Homework 3 - Understanding Samplers
DS6040 Bayesian Machine Learning
(Fall 2024, Instructor: Teague Henry)

1 The Problem

This assignment will ask you to implement two simple samplers for Bayesian change-point detection problem: a Gibbs sampler and a random walk Metropolis sampler, and to discuss the performance of each. Additionally, this problem will ask you to apply Bayesian logistic regression using the brms package to the white wine dataset you worked with in HW2.

What you will need:

- `coaldisasters-ds6040.csv` - Found in on the Collab site.
- `hw3samplerstubs-ds6040.Rmd` - Jupyter notebook containing code stubs for the samplers I am asking you to implement.
- `whitewine-training-ds6040.csv`
- whitewine-testing-ds6040.csv (if you are interested in the extra credit)

Prepare your own RMarkdown for submission. You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below. **Include an honor pledge with your submission.**

2 Part 1: Changepoint detection and samplers (50 points)

The term *changepoint*, used in a data science or statistics context, refers to the point in time when a data generating process changes in a drastic fashion. For example, changepoint detection is used quite often in financial analysis to estimate when sudden changes in certain markets occur. Another example of a changepoint would be the COVID-19 pandemic, and its impact on the global economy. The two most important qualities of a changepoint is that it is a point in *time*, and that the change is, in a vague sense, *drastic*[1]. In Part 1 of this HW, you are asked to analyze a famous dataset on coal mining disasters in the UK between 1851 and 1962 (112 years in total). The model you will use is a Poisson changepoint model, which is fully described below:

Let X_i be the number of coal mining disasters occurring in year i. The goal of this analysis is to estimate the year when *something* occurred that drastically altered the number of coal mining disasters. To model this, we first specify our model:

$$X_i \sim \text{Poisson}(\mu),\ i = 1, \ldots, k$$
$$X_i \sim \text{Poisson}(\lambda),\ i = k+1, \ldots, 112$$

with the overall likelihood being:

$$L(\mathbf{X}|\mu, \lambda, k) = \prod_{i=1}^{k} \frac{\mu^{X_i} e^{-\mu}}{X_i!} \prod_{i=k+1}^{112} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

where e is the standard Euler's number, and ! indicates a factorial operation. Note that the likelihood is just the product of two Poisson likelihoods with differing parameters.

[1] For the interested reader, this notion of drastic change is attempting to separate out phenomena that demonstrate changepoints from phenomena that demonstrate slower, continuous change in the data generating process

Translated, this model says that the number of coal mining disasters is distributed as a Poisson random variable (as numbers of events during a given time interval usually are), but the rate of coal mining disasters changes from μ to λ between year k and $k+1$[2].

The posterior distributions we are interested in obtaining are those for μ, λ and k. We can specify the priors for our parameters as follows:

$$\begin{aligned}\mu &\sim \text{Gamma}(a_\mu, b_\mu)\\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda)\\ k &\sim \text{Discrete Uniform on } 1, \ldots, 112.\end{aligned}$$

with a discrete uniform distribution just being a distribution of equal probabilities on integer values in a certain range (here, 1 through 112).

Fortunately (for you!), the conditional posterior distributions for these parameters are quite simple:

$$\mu|\lambda, k \sim \text{Gamma}(a_\mu + \sum_{i=1}^{k} X_i, k + b_\mu)$$

$$\lambda|\mu, k \sim \text{Gamma}(a_\lambda + \sum_{i=k+1}^{112} X_i, 112 - k + b_\lambda)$$

$$p(k = \hat{k}|\mu, \lambda) = \frac{L(\mathbf{X}|\mu, \lambda, \hat{k})}{\sum_{i=1}^{112} L(\mathbf{X}|\mu, \lambda, k = i)}$$

1. (50 points) With the above information, complete the Gibbs sampler in the accompanying notebook. You should only need to complete the update steps for the μ and λ (called `lambdap` in the notebook) parameters. Run the Gibbs sampler, plot the posterior densities and calculate the EAP estimates with 95% (equal tailed) credible intervals for μ and λ. Provide the top 5 most probable values of k. Then do the the following:

   (a) Describe your findings. What do these EAP and credible intervals imply? And what was the most likely year of the changepoint?

   (b) Why is an EAP or credible interval not necessarily the most appropriate thing to report for the year of the changepoint?

2. **(EXTRA CREDIT 10 points)** Now, instead of using a Gibbs sampling step to obtain your posterior estimate of k, change it to a Metropolis sampling step:

   - Sample a proposal k^* from a Discrete Uniform on $1, \ldots, 112$.
   - Calculate your acceptance ratio as

     $$a_k = \frac{L(\mathbf{X}|\mu, \lambda, k = k^*)}{L(\mathbf{X}|\mu, \lambda, k = \hat{k})}$$

     where $\hat{k}$ is the current value of k. Note that this is the full Metropolis acceptance ratio, I've already cancelled out the uniform priors.
   - Sample $r \sim \text{Uniform}([0, 1])$
   - If $r < a$, $\hat{k} = k^*$, else, $\hat{k} = \hat{k}$

   Run your new Metropolis-within-Gibbs sampler for 1000 iterations, then do the following:

   (a) How are the results similar/different than the fully Gibbs sampler?

   (b) What is the issue with this particular implementation?

3. **(EXTRA CREDIT 10 points)** Now, modify your Metropolis-within-Gibbs sampler by changing the proposal distribution for k from a discrete uniform on $1, 112$ to a discrete uniform on $\hat{k} - 1, \hat{k} + 1$. Put in an error check to make sure that your proposed k^* is not less than 1 or greater than 112. Run your modified sampler and compare the results to your previous Metropolis-Within-Gibbs sampler. (Note, only your proposal sampling step will change, you do not need to change your acceptance ratio calculation as the proposal distribution is still symmetric).

[2]Note that this changepoint is happening between years, not on a year. It's a small modelling choice, but as it turns out it is much more difficult to model the changepoint as happening during the year itself

3 Part 2: Bayesian Logistic Regression with brms (50 points)

In this section, you will be applying Bayesian logistic regression to the white wine data you worked with previously. The main goal of this part is not to test your programming abilities, but more to let you practice your applied interpretation/communication skills. Here is what I want you do to:

1. (40 points) Load the data from `whitewine-training-ds6040.csv` and create a new 0/1 quality variable, where the new quality of the wine is 0 if the wine received a C or F, and 1 if the wine received an A.
2. Using the brms R package, fit logistic regression models to determine the following
   - The set of 3 variables that give you the best classification rate of wine overall in the training data.
   - The set of 3 variables that give you the best classification rate of the A rated wines.

   Once you have determined your two models, plot both traceplots and forestplots.

   Discuss your findings. Note that I am not asking you to calculate miss-classification rates, rather to examine your parameter distributions and explain what those estimates mean, and how they compare to one another. Don't forget to interpret the intercept as well.

3. **(Extra Credit 10 Points)**: Figure out how to obtain the classification/misclassifcation rate from the two logistic regression models on the testing data. Compute and compare the performance of the two models using the classification/misclassification rate.