MATH501 Modelling and Analytics for Data Science Coursework

Drs Malgorzata Wojtys and Luciana Dalla Valle

Academic Year 2019/20

1 Coursework Information

Please read the following points before attempting the coursework:

- The deadline for this assignment is **10 am** on **Thursday, 23rd April, 2020**. 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 three 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 Luciana Dalla Valle** by **Wednesday, 22nd April, 2020** 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. In particular, you will get full feedback on your work by Friday, 22nd May, 2020.
- 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

Assessment Area	Maximum Mark	Awarded Mark	Feedback
Machine Learning Task: correct choice of technique, depth of understanding and critical reporting of insights gained	45		
Machine Learning Report: display inferential and computational results clearly, concisely, systematically and professionally, using correct spelling and grammar	5		
Bayesian Statistics First Sub-Task: correct choice of technique, depth of understanding and critical reporting of insights gained	15		
Bayesian Statistics Second Sub-Task: correct choice of technique, depth of understanding and critical reporting of insights gained	20		
Bayesian Statistics Third Sub-Task: correct choice of technique, depth of understanding and critical reporting of insights gained	10		
Bayesian Statistics Report: display inferential and computational results clearly, concisely, systematically and professionally, using correct spelling and grammar	5		
Total[1]	100		

[1]This mark is provisional. Like all marks it is considered by an External Examiner and an Assessment Panel.

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**.

You are approached by a biologist who wishes to obtain a classification rule to distinguish between orchids that grow in three different locations based on their measurements.

Data on 270 orchids are available with the following characteristcs recorded

- `loc` - indicator of the orchid location: A, B or C;
- `X1` - petal length (mm);
- `X2` - leaf width (mm);
- `X3` - petal width (mm).

The data are stored in the file *orchids.txt* (you can read in the data using the function `read.table()`; please see the help file).

Machine Learning Part (a):**

Present the data visually using bivariate scatter plots and colour-coding to distinguish between three locations of orchids. You may use either base `R` or `ggplot2` to do this.

Choose two of the three characteristcs `X1`, `X2` and `X3` that should be used as predictors for orchids' locations. Justify your choice using graphs. Comment on the data in the context of the problem.

Machine Learning Part (b)*: Create a training set consisting of 210 randomly chosen data points and a test set consisting of the remaining 60 data points.

Machine Learning Part (c)*:** Using the training data set apply the K nearest neighbours method to construct a classifier to predict `loc` based on the two predictors that you identified in (a). Find the optimal K using leave-one-out cross-validation for the training data set.

Visualize the resulting classification rule in the scatter plot of the two predictors.

Machine Learning Part (d)*:** Using the training data set apply the random forest (bagging) method to construct a classifier to predict `loc` based on the two predictors that you identified in (a).

Visualize the resulting classification rule in the scatter plot of the two predictors.

Machine Learning Part (e)*:** Using the training data set apply the Support Vector Machines to construct a classifier to predict `loc` based on the two predictors that you identified in (a). Consider linear and polynomial kernels and find the best values of the `cost` parameter and the degree of the polynomial. Decide which kernel is more suitable.

Visualize the resulting classification rule in the scatter plot of the two predictors.

Machine Learning Part (f):** Find the test error for the three classification rules constructed in (c), (d) and (e), respectively, using the test set created in (b). Comment on your results. Which of the rules would you recommend as the best one for these data and why?

3.2 Bayesian Statistics Task (with some frequentist analysis)

For questions about this task, please refer to **Dr Luciana Dalla Valle**.

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

A farming consortium conducts an experiment to test four types of fertilizer, using 20 fields that have similar soil types and weather conditions. Each fertilizer is assigned at random to five fields. A crop is grown and the yield in tonnes per hectare is recorded in the following table:

Fertilizer i	Yield
1	3 2 4 3 5
2	5 4 2 6 6
3	7 6 4 6 4
4	7 5 5 6 9

Bayesian Statistics Part (a)*: Use `ggplot2` to visualize insightfully these data.

Bayesian Statistics Part (b)*: Let $y_{i,j}$ be the value of the j-th yield, when the i-th fertilizer is applied; with $i = 1, \ldots, 4$ and $j = 1, \ldots, 5$. The following one-way Analysis of Variance model has been suggested for these data:

$$\begin{aligned} y_{i,j} &\sim N(\mu_{i,j}, \sigma^2), \quad i = 1, \ldots, 4, \; j = 1, \ldots, 5 \\ \mu_1 &= \mu_1, \qquad\qquad\qquad\qquad\quad j = 1, \ldots, 5 \\ \mu_2 &= \mu_1 + \alpha_2, \qquad\qquad\quad j = 1, \ldots, 5 \\ \mu_3 &= \mu_1 + \alpha_3, \qquad\qquad\quad j = 1, \ldots, 5 \\ \mu_4 &= \mu_1 + \alpha_4, \qquad\qquad\quad j = 1, \ldots, 5. \end{aligned}$$

Provide in words an interpretation of the parameter α_2.

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 underlying average crop yield is different when a different fertilizer is applied 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 underlying crop yield level μ_4 obtained using the fourth fertilizer more than 0.5 units greater than the average of the underlying crop yield levels obtained using the other three fertilizers, μ_1, μ_2 and μ_3 respectively?

State your hypotheses and conclusions carefully.

3.2.2 Second Sub-Task: Bayesian One-way Analysis of Variance

Bayesian Statistics Part (f):** Write jags/BUGS code to perform inference about the following related Bayesian one-way Analysis of Variance model (that we name the full Bayesian model). Run your code.

$$
\begin{aligned}
y_{i,j} &\sim N(\mu_i, \text{precision} = \tau), && i = 1, \ldots, 4,\ j = 1, \ldots, 5 \\
\mu_i &= \mu + \alpha_i, && i = 1, \ldots, 4, \\
\mu &\sim N(0, \text{precision} = 0.0001) \\
\alpha_1 &= 0, \\
\alpha_i &\sim N(0, \text{precision} = 0.0001), i = 2, \ldots, 4 \\
\tau &\sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001)
\end{aligned}
$$

standard deviation $\sigma = \dfrac{1}{\sqrt{\tau}}$.

Bayesian Statistics Part (g):** Include a graphical representation of the posterior densities of α_0, α_1, α_2, α_3 and α_4 in your report and discuss your results.

Bayesian Statistics Part (h):** Include a graphical representation and the numerical values of 95% credible intervals for the parameters α_i and μ_i, with $i = 1, \ldots, 4$. Based on your results, discuss whether the underlying crop yield is different when a different fertilizer is applied and report your conclusion with justification.

Bayesian Statistics Part (i)*: Compare the 95% confidence intervals of μ_1, μ_2, μ_3, μ_4 and α_2, α_3, α_4 obtained in part (c) using the frequentist analysis, with the corresponding 95% credible intervals obtained in part (f) using the Bayesian analysis.

Bayesian Statistics Part (j)*:** Modify your code to perform posterior inference about the differences between:

- α_3 and α_1;
- α_3 and α_2;
- α_4 and α_3;
- α_4 and α_1.

Provide an interpretation of your output. Compare these results with those obtained with the frequentist Follow-up analysis in part (d).

What is the posterior probability that the underlying crop yield level μ_4 obtained using the fourth fertilizer is more than 0.5 units greater than the average of the underlying crop yield levels obtained using the other three fertilizers, μ_1, μ_2 and μ_3 respectively? Provide an interpretation of your output. Compare this result with that obtained using the corresponding frequentist test implemented in part (e).

For this part, you should base your inference on one run of a very long chain (a million iterations, for example).

3.2.3 Third Sub-Task: Simpler Bayesian model

Bayesian Statistics Part (k):** Consider the following simpler Bayesian model for the crop yield data and perform inference about its parameters writing appropriate jags/BUGS code.

$$\begin{aligned} y_i &\sim N(\mu, \text{precision} = \tau), && i = 1, \ldots, n \\ \mu &\sim N(0, \text{precision} = 0.0001) \\ \tau &\sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001) \\ \text{standard deviation } \sigma &= \frac{1}{\sqrt{\tau}}. \end{aligned}$$

Bayesian Statistics Part (l)*: Include graphical representations and the numerical values of 95% credible intervals for μ and σ for the simpler Bayesian model and briefly comment on them.

Bayesian Statistics Part (m)*: Which of the two Bayesian models considered in parts (f) and (k) (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**.