real-world observations. Validation of the effectiveness of the implemented methods will be performed 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 (`https://ecs.victoria.ac.nz/Courses/COMP309_2018T2/RabbitHole`) will provide useful jumping off points for further exploration.

2 Motivation

"Wellington's skyrocketing rental prices are not an issue in some of the provinces 10:51, Jan 25 2018 Stuff"
"Wellington rental prices match Auckland as property listings plummet 21:44, Feb 26 2018 Stuff"
"Wellington's rental prices cool, national rent stays at record high 08:02, Mar 26 2018 Stuff"
"Rents rise around the country as cost crunch goes on landlords 14:07, Apr 16 2018 Stuff"

Study at Victoria Day is August 31, 2018, where many future scholars and their parents will have concerns regarding the reported high rental costs in Wellington. Anecdotal evidence suggests that prospective students will be put-off from enrolling due to this perceived issue, which has major financial consequences for the University.

The task is to analyse real-world data using 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 rental costs?
Thirdly, what trends and predictions can be made from the data?
Finally, what you would you tell a prospective student & parent, based on your analysis [i.e. consider the consequences of publicising the information found]?

This assignment introduces a range of data handling techniques. It demonstrates the ease of modern ML tools to produce knowledge from datasets. The task is also to determine how viable and trustworthy this knowledge is to a user.

2.1 Part 1: Evidence related to rental costs in Wellington [30 marks]

The first part of this assignment is to explore data manipulation. You will need to obtain data from publicly available repositories. The place to start is Data NZ (https://www.data.govt.nz/). Search for [datasets] associated with 'rent' and/or other key phrases.
The task is to use CRISP-DM and data manipulation to obtain data observations within a pipeline for future analysis. One dataset will be insufficient for this task. It is advised to use at least three datasets from various resources [that must be publicly available]. At least one dataset must be from Data NZ, with kudos for finding alternative sources of relevant public data (the more unique the better).

Requirements

Using existing ML tools, such as WEKA, a pipeline is to be constructed, such that proper data processes can be utilised. The methods employed need to be identified, i.e. a short report (1-2 pages) with illustrative examples of the methods employed and links to the datasets used.

This pipeline should be used to produce results, such as classification models or clustering results, that illustrate attributes of the Wellington rental market, e.g. most expensive NZ rental markets, costs compared with other university towns/cities and so forth.

- (10 marks) Describe why you selected the datasets, stating the criteria used.
- (10 marks) Describe how the datasets were manipulated (integrated) in the pipeline.
- (10 marks) Describe the important aspects of the data that were used to create the ML models, i.e. justify how features are transformed to attributes for the ML algorithms.
  Specifically identify any feature selection and/or feature manipulation of methods employed.

Snapshots of the data and its manipulation should be presented, i.e. illustrate and describe before and after the transformations throughout the pipeline prior to the final model being executed. Note, intermediate models that generate feature construction and feature selection should be described.

2.2 Part 2: Feature importance to rental costs in Wellington [20 marks]

Identify features that are important to determining the rental costs in Wellington. For example, is it the supply of properties to buy, the number of students at the University, the GDP of New Zealand, or other factors that determine (predict) the rental costs in Wellington? Show how the ML tools can identify such features, where 1-2 pages of description (plus results) is required.

Requirements

- (10 marks) Utilise dimensionality reduction technique(s) to identify which attributes are irrelevant to the selected tools' performance. The tool(s) can be from any tribe and any task, e.g. classification, regression and so forth. However, their use must be fully justified.
- (10 marks) Analyse the output of the ML tool to identify which attributes are important to selected tools' output, e.g. which attributes are high on the decision tree for classification, or which attributes are important in regression and so forth. Explore how this changes with dimensionality reduction included/excluded from the pipeline.

2.3 Part 3: Visualisation of results [30 marks]

ML Tools include visualisation methods as part of the pipeline. The task is to use such tools to highlight important findings from this work. 形象化

Problem Description

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 report electronically.

- (30 marks) Submit a one slide 'poster' in PowerPoint (or equivalent) package. This can include gifs and animated visuals to illustrate important points regarding the data and findings. [Animation should not be used to make more space on the slide only!]. A selection of the posters could be displayed at Study at Vic day.

2.4 Part 4: Consider the consequences and ethics of reporting your findings [20 marks]

The work in the first parts of this assignment should indicate hypotheses on the rental market, which are to be explored further here.
For example, if it was determined that the cost of living in Wellington, due to high rents, was 20% higher than alternatives, which results in say a 5% drop in students, then what would be the consequences? Alternatively, if it was determined that the cost of living in Wellington was 20% lower than what is predicted to be affordable, then how would the rental sector react to this information?
As well as using anecdotes and suppositions, use constructed ML models to determine the effect of any new information on the analysis. For example, what would happen if the University built a new 400 room Hall of Residence instead of a new building for the Faculty of Engineering? Please use your ML tool to investigate the effect on the data, e.g. investigate what size of hall of residents could make a difference if this was determined to be an important factor.

Problem Description

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.

Requirements

Hypotheses are to be described and then tested in the ML tools.
Produce models and evidence that supports hypothesised futures. Realistic and pragmatic concerns should be discussed, e.g. "the model predicts a 10,000 room Hall of Residence will solve the rental problem for students, but this is impractical due to space and cost considerations."
You should submit the following files electronically.

- (20 marks) `A short report` in PDF or Word format (1-2 pages). The report should include:
  1. A short description on the discovered hypothesis(es), including evidence.
  2. Interpretation of the consequences of these hypotheses.
  3. The use of ML techniques to determine if there is any support for the described consequences. That is, utilise the pipeline to manipulate the observations to see if the results can be changed. Students can change and create 'fake' data to feed into existing models.

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 from the course homepage.

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. 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 (if applicable). 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. Datasets below 1 Mb .csv can be included, else include comprehensive download instructions.

2. A document that consisting of the report of all the individual parts. The document should mark each part clearly. The document is preferred to be submitted in PDF, rather than text or .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.