CITY UNIVERSITY LONDON

EST 1894

INM701: Introduction to Artificial Intelligence
Coursework

Submission deadline: Sunday 20th December 2020, 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.

4. Give an individual presentation of your work and your findings.

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 9 pages)
- a written reflection of your individual contribution to the work (max 1 page)
- your practical implementation (code)
- an individual presentation (12 minutes including questions) given in January

Module mark

This coursework is worth 100% of the mark for INM701: Introduction to Artificial Intelligence.

Teamwork

This coursework should be completed either in **groups of two** or **individually**; I am suggesting that you work in pairs to help you better understand the concepts and learn the skills. You can form your own pairs and you must email me (j.m.howe@city.ac.uk) your teams by 5pm Monday 2nd November. All members of the team will receive the same mark for the nine 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.

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 9 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.

In addition:

- your code must be developed and available on a git server (github or otherwise), with a full revision history indicating who has pushed what code. This repository should be available to Jacob Howe 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.

Presentation

You will be giving an individual presentation, of a maximum of 12 minutes, including time for questions. This will be in the University's examination period and is provisionally scheduled for Monday January 11th.

Feedback

In the labs and surgeries, we can check your progress and give formative feedback. Evaluative feedback and marks on your coursework will be give out after the presentations.

Coding

- You should use Spyder or Jupyter Notebooks for your IDE.
- As noted above, your work should be maintained on a git server (github or otherwise), 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 viable plan: which data processing steps you will need to perform, how you will transform the data to make it useable, which artificial intelligence techniques you can potentially use and what sorts of potential observations these can lead to.

- 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).

- 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.

- 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?)

- 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 9 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). Graphically 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 data? If so, how did you cope it?
- Did you do apply techniques to understand your dataset?
- Have you omitted some data? If so, why?
- 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. You might also report further conclusions and discussion of your work from your individual perspective in this section.

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.

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 other possible sources of data:

Images: Labelme, ImageNet, LSUN, Google's Open Images, COCO
Text: Project Gutenberg
Clinical: MIMIC.

There are many, many other sources of data available. At the cost of some effort, you might even 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, 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). 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), and you will receive an overall mark. An indication of the relative importance of aspect 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) (10%)
- *Report, methodology*: summary of the models used, with their pros and cons, a hypothesis statement, description of choice of training and evaluation methodology (15%)
- *Report, results*: description and presentation of the output. The code acts as an appendix to this section, and code quality (e.g. commenting) contributes. (25%)
- *Report, evaluation*: analysis and critical evaluation of results. (10%)
- *Report, conclusions and referencing*: lessons learned, references (using Harvard format) and future work. (5%)
- *Reflection* (10%)
- *Presentation* (30%)

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.

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.

Appendix: University Grade-related criteria (Postgraduate)

>80 Distinction: Excellent or 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 and strong critical ability. It will be well-constructed and demonstrate a professional approach to academic practice. It will be of a professional standard.

70-79 Distinction: 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 and critical ability. It will be clearly written and adhere to the principles of good academic practice**.**

60-69 Merit: 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 some well. There will be evidence of thorough research of the topic(s) but some answers may not be complete or arguments sufficiently explored. It will be well-structured and logically written and will demonstrate good academic practice. Some critical ability will be evident.

50-59 Pass: Satisfactory. Work that demonstrates 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 fully accurate. Where relevant, development of ideas is limited but attempts will be made to analyse materials critically. Expression and structure may lack clarity and evidence of academic practice will be limited.

40-49 Fail: Poor. Unsatisfactory work that demonstrates very limited knowledge of the subject area and which 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 critical analysis will be very limited. Presentation is confused or lacks coherence.

<40 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. No critical ability will be displayed.