MAS61007OCT Machine Learning, Autumn 2020: Assignment 2

Classification

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 second half of the assessment for MAS61007.

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 `MAS61007-`*registration number*`-Assignment2.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 February 2nd 2021**.

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 MAS61007OCT, please send them through Dr Kostas Triantafyllopoulos, rather than directly to me.

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.

Note that I expect to give an average mark in the low-merit region. 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 whole-class feedback as soon as possible, but may make only limited individual feedback. You should feel free to ask me for more information.

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.

MAS61007SEP Machine Learning, Autumn 2020: Assignment 2

Classification

The data for this project are contained in the file `gamma.csv`, extracted from the MAGIC Gamma Telescope data set, at `https://archive.ics.uci.edu/ml/datasets/magic+gamma+t`

The data set consists of simulated data on high energy gamma particles in an atmospheric Cherenkov telescope. As particles pass through the telescope, a shower of electromagnetic radiation is produced, which are approximated by elliptical shapes. The goal is to distinguish the showers which arise from gamma particles from those which come from hadrons.

The data has 11 columns, 10 continuous variables in the first 10 columns, and a class label, `g` or `h`, in the final column.

1. `Length`: the major axis length of the ellipse;
2. `Width`: the minor axis length of the ellipse;
3. `Size`: the (log of the) total brightness of the ellipse;
4. `Conc`: a measure of concentration of the brightness;
5. `Conc1`: a measure of the maximum brightness to the size;
6. `Asym`: a measure of how far the brightest pixel is from the centre;
7. `M3long`: a measure of the concentration along the major axis;
8. `M3Trans`: a measure of the concentration along the minor axis;
9. `Alpha`: the angle of the major axis to the axis of the telescope;
10. `Dist`: the distance from the central point of the telescope to the ellipse;
11. `class`: either `g` (for a gamma particle) or `h` (hadron).

Distance variables (`Length`, `Width`, `Asym`, `M3Long`, `M3Trans` and `Dist`) are measured in millimetres; `Size` is measured in photons; `Alpha` is measured in degrees, and the other two numerical variables are dimensionless.

In order to make a balanced classification problem, I have randomly chosen a training set of 5000 of each of gamma particles and hadrons. I have a further test set, `gamma_train.csv`, consisting of 1000 observations from each type, which I may use to test your code. I will not release this latter set.

The main task (maximum 4 pages)

Using classification techniques from the course, find a method which you think will predict as successfully as possible the `Class` variable from the other variables in my test set.

Explain why your method is the best you have found, and predict the accuracy rate of your method on my test set.

You will be expected to try some or all of the classification techniques from the module: logistic regression, discriminant analysis, decision trees and variants, support vector machines and neural networks.

Please let me know if you have any computer issues which make this task awkward. I have not had many problems with the software mentioned on the Lab Sheets, but computer environments differ!

Marks will be awarded for the quality of your report and explanation of how you have come to the conclusion that your method is the best of those that you have tried.

If time permits, I may run your code on the test data, and see how close you are!

Extra task for MAS61007 (maximum 1 page)

For the final part of the task, choose a random sample of 50 of each of the `g` and `h` particles. Using nonmetric multidimensional scaling, produce a plot in 2 dimensions of your data, on which the two classes are distinguished.

Does it seem that 2 dimensions is an appropriate dimension to visualise the data? Justify your answer, and suggest alternatives if appropriate.

What is the mean of the `g` observations on your plot? What is the mean of the `h` observations on your plot?