UNIVERSITY OF EXETER

COLLEGE OF ENGINEERING, MATHEMATICS, AND PHYSICAL SCIENCES

MTH3041

Bayesian Statistics, Philosophy and Practice

Continuous Assessment 1

Date set:	Friday 6th December 2019
Hand-in date:	Friday 17th January 2020 by noon
Return date:	Friday 7th February 2020

This assessment comprises 20% of the overall module assessment.

This is an *individual* exercise, and your attention is drawn to the College guidelines on collaboration and plagiarism, http://intranet.exeter.ac.uk/emps/subjects/mathematics/assessment/academicmisconduct/. **You must not collaborate when working on this assessment.**

There are **3 parts to hand in**.

1. Your answers to each question, including any plots and summary `R` output. I recommend using RMarkdown to produce this document.
2. Please hand in an electronic `R` script (a .R file) **entitled with your anonymous student code and then '.R'** or a .Rmd (R markdown file used to generate your report), and any .stan scripts you used via the college electronic submission system (a link and submission instructions will appear on ELE in due course), containing the code you used to answer each question, and clearly annotated. This script may be run in R to help assess your mark, so please ensure that when the command `source("YourScript.R")` is run in a new R session, all of your code runs. If you are submitting a .Rmd instead of .R file, ensure that all code blocks run in a blank R session and the markdown knits.
3. A hard copy of the above `R` code and any Stan scripts for marking.

1. The data `pulsar_stars.csv` contains 8 statistics on the radio emission patterns of 17898 potential pulsars detected during a project called the High Time Resolution Universe Survey, and a final column indicating whether the actual emission was caused by a real pulsar. The first 4 variables represent statistics on the time series of the radio signal and the next 4 represent statistics from a curve fitted between the Dispersion Measure (DM) and the Signal to Noise Ratio (SNR), for each signal.

   Most detected signals come from radio frequency interference (almost 90% of this data), and independent verification of a pulsar is time consuming.

   (a) Fit an appropriate Bayesian model to this data in order to be able to accurately classify pulsars based on 8 radio emission pattern statistics. You must:
   - Fit your model using Stan.
   - Reserve an appropriate portion of the data for out of sample validation.
   - Justify your model and your priors appropriately.
   - Ensure that your model converges and present convergence diagnostics.

   [40 marks]

(b) Denoting the training data $\boldsymbol{y}$ and a new potential pulsar $\tilde{y}(x)$, suppose that our classifier labels a signal with attributes x as a pulsar if $\pi(\tilde{y}(x) = 1 \mid \boldsymbol{y}) > 0.5$. Show that the probability in this classifier can be written as an expectation and hence explain how samples from your model's posterior distribution can be used to classify pulsars by Monte Carlo. *Note that if you have used the generated quantities block in your Stan code, the code in this block may be performing part of the calculation that you must describe here.*

[5 marks]

(c) Use the estimates of the expectation described in the previous part to present a confusion matrix for your reserved data and give the overall success rate and the success rates given the true origin of the signal (pulsar or not) of your Bayesian classifier.

[15 marks]

(d) Calculate the average Monte Carlo error for the posterior probabilities that

i. you classified as belonging to pulsars,

ii. were edge cases (within probability of 0.1 of being classified another way),

justifying the Monte Carlo sample size that was used. Is your simulation long enough?

[10 marks]

(e) Using appropriate calculations and figures from your Bayesian analysis explain what you feel are the important indicators that a signal is in fact from a pulsar and not from radio frequency interference.

[15 marks]

(f) Suppose that, for any given validation data set of size M, our utility function for the performance of our classifier is

$$U(d) = 3M - \sum_{i=1}^{M} \left[2I(FN_i(d)) + I(FP_i(d))\right],$$

where $I(FN_i(d))$ is 1 if, under the decision d driving our classifier, the ith validation point was given as a false negative (it was classified as radio frequency noise but really a pulsar) and 0 otherwise; and similarly $I(FP_i(d))$ is 1 if the ith validation point was given as a false positive (incorrectly classified as a pulsar) and 0 otherwise.

Let your classifier in part (b) be replaced by

$$\pi(\tilde{y}(x) = 1 \mid \boldsymbol{y}) > d \qquad\qquad d \in [0, 1].$$

Find your optimal d using your validation set, interpret the answer in terms of the pulsar problem.

[15 marks]