ISSU0053 Data Science and Big Data Analytics

UCL International Summer School for Undergraduates 2019

Assessment I: Computer Practical Work and Write-up (50%)

Upload report by: Wed 17th July 2019 11:00pm

This assessment requires you to demonstrate how to apply data analysis techniques using a dataset of your choosing. The assessment should be submitted in the form of an R Notebook (.Rmd file)

TASKS: The investigation should include:

i)
- exploratory data analysis including appropriate figures to illustrate features that are relevant to your investigation.
- an example showing how to use t-tests

ii)
- multivariate linear analysis **or** logistic regression **or** LDA
- model selection, cross validation and evaluation
  (this can be a regression or classification problem).

iii)
- (at least) two tree based approaches to analysing the data:
  pruned tree / bagged trees / random forests / boosted trees
- an analysis showing how the optimal model is selected by cross validation.
  (we will cover these methods next week)

iv)
- an summary including a comparison between the performance of the different models developed in part **ii** and **iii** and a brief discussion of potential next steps in the analysis.

Your write up should include brief comments and notes, so that it forms a step-by-step tutorial that could be followed by a fellow student. i.e.

- Your report should have a brief introduction setting out the context and overview of the work.
- For each investigation step you should include a brief comment on what you are doing, and why.
- Where appropriate you should ensure you critically evaluate and explain your results fully.

The expected word count for the write-up ~1000-1500 words. However a formal word limit will not be enforced.

Marks will be given for the following components:

Component (details overleaf)	**Marks**
Task completion	40
Use of Figures / Tables	30
Introduction, Commentary, Summary	40
Coding technique	10
Writing Standard	10
Presentation	10
Total	**140**

During the practical sessions you may work as usual and you may use textbooks, web resources, and request help from tutors. You may collaborate on a data investigation as a pair, but your report must be written on your own.

Pair work: choosing, preparing and exploring the datasets, designing and carrying out the analysis.

Individual work: generation of figures for report, write up and evaluation of work completed.

Grading Criteria:

Grading is carried out in accordance with the key indicators in the standard UCL report mark scheme. Additional guidance on the individual component grading, in particular reference to this assessment is provided below:

i) Task Completion

Marks will be assigned in terms of tasks completed.

0-7	Incomplete tasks, or tasks for which code which is non-functioning code or bugs will be awarded partial marks.
7.5	Full completion of all tasks with no significant omissions or errors.
7.5-10	If work has been done that exceeds the task specification to develop the investigation in a meaningful way higher grades may be awarded (see criteria below).

ii) Use of Figures / Commentary / Summary and Evaluation / Writing Standard / Presentation

9-10	Work handed is of a publishable standard. (nothing substantial to object to).
8-9	Work handed in exceeds requirements of the first class, showing evidence that students understanding and ability has been developed beyond the scope of the core materials and objectives. May need minor edits to reach publishable standard. Discussions are rounded and complete, and accurately evaluate the work completed in the context of the problem.
7-8	Work handed is of a very high standard with no omissions and shows a thorough understanding of the work carried out with respect to the materials covered in lectures and activities. Discussions are rounded and complete with a critical evaluation of the work carried out.
6-7	Work is complete and of a good standard. Some misunderstandings or omissions with respect to core lecture materials or activities may be present but arguments made are broadly correct.
5-6	Work submitted is appropriate to the investigation and is of a satisfactory standard, but contains significant (minor) flaws. The work may include multiple instances where mistakes or omissions are made.
4-5	Works submitted contains some satisfactory elements, and displays some understanding of relevant facts but includes numerous instances where mistakes, omissions or irrelevances are made, with respect to core materials covered in lectures.
2-4	Works submitted contains few satisfactory elements with major omissions, irrelevancies and/or misunderstandings made with respect to core materials covered in lectures. Work may be unfocussed but at least contains evidence of structured attempt.
<2	Little or no evidence presented of work that is relevant to the investigation. Writing is unstructured, unfocussed or unacceptably brief.

iii) Coding Technique

9-10:	Code for completed tasks demonstrates techniques that are above and beyond lecture material.
6-8:	Code for completed tasks is all fully functional. Code is presented in understandable form, and generally adheres to good coding practises, but may not be fully optimised.
4-5:	Majority of code for completed tasks is functional. However code is difficult to interpret e.g. may be poorly formatted, structurally badly organised, inefficient, and uncommented.
0-3:	Majority of code included in the submission is missing/non-functional