CITY UNIVERSITY LONDON

EST 1894

IN3062: Introduction to Artificial Intelligence
Coursework

Submission deadline: Wednesday 29th December 2021, 5pm

Introduction

On completing this coursework, you should be able to:

1. Describe a machine learning problem and apply artificial intelligence techniques to that problem.

2. Describe the systematic application of your chosen artificial intelligence methodology to the chosen problem (for example, data preparation, parameter tuning).

3. Apply, compare, contrast and critically evaluate at least **two** ways of analysing your problem data.

This coursework builds on the material covered in the tutorials and lecture. Python should be used for all implementations. Deliverables are:

- a written report of your work (max 8 pages)
- a written reflection of your individual contribution to the work (max 1 page)
- your practical implementation (code)

Module mark

This coursework is worth 100% of the mark for IN3062: Introduction to Artificial Intelligence.

Teamwork

This coursework should be completed in **groups of two or three**.

- You can form your own groups and you must email Jacob Howe (j.m.howe@city.ac.uk) your teams by 5pm Monday 8th November.
- Anyone not forming a team will be assigned to a team.
- If you would like help forming a team, please let Jacob Howe know as soon as possible.
- Students on the Professional Pathway route may, if they choose, work individually.
- Any student studying overseas because of covid may, if they choose, work individually.
- All members of the team will receive the same mark for the eight page main report, and the one page reflection will moderate this mark. Whilst this is a group project, all team members are expected to contribute to all parts of the work: both the coding and the report.

Time Allocation

This 15 credit module represents 150 study hours. 30 of these hours are on your timetable. The remaining time is for your self-direct study and for completing this assignment. Whilst different people might have different views on the balance between what is study for learning the material and what is time spent on the assignment, this coursework is designed to take *40-50 hours of study per team member*.

Submission

Submission is through Moodle, and no other method of submission will be accepted. You should submit three files:

- your report in pdf format, single column, standard margins, font Arial 11, maximum 8 pages (see below for details), including all figures
- your one page reflection (again, in pdf format, single column, font Arial 11)
- a single zip file (or single python file), including all files needed to run your code. Either submit the dataset as part of the zip file or include an http link in the report stating where the dataset can be downloaded from.

Do not zip up, or otherwise compress, your report.

In addition:

- your code must be developed and available on github, with a full revision history indicating who has pushed what code. A single push at the end of the work is liable to score 0. This repository should be available to Jacob Howe, Daniel Chicharro and Atif Riaz.

Late submissions will score 0. You can upload work to Moodle more than once, so there is no need for last minute submission. Don't leave final submission to the last minute.

Feedback

In the labs and surgeries, we can check your progress and give formative feedback. Evaluative feedback and marks on your coursework will be given out once all assessment for IN3062/INM701 is complete and the work marked.

Coding

- You should use Spyder or Jupyter Notebooks for your IDE.
- You should work with the libraries studied in the module. For neural networks: *Tensorflow/Keras*, for other techniques *scikit-learn*. You might possibly use other libraries to stretch your work, but those studied should form the core of your work.
- Your code should be well commented.
- As noted above, your work should be maintained on github, with a full revision history indicating who has pushed what code.
- It is possible that some of your experiments may take significant computation time, so it is in your interest to start running experiments for this coursework as early as possible.

The Task

In this coursework, you are expected to *demonstrate what you have learned in the module* by applying artificial intelligence techniques as covered in the module to a dataset and domain of your choice. This will include some or all of:

- Define the domain and dataset(s) (you are free to choose the domain and the dataset that you want to investigate).

- Define questions and analysis tasks (a brief overview of the domain, analytical questions that are being asked, a list of your objectives and the expected output(s) of your analysis).

- Perform an initial investigation of the dataset and the characteristics of the data. Develop a plan as to how you might transform the data to make it useable.

- Develop a plan as to which artificial intelligence techniques you might use and what sorts of potential observations these can lead to, and how you will evaluate these.

- Split your dataset (train/validate/test, some datasets come pre-split). If you have a holdout test set then you most likely don't want to use this until the near the end of your work.

- Perform the analysis. Get the data ready for analysis, carry out your analysis/modelling as needed, validate your results and communicate observations, iterating through this process. Analytical operations can include data processing to an extent that is needed (not all datasets are messy) to prepare a useful and robust dataset to work within, and data derivation (such as feature engineering).

- You might establish a baseline result first, computing metrics on training and validation sets, analyse errors, work on succeeding iterations, and alternative models. (If initial metrics are amazing and there are no errors is the problem too easy?)

- Generally, be close to your data (visualise the dataset, collect summary statistics, look at errors, analyse how different parameters affect performance, try out different model variants).

Report

Your final report should be a maximum of 8 pages from the start of the Introduction to the end of the Conclusion. You may in addition have a title page, and as much space for a references section as needed.

Your final report should cover each of the aspects above (and any other element of your work that you believe should be reported). Graphical illustration of your results is expected (perhaps training/testing error curves, confusion matrices, algorithm outputs, etc), as well as results. Following the above analytical process, make sure that in your report you answer the following questions (where appropriate):

- What is your dataset, problem domain?

- Is your model classification or regression?
- Did you have any missing, corrupt or misleading data? If so, how did you cope it?
- Have you omitted some data? If so, why?
- Did you apply techniques to understand your dataset?
- What models did you use?
- How did you encode the input variables?
- What are the criteria for selecting model performance evaluation tools?
- What were your outputs?
- Did you have any problems or difficulties working with the dataset?

You should present the results clearly and concisely and provide a discussion of the results, with conclusions related to problem being addressed. The conclusions section might propose some further work based on the results of this coursework.

I hope that you will have a lot of work to report, maybe more than you can fit into the page limit. In this case you will need to display good editorial judgement as to what to report: what was most important, what was most interesting.

Reflection

The most important point to be addressed in the reflection is who did what. However, this is also a chance to report further conclusions and discussion of your work from your individual perspective.

Datasets

You are free to choose the domain and the dataset that you want to investigate. Here are some suggestions and sources for datasets.

You ***cannot*** use the datasets that come with scikit-learn, or others used in the exercises. Other heavily used datasets make bad choices.

Some of these sources will come with code. Whilst you may use this code if referenced, you get little credit for this.

General:

- Kaggle is Google's online data community, and contains thousands of datasets.
- UCI Machine Learning Repository. The University of California, Irvine has a collection of several hundred datasets (some of them a little small).

Some other possible sources of data:

Images: Labelme, ImageNet, LSUN, Google's Open Images, COCO
Text: Project Gutenberg
Clinical: MIMIC, The World Health Organisation (https://www.who.int)
General: FiveThirtyEight (https://data.fivethirtyeight.com)

There are many, many other sources of data available. At the cost of some effort, you might also collect or create your own dataset.

Note: You are not necessarily being marked on how good the results are. What matters is that you try something sensible and clearly describe the problem, method, what you did, and what the results were. Don't pick a dataset that is way too hard for your experiments. Don't pick a dataset that is too straightforward (too small) to produce interesting results. Be careful not to do foolish things like data snooping, testing on your training data, including plots with unlabelled axes, using undefined symbols in equations. Do sensible cross-checks like running your models several times, varying your random seed, leaving out small parts of your data, adding a few noisy points, etc. to make sure everything still works reasonably well. If you pick something you think is interesting it will make the process of getting it to work more enjoyable.

Coding & Referencing

This is, in large part, a coding assignment. If you use code (or other materials) written by someone else, you should ***cite*** that code (or other material) in Harvard format. If you do not cite work appropriately you will have committed academic misconduct. Making superficial changes to the code does not make it yours. You are also expected to make a coding contribution, so if you use a large amount of code written by someone else, and cite it appropriately, your contribution will be low and your work marked accordingly.

Grading

Your work will be graded in accordance the University's assessment criteria (see Appendix A), and you will receive an overall mark. An indication of the relative importance of aspects of your work are given below (these are not section marks):

- *Report, introduction*: description and motivation of the problem, description of the dataset including data types (e.g. discrete, continuous) (15%)
- *Report, methodology*: summary of the models used, with their pros and cons, a hypothesis statement, description of choice of training and evaluation methodology (20%)
- *Report, results*: description and presentation of the output. The code acts as an appendix to this section, and code quality (e.g. commenting) contributes. (35%)
- *Report, evaluation*: analysis and critical evaluation of results. (10%)
- *Report, conclusions and referencing*: lessons learned, references (using Harvard format) and future work. (10%)
- *Reflection* (10%)

Extenuating Circumstances

If you are not able to submit your coursework on time for unforeseen medical reasons or personal reasons beyond your control you should contact the Programmes Office asap and fill an Extenuating Circumstances form. Strong evidence in the form of, for instance, medical certificates or legal statements will have to be produced.

Academic Misconduct/Plagiarism

If you copy the work of others (either that of another team or of a third party), with or without their permission, you will score no marks and further disciplinary action will be taken against you. The same applies if you allow others to copy your work. This is a group responsibility.

Appendix A: University Grade-related criteria (Undergraduate)

>85 First class: Outstanding. Work that demonstrates a comprehensive knowledge of the subject area and addresses the learning outcomes/assessment criteria in full. Where relevant, it will show evidence of independent reading, thinking and analysis. It will be well-constructed and demonstrate a professional approach to academic practice. It will be of a professional standard.

70-84 First class: Very good. Work that demonstrates strong knowledge of the subject area and addresses the learning outcomes/assessment criteria well. Where relevant, it will show evidence of wide and comprehensive reading. It will be clearly written and adhere to the principles of good academic practice.

60-69 2:1: Good. Work that demonstrates a sound level of knowledge of the subject area and makes a good attempt to address the learning outcomes/assessment criteria, realising all to some extent and most well. Where relevant, there will be evidence of thorough research of the topic(s). It will be well-structured and logically written and will demonstrate good academic practice.

50-59 2:2: Fair. Work that demonstrates knowledge of the subject area and attempts to address the learning outcomes/assessment criteria, realising all to some extent and some well but perhaps also including irrelevant or underdeveloped material. Where relevant, answers will provide some evidence of analysis but may be largely descriptive. It will have structure but this may not always be clear. Attempts to demonstrate academic practice will be evident.

40-49 Third class: Satisfactory. Work that demonstrates basic knowledge of the subject area and provides some level of response to the learning outcomes/assessment criteria but only realises these outcomes and criteria to some extent and may not include important elements or information that is completely accurate. Where relevant, development of ideas is limited. Expression and structure will lack clarity and evidence of academic practice will be limited.

34-39 Fail: Poor. Unsatisfactory work that demonstrates very limited knowledge of the subject area and does not succeed in grasping the key issues. Learning outcomes/assessment criteria will not be realised. There will be no real development of ideas and few sources will be used or used correctly. Presentation is confused or lacking in clarity.

<34 Fail: Very Poor. Work that demonstrates no real knowledge of the subject area and which demonstrates a totally inadequate attempt to address the learning outcomes/assessment criteria.

Appendix B: Mapping of Assessment to Programme Learning Outcomes

This assignment assesses the following programme (degree) learning outcomes.

Knowledge and understanding:

- ✓ Use and explain the core concepts and theories of computer science and computer applications
- ✓ Discuss scientific and engineering practice and theory in computing and extend your knowledge through self-led study
- ✓ Discuss management issues concerning the planning, design and delivery of computer-based systems
- ✓ Identify and model requirements for specialised computing systems and propose and evaluate solutions to fulfil them
- ✓ Use appropriate theories, practices and tools for the specification, design,
- ✓ Implementation and evaluation of computer-based systems
- ✓ Explain the concepts of computer programming and critically evaluate and predict their utility in models, tools and applications
- ✓ Demonstrate advanced, specialist theoretical and practical knowledge in a range of computer science sub-fields

Skills:

- ✓ Analyse, develop and select algorithms for computational tasks
- ✓ Analyse and solve problems based on theoretical considerations
- ✓ Analyse and abstract problems and propose and apply effective solutions
- ✓ Synthesise information from disparate sources to compose systems and documents
- ✓ Apply controlled compromise in meeting requirements
- ✓ Apply techniques and tools for modelling and managing information
- ✓ Design and execute methodologically sound scientific and engineering studies
- ✓ Plan work
- ✓ Manage personal time
- ✓ Present and communicate complex ideas
- ✓ Apply sound research methods
- ✓ Understand, evaluate, synthesise and apply complex ideas

Values and attitudes:

- ✓ Assess the nature of intellectual property and its ownership, and respect it accordingly