CGS698C: Bayesian Data Analysis and Modelling
Assignment #2:

`Max marks: 200`

`Due on/before: 11.59, 24-Apr-2021.` `11-Apr-2021`

Use Jupyter lab/notebook and submit a notebook as your submission. Use libraries from the Python ecosystem pymc3, arviz, matplotlib.pyplot, numpy, scipy etc. Submit on mooKIT.

Credit will be based on how thorough and convincing your analysis is and how you support your conclusions.

1. In this assignment you will fully analyze data from an old experiment on categorization. The experiment involves 8 stimuli that have been categorized into two classes A and B. The categorization is based on two stimulus attributes - height of the rectangle and position of a vertical line inside the rectangle. The stimuli, their dimensions and categories are shown below in Figure 1:

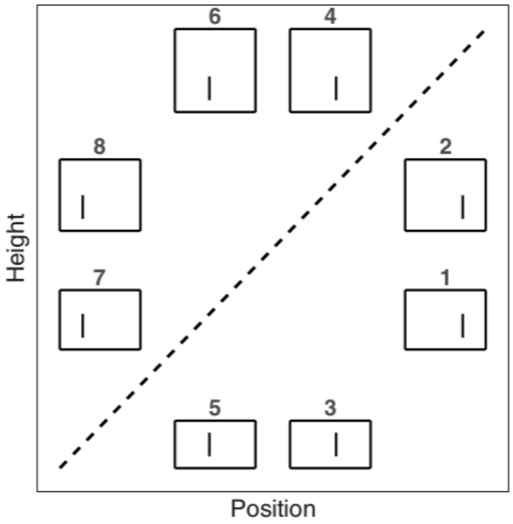

Figure 1: Eight stimuli, two stimulus dimensions (height, position) and two categories A: above diagonal, B: below diagonal.

The experiment had 40 participants. In each block each participant was shown each stimulus once in random order where the particpant had to categorize it as belonging to category A or B. There were 8 blocks. The data is as follows: y_{ij} - the number of times the i^{th} stimulus was categorized by the j^{th} subject as category A; then $y_i = \sum_j y_{ij}$ the total number of times i^{th} stimulus was categorized as A in $t = 320 = 8 \times 40$ in a total of t presentations across all subjects.

The data from the experiment is in the file data.txt as a Python dictionary. The important component is the one indexed by ‘x’ which gives the data for each participant - list of 8 entries one for each stimulus. The other entries in the dictionary are self-explanatory and relate to y_i; stimulus dimensions d_1, d_2; number of stimuli; number of blocks.

Since there are two stimulus dimensions let stimulus i be represented as the pair (p_{i_1}, p_{i_2}) in psychological space. Let the weight given to the first dimension be $0 \leq w \leq 1$ and that of the second dimension be $(1 - w)$.

The weight gives the attention/ emphasis given to the respective dimension. The distance between stimuli i and j can be written as $d_{ij} = (w|p_{i_1} - p_{j_1}| + (1-w)|p_{i_2} - p_{j_2}|)$ and the similarity can be modelled by $s_{ij} = e^{-cd_{ij}}$ which will vary between 0 and 1. Here, c is a parameter. Let $s_{iA} = \sum_{j \in A} s_{ij}$ be the overall similarity of stimulus i to category A. The probability that the i^{th} stimulus will be categorized as A is $r_i = \frac{bs_{iA}}{bs_{iA} + (1-b)s_{iB}}$ where b is a bias parameter which is chosen to be 0.5 for an unbiased choice. Then $y_i \sim binomial(r_i, t)$.

The categorization process is supposed to differentially attend to the different dimensions of the stimuli that are stored as exemplars in memory and then categorize based on the value of r_i.

You have to analyse the experimental data and write an analysis. Here are some major points you must cover.

(a) Construct a dependency diagram for the model choosing relevant priors and justifying your choice of priors.

(b) Run MCMC simulations, plot the distributions and 95% HDI intervals, choose credible values for parameters and draw conclusions.

(c) Are all 40 participant behaviours similar or are there individual variations that can be grouped such that the 40 participants can grouped into two or more categories based on the performance data? Use visualisation to see if you spot any patterns then do an analysis based on the patterns you see. In particular you may need a new dependency model.

(d) We chose the bias parameter $b = 0.5$ assuming no bias between the two categories. Is that a good choice? Check to see what happens if b is allowed to be a free parameter that is estimated by the data.

(e) Do a post predictive check to see if your chosen model parameters fit the data well. Can you think of other models for the data? If yes, then fit these other models to the data and pick a 'best' model. On what basis do you say it is best?

[200]

Note on how to use the data:

Actually, the raw data has been re-organized so that you can directly use the data in a hierarchical Bayes model based on the descriptions given. **It does not correspond to measurement data of the experiment.**

Experimental procedure:
There are only 8 stimuli in all. To each subject in one block they are shown in random order. There are 8 blocks so a subject sees each stimulus 8 times. Feedback on the category is given after each stimulus exposure. So, each stimulus is categorized a total of $320 = 40 \times 8$ times.

Calculation procedure:
The model has 3 parameters: w, c and b for which you will have to choose priors. The distance $d_{ij} = d^1_{ij} + d^2_{ij} = |p_{i1} - p_{j1}| + |p_{i2} - p_{j2}|$. In the data the $d1$ and $d2$ values correspond to d^1_{ij} and d^2_{ij}. It gives us the L_1 distance between stimulus i and j (the reason we see 0.0 values). The similarity will be: $s_{ij} = e^{-c(wd^1_{ij} + (1-w)d^2_{ij}))}$. To calculate r_i we need to know the labels of each stimulus. The list 'a' in the data gives the labels. 1 corresponds to category 'A' and 2 to 'B'. This will allow us to calculate $r_i = \frac{bs_{iA}}{bs_{iA} + (1-b)s_{iB}}$.

With the above explanation it should be straightfoward to set up the model and answer the questions.

The data is from an experiment done by Kruschke.

John Kruschke, Human Category Learning: Implications for Backpropagation Models, Connection Science, 5(1), 3-36, 1993. DOI: 10.1080/09540099308915683