HOMEWORK 05

MATH 295: DATA SCIENCE II

Assigned: 2014-11-04
Due: 2014-11-11

Instructor: James P. Bagrow
bagrow.com/ds2/

Instructions Submission, showing your work, etc. are as per previous homework. (Upload `HW05_<yourname>.zip` 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, one of the more terrifying remnants 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 still remember the vaunted Age of Light, you sit in a blown 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.

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

1B 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.)

Hint There are four numbers given, but you should consider these as defining two datasets, one with $26,751$ datapoints, and the other with $27,079$. But what are the datapoints?

Problem 2

In class 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.

2A Implement this inference problem yourself (taking PyMC code from the 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?

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

[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.

2B The "switchpoint" model discussed in class 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?
- Perform Bayesian Inference with f replacing the deterministic switchpoint function we used in class 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?

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

If you are using **Enthought Canopy** and wish to use the PyMC package you will have to install it in a special way (since PyMC is not one of the free/academic Canopy packages). In Canopy, select the `Tools > Canopy Terminal` menu option. This should open a separate terminal (command line) window. Confirm that your terminal is using the Canopy python, and not another version of python, by entering

`which python`

This displays the path to the python program, and it should contain references to Enthought. For example, on my system it reads

`/Users/bagrowjp/Library/Enthought/Canopy_64bit/User/bin/python`.

Finally, to install PyMC, run `pip install pymc` at this prompt. You can now close this terminal window.

For more details, see: Installing Packages into Canopy User Python from the OS command line.