MSc Data Science, DSM110 R for Data Science

Endterm Coursework: Data Modelling

Introduction

During the course so far we have covered programming in R and wide range of topics in statistics and data modelling. In this coursework you are tasked to build a statistical model and predictor for a cleaned dataset. The assignment is divided in to two parts: you must write R code to build and optimise one or more statistical models for your dataset, the code will also include everything needed to generate your results including images and data for tables. The second part of the assignment is to prepare a short report in the style of an academic paper about your statistical model

This assignment is worth 70% of the total mark for this module

Task

In the first coursework you were tasked to find an appropriate and interesting dataset and then prepare it for a statistical modelling task. In this coursework you should return to this dataset and perform an appropriate investigation in to building a statistical model, predictor or classifier for your dataset. For the purposes of this coursework we are interpreting ‘model’ liberally to include statistical models, predictors and classifiers. The you should aim to built as optimal model as you can and generate sufficient results that you can show how optimal your modelling is. You should write appropriate code to handle the statistical modelling, generate any results, figures and any values needed for tables.

Your report should be formatted as per a scientific paper that deals with statistical modelling or machine learning. Your report should have the following sections:

SECTION	CONTENT
Abstract	A brief 200 word summary of your report highlighting the main result or conclusion
Introduction	Discuss the context of your work. Use this section to write a brief literature review that describes any prior academic work has been completed with your dataset or the modelling techniques you use in your study. You should also summarise the statistical methods or algorithms used in your methods. This will demonstrate you understand the methods you are using. Ensure you use appropriate references in this section.
Methods	A clear description of how to complete your analysis and modelling. You should briefly summarise where the data is from and how it was prepared (i.e. briefly summarise in one paragraph how any cleaning and missing values were handled, there is no need to repeat the content of your first coursework). The bulk of this section should be concerned with describing all the steps in your modelling any related data experiments. There should be sufficient detail here that a motivated reader and R programmer can replicate your modelling. Feel free to use code snippets if they will aid the reader's understanding. You should also ensure you clearly articulate and justify all the choices in your modelling from dataset selection to the type of modelling and benchmarking you have chosen
Results	Use figures and tables to present the outcome(s) of your modelling, benchmarking and any related investigations and data experiments. For instance; you may choose to compare how changes in hyper-parameters affect the performance of your modelling, you might compare the performance of one or more different types of model/predictor, or you may compare how size or make up of the dataset affects the performance of the model. Your dataset, modelling or project may suggest or lead you to perform other types of analysis
Discussion & Conclusions	In this section you should summarise your results. The purpose of this section is to synthesise the information in the the results section in to the new knowledge your study has discovered. You should also comment on the benefits and limitations of your dataset, methods and modelling. You should additionally comment on what future direction you might take your modelling.
Bibliography	A list of all the papers and material you have cited in your paper.

In research papers such analyses often take one of two main forms, you should choose which of these analyses you would like to do (see Appendix below);

A) A report on optimising a single predictor or model. This would include detailed investigation on how each hyper-parameter affects the final model and often detailed investigation on how the quality of the input data affects the final model (i.e. the performance impact on the amount of data used for train, the impact of cleaning or imputing variables, the impact or omitting or including variables, and so on...).

B) A report that compares the performance of multiple predictors, so that the best or most appropriate modelling methods can be chosen for the dataset. Typically there will be less work on model optimisation but 3-5 different algorithms will be trained and compared. Sometimes a meta-predictor is created the combines the output of the best performing methods to make a more accurate predictor.

In either type of study it is common to compare your best predictor to some naive method. For instance if you build a neural network for a regression task you might wish to show that your neural network can outperform a trivial linear regression. If you choose to compare your model to a naive method you should ensure you research what may count as naive for your type of modelling or dataset.

For instance, in the field of Bioinformatics researchers will often compare their methods to predictions made using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) so they can show their method outperforms this common method in the field.

Negative results are valid. Not all investigations or data experiments are successful. Perhaps you find you can not beat the naive method. Perhaps you find that performance of your model is invariant to some hyperparameter change. These are interesting results and should not be excluded.

Deliverables

1. A single input CSV file of the primary data. This should typically represent the final data.frame that was prepared in the first coursework. Submitted data should be no more than the 100Meg, if you datasets is very large submit only the 1st 100Meg of data should be submitted. Sufficient data must be provided to execute your code without error.
2. One or more R scripts that conduct your modelling and generate the results for your report. These scripts should be functional over the dataset you submitted.
3. A report in the style of a research paper that describes and explains your statistical modelling. This should be between 4,000 and 6,0000 words, about the size of a typical research paper.

Requirements

We will assess your work based on the following requirements and criteria:

Functional Requirements

R1: Code correctly loads data
R2: Code correctly handles dataset preparation as described in report
R3: Code appropriately handles statistical modelling and data experiments as describe in the report

Code style and technique

Your code should be written according to the following style and technique guidelines:

C1: Code is clearly organised
C2: Appropriate comments are included to ensure the code is clear and readable
C3: Code is laid out clearly with consistent formatting
C4: Code is organised into appropriate functions with clear, limited purpose
C5: Functions, classes and variables have meaningful names, with a consistent naming style
C6: Code runs without fault using either the R-Studio ‘Run’ button/command or the Base R source() function

Modelling

Your modelling and statistical analysis will be assessed on the following basis

M1: Appropriate modelling is selected for the dataset and modelling taskand the selection is well justified
M2: Appropriate data experiments investigating and optimising the model(s) are chosen and justified
M3: Results are clearly explained and appropriate plots, and tables of data are produced to prove and demonstrate the points made
M4: The discussion & conclusion appropriately summarises the results and puts them in the broader context. Limitations of the study, it’s achievements and possible future improvements are clearly discussed

Submission

You should write a report and submit your source code. The submission should contain the following items and information:

D1: Code in the data CSV standard ZIP format (7z, gz, tar or rar filkekaes will be permitted)
D2: A report in PDF format in the style of a computational modelling research paper. No more than 6,000 words

Marking Criteria

We will mark your work according to the set of criteria shown below, which consider the requirements, your programming technique and style and the documentation you have provided:

Category	**Not addressed**	**Attempted but did not meet requirements**	**Met described requirements**	**Met requirements and went significantly beyond them**
R1				
R2				
R2				
C1				
C2				
C3				
C4				
C5				
C6				
M1				
M2				
M3				
M4				
D1				
D2				

Appendix

A) An example research paper about developing and optimising a single (neural network) predictor
https://pubmed.ncbi.nlm.nih.gov/10493868/

B) An example research paper where multiple predictors are compared to one another
https://www.embopress.org/doi/full/10.15252/msb.20199380