Machine Learning Assignment 1

Data handling, Visualisation and PCA

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 the first half of the assessment of MAS369/469/61007 and of the first part of MAS6019. The remaining half will come from another project later this semester.

   The second half of MAS6019 will have a project and a final exam.

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.

   Please ensure that your *registration number* is used on the front page instead of your name, so that marking is done anonymously.

4. The total length of the project is **BETWEEN THREE AND FOUR PAGES**, including all tables, diagrams, references etc. Sensible sized fonts and margins should be used and diagrams should be legible to the naked eye. **The report should not exceed four pages.**

   Note that MAS61007 students have an additional task, which may add another page.

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 `MAS369-`*registration number*`-Assignment1.pdf` (or `MAS469-...` or `MAS6019-...` or `MAS61007-...` if appropriate), 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 **12:00 (noon) Friday November 4th**. for MAS369/469/6019, and **12:00 (noon) Tuesday November 15th** for MAS61007.

   The standard University penalties for late submissions applies.

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

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

Marking

General feedback will be available, hopefully well within the University guidelines of three working weeks after submission. After marking, the projects will be checked by a second marker, and the two markers will meet to agree a final mark. This process of second marking, and agreeing marks, might take more time, but we hope to release marks within three working weeks of submission.

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.

MAS369/469/6019/61007: ASSIGNMENT 1

DATA HANDLING, VISUALISATION AND PCA

In this project you will be looking at data concerning wine. There are different attributes which contribute (potentially) to the quality of wine. I have downloaded data for several variables, and merged them into two files: `winequality-red` (for red wines) and `winequality-white` (for white wines). For convenience, there is also a file `winequality-all` with the combined data (and an extra column for the colour).

The data are obtained from laboratory tests and are related to the chemical composition of the wines. The *quality* attribute is a subjective measurement.

The data are:

- `fixed acidity`
- `volatile acidity`
- `citric acid`
- `residual sugar`
- `chlorides`
- `free sulfur dioxide`
- `total sulfur dioxide`
- `density`
- `pH`
- `sulphates`
- `alcohol`
- `quality`: a score between 0 and 10 (with 10 perceived as an excellent wine and 0 a terrible wine)
- `colour`: 0 for white and 1 for red (only in the `winequality-all` file).

Remark. The data files are included in the Blackboard folder, and are taken from `https://archive.ics.uci.edu/ml/datasets/wine+quality`.

THE TASK

Your task is to try to say interesting things about aspects of this data, and to produce some visualisations.

- Your report should contain an introduction with an overview of the data (not more than 1/2 page).
- Your first task is to perform a PCA on the combined file. Make some plots of principal components. What can you say about the main *chemical* differences between red and white wines?
- Including the `quality` variable, can you suggest particular features which are likely to be present in wines of good quality? Are these different for the red and white wines?

- After this, *and if space allows*, you could make some further exploratory analysis, including some visualisations of aspects of the data. To give you examples: are there trends between variables? Do you find something interesting if you consider the data sets separately?

Be sensible of which figures to include in your analysis. For example, depending on your investigations a scree plot might not be necessary.

Please let me know if you have any computer issues which make this task awkward. I have had no problems installing the software mentioned on the Lab sheets onto my computer at home and the one at work, but computer environments differ!

Marks will be awarded for the quality of your statistical methodology, the quality of your visualisations, and the quality of your write-up.

Extra for MAS61007

Read the first chapter of the Statistics handout, especially the theory and examples relating to Hotelling's T^2-test.

Perform a Hotelling's T^2-test to test the hypothesis that the red and while wines have the same acidity means (the variables `fixed acidity`, `volatile acidity` and `pH`).

Next, select some variables, and compute the means μ_W of these variables for the white wine data set. Perform a 1-sample T^2-test to check whether the corresponding means for the red wine data set are equal to μ_W.

Add an additional section to your project (no more than 1 page) containing the details of this investigation. State your question clearly, explaining the statistical test you are using to consider the problem, and give the solution (and possibly a visualisation, if appropriate).