STAT 237: FINAL PROJECT DESCRIPTION

COLIN REIMER DAWSON, SPRING 2022

Overview

The goal of this project is to identify some questions of interest that can be investigated using data and a parametric model, using Bayesian updating to arrive at posterior distributions over unknown quantities of interest, carrying out and documenting your investigation in an RMarkdown writeup which is structured like a short research article.

You can work either individually or in a small group of no more than three students. If you want to work in a group but don't have one identified, let me know what general subject areas you are interested in investigating and I can match you up with like-minded students.

Broad Outline

The investigation process should consist of

1. **Formulating a topic** which is of interest to you
2. **Finding a set of data** with variables that are informative for your question
3. **Defining a parametric probabilistic model** of the data, represented as a directed acyclic graph, where nodes represent parameters or data variables, and edges define dependencies needed for conditional distributions of the nodes.
4. **Performing prior predictive checks** trying to ensure that the prior hyperparameters you choose produce distributions over parameters and/or observables in which all plausible values/intervals have probability mass placed on them, and where most of the probability mass is placed on conceivable values.
5. **Carrying out posterior inference via MCMC** to obtain a set of samples from the joint posterior distribution over all of the unknown parameters
6. **Reporting estimated posterior expected values of some intuitively interpretable quantities of interest**. What these are will vary from application

Date: Last Revised May 20, 2022.

to application, but some examples might be the predictive probability of some event of interest as it applies to future data, the expected utility of some action whose utility varies depending on the value of one or more parameters, or on the value of some future observation, or simply the value of a particular parameter, difference of parameters, etc.

7. **Write up your analysis** using RMarkdown, in the form of a research paper with the following sections:

   A. An **introduction** section, where you describe your goals, give some background context, and describe how your investigation fits into a bigger picture

   B. A **data** section, where you describe your dataset: where does it come from, what variables does it contain that you are using, what kinds of values do those variables take on, etc.

   C. A **model** section, where you define your model, preferably both with a visual representation in the form of a directed acyclic graph, and with a factorization of the joint distribution over parameters and data using conditional distributions. Make sure all symbols are defined clearly

   D. A **results** section, in which you report the results of your prior predictive checks (preferably with some graphs of key quantities), followed by your posterior inference

   E. A **discussion** section, in which you connect the results back to the broader context of your investigation

Honor Code

The central principle of the honor code as it applies here is that **your work is original**, and any **outside sources** that are used must be **cited**.

If you are working in a group, the honor pledge also affirms that **every member** of the group **contributed at every stage** of the process.

Format of the Final Written Report

The final writeup should be in the form of a report whose **target audience is a fellow student in this class**: basically, a reader who knows some basic things about Bayesian statistics, but is not a statistics expert. Err on the side of a more basic presentation: perhaps your ideal audience should be a peer as they were a few weeks ago: less knowledgeable than you are now, but still having been immersed in the content of this course.

It should be in the form of an RMarkdown document. All plots and other results should be produced from code embedded in code chunks in the document, with all auxiliary files (dataset, `.stan` program) submitted alongside it.

I encourage you to take one of the lab documents as a template: delete the text paragraphs and the code chunks after the initial set-up chunk, but leave the header section and first setup chunk.

Formatting Note: Before Knitting the document to turn it in, you should set `echo=FALSE`, `message=FALSE`, `results='hide'`, and `warning=FALSE` in the chunk options in the initial setup chunk. This will hide the code, raw text output, and various messages from R from the Knitted report, while leaving plots and object definitions that you might want to refer to in text or tables.

The idea is to make the report **looks like a research paper**, without code and unformatted output of code displayed in the writeup). In other words, someone reading it should **not necessarily be able to tell by looking at it that you used RMarkdown** to write the report.

This also means you will need to describe your methodology in enough detail that it is clear what you did **without seeing the actual code**. I will ask you to turn in the `.Rmd` file alongside your writeup, but **the paper must stand on its own**: all necessary information must be in the text and figures so that someone can **read the Knitted document without seeing the code** and get all of the necessary quantitative and qualitative results.

Once again, the Markdown document **must be self-contained** so that it can be run successfully, **including the data import step**, from a fresh RStudio session with nothing in the workspace. If your data is being read in from a `.csv` file or similar, this file should be placed in the same folder as the `.Rmd`, and then read in using a `read_csv()` (or similar) command **within the Markdown document**.

Grading Criteria

The project will be graded on the following five dimensions, each weighted equally:

1. **Suitability of the model and the data** for the substantive questions of interest
   (a) To what extent does the data shed light on the phenomena of interest?
   (b) To what extent does the parameter space of the model allow useful distinctions between different ways the phenomena of interest might work?

2. Suitability and accuracy of the **mathematical representation** of the model (as a graph and collection of probability distributions) given the verbal description and nature of variables
   (a) To what extent does the graphical representation match the verbal representation?
   (b) To what extent does the collection of conditional distributions provided correspond to the graphical representation?
   (c) To what extent are the distributional forms appropriate to the nature of the data and the parameters?
3. Accuracy of the **implementation** of the model in code and the sampling algorithms used
   (a) To what extent does the specification of the model in Stan reflect the mathematical representation?
   (b) To what extent is the prior predictive process implemented accurately
   (c) To what extent is the posterior sampling process implemented accurately?
   (d) To what extent are the results extracted accurately?
4. Soundness of the logic of the **investigative process** (that is, results guide decisions appropriately)
   (a) To what extent does the information revealed by the prior predictive check appropriately guide the modeling process
   (b) To what extent are the results obtained from the posterior distribution used effectively to answer questions of interest?
   (c) To what extent are the results interpreted sensibly with respect to the substantive phenomena of interest?
5. **Clarity and organization** of the writeup
   (a) To what extent is the layout of the writeup well organized and easy to follow?
   (b) To what extent is the written prose clear and easy to understand?
   (c) To what extent do the included visualizations aid in understanding the analysis?
   (d) To what extent is the Knitted writeup clean and polished-looking?

Turning In Your Work

Your submission should consist of **at least four files** (could be three if your data is read in directly from the web):

1. The `.Rmd` **source document**
2. **A Knitted PDF** output file. For mathematical notation, use the dollar sign syntax that we've seen in various labs.
3. Your **dataset** as a plain text `.csv` file (unless it is read in directly from the web)
4. Your `.stan` file (or files, if you implemented your prior predictive check in a separate program)

Copy these files to the turnin folder at `~/stat237/turnin/project/`. I should be able to Knit the Markdown file from there and reproduce your report.