CGS698C: Bayesian Data Analysis and Modelling

Assignment #1:

Max marks: 150

Due on/before: 11.59, 17-Feb-2021. 1-Feb-2021

This assignment is to get you started on using Python and its libraries for doing Bayesian analysis. Use the PyMC3 library for the Monte Carlo part and numpy, scipy, matplotlib, pandas, seaborn etc. where necessary. Please see the 'Useful links' page on the course site for more information and documentation on these modules.

Use Jupyter lab/notebook and submit a notebook as your submission. A submission link will be sent closer to the submission deadline.

1. Consider the first example we did in class where we wanted to find the probability p (for Heads) of a coin. Reproduce what we discussed in class as follows:
   (a) First construct a sampler with a known p value. You will use this to generate data for your Bayesian analysis for different values of p.
   (b) Using the following 4 priors i) non-informative ii) prior with a peak at $p = 0.25$ iii) prior with a peak at $p = 0.5$ and iv) prior with peak at $p = 0.75$ and the likelihood (you can use the one which was discussed in class) plot the posterior distribution as your generate data. In particular show plots for what happens after 1, 2, 3, 4 data points and after 1K and 4K data points. You can use a β-distribution to create the peaked priors.
   (c) Also, check what happens when you do the first 4 data items one-by-one versus after all four items are available - whether you get the same posterior.

   [10,50,10=70]

2. Consider the height-weight correlation example. Build a Bayesian model for the correlation by finding posteriors for β_0, β_1, σ as follows:
   (a) First create a data generator by assuming reasonable values for the three parameters β_0, β_1, σ. To generate a data point w first draw a height value $h \sim \mathcal{N}(\mu_h, \sigma_h)$ (choose μ_h, σ_h suitably), then use the regression equation to generate a predicted value, say $\hat{w}$, draw the sample data point using $w \sim \mathcal{N}(\hat{w}, \sigma)$.
   (b) For two priors given below compute the posterior distributions for the three parameters for sample sizes of 10, 100 and 1000 and plot the distribution and the 95% HDI.

   The two priors are: i) a non-informative prior (done in class) and ii) a more informed prior created by drawing 10 data points using the generator in part (a) and then using $\mathcal{N}(\mu', \sigma')$ as the prior where μ', σ' are the sample mean and sample standard deviation of the ten-sized sample.

   [15, 65=80]