University of Aberdeen

School of Natural and Computing Sciences

Department of Computing Science

MSc in Artificial Intelligence

2020 – 2021

****Please read all the information below carefully****
Assessment Item 1 of 2 Briefing Document – Individually Assessed (no teamwork)

Title: CS5062 – Machine Learning	Note: This assessment accounts for 50% of your total mark of the course.

Learning Outcomes

On successful completion of this component a student will have demonstrated competence in the following areas:

- Understanding and practice machine learning principles by applying machine learning theories and methodologies into benchmark data

Information for Plagiarism: The source code and your report may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

Report Guidance & Requirements

Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report, along with the corresponding code, containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken.

This assessment includes two tasks. The first task focuses on a regression problem. The main purpose of this task is to understand that when analysing a data set, there are often a number of machine learning models available. Evaluating these models and choosing the best possible models are the core efforts when using machine learning. The second task will focus on an image classification which provides you an opportunity to employ the state-of-the-art machine learning tools to analyse a relatively big data set, providing you a taste of using machine learning tools in real-world problems.

The following provides a detailed description over the two tasks. To complete these tasks, you are allowed to use any machine learning frameworks including TensorFlow and PyTorch.

Both datasets needed to fulfil the requirements of this assessment can be found in MyAberdeen.

****Please read all the information below carefully****

Task 1: Regression (25 marks) [~ 1000 words]

Data: This data contains the level of a specific cancer antigen and as well as several clinical measures of patients who were due to receive radical treatments. The purpose of this experiment is to explore the relationship between the level of cancer antigen and those clinical measures. The data has 97 subject records and the last column `train' indicates which will be used as training set with label `T' and which will be test set with `F'. The variables are the following:

- logCancerVol: the log cancer volume
- logCancerWeight: the log cancer weight
- age: the age of patients
- logBenighHP: the log value of benign cancer amount
- svi: seminal vesicle invasion
- logCP: the log capsular penetration
- gleasonScore: the Gleason score
- gleasonS45: the percentage Gleason scores 4 or 5
- levelCancerAntigen: the log cancer antigen

Objectives: The main objectives of using this data for clinical purposes could be summarized as follows:

1. Prediction: to predict the level of cancer antigen given these clinical measurements
2. Inference: to infer which clinical factors would potentially influence the level of cancer antigen

In order to achieve these objectives, we would like to accomplish the following subtasks using machine learning.

Subtasks:

1. Data import: Please provide a short description of the data provided and import the data into your programming environment; provide snippets of code for these purposes. (**3 marks**)
2. Data preprocessing: If you did any preprocessing over the data, e.g., normalization, please explain it and the reasons why you did that preprocessing; if you did not do any preprocessing, also please explain it. (**3 marks**)
3. We choose multiple linear regression models for our prediction and inference purposes. Although there are enormous number of regression models available, we choose the following linear models:
    - Least squares estimates
    - Ridge regression
    - Lasso regression

    In this subtask, train each of these models. For ridge regression and Lasso, you should use model selection methods to learn the complexity parameters inherent in these two models. You should produce the test errors of each method. Using the model selection methods to choose the best model for prediction purposes. Using machine learning principles to explain your results with graphs and/or tables. (**10 marks**)

****Please read all the information below carefully****

4. As you have chosen the best model for the prediction previously, in this subtask you need to use the chosen model to infer the clinical measures mostly influencing the cancer antigen. Explain your results indicating why these are the most important clinical measures. (**9 marks**)

Task 2: Classification (25 marks) [~ 1000 words]

In this task, you are given a set of images which contain either a dog or a cat. The aim is to train machine learning classifiers to classify whether an image contains either a dog or a cat. This data was originally taken from the Kaggle competition (https://www.kaggle.com/c/dogs-vs-cats). Both training and test data sets will be made available on MyAberdeen. Note that the training data will be used to train the classifiers and the test data used for evaluations. The class labels are contained in the filenames with the words "cat" and "dog".

To accomplish this task, you are expected to explore a range of machine learning classification algorithms introduced in the module and other materials from literature. You will then investigate which classifier would be recommended as the best for this classification task providing critical comparisons and justifications. As this is a binary classification problem, accuracy is chosen as the error metric to evaluate the performance of the classifier.

In this assignment you are free to choose any classifiers, however, the marking will be reasonably relying on the accuracy on the test data computed by your algorithms. Therefore, you need to report the best mean accuracy value over the test images, which will be ranked against other students.

When working on this assignment, you must analyze and report the points including but not limited to,

- Data preprocessing: What data preprocessing strategies have you applied to the data before applying classification models? Explain why or why not you have made data preprocessing. **(3 marks)**
- You may have to use convolutional neural network (CNN) for this task. Explain why CNN could be the appropriate model for this particular task. **(3 marks)**
- Explicitly demonstrate and justify the training process. For example, you may have to use and explain the early-stopping technique to monitor when to stop training when using deep learning models; you may want to monitor the convergence process by plotting the accuracy against the training iterations. **(9 marks)**
- Compare and report the performance results of those algorithms you have chosen. You may have to use tables and graphs to demonstrate the results. Report the mean accuracy values of the trained classifiers applied to the test data. **(10 marks)**

Useful Information

- Please describe and justify each step that is needed to reproduce your results by using code-snippets, screenshots and plots. When using screenshots or plots generated in Python please make sure they are clearly readable.
- If you use open source code, you must point out where it was obtained from (even if the sources are online tutorials or blogs) and detail any modifications you have made to it in

****Please read all the information below carefully****

your tasks. You should mention this in both your code and report. *Failure to do so will result in zero marks being awarded on related (sub)tasks.*

Marking Criteria

- Quality of the report, including structure, clarity, and brevity.
- Reproducibility. How easy is it for another MSc AI student to repeat your work based on your report and code?
- Quality of your experiments, including design and result presentation (use of figures and tables for better reporting).
- Configured to complete the task and the parameter tuning process (if needed).
- In-depth analysis of the results generated, including critical evaluation, insights into data, and significant conclusions.
- Quality of the source code, including the documentation of the code.

Submission Instructions

You should submit a PDF version of your report along with your code via MyAberdeen by XXX 2020. The name of the PDF file should have the form "CS5062_Assessment1_< your Surname>_<your first name>_<Your Student ID>". For instance, "CS5062_Assessment1_Smith_John_4568985.pdf", where 4568985 is your student ID.

You should submit your code and any associated files along with your report. If you have additional files that you wish to include then these should also be included in your submission.

If you have more than two files to submit, please compress all your files into one "zip" file (other format of compression files will not be accepted). Please try to make your submission files less than 10MB as you may have issues when uploading large files to MyAberdeen.

Any questions pertaining to any aspects of this assessment, please address them to the delivery team Mingjun Zhong (mingjun.zhong@abdn.ac.uk) or Dewei Yi (dewei.yi@abdn.ac.uk).