MAS61007OCT Machine Learning, Autumn 2020: Assignment 1

Data handling, Visualisation, PCA, Logistic regression, Discriminant Analysis

Instructions

Notes

1. This assessment is a University examination, and as such is subject to the University regulations governing examinations. In particular, *all work submitted for assessment should be the candidate's own work*. However, you are permitted to ask for help regarding inputting the data into Python or R (or whichever program you prefer), but the analysis and the write-up must be your own work.

2. This assessment constitutes one-half of the assessment for MAS61007. The remaining half will come from another project later in the semester.

3. Work submitted for this assessment should be word-processed, ideally with LaTeX (possibly via `Rmarkdown` and `knitR` for projects done in R). However, Microsoft Word is perfectly acceptable also.

4. The total length of the main body of the project **SHOULD NOT EXCEED FIVE PAGES**, including all tables, diagrams, references etc. Sensible sized fonts and margins should be used and diagrams should be legible to the naked eye.

5. The main report should be submitted as a PDF file electronically through Blackboard. It will go through Turnitin, which is plagiarism-detecting software.

   Please name your file `MAS61007OCT-`*registration number*`-Assignment1.pdf`, and use the same name for a Python or R script file, or Python notebook, with a different extension (`.py`, `.R` or `.ipynb`).

   The deadline for submission of the work is **12 noon** on **Tuesday December 15th**.

6. Please submit all code separately online through Blackboard. You may also wish optionally to submit the code as an appendix to your project; it will not count towards the page limits above or to the final mark. However, it is sometimes useful for the marker to clarify exactly what you are doing, and we will select a fairly small random sample of code to test that it seems to be original.

7. Reasoned requests in advance for extension of this deadline will be considered. For MAS61007, please send them through Dr Kostas Triantafyllopoulos, rather than directly to me. Note that computer failure is not an acceptable excuse for late submission.

Marking

I shall mark the work on the scale “high distinction” down to “low pass” and “fail” for MAS61007. That is, a good pass-level project will get a scores in the high 50s, and so on.

Every reasonable attempt will pass.

Given the number of students taking the module, in order to mark and give feedback in a timely way, I will release as much whole-class feedback as possible in time for the second assessment, but may make only limited individual feedback. But you should feel free to ask me for more information.

I shall be most interested in the questions you come up with to think about; how well you write up your answers; the quality of your visualisations, and the quality of discussion around your PCA.

No marks will be available for your code, although I shall skim through it to see what you are doing if your explanations aren’t sufficiently clear, and I may run a small number of randomly chosen students’ scripts.

MAS61007OCT Machine Learning, Autumn 2020: Assignment 1

Data handling, Visualisation, PCA, Logistic regression, Discriminant Analysis

PART I

The file `ass1.csv` contains 10 observations on 9 variables. Add an 11th observation, consisting of the digits of your registration number. That is, if your registration number is 170123456, then the eleventh observation should have a value of 1 for variable `V1`, 7 for variable `V2`, 0 for variable `V3`, and so on, up to 6 for variable `V9`.

1. Give the correlation between variables `V8` and `V9`.
2. Which pair of variables has the greatest correlation?
3. What is the trace of the variance matrix of your data set?
4. Plot the points on the first two principal components, marking your registration number, `Ob11`, with a distinguished colour and symbol. Make your plot as visually appealing as possible.
5. What proportion of the information in the data set is given by the first 4 principal components?
6. Suppose that `V1` is a class variable. Regard the observations {`Ob2`, `Ob3`, `Ob4`, `Ob5`, `Ob6`, `Ob7`, `Ob8`, `Ob9`, `Ob10`, `Ob11`} as a training set. Perform a linear discriminant analysis, and see whether `Ob1` is correctly predicted, explaining your answer.

PART II

The data for this project are contained in the file `rubies.csv`.

The data consist of 9 variables measured on a number of rubies. The variables are

- `where`: country of origin: 1 Burma; 2 Thailand; 3 Cambodia
- `price`: guide price
- `color`: degree of "redness" from 1 [pink] to 8 [medium-dark red] – higher numbers may be more prized
- `diameter`: largest distance between extremities of ruby
- `thickness`: largest distance within plane orthogonal to diameter
- `angle`: measure of sharpness of cut
- `cut`: expert score of quality of cut
- `clarity`: measure of absorbency of light: how transparent ruby is
- `caratwt`: carat weight (the carat is a unit of mass equal to 200mg)

The task

1. Determine an appropriate number of principal components to summarize the sample variability of the quality of the rubies (i.e., excluding location and price).

2. Give a description of the main sources of variation of the quality of the rubies.

3. What are the characteristics that distinguish the three countries of origin?

4. Are there any outliers? If so, what distinguishes them in terms of their characteristics?

5. Do a logistic regression between the rubies from Burma and Thailand. How successful is the classification?

6. If these data are used to form a classification rule for determining the country of origin of a ruby (assumed to be one or other of these three rubies), how successful do you estimate the rule to perform?

7. If a ruby has `color` 4, `diameter` 20, `thickness` 5.00, `angle` 32.5, `cut` 3, `clarity` 0.42 and `caratwt` 0.900 where would you classify it as originating from? How would you predict a guide price?

Your report should present an account of your analysis and conclusions. You may use any computer package as an aid in your analysis. You may include graphical and textual output from this computer package (suitably edited and formatted) within your report to justify your conclusions. You may include your code as an appendix, but please also submit it separately as in the instructions at the start of the assignment.