COURSEWORK: Classification

Prof. Lorenzo Pellis

Deadline: Tuesday 18 March 2025 at 3pm. Length: 2000 words.

Data

The coursework consists in analysing the data set `vertebral_column_data.txt` (with its associated metadata `vertebral_column_metadata.txt`), which lists biomechanical attributes together with whether a patient was classified as normal or abnormal. Both files are available on Blackboard in the coursework folder.

Groups logistics

This is a *group project*, with each group consisting of *5 students* (where possible). Some logistic details:

- You can choose how to gather in groups.
- Once you have formed a group, please send an email to me (cc'ing all other members) at lorenzo.pellis@manchester.ac.uk, to tell me who is in your group.
- You have 1 week from the moment the coursework is announced. After one week, I'll proceed to randomly assign the remaining students to groups
- If you can only form a group of 3, say, try to merge with another group of 2. Please try to do it yourself, rather than relying on me pairing smaller groups.
- Remember that you are jointly responsible for the result, so you should all check (and challenge) the contribution of each other to the project – i.e. you have to read what other people have written, give constructive feedback to improve/correct it, make sure the code works, suggested different visualisations if you think they are more informative, discuss with the others if you think a different insight should be drawn, correct citations if they are wrong or inappropriate, etc. This should also allow you to share the workload more evenly.
- If there are issues in terms of people's contribution to the project, please try to resolve them among yourself first, keeping me informed. This is an important skill to develop in life. Only if they cannot be resolved, do let me know and I will see what I can do.

Coursework instructions

Using the vertebral column data, apply at least one of the unsupervised and one of the supervised classification procedures you have encountered during the first 4 weeks of this module. Produce a report, which should contain the following sections:

1. A description of the unsupervised clustering method, using your own words, including equations and citations as appropriate. **[3]**
2. A description of the supervised classification method, using your own words, including equations and citations as appropriate. **[3]**
3. Exploratory analysis of the data and any processing / transformations performed on the basis of this. **[3]**
4. Results of the analyses, including appropriate figures and tables to support the conclusions, and a discussion of how the supervised and unsupervised analyses inform each other. **[8]**
5. R or Python code used to produce the analysis. Note that it is expected that you will use packages such as *scikit-learn* rather than code from scratch. **[3]**

This gives a total of **20** marks.

Notes:

- Do read the instructions above carefully: everything that is asked will be marked.
- You do not *have to* discuss more than one method per type (you will not lose marks for not discussing more than one method), and it is indeed better to focus only on one method well than do other things badly. However, using more methods and comparing them might lead to more insight in the data, which might allow for a richer discussion. You do not need to present in detail every method you use, just one per type. However, briefly discussing the methods used might allow you to discuss why you might get a different insight from different methods.
- Please do not discuss (only) methods that we have not encountered in the module – I am trying to test your understanding of the material covered.
- Do not go over the word limit (tolerance $\pm 10\%$). This includes footnotes and figure captions, but not the code or the equations. It also does *not* include the appendix, but I should not be expected to read it to mark the work.
- Please put the full code in the appendix, but you are encouraged to use snippets of code in key places in the main text (it still does *not* count against the word limit) and explain the rationale for what you are doing (e.g. problems you are facing, solutions you are taking, etc.). You are supposed to understand your code (rather than having asked AI to write it for you!).
- Make good use of citations to support what you say – you are supposed to integrate the concepts in the slides with your own reading of other sources.
- Make the report readable (e.g. an equation chucked there, not part of a sentence, with no explanations around it, is not great... is this how they are presented in your citations?).

Due Date: 3:00pm on Tue 18 March 2025, uploaded to BlackBoard as a PDF. Length: 2000 words.