MATH501 Modelling and Analytics for Data Science Coursework

Dr Malgorzata Wojtys and Stephanie Riley

Academic Year 2023/24

1 Coursework Information

Please read the following points before attempting the coursework:

- The deadline for this assignment is **2 pm** on **Tuesday, 30th April, 2024**. You should submit your work through the MATH501 Modelling and Analytics for Data Science DLE site. Your submission will be marked anonymously.
- **This is a group coursework. Please work in self-assigned groups of up to four people.** Each member of the group will receive the same mark, unless any member chooses to make use of the Peer Assessment option. You should keep notes of all your group meetings to use as evidence in case you choose to make use of the Peer Assessment option. If you wish to make use of the Peer Assessment option, you will need to contact the Module Leader **Dr Malgorzata Wojtys** by **Tuesday, 16th April, 2024** to make an appointment.
- This assignment counts for all of your mark on this module. Marks will be assigned according to the marking grid on page 3.
- Marked scripts will be returned within **20 working days** of the submission date, that is by Thursday 30th May, 2024.
- The necessary **data files** are available from the MATH501 Modelling and Analytics for Data Science Coursework DLE site.

- You are reminded of the **University's Academic Regulations**:

  Academic offences occur when activity is undertaken which could confer an unfair advantage to any candidate(s) in assessment. The University recognises the following (including any attempt to carry out the actions described) as academic offences, regardless of intent:

  a. Plagiarism, which is copying or paraphrasing of other people's work or ideas into a submitted assessment without full acknowledgement. More information on plagiarism is available here:
     https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations/plagiarism

  b. Collusion, which is unauthorised collaboration of students (or others) in producing a submitted assessment. The offence of collusion occurs if a student copies any part of another student's work, or allows their own work to be copied. Collusion also occurs if other people contribute significantly to work that a student submits as their own.

  The complete list of regulations can be found here:
  https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations

  By submitting this coursework, all group members confirm that they have understood the University's policy on plagiarism and collusion.

We now state the relevant MATH501 Modelling and Analytics for Data Science Assessed Learning Outcomes (ALOs) for this assignment.

At the end of the module the learner will be expected to be able to:

ALO1 Display an in-depth understanding of a broad range of up-to-date modelling and analytics techniques for Data Science and a critical awareness of their limitations;

ALO2 Critically choose and evaluate appropriate modelling or analytics techniques in new and complex practical situations to yield insight and innovation;

ALO3 Present results professionally and systematically to technical and non-technical audiences.

You should keep these ALOs in mind when doing this coursework.

2 Marking Grid

MATH501 Modelling and Analytics for Data Science: Coursework Marking Grid

Mark Band	Analysis	Report Style
Above 80	Correct choice of techniques with exceptionally insightful and clear justification. Outstandingly clear, correct, critical understanding of the use of appropriate techniques. Broad, highly insightful and critically reflective discussion. Very well justified conclusions. Precise explanations of the analysis concepts. Almost no technical errors. R code: exceptionally well written, correct, very clear, tidy and very well commented.	Exceptionally well structured and exceptionally well written report, with outstanding figures. Almost no presentational or grammatical errors in the text. Well formatted references used where needed.
70 to 80	Correct choice of techniques with very clear justification and correct, critical understanding of the use of appropriate techniques. Insightful explanations of concepts and interpretations of results. Well justified conclusions. Very few minor technical errors. R code: very well written, correct, clear, tidy and well commented.	Very well written and very well structured report, with very good figures. Minor grammatical or presentational errors in the text.
60 to 70	Correct choice of techniques with good justification and clear, correct, critical understanding of techniques. Mostly correct analyses, with some comparisons. Some critical and insightful commentary, but perhaps not so deep or lacking detail in some places. Well justified conclusions, with some limitations. Clear explanations of some analysis concepts. Some less minor technical errors may be permitted. R code: mostly correct, tidy, commented on, a few minor errors permitted.	Well written and well structured report, with good figures. Some less minor grammatical or presentational errors in the text.
50 to 60	Mostly correct choice of techniques, perhaps lacking in good in-depth justification. Generally correct analyses, with limited comparisons. Some discussion, but lacking insight or critical reflection. Limited or poorly justified conclusions. Acceptable explanations of some analysis concepts. Generally correct critical understanding of the methods and techniques. Some more serious technical errors present. R code: generally correct, may contain some serious errors.	A report with logical structure, mainly correct English and some good figures. Some more serious grammatical or presentational errors in the text. Some spurious or unnecessary R output included in the report.
Below 50	Incorrect choice of methods with lacking or flawed justifications. Poor analyses and muddled discussion. Unclear or very limited conclusions. Very limited or incorrect understanding of the use of methods and techniques. Many technical errors. R code: mostly incorrect, contains many major errors.	A report with poor structure, poor English or badly produced figures. Many grammatical or presentational errors in the text. A lot of spurious or unnecessary R output included in the report.

3 Your Tasks

This coursework comprises a machine learning task and a Bayesian statistics task (which also contains some frequentist analysis). You need to produce a report of your work following the instructions below. Please note that one report is required. Please do **not** submit a separate report for each task. Your single report should contain a description of your work for both tasks.

Your report should contain well presented and annotated **R** or BUGS type code for all of your analyses.

The **page limit is thirty pages**, including code and figures. Please do not submit an additional appendix as it will not be considered. Reports that contain irrelevant or uninteresting discussion or code will be penalized.

It is not necessary to repeat figures or code that are very similar.

Stars indicate the relative importance of the individual parts, with more stars indicating that the part is more important. They are included as an indicative guide only.

3.1 Machine Learning Task

For questions about this task, please refer to **Dr Malgorzata Wojtys**.

Carrying out underground explosions is a standard way of testing nuclear weapons. Such explosions generate seismic waves similar to those generated by earthquakes which can be detected by seismic stations thousands of kilometres away.

A rule that accurately distinguishes signals generated by nuclear explosions from those generated by earthquakes enables scientists to learn if a foreign power is developing nuclear weapons.

Your task is to construct such a rule using the data set `earthquake.txt`, available from the MATH501 DLE site. The data set contains information about two seismologic features:

`body`	-	body-wave magnitude (m_b), that is the magnitude of the wave that travels through the interior of the Earth, and
`surface`	-	surface-wave magnitude (M_s), that is the magnitude of the wave that travels along the Earth's surface.

The data are collected for 26 instances of earthquakes and 11 nuclear explosions, as recorded in the variable `type`.

Machine Learning Part (a)**: Explore and visualize the data to gain insights into the relationships between the two predictors and the output variable `type`. Comment on the numerical summaries and graphs in the context of the problem and justify your statements.

Machine Learning Part (b)***: Select two supervised classification methods that you studied on this module that would be suitable for these data. Briefly justify your choice.

Apply the two selected methods to the data to build a classifier to predict the type of explosion (`type`) based on `body` and `surface`.

For each method, include the following aspects:

- Model tuning: where appropriate, tune the hyperparameters of your classifiers to optimize their performance by minimising the validation error. For this, you may apply the cross-validation method.
- Model visualisation: present the two resulting classification rules visually on the scatter plot of `body` and `surface`.
- Model evaluation: evaluate the final performance of the two resulting classifiers using leave-one-out cross-validation to estimate the classification error.

Comment on all the results and graphs.

Machine Learning Part (c)*: Compare the performances of the two classifiers that you developed in part (b). Comment on the advantages and disadvantages of each method. Which one of them would you recommend as the best for these data? Provide justification for your recommendations.

Machine Learning Part (d)**: Suppose that the information on the type of explosion is not available and only the measurements for body and surface wavelengths are known. Apply the K-means algorithm using the two variables `body` and `surface`, and ignoring the variable `type`, to cluster the data. Consider different numbers of clusters and explore if the resulting clusterings could be useful to distinguish between earthquakes and explosions. Comment on your findings.

Important notes for parts (a) - (d)

Your solutions should be implemented using R. You should provide a clear and concise report that summarises your approach, results, and conclusions. The report should include code snippets with comments, plots, and tables to support your findings, as well as your narrative, commentary and interpretation of all results.

You will be assessed on the effectiveness of your data exploration and visualization techniques, the choice of classification algorithms and their implementation, the quality of your model evaluation, and the clarity and conciseness of your report.

3.2 Bayesian Statistics Task (with some frequentist analysis)

For questions about this task, please refer to **Stephanie Riley**.

3.2.1 First Sub-Task: Frequentist One-way Analysis of Variance

Four different airlines asked their customers how satisfied they were with the service provided by the airline.

The data are stored in the file `airline.csv`. The variable `airline` indicates the which airline the customer flew with; (`A`, `B`, `C` and `D`). The variable `satisfactionscore` indicates the score the customer gave the airline from 1 - 10.

Bayesian Statistics Part (a)*: Use `ggplot2` to visualize insightfully these data. What can you conclude from the plot(s)?

Bayesian Statistics Part (b)*: Let y_{ij} be the score given by the j-th customer using the i-th airline; with $i = 1, \ldots, 4$ and $j = 1, \ldots, 15$. The following one-way Analysis of Variance model has been suggested for these data:

$$\begin{aligned}
y_{ij} &\sim N(\mu_{ij}, \sigma^2), \quad i = 1, \ldots, 4, \; j = 1, \ldots, 15 \\
\mu_{1j} &= \mu_1, & j = 1, \ldots, 15 \\
\mu_{2j} &= \mu_1 + \alpha_2, & j = 1, \ldots, 15 \\
\mu_{3j} &= \mu_1 + \alpha_3, & j = 1, \ldots, 15 \\
\mu_{4j} &= \mu_1 + \alpha_4, & j = 1, \ldots, 15.
\end{aligned}$$

Provide in words an interpretation of the parameter α_4.

Bayesian Statistics Part (c):** Fit this model in the frequentist framework and report $\hat{\mu_1}$, $\hat{\alpha_2}$, $\hat{\alpha_3}$ and $\hat{\alpha_4}$.

Perform a frequentist hypothesis test of size 0.05 of whether the satisfaction score is different for each of the airlines and report your conclusion with justification.

Bayesian Statistics Part (d):** Perform a Follow-up Analysis using Tukey Honest Significant Differences. State your hypotheses and conclusions carefully.

Bayesian Statistics Part (e):** Is the satisfaction score for airline `D` more than 3 points higher than the average satisfaction score for airline `B` and `C`?

State your hypotheses and conclusions carefully.

3.2.2 Second Sub-Task: Bayesian Two-ways Analysis of Variance

Bayesian Statistics Part (f)***: A farmer wants to test the level of carbon sequestration in their fields. There are five possible techniques to capture carbon: T_1, T_2, T_3, T_4 and T_5. The farmer suspected that there may be variation because of slight differences in the locations of the fields. To allow for this the five possible types of treatment were used on each of three different fields. The following values of total carbon in the soil samples taken (the units of which are not important) were obtained:

	Type of treatment				
Field	T_1	T_2	T_3	T_4	T_5
1	208	216	220	226	209
2	194	212	218	239	224
3	199	211	227	227	221

The following Bayesian two-ways Analysis of Variance model (that we name the full Bayesian model) is postulated for these data:

$$\begin{aligned}
y_{ij} &\sim N(\mu_{ij}, \text{precision} = \tau), && i = 1, 2, 3,\ j = 1, \ldots, 5\\
\mu_{ij} &= \mu + \alpha_i + \beta_j\\
\alpha_1 &= 0\\
\beta_1 &= 0\\
\mu &\sim N(0, \text{precision} = 0.0001)\\
\alpha_i &\sim N(0, \text{precision} = 0.0001),\ i = 2, 3\\
\beta_j &\sim N(0, \text{precision} = 0.0001),\ j = 2, \ldots, 5\\
\tau &\sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001)\\
\text{standard deviation } \sigma &= \frac{1}{\sqrt{\tau}},
\end{aligned}$$

in which y_{ij} is the carbon sequestration for field i and type of treatment T_j.

Write jags/BUGS code to perform inference about the model. Run your code. Summarize the posterior probability density functions of μ, α_i, β_j and τ, indicating the numerical values of the posterior means and medians and the 95% credible intervals. What are your results telling you?

Bayesian Statistics Part (g)**: Include a graphical representation of the traceplots and posterior densities of α_i and β_j in your report and discuss your results.

Bayesian Statistics Part (h)***: Include a graphical representation of 95% credible intervals for the parameters α_i, β_j and μ_{ij}, with $i = 1, 2, 3$ and $j = 1, \ldots, 5$. Based on your results, discuss whether the underlying total carbon value is different when a different treatment is applied and/or when a different field location is used and report your conclusion with justification.

Bayesian Statistics Part (i)***: The farmer thinks that treatment T_4 should yield a higher level of carbon sequestration.

Modify your code to perform posterior inference about the differences between:

- β_4 and β_1;
- β_4 and β_2;
- β_4 and β_3;
- β_4 and β_5.

In the light of what the farmer thinks, explain what these quantities represent. Produce appropriate graphical representations of your output and provide an interpretation.

3.2.3 Third Sub-Task: Simpler Bayesian model

Bayesian Statistics Part (j)**: Consider the following simpler Bayesian model for the carbon sequestration treatment data and perform inference about its parameters writing appropriate jags/BUGS code.

$$\begin{aligned}
y_{ij} &\sim N(\mu_j, \text{precision} = \tau), & i = 1,2,3,\ j = 1,\ldots,5\\
\mu_j &= \mu + \beta_j\\
\beta_1 &= 0\\
\mu &\sim N(0, \text{precision} = 0.0001)\\
\beta_j &\sim N(0, \text{precision} = 0.0001),\ j = 2,\ldots,5\\
\tau &\sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001)\\
\text{standard deviation } \sigma &= \frac{1}{\sqrt{\tau}}.
\end{aligned}$$

Report the numerical values of the posterior means, medians and 95% credible intervals of the parameters.

Bayesian Statistics Part (k)*: Include graphical representations of the posterior densities and 95% credible intervals for β_j for the simpler Bayesian model and briefly comment on them.

Bayesian Statistics Part (l)*: Which of the two Bayesian models considered in parts (f) and (j) (the full or the simpler model) do you prefer? Why?

3.3 Report Production

You should write a single report using RMarkdown that, as a minimum:

- discusses in detail and in a **reproducible** way the above analyses for
    - the machine learning task, and
    - the Bayesian statistics task.

You should specify your Student Identification Numbers as the authors of your report. You can do this in RMarkdown by including the following line as the second line of text of the header of the document:

```
author: "11034023, 2045043"
```

for example for a group of two people.

4 What You Need to Submit

One member of your group needs to submit the following files electronically using the DLE:

- A Portable Document Format file containing your report produced by RMarkdown `Report_First_Second_Third_Fourth_Student_ID.pdf` where you substitute in the Student Identification Numbers of all the group members. For example, `Report_11034023_12045043.pdf` for a group of two people.
- The RMarkdown file that produces your report `Report_First_Second_Third_Fourth_Student_ID.Rmd` where you substitute in the Student Identification Numbers of all the group members. For example, `Report_11034023_12045043.Rmd` for a group of two people.

If anything is unclear, you should ask **without delay**.