Assignment 1A

CAB420, Machine Learning, Semester 1, 2021

This document sets out the two (2) questions you are to complete for CAB420 Assignment 1A. The assignment is worth 8% of the overall subject grade. All questions are weighted equally. Students are to work either individually, or in groups of two. Students should submit their answers in a single document (either a PDF or word document), and upload this to TurnItIn. If students work in a group of two, only one student should submit a copy of the report and both student names should be clearly written on the first page of the submission.

Further Instructions:

1. Data required for this assessment is available on blackboard alongside this document in *CAB420_Assessment_1A_Data.zip*. Please refer to individual questions regarding which data to use for which question.

2. Answers should be submitted via the TurnItIn submission system, linked to on Blackboard. In the event that TurnItIn is down, or you are unable to submit via TurnItIn, please email your responses to `cab420query@qut.edu.au`.

3. For each question, a short written response (approximately 2-4 pages depending on the nature of the question, approach taken, and number of figures included) is expected. This response should explain and justify the approach taken to address the question (including, if relevant, why the approach was selected over other possible methods), and include results, relevant figures, and analysis.

4. MATLAB or Python code, including live scripts or notebooks (or equivalent materials for other languages) may optionally be included as appendices. **Figures and outputs/results that are critical to question answers should be included in the main question response, and not appear only in an appendix**. Note that MATLAB Live Scipts, Python Notebooks, or similar materials will not on their own constitute a valid submission and a written response per question is expected as noted above.

5. Students who require an extension should lodge their extension application with HiQ (see `http://external-apps.qut.edu.au/studentservices/concession/`). Please note that teaching staff (including the unit coordinator) cannot grant extensions.

Problem 1. Regression. The data in `Q1/communities.csv` contains socio-economic data from the 1990 US census for various US communities, and the number of violent crimes per capita (in the column `ViolentCrimesPerPop`). The purpose of the data is to explore the link between the various socio-economic factors and crime.

Given the provided data, you are to:

- Split the data into training, validation and testing sets.
- Train a linear regression model to predict the number of violent crimes per captia from the socio-economic data.
- Train a LASSO regression model to predict the number of violent crimes per captia from the socio-economic data.
- Train a Ridge regression model to predict the number of violent crimes per captia from the socio-economic data.

For your analysis, you should disregard the first five columns (`state`, `county`, `community`, `communityname string` and `fold`). Note that the provided data may also contain missing values, and may need to be sanitized in some way prior to model development.

For LASSO and Ridge models, the validation dataset should be used to select the optimal value of λ. The performance of all models should be compared on the separate test set, and should consider the predictive power of the model, the model complexity, and the model validity.

Your final response should include:

- Discussion of how the data is handled, including any data cleaning or removal, and how the data is split into training, validation and testing.
- Details of the three trained models, including details such as values for λ for the LASSO and Ridge models, and a discussion of how these were selected.
- An evaluation comparing the three models, considering model accuracy and model validity.

Problem 2. Classification. Land use classification is an important task to understand our changing environment. One approach to this involves the use of data from aerial sensors that captures different spectral reflectance properties of the ground below. From this data, the land type can be classified.

You have been provided with training and testing data (`Q2/training.csv` and `Q2/testing.csv`) that include 27 spectral properties and an overall classification of land type, which can be one of:

- *s*: 'Sugi' forest;
- *h*: 'Hinoki' forest;
- *d*: 'Mixed deciduous' forest;
- *o*: 'Other' non-forest land.

You are to use the provided training set as-is, and split the provided test data (`Q2/testing.csv`) into validation and testing sets.

Using this data you are to train three multi-class classifiers to classify land type from the spectral data. These classifiers are to be:

1. A K-Nearest Neighbours Classifier;
2. A Random Forest; and
3. An ensemble of Support Vector Machines.

Model hyper-parameters should be selected using the validation set you create by splitting the provided test data. Note that using random automatic hyper-parameter optimization is not an acceptable way to choose hyper-parameters for this question. Instead you should perform a series of evaluations or a grid-search with a clear rationale to evaluate and select appropriate parameters. The resultant models are to be evaluated on the testing set and compared.

Your answer to this question should include:

- Details on how the data was split into training, validation and testing sets.
- Details of hyper-parameter selection, including justification for the approach taken and any intermediate results that led to the final models.
- An evaluation and comparison of the final three models, including a discussion exploring any difference in performance between the models. This should also highlight any major differences in the models in terms of classification performance or capabilities.