Homework 1 STAT/CS 387: Data Science II
Bayesian inference

Instructions

Please read these instructions carefully before you begin!

Included with this file is a directory called `work/`. This is where you will include any code you have written to answer the questions. Inside `work` is a directory called `data/`. Your code should sit in the `work` directory and read those files from the `work/data/` subdirectory.

- This assignment follows directly from Data Science 1. Consult those lecture notes as needed.

Your writeup

Please make a file `HW01_[NETID].pdf` (e.g., `HW01_jbagrow.pdf`) containing your written answers to the below problems. Include the assignment number, date, your name, and label each problem clearly. Any figures you provide should also be in this file including figure captions. You may use any tool you wish to make this file but you must convert the final output to a PDF. If you use Microsoft Word, please export the file to PDF.

Show your work! This means (i) provide the code you have produced by placing it in `work/`, (ii) next to any answers in your writeup include a parenthetical statement pointing out where in your code the answer was computed. For example: "(See plot_timeseries.py, lines 100-110.)"

To submit

With your final writeup in `HW01/` and your code in `HW01/work/`, rename `HW01` to `HW01_[NETID]`, compress (zip) the directory and upload the `HW01_[NETID].zip` file to Blackboard.

Problem 1. The year is 2079. The devastating water wars of the 2060s have pushed humanity to the brink. Bayesian Hunter Killers (BHKs)—intelligent, autonomous death machines—roam the land, a terrifying remnant of the wars and killing at will. With our very survival at stake, the remaining tribal warriors spend their days ambushing and destroying BHKs.

As a tribal elder, and one of the few who can still remember the Time Before, you sit in a burned out bunker, spending precious electric power to crunch available data on BHK movements and strategies. Your warriors have brought you a new dataset! They have identified two primary BHK models, designated BHK-Mk1 and BHK-Mk2:

- Your warriors have fought 26,751 Mk1s and 27,079 Mk2s. In their respective battles, the Mk1s have killed 183 warriors, while the Mk2s have killed 222.
- No BHK has been fought more than once; each battle is different. BHKs are always fought one at a time, at most one warrior is lost per battle, and assume the probability of losing a warrior is the same for all battles.

(1) What probability distribution (or type of random variable) characterizes these data and why?

(2) Using Bayesian Inference, as the frequentists have been hunted to extinction, what is the probability that the Mk2 model type is deadlier than the Mk1[1]? (Show your work, else a warrior will challenge you for leadership of the tribe.)

[1] Here "Bayesian Inference" means sampling from appropriate posterior distributions using MCMC, as in class.

Problem 2. In Data Science 1 we used Bayesian Inference to determine whether the rate of text messages received per day changed at some time. We did this by building a model of two poisson distributions, with rates λ_1 and λ_2, along with a critical time τ where the rate instantaneously switched from λ_1 to λ_2[2]. We then used MCMC to generate samples from the (unspecified) posterior distribution and plotted the distributions of λ_1, λ_2, and τ.

(1) Implement this inference problem yourself (taking PyMC code from Data Science 1 lecture notes if you wish). The count data (C_t = # of messages received on day t) is provided in the file `txtdata.csv`. Construct an argument with plots as needed to demonstrate that your sample has converged in distribution to the underlying posterior. What steps did you take to ensure you have converged? What did you calculate to show convergence and why?

(2) The "switchpoint" model discussed in Data Science 1 seems unrealistic to me, because I do not think there exists a single privileged day where the rate of messages suddenly jumped from λ_1 to λ_2.

- Propose a function of time $f(t; \lambda_1, \lambda_2, \phi_1, \phi_2)$ that smoothly changes the (time-dependent) poisson rate $\lambda(t) = f(t)$ from λ_1 to λ_2. (Note that t remains integer-valued from the count data but it is best if f be defined for all $t \in \mathbb{R}$.) The parameters ϕ_1 and ϕ_2 control the transition from λ_1 to λ_2. What is f and why did you choose it? How do you interpret the parameters ϕ_1 and ϕ_2? (Hint: we discussed a function in previous lectures that may help you choose $f(t)$.)
- Perform Bayesian Inference with f replacing the switchpoint function we used in Data Science 1 to define $\lambda(t)$. What are appropriate priors for ϕ_1 and ϕ_2? Demonstrate converge of your sample to the posterior as per question 2A.
- Average over your posterior samples and plot the expected poisson rate $\lambda(t)$ as a function of time t. Include with this plot a 95% CI for λ, again taken from the distribution of posterior samples. Inspecting this plot, does it support or contradict our earlier switchpoint model? Why or why not?

(3) This new model is more complex than the one studied in Data Science 1 because one parameter (τ) has been replaced with two (ϕ_1 and ϕ_2). Which model is better justified and why?

[2] Remember the prior distributions were $\lambda_1 \sim \text{Exp}(\alpha)$, $\lambda_2 \sim \text{Exp}(\alpha)$, $\tau \sim \text{DiscreteUniform}(0, T)$, and α was fixed by the data.