Stat 610 Homework 9

Due Thursday, December 1, 11:59pm

Background

Contagion models describe the spread of an attribute in a network: diseases in contact networks, behaviors in friend networks, ideas in discussion networks, and so on. We will look at data from a study on how doctors adopted a new drug. Coleman, Katz, and Menzel [1] collected data on doctors in several cities in Illinois, including the first date on which they prescribed tetracycline (a new drug at the time) and which other doctors they went to for advice or discussion of treatment options. One of the theories investigated in the paper was that doctors who knew other doctors who had already prescribed the drug adopted it more quickly than those who didn't. This behavior can be described as a contagion model. In this assignment, we will simulate from a contagion model and use the simulations to obtain an approximate posterior distribution of the contagion parameter using approximate Bayesian computation.

The contagion model we will consider is the susceptible-infected (SI) contagion model. In this model, each of the nodes in a network can be either susceptible to infection or already infected. At each time step, each infected node picks one of its neighbors uniformly at random. If the chosen neighbor is not infected, it becomes infected with probability p. For us, each of the nodes is a doctor, an "infected" node is a doctor who has prescribed tetracycline, and the links in the network are doctors who discuss cases with each other.

Assignment

Part 1: Simulating from the contagion model

1. Download network and attributes datasets from `https://jfukuyama.github.io/teaching/stat610/assignments/network.csv` and `https://jfukuyama.github.io/teaching/stat610/assignments/attributes.csv`.
   - The network is a matrix describing a directed network: element (i, j) is 1 if doctor i listed doctor j as someone he discussed patients with.
   - The attributes dataset has more information about the doctors, but the important column for us is `adoption_date`, the month in which the doctor first prescribed the drug.

   Note: The data on the website is a subset of the data sets that used to be available at `http://moreno.ss.uci.edu/data.html#ckm`, which you can find at `https://web.archive.org/web/20200705155408/http://moreno.ss.uci.edu/data.html`. You can find more information about the data attributes there. For this assignment, we are only using the doctors from Peoria.
2. Write a function that simulates from a contagion model. Your function should take the following parameters:
   - A network,

- A set of initially infected nodes,
- The infection probability p,
- The number of time steps to simulate

and it should return a vector giving the time point at which each node was infected.

3. Simulate from the model given the real data parameters for the network (the doctor discussion network), the initially infected nodes (the doctors who prescribed tetracycline in month 1), and the number of time steps (18 months). Do one simulation each for $p = .1, .5, .9$, and show what one realization of the simulation looks like.

Part 2: Approximate Bayesian Computation

4. Using a uniform prior over p, the infection probability, use ABC to get samples from the approximate posterior distribution of p given the observed data.

   Recall that for ABC, you need to specify

   - A prior distribution for p. You may use a uniform distribution.
   - Summary statistics. If there are n nodes/doctors, your summary statistic should be $x \in \mathbb{R}^n$ such that x_i is the inital infection time for node i.
   - A tolerance ϵ. In class, we said that if $x^{(i)}$ is the summary statistic resulting from simulation i and $x^{(obs)}$ is the summary statistic for the observed data, we keep the parameter from simulation i if $\|x^{(i)} - x^{(obs)}\| \leq \epsilon$.

     We can use a slightly different strategy: draw N times from the prior, and retain the parameters corresponding to the ϵN closest matches between $x^{(obs)}$ and $x^{(i)}$. Take $N \geq 1000$ and $\epsilon \leq .05$, so that you retain at least 50 samples for the posterior.

5. Make a histogram of your posterior samples. Report the posterior mean, the .025 quantile of the posterior, and the .975 quantile of the posterior.

Submission parameters

Submit two files:

- A pdf writeup containing the answers to the questions.
- A file containing the code you used.

References

[1] Coleman, J., Katz, E., and Menzel, H. The diffusion of an innovation among physicians. *Sociometry 20*, 4 (1957), 253–270.