

CMM535 Coursework Instructions 2019-20

Issue Date: 11th February 2020

Submission Deadline: 23rd April 2020 at 1600

Submission Method: Dropbox on CampusMoodle page

Responsible Academic: Benjamin Lacroix

Weighting: 100% - this is the only assessment for this module

Overview

You are required to carry out the data science process on a public dataset of your choice. Some relevant data sources were indicated in lectures.

The process should be carried out using RStudio and be reproducible.

Your data science process should include (at least):

- Any data preparation required
- An exploratory analysis of your data
- A supervised learning experiment (regression or classification)
- Evaluation
- Presentation of your results

Data Set Selection

Although you have freedom to choose a dataset, bear in mind the following constraints.

You need to know what the data represent, so you must have access to an adequate codebook or data dictionary.

Your main analysis must be a supervised learning experiment, either a regression or classification. In order to carry out that analysis, you will need to form a rectangular multivariate dataset, i.e. a table of data where rows are instances and each column is a feature. In R terms, the data can be placed in a dataframe.

If it takes a lot of effort to construct the dataframe from the original data, this will be taken account of in marking, especially if significant extra skill is required. However, you should not allow this phase to dominate the project. In particular, it is very risky to embark on a project without having a clear process for generating the target format.

The correct size is big enough to be interesting, but not so large that you struggle to store and process the data. For example, we were able to do some interesting things with 'mtcars', but a good ML experiment probably needs more instances (rows) and may use more features (columns).

You may not use a dataset that you have used or are currently using for coursework in other modules.

Exploratory Data Analysis

The typical contents of an exploratory data analysis are described in the relevant lecture. For a modest number of features, an analysis of every feature should be presented. If the number of features is larger, then you will need to be more selective.

You may include some investigation related to the questions that your main experiment will seek to answer. This element will be larger for some projects than others and credit will again be given as appropriate. Once again, you are advised not to let this get out of hand and prevent you from spending enough time on your main analysis.

Supervised Learning Experiment

You should define a supervised learning problem, either classification or regression. To do so, you will need to declare one of your features to be the target label. You will also need to specify your performance metric(s) e.g. accuracy, MSE.

Choose at least two algorithms and run an experiment to compare their performance on the data. The experiment will normally involve cross-validation or bootstrapping to estimate performance metrics. If you choose to use a simple train-test split, you should explain why (it is possible that this is the best you can do for your specific experiment).

As well as comparing different algorithms, you should try to achieve optimum performance for your chosen algorithms. For example: feature selection or extraction prior to running the algorithm; tuning the parameters of the algorithm.

Finally, if the model provides any forms of insight on the data and the classification, the student must report it and reflect on it.

Evaluation

This should include a discussion of performance as measured by your chosen metric(s). Other questions you should address include: What did you learn from your data and which algorithm was more effective? Can you explain why your methods were effective or ineffective? Which features were the most important predictors of the target label?

Presentation of Results

You should produce a technical report explaining your process from start to finish. The report should be generated by 'knitting' an R script in .Rmd or .Rnw format. The document should give enough information to allow the reader to exactly reproduce your analysis (typically by citing the data source and including all processing commands).

You should also produce a short report on your project for a non-technical audience. The intended audience may be the general public or a specific group with a presumed interest in the data. This should briefly explain what the data is and why your audience should be interested in it, as well as summarising your conclusions.

Deliverables

Dataset choice 24th March

A short (max 1 page) presentation of the dataset chosen, containing:

- Description of the data
- Code book of the variables
- Objectives (classification or regression)

The purpose of this exercise is to make sure that the dataset chosen by the student is interesting. Feedback will be purely informative and will not be marked.

Final report submission 23rd April

An R file containing all commands and discussion. (.Rmd or .Rnw)

The document generated by knitting the R file. (.docx or .pdf)

A non-technical report: maximum one page text plus appendix with one or two figures that support the text (.docx or .pdf).

A drop box for submission will be provided on the module's CampusMoodle page. The three files should be uploaded separately. Files should not be zipped.

Grading Guidelines

Coursework submissions will be given an overall grade from A-F and that grade will be the final grade for the module. The overall grade will be calculated from three equally weighted component grades, described as follows:

- 1) Preparation of data and exploratory analysis.
- 2) Supervised learning experiment.
- 3) Description of process and presentation of results.

The criteria for each grade are outlined in the grid below. (Since all projects are different, this is indicative and may be slightly altered to make sure that all project effort is rewarded appropriately.)

	Preparation of Data and Exploratory Analysis	Supervised Learning Experiment	Description of Process and Presentation of Results
A	An interesting dataset selected. All necessary pre-processing steps correctly executed. Thorough exploratory analysis.	A well-designed experiment, implemented correctly. Chosen algorithms justified. Tuning or feature selection used to improve performance. Appropriate metrics reported.	Clear description of data and purpose of investigation. Clear description of process, results and conclusions. Analysis reproducible as instructed. Literature and data sources referenced appropriately. Non-technical report is a clear summary with appropriate emphasis and level of detail. High level of document presentation (spelling, grammar, neat layout).
B	As A, but with minor weakness, such as a less thorough exploration or non-critical technical errors.	As A, but with minor weakness(es) in design, implementation or evaluation.	As A, but with minor weakness(es), such as lack of clarity or detail, poor document presentation, lack of reproducibility.
C	A major weakness, such as a very brief exploration or multiple minor weaknesses.	A more serious flaw in the experiment or several minor weaknesses.	Many minor weaknesses or a major problem such as omission of important details or a discussion that displays a lack of understanding.
D	There are some serious errors, but the dataset can still be used for the main experiment.	The design and implementation of the experiment are poor, or the experiment omits major parts of the requested analysis.	Both reports adequately describe the data science project, but do not meet the standard for a higher grade.
E	The prepared data does not really support the main experiment. Alternatively, little preparation or exploration is presented.	The attempt to run an experiment falls short of adequacy. For example, a failure of implementation or a thoroughly misguided design.	The technical report is not an adequate description of the project process or there is little to say about a poor project. Non-technical report poor or omitted.
F	Minimal Effort Made	Minimal Effort Made	Minimal Effort Made
NS	No submission		