Assignment 1A

CAB420, Machine Learning

This document sets out the three (3) questions you are to complete for CAB420 Assignment 1A. Problem 1 highlights the content from weeks 1 and 2, Problem 2 from week 3, and Problem 3 from weeks 4 and 5; though material from all weeks will aid in addressing these problems. This assignment has been designed such that it can be completed solely using the resources covered in class, though students are encouraged to refer to external sources and conduct further research into the tasks and cite relevant sources.

The assignment is worth 15% of the overall subject grade. All questions are weighted equally. Students are to work individually. Students should submit their answers in a single document (either a PDF or word document), and upload this through the Canvas submission link. Further Instructions:

1. Data required for this assessment is available on canvas alongside this document in *CAB420_Assessment_1A_Data.zip*. Please refer to individual questions regarding which data to use for which question.
2. Juptyer Notebook python templates have been provided for each question on Cavnas.
3. Answers should be submitted via Cavnas. In the event that Cavnas is down, or you are unable to submit via Cavnas, please email your responses to `cab420query@qut.edu.au`.
4. For each question, **a concise written response**. Guidelines for the length of the written responses have been supplied in this assignment brief. These guides pertain to the **written** component of each section (ie. excluding figures), though figures should be included to enhance the response only. Figures without sufficient discussion show no evidence of understanding of the content or the task. The length of responses may exceed these guides, but should do so sparingly and only when believed to be suitable. Note that the CRA requires responses to be "clear and concise", so verbose and lengthy submissions that needlessly disregard these guides will be penalised according to the CRA. This similarly applies to the excessive use of figures without sufficient discussion or evaluation.
5. Responses 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. **Python Notebooks, or similar materials will not on their own constitute a valid response to a question and will score a mark of 0.**

6. Responses should highlight where the modelling approach has been successful, and to highlight the limitations apparent. When limitations are encountered, students are encouraged to investigate and identify what the underlying cause of any failure or limitations are.

7. 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**.

8. Students are encouraged to use **any code presented in lectures/tutorials/examples** from CAB420. Use of external code is also permitted, though requires to be correctly cited and referenced.

9. 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 the `Q1` directory in `CAB420_Assessment_1A_Data.zip` 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.

The provided data has been split into training (`communities_train`), validation (`communities_val`), and testing (`communities_test`) sets.

Your Task: Given the provided data, you are to:

- Train a linear 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.
- Train a LASSO regression model to predict the number of violent crimes per captia from the socio-economic data.

For your analysis, you are to use all provided data (i.e. **DO NOT** perform any filtering or selection to remove columns and/or rows). For LASSO and Ridge models, the validation dataset should be used to select the optimal value of λ. All models should be evaluated on the test set. Evaluations should consider the predictive power of the model, the model complexity, and the model validity.

Your response must include sections that address the following:

- Discussion and justification of any pre-processing performed, such as standardisation. **(1/3 page)**
- Details of the three trained models, including key parameters such as values for λ for the LASSO and Ridge models, and a brief discussion of how these were selected. **(1/3 page)**
- An evaluation comparing the three models. Your evaluation should consider model accuracy and model validity. Discussion of accuracy/validity should also consider the socio-economic nature of this data, and if or how this impacts the the required accuracy and/or utility of the resultant models. You may also wish to discuss the relevance of terms within the model. The use of tables to summarise results, and figures to highlight details such as model validity, is strongly recommended. **(1 page)**
- Socio-economic modelling is incredibly complex and often dealing with sensitive data and variables. Briefly discuss the **ethical concerns** that need to be considered when evaluating the models developed for this problem. This discussion may include reference to potential confounding variables, or to limitations in the modelling schemes used. Students are encouraged to highlight these ethical considerations and limitations to help avoid misinterpretation of the results. **(1/3 page)**

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, validation and testing data (`Q2/training.csv`, `Q2/validation.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.

Your Task: Using the provided data as-is, 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 a grid search operating over the validation set. The resultant models are to be evaluated on the testing set and compared.

Your response must include sections that address the following:

- Discussion and justification of any pre-processing performed, such as standardisation. **(1/3 page)**
- Details of the hyper-parameter selection method, the final model parameters selected, and a discussion of these in relation to characteristics of the data. Discuss the effect of your hyper-parameters will have on the model (ie. if you have found a small value for a hyper parameter, what effect will this have on the model compared to a large value?) **(2/3 page)**
- An evaluation and comparison of the final three models, including a discussion exploring any difference in performance between the models. **(1 page)**

Problem 3. Training and Adapting Deep Networks. When training deep neural networks, the availability of data is a frequent challenge. As such, methods including fine tuning and data augmentation are common practices to address data challenges.

You have been provided with two portions of data from the Street View House Numbers (SVHN) dataset, a 'real world' MNIST style dataset. The two data portions are:

1. A training set, `Q3/q3_train.mat`, containing $1,000$ samples total distributed across the 10 classes.
2. A testing set, `Q3/q3_test.mat`, $10,000$ samples total distributed across the 10 classes.

These sets do no overlap, and have been extracted randomly from the original $SVHN$ testing dataset. Note that the training set being significantly smaller than the test set is by design for this question, and is not an error. The template code provided for this question implements an SVM to classify this data, and to time how long the SVM takes during training and testing.

Your Task: Using these datasets and the provided code, you are to:

1. Design/select a DCNN architecture, and using this network:
   (a) Train a model from scratch, using no data augmentation, on the provided abridged SVHN training set.
   (b) Train a model from scratch, using the data augmentation of your choice, on the provided abridged SVHN training set.
2. Compare the performance of two DCNNs (with and without augmentation), considering the accuracy, training time and inference time (i.e. time taken to evaluate the models), using the provided test set.

All models should be evaluated on the provided SVHN test set, and their performance should be compared. DCNN architectures may be taken from lecture examples or practical solutions. Your selection of model may take computational constraints into consideration.

Your response must include sections that address the following:

- Discussion of neural network design (i.e. the network topology), and training approach. Discussion of training should include details of how long the network was trained for, and if training converged. If the design was constrained due to limited computational resources, this should be clearly stated. **(2/3 page)**
- Discussion of the data augmentations used, including a brief justification as to why these were chosen. **(1/3 page)**
- Comparison between the two DCNNs (with and without augmentation) and the SVM which considers both performance, training time, and inference time. Evaluation should us appropriate metrics and/or visualisations to highlight any differences in performance between the models. **(1 page)**