University of Aberdeen

School of Natural and Computing Sciences

Department of Computing Science

MSc in Artificial Intelligence

2022 – 2023

****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:

- Have knowledge & understanding of the core concepts of, and common practices, in Machine Learning.
- Have knowledge and understanding of fundamentals of machine learning, including a range of popular machine learning algorithms.
- Be able to use existing machine learning tools, frameworks, and libraries to build solutions for real-world or benchmark problem solving.
- Be able to perform data pre-processing for machine learning.
- Be able to systematically evaluate the built machine learning solutions.
- Be able to critically examine the strengths and limitations of common machine learning algorithms when solving a specific problem.
- Be able to write reports for machine learning solutions.

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 applying a machine learning model to analysing a data set,

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

model hyper-parameters may impact the performance of the model and so it is very important to select these hyper-parameters. 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.

Task 1: Regression (27 marks) [~ 1000 words]

Data: This data contains percentage of body fat, age, weight, height, and ten body circumference measurements of 252 men. The purpose of this experiment is to explore if it would be possible to fit body fat to other measurements using multiple regression, which could provide a convenient way of estimating body fat for men using only a scale and a measuring tape. This is described in the document provided.

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

1. Prediction: to predict the body fat using other measurements
2. Inference: to infer which measurements would impact body fat of men

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 or standardisation, please explain it and the reasons why you did that preprocessing; if you did not do any preprocessing, also please explain why preprocessing was not necessary. (**5 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 to use Lasso regression. Remember Lasso regression is to optimize the following problem:

   $$\min_{\beta} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

   where λ is a hyper-parameter for Lasso, and is called complexity parameter.

   When using Lasso regression model, the complexity parameter λ would influence the performance greatly. Therefore, it is important to choose the best complexity parameter. In our experiments, we want to choose the best λ from $0.00, 0.01, 0.02, \cdots, 0.98, 0.99, 1.0$. Cross validation (CV) should be used for choosing λ. Write a Python programme to implement cross validation for choosing the best complexity parameter (**8 marks**); explain how CV was used (**4 marks**) and discuss the results how λ was chosen (**4 marks**).

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

After the best complexity parameter was chosen, apply the Lasso model to the data for regression purpose. What would be the measurements which could impact body fat of men? Explain your conclusions. (**3 marks**)

Task 2: Image classification (23 marks) **[~ 1000 words]**

In this task, you are given a set of blood cell images which contain four different cell types. The aim is to train a machine learning classifier to classify cell types using blood cell images. 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 folder names indicate class labels which are Eosinophil, Lymphocyte, Monocyte, and Neutrophil.

In this assignment you will use convolutional neural network for classification. You will report the classification results based on the test data.

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)**
- Report the model architecture that you have used for your model and explain how this architecture was chosen (**3 marks**). Explain why CNN could be the appropriate model for this particular task. **(3 marks)**
- Explicitly demonstrate and justify the training process. You may have to use early-stopping for training your model. Explain how early-stopping was used (**5 marks**) and what is the purpose of using early-stopping (**4 marks**)?
- Report your classification results against the test data. You may have to use tables and graphs to demonstrate the results of some accuracy metrics. **(5 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 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.*

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

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 October 31, 2022. 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).