

Aston University Machine Learning

Portfolio Task 5

Released: 14/12/20

Due: 22/01/21, before 23:59

Instructions:

In this assessed task you will be given a problem, 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 used. The task will be worth 40% of the portfolio mark.

Details:

Follow the instructions below to complete the portfolio tasks. The task requires you to carry out some implementation in Python and to provide written justifications of your choices (maximum 1000 words). The recommended format for submission would be a Jupyter notebook, integrating your code and written justification, but a zip file containing Python source code and written justification in either document or pdf format is also acceptable.

Marking:

The mark scheme for these tasks is as follows:

- **50-59** The given problem has been reformulated as a machine learning problem and a solution has been proposed, implemented and evaluated. The solution has some value, but 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 to ensure good performance for both fraudulent and non-fraudulent transactions. The written justification shows clear understanding of why these methods 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 with 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.

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.

Portfolio Task 5 (40% of module mark)

You have been contracted by a bank, who are interested in whether machine learning could be used to help them to detect fraudulent payments. As a starting point they have asked you to create a proof-of-concept system which, given some data about a payment, can predict whether or not it is fraudulent.

They have provided you with a dataset (bs140513_032310.csv, available on [kaggle](#)) which contains information about approximately 600k payments. The first nine columns contain details of the transactions including the customer's ID and the amount and category of payment. The tenth column tells you whether or not the payment was fraudulent.

From your client's perspective, a good solution to this problem would accurately predict whether or not a new payment (i.e. one not in the dataset) was fraudulent with reasonable accuracy. It is also important to them that the solution performs well for both genuine and fraudulent transactions. 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 write-up. Also include information on solution methodologies that you tried and rejected.