COMP 309 — *Machine Learning Tools and Techniques*

Assignment 3: Kaggle Competition

15% of Final Mark — Due: 11:59pm Monday 10th September 2018

1 Objectives

The goal of this assignment is to help you tie together all the concepts you have learnt in the first half of this course in the lectures and assignments. To aid you in completing this assignment, you should review the major aspects of the course that have been explored so far, such as:

- Data understanding, cleansing, and pre-processing,
- Machine learning concepts,
- CRISP-DM and pipelines in general,
- Feature manipulation, including feature selection, feature construction and imputation,
- Statistical design and analysis of results.

These topics are (to be) covered in lectures 01–12. Research into online resources for AI is encouraged, where the rabbit-hole[1] will provide useful jumping off points for further exploration.

2 Question Description

"A cross-government review is underway to increase the transparency and accountability of how government uses algorithms to help improve the lives of New Zealanders." [https://www.data.govt.nz/]
As bright and eager COMP309 students, you have been tasked with showcasing how ML tools can be used for the good of society.
The overall aim of this assignment is to develop the best possible machine learning system to **predict the household income of families of newborn children**. The hope is that officials will be able to use your model to better understand the factors behind inequality in New Zealand.

We have set up a Kaggle InClass Competition[2] to facilitate finding the best machine learning system for officials to use. You will be expected to analyse the provided census data, design and improve your own machine learning pipeline, and consider the consequences of applying your pipeline to this data. [Note the data is synthetic, as the ethics application process for using the equivalent real data is time consuming, where the data has been constructed by a government agency to model actual observations]

2.1 Preliminary: Accessing the Kaggle InClass Competition

To access the class competition, you must use the below url. **Please do not share this publicly** as it will allow anybody to access our competition, which will make the experience less enjoyable for your classmates. Deliberate cheating is a disciplinary matter, so please don't go there.

Competition link: `https://www.kaggle.com/t/905957a761ff438e90b93ea6932a6545`

You will need to register a Kaggle account. It is perfectly fine (and expected) to use a pseudonym as your Kaggle username so your classmates do not know your real-life identity. However, **you will need to fill out the following**

[1] `https://ecs.victoria.ac.nz/Courses/COMP309_2018T2/RabbitHole`
[2] `https://www.kaggle.com/about/inclass/overview`

form so that the lecturers and tutors can link your Kaggle result to your ECS account. No other people will have access to this information!

Please fill out the following form: `https://goo.gl/forms/Hvj2AQqf6o3zRDfQ2`
Please submit as part of your report.
Once you have completed the above steps, please verify that you can access the following page:
`https://www.kaggle.com/c/comp309-2018t2/overview` (when logged in).
Once successful, you may proceed to the rest of the assignment!

2.2 Part 1: Exploring and understanding the census data [25 marks]

It is often much more effective to first learn about the properties of a dataset (business and data understanding) before applying machine learning to it. You should begin by familiarising yourself with the dataset by reading the "Overview" and "Data" tabs of the Kaggle competition. Please download the dataset from the Data tab (in .csv format). You should now spend some time examining the data and taking notes of any interesting patterns you find.

Requirements

Using any tools you find useful, you should explore and analyse the dataset. You should draw upon your previous experiences and what you have learnt in this course to find a number of interesting patterns (*at least four*). You may wish to start by examining the quality, completeness and representation of individual features.

You should submit a short (2 page) report electronically:

- (25 marks) The report should highlight the findings of your dataset exploration. You should identify each pattern clearly using examples, and discuss the potential consequence this may have on your results. To achieve a high mark, you should consider more complicated patterns, such as feature interactions, or provide a particularly insightful discussion that is convincing to the reader. Visualisation is an important aspect of this task.

potential consequence - 潜在因素

2.3 Part 2: Developing and testing your machine learning system [50 marks]

Now that you have some initial understanding of the census dataset, you should design an initial system (model) that you consider has the potential to accurately predict the total household income of a newborn. You may use any ML tools you wish, but a good solution will consider a number of factors, such as: pre-processing steps, the properties of the dataset and generalisation/over-fitting. Decisions around how to split your labelled data into training/testing/cross-validation set/s are your choice, which are important and should be explained.

In this part, you should not remove features that you think may be "ethically questionable" — these concerns will be discussed later.

The final output of your system should be a single csv file containing two columns that represent an instance's unique ID and your predicted class label. The csv should include your predictions for all 7621 instances in the dataset, plus a header line (7622 lines). For example:

```
random_ID, total_income_hhld_code
1, 3
2, 0
3, 9
...
10000, 2
```

Once you are satisfied with your initial attempt (do not spend too long on it!!), you should upload your output csv to the submissions page of the competition: `https://www.kaggle.com/c/comp309-2018t2/submissions`. Once your submission has been processed, you will be able to see your classification accuracy on the public leaderboard. You

should use this feedback to further improve your system. For example, if your leaderboard performance is much lower than on your own test set, you have over-fitted your model. You should use your judgment to decide how extensively to change your system. This may only be tweaking parameters, or you may decide to try a completely different algorithm. Note that you are limited to 5 submissions per day, but submitting this many may be to your detriment as you may "over-train" on the public leaderboard!!
Do not be discouraged if your performance appears low on the leaderboard: we are interested in novel/interesting solutions even if they have lower performance, and you may come out on top on the private leaderboard anyway...

Requirements

You should refine your machine learning system a number of times (*at least 3*, including the initial system) based on the performance you achieve on the public leaderboard. You should submit a 3-4 page report electronically:

- (20 marks) Discuss the initial design of your system, i.e. before you have submitted any predictions to the Kaggle competition. Justify each decision you made in its design with reference insight you gained in Part 1, and any other factors.
- (20 marks) Discuss the design of one or more of your intermediary systems. Justify the changes you made to the previous design based on its performance on the leaderboard, and from any other additional investigation you performed.
- (10 marks). Use your judgement to choose the best system you have developed — this may not necessarily be the most accurate system on the leaderboard. **Make sure you select this submission as your final one on the competition page before the deadline**. Explain why you chose this system, and note any particularly novel/interesting parts of it. You should submit the source and executable code required to run your chosen submission so that the tutors can verify its authenticity.

2.4 Part 3: Reflecting on your findings [25 marks]

Until now, we have been focusing on achieving the best performance possible — but consider whether this is all that ML tool users should consider?[3]

The census dataset has a number of features that could be used to produce a machine learning system that, while accurate, has a number of biases towards certain population groups. Should the officials be worried that using your model in their analyses could be harmful to society?

Requirements

You should consider the interpretability of your final chosen model from Part 2, and analyse any ethical concerns associated with its structure.
You should submit a short (1-2 page) report electronically that answers the following questions:

- (10 marks) How easy is it to interpret your chosen machine learning model? Discuss any ethical consequences of how it uses the chosen features to make a prediction.
- (15 marks) Were your other systems more or less interpretable? Discuss any relationship between the performance of your systems and how interpretable they are. If you were told to produce a system that was not biased towards certain population groups, what decisions would you make to achieve this?

[3] `https://towardsdatascience.com/can-a-machine-be-racist-5809b18e5a91`

3 Relevant Data Files and Kaggle Information

The census dataset, and additional information about the Kaggle competition can be found online:

`https://www.kaggle.com/c/comp309-2018t2/`

4 Assessment

We will endeavour to mark your work and return it to you as soon as possible, hopefully in 2 weeks. Your position on the private leaderboard will be influencial on the final grade, but will not be the only consideration. 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. 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. *Ensure you submit your chosen solution in a form that the tutors can understand and run easily.*

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.