DATA MINING W4240

HOMEWORK 4 QUESTIONS

November 23, 2010

Professor: Frank Wood

Preliminary Instructions

1. Download the skeleton code for the assignment at http://www.stat.columbia.edu/∼fwood/w4240/Homework/index.html
2. Unzip the downloaded material in an appropriate folder, something like w4240/hw5/
3. Open MATLAB and navigate to the folder containing the downloaded material

In this homework you need to simulate from the posterior distribution of two different models using a Markov chain. The first model is a Bayesian logistic regression model and the second is a model known as latent Dirichlet allocation (LDA). The deliverable for the homework includes both code and a short write up of the results.

1. (50 points) For this problem you need to analyze the results of a study on arthritis using a Bayesian logistic regression model. The model is

$$\begin{aligned} \frac{1}{\sigma} &\sim \text{Gamma}(a, b) \\ \beta &\sim \text{Normal}(\vec{\mu}, \sigma \mathbf{I}_8) \\ P(Y = 1) &\sim \frac{e^{X\beta}}{1 + e^{X\beta}}. \end{aligned}$$

Answer the following questions using a sample from the joint posterior distribution of the latent parameters generated using a MCMC sampler.

1. What is the expected probability of improvement for an 80 year old male receiving treatment?
2. What is the expected probability of improvement for an 80 year old male not receiving treatment?
3. What is the expected treatment effect, on the probability scale, for an 80 year old male?

4. What is the posterior probability that the treatement is beneficial for 80 year old males?

You need to answer these questions in a seperate document, preferably in LaTeX. Regardless of what you use to write the report your submission will need to be a PDF file. You should include diagnostic plots to demonstrate that your Markov chain converged as well as what burn in you chose to use and why.

2. (50 points) For this problem you will estimate the LDA model on a corpus consisting of abstracts from NIPS papers over the last several years. After you estimate the model you again need to use the output to answer questions about the corpus. The data for this model are in bagofwords_nips.mat. When you load the data you will load two variables, **DS**, and **WS**. The variable **WS** contains a row vector of numbers, which correlate to words. The variable **DS** contains a row vector of the same length with a number indicating which document number the word comes from. The datasets title_nips.mat contains the list of document titles and words_nips.mat contains a translation of integers to words. Fit the model with 20 topics using an MCMC sampler in the collapsed representation and choose the sample with the highest joint log likelihood to answer the following questions :

1. What are the top ten most probable words for each of the 20 topics?

2. One can consider the distribution over topics as a low dimensional representation of a document. We can use the dot product between topic distributions for two documents as a similarity metric. What are the ten most similar documents to the first document, " Connectivity Versus Entropy"?

Again, submit your answers in a seperate LaTeXwrite up. You should include a trace plot of the joint log likelihood of the model to show convergence of the sampler. **BE AWARE THAT RUNNING THE SAMPLER ON THE ENTIRE CORPUS WILL BE SLOW AND YOU NEED TO START WORKING ON THIS HW EARLY BECAUSE YOU MAY NEED TO RUN THE SAMPLER FOR A COUPLE OF DAYS**. You can save the output once the sampler has run using the save command in **MATLAB** so you don't have to re-run the sampler once you have good results.

Submitting your HW

You must complete this HW assignment on your own, you are not permitted to work with any one else on the completion of this task. Your grade will reflect your ability to implement a working version of the procedure. The write up you submit **MUST BE A PDF FILE**.

1. Send an email to w4240.fall2010.stat.columbia.edu@gmail.com
2. Attach updated Matlab files and your LaTeXwrite up
   (a) write_up.pdf
   (b) joint_log_lik.m
   (c) sample_topic_assignment.m
   (d) log_liklihood_prior.m
   (e) logistic_log_likelihood.m
   (f) sample_beta.m
   (g) sample_sigma.m

   It is imperative that the names be exactly as described here. There should be no folders attached. You may attach other MATLAB code files if they act as utility functions for the other programs.
3. The subject will be exactly your Columbia UNI followed by a colon followed by hw5. For example, if the TA were submitting this homework the subject would read **nsb2130:hw5**
4. If you submit hw more than once, later files will overwrite earlier files.