Aston University
BIRMINGHAM UK

College of Engineering & Physical Sciences
Assignment Brief

CS4730 Machine Learning AM41ML Statistical Machine Learning	Coursework 1: Supervised Learning
Dr Harry Goldingay h.j.goldingay1@aston.ac.uk	

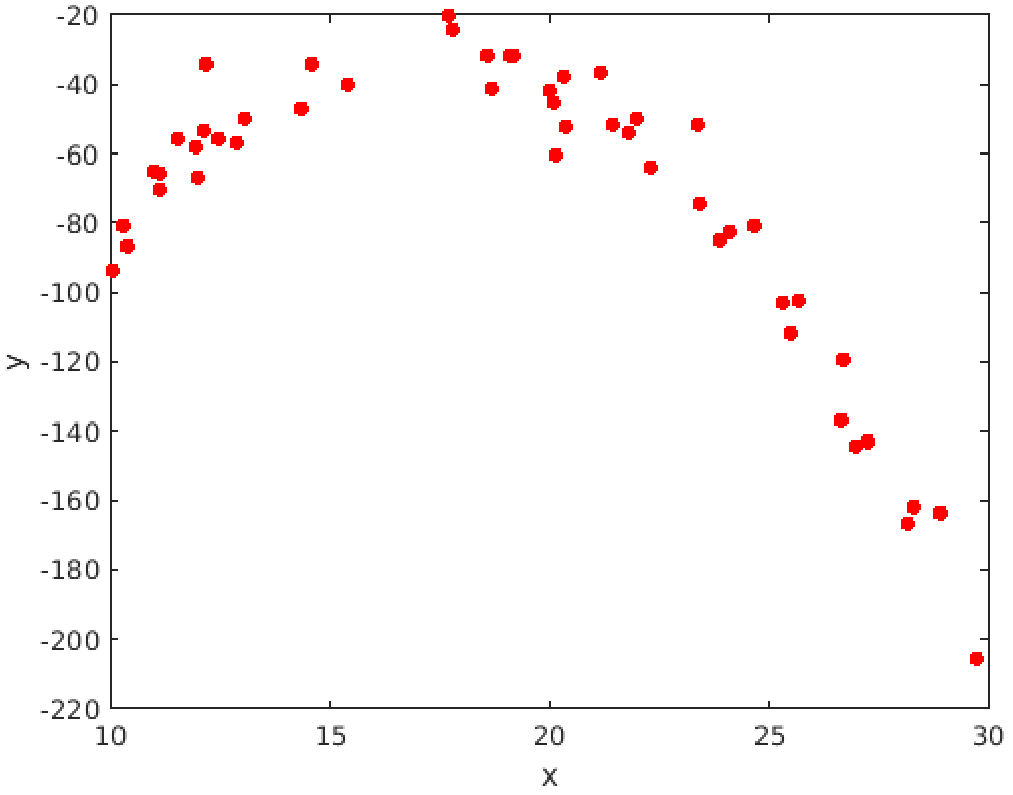
Assignment Brief/ Coursework Content:

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 no more than 250 words for each sub-task. The required format for submission is a Jupyter notebook, integrating your code and written justification.

Sub-task 1:

A colleague of yours is trying to solve a regression problem of one independent variable (x) and dependent variable (y). They aim to find a model which performs well on unseen data in terms of Mean Squared Error (MSE). They have decided on an approach to solve the problem and want your feedback on it. They have a dataset containing 50 data points. For context, they have provided you with the following graph of their data.



Their proposed approach is, to:

- Fit a linear model to the data by training it on their 50 data points.
- Estimate the generalisation performance of their trained model by calculating its MSE on what they are calling a "test set" containing 20 data points, drawn at random from the original 50.

They have asked for your feedback on their approach. State your opinion of your colleague's approach and, if applicable, make brief suggestions to improve it.

Sub-task 2:
Download the file `classification.csv` from Blackboard. This dataset represents a binary classification problem. The dataset contains 250 samples, each with 100 independent feature values. The final column in the dataset contains the dependent variable values (0 or 1).

Using the techniques introduced in units 1-6 of the module, design and test a variety of classifiers for this problem with the ultimate aim of finding a good classifier and estimating its generalisation performance.

Test two distinct classification algorithms on the problem, stating why you believe that they may be appropriate for the problem. For one of these approaches, you should try and parameterise it to try and find a good balance between bias and variance. You may also address the dimensionality of the dataset if you believe that it is necessary.

State, with evidence, which of the approaches you have taken (combination of classifier type, hyperparameters, method of addressing dimensionality if applicable) you believe would generate best to unseen data

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 1-6 of AM41ML/CS4730

Key Dates:

Any key dates regarding the coursework. For example:

28/10/2024	Coursework set
11/11/2024	Submission date
09/12/2024	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 through the “Assessment Submission” tab.

Marking Rubric:

The mark scheme for the task is as follows:

- **0-39** Brief, irrelevant, confused, incomplete. Does not come close to meeting the required learning outcomes.
- **40-49** Evidence that some learning outcomes have been achieved or most learning outcomes achieved partially. Although work may include brief signs of comprehension, it contains basic misunderstandings or misinterpretations, demonstrates limited ability to meet the requirements of the assessment.
- **50-59** In sub-task 1 the answer shows good awareness of core concepts, but may include small errors or omissions. Solution approaches have been applied to the second sub-task 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 show limited understanding.
- **60-69** In sub-task 1, the answer correctly identifies all errors in the proposed approach. The approach taken in sub-task 2 shows understanding of how to estimate model generalisation ability. Justification for the selected approach is evidence-based and well presented.
- **70-79** The discussion in sub-task 1 shows clear understanding of how to resolve errors in the proposed approach. Experimentation for sub-task 2 is comprehensive and well designed to lead to a robust conclusion.
- **80+** As above, but with additional evidence of some or all of: attention to quality in the implementation, thorough understanding in experimental design, excellent justification.