Coursework

Question 1 (25 pts)

The idea for this question is to cluster languages using linguistic features. For the question, you will need data from the World Atlas of Language Structures (WALS), available at

`https://wals.info/`

Download the dataset, select approximately 20 linguistic features, and create a matrix of languages against these features [m1].

Once you have selected and prepared your features [m2], you should define a similarity or distance metric between pairs of languages. Specifically, I suggest using a simple metric based on the count of identical feature values: the similarity between two languages is the number of features on which they share the same value [m3].

Use this similarity measure to perform unsupervised learning—employing any clustering algorithm you choose—and interpret your clustering results. Explain clearly the patterns or linguistic groups you identify [m4].

Throughout the assignment, there are a series of marking waypoints indicated with numbers in square brackets ([m1] to [m4]). These will guide you on what needs to be included in your submission and how your work will be assessed.

The distance between two languages may be greater than distance you get if you go by way of a third language: for example, and this example is, of course, completely made up, imagine

$$d(\text{Bristolian}, \text{Bathish}) = 34.4 \tag{1}$$

$$d(\text{Bristolian}, \text{Saltfordy}) = 2.8 \tag{2}$$

$$d(\text{Saltfordy}, \text{Bathish}) = 4.6. \tag{3}$$

Now it is quicker to get from Bristolian to Bathish if you go by way of Saltfordy and it might be more correct to regard

$$d(\text{Bristolian}, \text{Bathish}) = 2.8 + 4.6 = 7.4 \tag{4}$$

Use Dijkstra's algorithm to find the shortest distance between an pair of languages in L [m5] and repeat the unsupervised learning. Is it any different? [m6]

The mark waypoints [m1]-[m6] are there to give an idea of the marking scheme but is intended to be a rough guide, not a rigid scheme. At [m1] four marks will have been allocated, to get a particularly good mark here make sure the data has been cleaned. At [m2] an additional two marks have been allocated. At [m3] another four marks have been allocated, be careful to explain how you are calculating the distance, why you made that choice and what you see, for example, by noting which Languages are closest and which furthest apart. Six more marks are allocated by [m4], for a very good mark discuss the choice of algorithm and any meta-parameters, use graphs to compare different choices. Six more marks are allocated by [m5]; obviously the challenge here is implementing the algorithm. Explain clearly what you have done. Finally there are three more marks to be allocated by [m6].

Question 3 (15 marks)

Generate a data set in two dimensions with a division boundary of the form $y = ax^2 + x$, so points one side of this boundary belong to class A and to the other, class B. Investigate how well logistic regression works for these data as a is varied. What about a small neural network? How are these approaches affected if the number of points is varied, or the balance between class A and class B in the number of points? How does changing the size of the network change the performance.

Question 3 (10 marks)

Are large language models likely to make society more or less fair? How can we effect that outcome? Discuss this in an essay of about one page.

- Your essay should stay focused on that topic throughout.
- If you argue for or against a position, ensure your points directly support your stance. If you're presenting both sides, weigh the pros and cons.
- Start with a clear introduction that states your main argument or the scope of your discussion.
- Organise your points logically in the body of your essay. Each paragraph should advance your argument or analysis.

- Conclude with a summary of your argument or final reflection on the topic.
- Reference key studies or theories that support your argument. Provide brief explanations of why the cited studies are important for your argument.
- Don't just summarise; engage critically with the literature.
- For higher marks, address counterarguments or limitations in the approaches you're discussing.
- Aim for clear and precise writing. Avoid jargon unless necessary and explain technical terms.
- Ensure that your essay is free of grammatical and spelling errors, and that your ideas flow smoothly from one to the next.

You can use AI in helping develop your thoughts, or for finding and fixing errors, but I expect clarity, originality and incisive thought with strong, clearly held views while avoiding platitudes and weasel statements. None of this will be present in an essay which has, from the start, been taken from an LLM.

Report

Your report should be no longer than seven pages, excluding any references. Use an 11 or 12pt font and use a standard page layout, do not expand the page just to make it fit more text; if you are fiddling with margins to avoid cutting your submission length you would do better to spend the same time making your submission shorter.

Your report must be submitted in pdf and should be prepared in LaTeX; overleaf is a good approach, but not required as long as LaTeX has been used[1]. As always when using LaTeX, give yourself over to defaults, our expectation of what a document should look like has been conditioned on LaTeX, so it is best not to try to override the look of the document.

[1] R-markdown and some other notebook-based environments typeset using LaTeX, this is acceptable

Avoid code snippets in the report unless that feels like the best way to illustrate some subtle aspect of an algorithm; do always though consider a mathematical description if possible. You will be asked to submit code and it may be tested to make sure it works and matches your report. It will not, however, be marked in and of itself.

Submission

The deadline for report and code: 13h00 (GMT+1) on Monday 2025-07-21, there will be a submission point on Blackboard under the "assessment, submission and feedback" link. Please upload the following two files:

1. Your report as a PDF with filename ⟨student_number⟩.pdf, where the "⟨student_number⟩" is replaced by your student number, not your username. Upload this to the submission point "Introduction to AI Coursework (Turnitin)".

2. Your code inside a single zip file with filename ⟨student_number⟩.zip. Inside the zip file there should be a single folder containing your code, with your student number as the folder name. Please remove datasets and other large files to minimise the upload size - we only need the code itself. Upload this file to the submission point "Code for Introduction to AI Coursework".

We may review your Python code by eye but your marks will be based on the contents of your report, with the code used to check how you carried out the experiments described in your report. We will not give marks for the coding style, comments, or organisation of the code. Code written in Julia or R is also acceptable as is the use of a standard notebook format. If you are particularly keen on another programming language let me know and I will consider this.

Please do not include your name in the report text itself: to ensure fairness, we mark the reports anonymously.

Avoiding Academic Offences: Please re-read the university's plagiarism rules to make sure you do not break any rules. Academic offences include submission of work that is not your own, falsification of data / evidence or the use of materials without appropriate referencing. Note that sharing your report with others is also not allowed. These offences are all taken very seriously by the University and we have very little leeway within the

framework the University has set out. Do not copy text directly from your sources - always rewrite in your own words and provide a citation. Work independently – do not share your code or reports with others; you can, of course, discuss your work with your classmates, but do not share text or code.

Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the opportunity to defend your work. The plagiarism panel can apply a range of penalties, depending on the severity of the offence. These include a requirement to resubmit work, capping of grades and the award of no mark for an element of assessment. Again, we are not in a position to be lenient here, the academic offences procedure is not one we control.

Extensions and Exceptional Circumstances

If the completion of your assignment has been significantly disrupted by serious health conditions or personal problems, or other serious issues, you can apply for consideration in accordance with the normal university policy and processes. Students should refer to the guidance and complete the application forms as soon as possible when the problem occurs. Please see the guidance below and discuss with your personal tutor for more advice:

- `www.bristol.ac.uk/students/support/academic-advice/assessment-support/request-a-coursework-extension/`
- `www.bristol.ac.uk/students/support/academic-advice/assessment-support/exceptional-circumstances/`