Aston University
BIRMINGHAM UK

College of Engineering & Physical Sciences
Assignment Brief

CS4730 Machine Learning	Coursework 2
Dr Harry Goldingay h.j.goldingay1@aston.ac.uk Dr Mohammed Hadi m.hadi2@aston.ac.uk	

Assignment Brief/ Coursework Content:

This assessment consists of four sub-tasks. In the first three sub-tasks, you you will be applying techniques from the unsupervised learning, reinforcement learning and dimensionality reduction topics covered in this module. The aim of these tasks is to test your ability to apply these machine learning techniques to well-specified tasks and, where applicable, to evaluate their performance. These first three sub-tasks are each worth 10% of the overall module mark.

In the final sub-task, you will be required to solve a more complex task, described in natural language. You will be required to reformulate the problem and design a solution, justifying your design choices in terms of the properties of the problem and of the algorithms you use. This sub-task is worth 50% of the module mark.

Overall, this assessment is worth 80% of the overall module mark.

Follow the instructions below to complete the sub-tasks. The sub-tasks require you to carry out some implementation in Python and to provide a short written justification of your choices. Across the 4 sub-tasks, your justification should be of no more than 1500 words, excluding code. We recommend that you aim to use no more than a total of 500 words of justification for the first 3 sub-tasks, with the remainder used for sub-task 4.

For sub-tasks 1, 3 and 4 you should produce a Jupyter notebook file integrating your code and written justification. For sub-task 2, you should produce a document file (e.g. docx, pdf) containing your working and justification. The required format for submission is a zip file containing the solution files for all four sub-tasks.

Sub-task 1: Unsupervised Learning

Download the file `cluster2.csv` from Blackboard. It contains 1000 data points. Each data point has two features. You do not know anything about how many clusters to expect in this data.

You decide to take a model-based approach to finding clusters in this data set. You may choose between applying the k-means algorithm or fitting a GMM to the data. In either case, you will need to choose an appropriate number of clusters for your model.

Based on the principles discussed in lectures, design a methodology to choose an appropriate number of clusters for your chosen algorithm. Implement your methodology using Python. Explain your methodology and justify your choice of the number of clusters. You may want to produce graphs to help support some aspects of this justification.

Sub-task 2: Reinforcement Learning

Follow the instructions below to complete this task which requires you to carry out some calculations and to provide short written justifications of your choices while observing not to go over the word count limit. You are not expected to write any code to support your answers. In contrast to the other tasks, you are expected to submit your answers as either a Word document or a PDF.

The below grid is a variant of the grid world problem. It has three inaccessible states (the black-shaded squares) and four terminal states (the grey-shaded squares).

In each state, the agent:

- receives a reward of -0.2 in a non-terminal state or of the value indicated below if in a terminal state,
- ends the game if it is in a terminal state,
- otherwise, it must choose to try and move to one of the neighbouring states (the horizontally or vertically adjacent states. Diagonal movement is not permitted).

After attempting to move, the agent:

- reaches the state it was attempting to move to with a probability of 0.8,
- fails and makes a perpendicular move with a probability of 0.2 (each direction is equally likely),
- if, as a result of this, the agent attempts to move outside of the grid or to an inaccessible state, it instead remains where it is.

In the diagram below, the number in each state shows the (expected) utility of that state under some policy, rounded to two decimal places. Using this information, draw a diagram to show the policy used to derive these utility values. For each of the three states highlighted in green, show how you determined the policy action for that state.

6.52	6.80	7.08	7.33	7.58
6.30	6.52			7.86
6.02	5.82	-5.00	-5.00	8.11
5.74	5.46	4.30	6.30	10.00
5.46	5.24		-10.00	7.56

Sub-task 3: Dimensionality Reduction

Download the file `energydata_complete.csv`, available on Blackboard based (adapted from the Bank Marketing dataset from Candanedo [1] hosted in the UCI Machine Learning Repository) Download this file from Blackboard. The dataset contains approximately 20000 data points, each of which contains on a snapshot of data related to environmental conditions and appliance energy consumption in a 10 minute period. The online documentation contains fuller descriptions of the variables.

Use a dimensionality reduction method of your choice to visualise the data in three dimensions. Evaluate the success of your dimensionality reduction procedure to decide whether the three-dimensional projection of your dataset captures the important details of the original dataset. Justify your answer and, if you don't feel that a three-dimensional projection captures these details, design and implement a methodology to determine the appropriate minimum dimensionality to project the data to.

Sub-task 4

You have been contracted by a bank, who are interested in whether machine learning could be used to help them to make their marketing campaigns more efficient. They have run a campaign in which they contacted a large number of potential customers to try to sell them a given banking product, but have noted that their success rate – the proportion of customers who subscribe to the product having been contacted is low. They want to focus their effort on those most likely to subscribe. As a starting point they have asked you to create a proof-of-concept system which, given some data about a person, can predict whether or not they will subscribe to a given banking product.

They have provided you with a dataset (bank-full.csv, available on Blackboard and adapted from the Bank Marketing dataset from Moro et al. [2] hosted in the UCI Machine Learning Repository) which contains information about approximately 41k contacts they have had with potential customers. The last column ('y') contains the target variable – whether or not the contact subscribed to the product – and the others (columns one to twenty) contain other potentially relevant features which you can find a fuller description of here: https://archive.ics.uci.edu/dataset/222/bank+marketing.

From your client's perspective, a good solution to this problem would accurately predict whether or not a new marketing lead (i.e. a person in the dataset) would subscribe to the banking product if contacted as part of this campaign. It is also important to them that the solution performs well for both types of leads: those who both would sign up if contacted and those who wouldn't. As part of your work, they would like to know how well they could expect your solution to perform on both counts.

Use your knowledge of machine learning and of Python, supported by personal research where necessary, to design and implement a solution to the problem described above and to answer your client's question about performance. Note that, as described in the marking scheme, you will be marked primarily on your approach to the task and your understanding as evident from your written justification rather than the final performance of your solution (although a high performing solution may be evidence that your approach is a good one). As such, make sure to include more than just your final solution in your submission. Also include information on solution methodologies that you tried and rejected.

Note also that the content of the module is sufficient for you to complete this task well. However, as you will see from the mark scheme, marks in the higher ranges are characterised by use of citations and of independent research. To get you started, the following references contain overviews of topics which may be useful to you when preparing your solution. They are much too broad for you to implement all of their suggestions in your work, but if you find any part of them interesting or relevant then they, and the references within them, could give you a starting point for your own literature search:

- He, H. and Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), pp.1263-1284.
- Hutter, F., Kotthoff, L. and Vanschoren, J., 2019. Automated machine learning: methods, systems, challenges. *Springer Nature*.
- Dong, X., Yu, Z., Cao, W., Shi, Y. and Ma, Q., 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14, pp.241-258.

References:

1. Candanedo, Luis. (2017). Appliances Energy Prediction. UCI Machine Learning Repository. https://doi.org/10.24432/C5VC8G.
2. Moro, S., Rita, P., and Cortez, P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

Descriptive details of Assignment:

- Preferred Format: Jupyter Notebook
- Word Count: 1500 words (code does not count towards word limit)
- Preferred reference style: Harvard referencing

Recommended reading/ online sources:

- All units of CS4730
- Articles referenced and linked to in sub-task 4.

Key Dates:

26/03/2024	Coursework set
06/05/2024	Submission date
03/06/2024	Expected feedback return date.

Submission Details:

- As discussed at the start of the document, for sub-tasks 1, 3 and 4 you should produce a Jupyter notebook file integrating your code and written justification. For sub-task 2, you should produce a document file (e.g. docx, pdf) containing your working and justification. Submit your work as a zip file containing the solutions to each of the sub-tasks through the link on Blackboard.

Marking Rubric:

Sub-tasks 1-3 (10% of module mark each):

- **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%** The applied element of the task has been completed in a broadly correct fashion but may have some flaws in application or methodology. Written answers and justification have been attempted but are largely descriptive and show some limitations in understanding.
- **60-69%** The applied element of the task has been completed correctly. The written answers and justification show clear understanding of any models used and of the implications of the results of the applied element.
- **70-79%** The approach to the applied element of the task has been carefully designed or chosen to produce the evidence needed for the written element.
- **80%+** As above, but with additional evidence of some or all of: attention to quality throughout the implementation, thorough understanding in experimental design, excellent justification.

Sub-task 4 (50% of module mark):

- **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%** The given problem has been reformulated as a machine learning problem and a solution has been proposed, implemented, and evaluated. Some attempt has been made to pre-process the data appropriately. The

solution has some value, but the quality may be compromised by a range of factors including misunderstanding how to best approach the problem, flaws in implementation or in evaluation methodology.

- **60-69%** Appropriate machine learning methods have been used to address the given problem, including ensuring good performance for both fraud and non-fraud cases. The written justification shows a clear understanding of why the methods employed are appropriate given their properties and those of the problem. Experimental work to empirically support this justification has been undertaken.
- **70-79%** The approach for choosing solution methodologies is systematic. An appropriate range of options has been considered and critically analysed for suitability, both in terms of their properties and, where appropriate, through experimental comparison. This has led to a well-designed solution to the problem. The written justification documents the rationale for all choices with supporting evidence, including relevant references.
- **80%+** The chosen approach and written justification show insight into the problem. The work makes use of information from the student's independent research and draws on academic work outside of the texts discussed in the module.