Workshop 11: Stan Modelling Language

COMP90051 Statistical Machine Learning
Semester 2, 2017

1 Introduction to Stan

Stan is a modelling system that allows one to define a statistical model, and then use probabilistic and statistical inference to perform parameter estimation and predictions. In order to use Stan, one first creates a text file with model description. The description should follow a certain syntax (Stan file format). One then uses a common programming language, such as R to load the model and perform analysis. As the first step, briefly familiarise yourself with the syntax for model description using the first part of this tutorial (from Machine Learning Summer School 2014, Reykjavik). Don't spend too much time at this point, as you are going to learn Stan through exercisers below.

In the following, we will focus on using Stan via the interface to R language. This interface is called RStan. Don't worry if you are not familiar with R, we will only need minimum number of commands to get Stan working. Your tutor will be able to help you with R. Ask you tutor whether you need to go through installation instructions.

2 Eight Schools Examples

Use the Example 1 to learn how Stan can automatically estimate parameters from data. This model is related to a study of coaching effects in eight schools [1]. Model description is available in the $8schools.stan$ file.

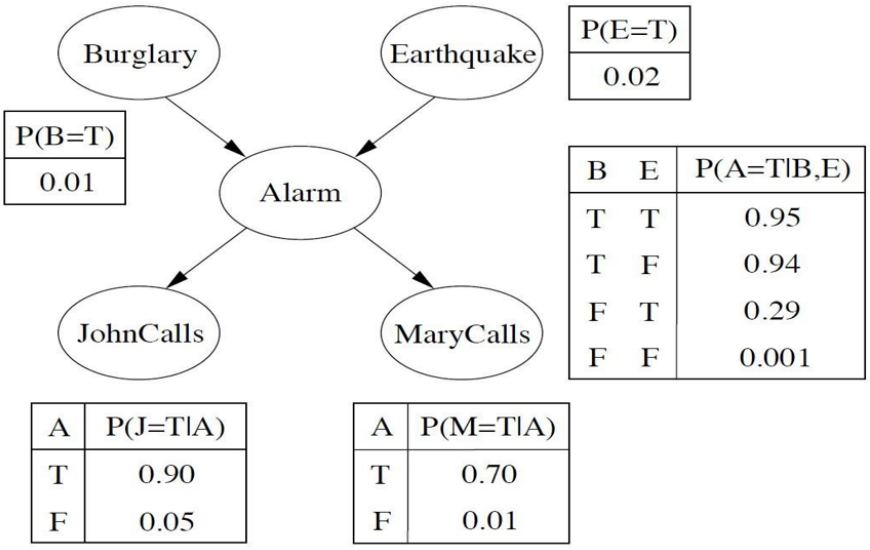
3 Linear Regression Stan Model

Use Stan to explore the Bayesian version of linear regression. Explore and interpret resultant distributions (posterior) for each parameter.

4 Nuclear Power Plant

Now you will use Stan to explore the Bayesian Version of the Nuclear Power Plant example from the lectures. In the Bayesian version probabilities for each node (i.e., model parameters) are randomly distributed rather than being fixed numbers. Each of these probabilities follows a Beta prior distribution. The Beta distribution is a popular choice for such type of parameters, among other reasons, because it is defined on the interval between 0 and 1.

Read model description (*nuclear.stan*) and make sure that you understand it. Load the data (*nuclear.txt*), and initialize the corresponding variables (from the data section in the model description). In R, *read.table* command can be used to read a table from a text file. Now execute the model similarly to the Eight Schools example. Explore and interpret resultant distributions (posteriors) for each parameter.



5 Variable Elimination Inference Example

Last workshop, we have covered query answering using the enumeration approach. Use the above alarm system graphical model to answer the following query: $P(E = T|J = T, M = T)$ by using Variable Elimination Approach.

6 Linear Mixed Models

If time permits, use an example from computational linguistics to see how Stan can fit parameters from two linear mixed models. Models description and the data are provided ($fixEf.stan$, $ranInt.stan$, and $gibsonwu2012data.txt$). These models are taken from a recent paper illustrating use of Stan for problems of computational linguistics [2]. The paper is rather lengthy, but you do not need to read everything for the purpose of this tutorial. You will only need to look at the beginning of Section 2 (Example 1, ...), as well as Subsections 2.1 and 2.2. Do not worry, if you do not understand concepts in "The scientific question" paragraph. Instead, focus on the data structure and mathematical model description.

Explore the description in the *.stan* files and relate it to the explanation given in the paper. Note that $fixEf.stan$ model is explained in Subsection 2.1, and $ranInt.stan$ is explained in Subsection 2.2.

References

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2 edition, 2003.

[2] T. Sorensen and V. Shravan. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint*, 1506:06201, 2015.