Aston University
BIRMINGHAM UK

College of Engineering & Physical Sciences
Assignment Brief

CS4730 Machine Learning	Coursework 1: Supervised Learning
Dr Harry Goldingay h.j.goldingay1@aston.ac.uk	

Assignment Brief/ Coursework Content:

In this assessed task, you will be applying algorithms from the first family of machine learning techniques covered in this module: supervised learning. The aim of this task is to test your ability to apply machine learning algorithms to well-specified tasks and to evaluate the performance of these algorithms and to use this evaluation to improve performance.

Follow the instructions below to complete the portfolio task. The task requires you to carry out some implementation in Python and to provide a short written justification of your choices, of maximum 250 words. The required format for submission is a Jupyter notebook, integrating your code and written justification.

Sub-task 1:

Download the file task1.csv from Blackboard. It contains 150 rows of data with a single row of headers at the top. The values in the three columns are independent variable (`x1`, `x2` and `x3`) values. The values in the fourth column are the corresponding dependent variable (`y`) values. The problem that the dataset represents is one of regression.

You believe that applying linear least squares regression with L2 regularisation would be a good way of solving the problem. You want to know how well this approach would perform on unseen data in terms of its coefficient of determination (the R^2 score). Using your knowledge of machine learning and of Python, estimate the performance of this approach.

Note that scikit-learn implements linear least squares regression with L2 regularisation in the class `sklearn.linear_model.Ridge`.

Sub-task 2:

Download the file task2.csv from Blackboard. This dataset contains 1000 rows of three values each with a single row of headers at the top. The values in the first and second column are independent variables and the values in the third column are the corresponding dependent variable values. The problem one is classification and the dependent variable values represent the class (either 0 or 1).

As a first step, you want to decide whether this dataset would be suitable for classification using a linear model. Plot a scatter graph of the data with the two independent variables on the x and y axes and colours of the markers representing the class values. State whether you think that the dataset is suitable for classification with a linear model, justifying your answer.

You decide to use a multi-layer perceptron (MLP) with a single hidden layer of 100 nodes (the default value within scikit learn) to model the data. To implement this, it is recommended that you use scikit-learn's `MLPClassifier`. Unless you have a good reason to do otherwise, use its default parameter values (except for alpha – see below – and, potentially, `max_iter` if your training process is not converging).

One of the parameters we can set in an MLP is the strength of the L2 regularisation term: alpha. Changing the value of this term can affect how well our model generalises to new data. Based on the principles discussed in the module, design a methodology to:

- choose an appropriate value for `alpha` applied to this classification problem,
- train an MLP with this value of `alpha` and estimate its generalisation error.

Compare the performance of the alpha value you found with a very high value (e.g. 20). Is there a difference in performance between the two and, if so, how would you explain it? It may help you to plot a scatter graph for both alpha values, similar to the one you plotted at the start of this graph, visualising the predictions of your two trained models.

Descriptive details of Assignment:

- Preferred Format: Jupyter Notebook
- Word Count: 250 words (code does not count towards word limit)
- Preferred reference style: Harvard referencing

Recommended reading/ online sources:

- Units 1-5 of CS4730

Key Dates:

Any key dates regarding the coursework. For example:

25/10/2022	Coursework set
08/11/2022	Submission date
06/12/2022	Expected feedback return date.

Submission Details:

- Submit your file, either as a plain Jupyter notebook file or as a zip file containing one or more Jupyter notebooks and data files, through the link on Blackboard.

Marking Rubric:

The mark scheme for the task is as follows:

- **50-59** Solution approaches have been applied to both sub-tasks and, where requested in the task, their performance measured. The approaches taken are broadly correct but may have some flaws in application or methodology. Model evaluation and a justification of chosen approaches have been attempted but shows limited understanding.
- **60-69** The approach taken in sub-task 1 shows understanding of how to estimate model generalisation ability. Multiple solution approaches (algorithms/models/parameter sets) have been applied to the problem in sub-task 2 and have undergone evaluation. Justification for the selected approach is evidence-based and well presented.
- **70-79** The methodology used to compare solution approaches for sub-task 2 is carefully designed and leads to well-supported conclusions. The solution and supporting text show clear understanding of the models used and their properties.
- **80+** As above, but with additional evidence (for both sub-tasks) of some or all of: attention to quality throughout the implementation, thorough understanding in experimental design, excellent justification.

No specific descriptors are provided for marks below the threshold of 50. Marks in the range **0-49** are allocated where the submitted work has not reached the expectation for the threshold descriptor.