COMP90051 Statistical Machine Learning
Semester 2, 2015

Workshop Week 6: Hands-on PGMs with Stan

August 29, 2015

1 Practical Bayesian Inference with Stan

Stan Tutorial Exercises

1. Follow the instructions at http://mc-stan.org/interfaces/rstan.html to install rstan on your computer. This will allow you to access Stan within R. *(Alternatively, if you wish to work in another language, there are a list of alternative interfaces to Stan at http://mc-stan.org/interfaces/)*

2. Follow the instructions to work through the **8schools example** in the "getting started" guide. See if you can work out what each part of the Stan file means. Make sure you can run this example in R (or in your alternative programming environment).

3. Download the **"nuclear.stan" file** on the LMS. This is a Stan file for the nuclear power plant model discussed in lectures. Explore this file, and try to work out how it relates to the graphical model. *Note: Unlike the model discussed in lectures, this model assumes the table probabilities for each node in the PGM is randomly distributed, rather than a fixed parameter. In other words, we have created the Bayesian version of the model, where each of these probabilities (a parameter in the unit interval [0,1]) has a Beta prior distribution. The Beta distribution is the "natural" distribution to place over probabilities since it is a distribution over the interval itself, but also because it is a mathematically-similar form to the Bernoulli distribution (they are what is called "conjugate" to each other).*

4. Download the **"nuclear-data-full.dat" file** from the LMS. This file contains a number of observations of the nuclear power plant variables. Each observation corresponds to a Boolean value (true or false) for each of the five variables. *(The file is comma-delimited, with a header row and five columns one per variable).*

5. Following a similar procedure as in the 8schools example, load the nuclear data observations into an appropriate R data structure (or respective data structure in language of choice), and execute the "nuclear.stan" model on this dataset. Try to interpret the resultant distributions for each parameter. *Hint: An example correctly-formated list structure in R holding three observations is list(N=3, HT=c(1,1,0), FG=c(0,1,0), FA=c(0,0,1), HG=c(1,0,0), AS=c(0,1,0), ALPHA=1, BETA=1). Second hint: we advice setting iter=100 and chains=1 in the call to Stan, as the default 4 separate chains of 1000 iterations each can be slow to converge.*

6. Try to interpret the results of fitting the model to data, and fill in the probability tables for each vertex in the PGM.

7. **Challenge Question** Now suppose we have another set of observations for the same model, but the HG (high gauge) observations are hidden. Investigate how the provided Stan model could be edited to deal with this new data format. Try running your edited Stan model on the provided dataset by deleting the HG observations. *Hint: This is a challenging question because Stan does not allow integer parameters. Therefore you must come up with a way to represent the model without introducing Boolean parameters for the latent variables. You may find the function "increment_log_prob" useful for doing this.*