LINKÖPING UNIVERSITY
Department of Computer and Information Science
Division of Statistics
Mattias Villani

2014-05-27
Advanced Bayesian Learning
PhD course

Bayesian Nonparametrics - Computer Lab

Deadline: The day before midsummer's eve (June 19)
Teacher: Mattias Villani
Grades: Pass/Fail
Submission: By email to Mattias Villani, use message header: 'ABL - BayesianNP'

The lab is to be reported in a concise report.
You have to write your own code to solve the problems, in whatever language you like.
You are not allowed to use existing toolboxes
Existing basic functions for matrix algebra, random number generators, pdfs/cdfs etc are allowed.
Attach your code to the email as separate executable files.

---

1. **Dirichlet process prior**. Let y_i be iid observations from some unknown probability distribution P. Let $P \sim DP(\alpha P_0)$ a priori, where P_0 is the $N(0,1)$ base measure.

   (a) Plot the empirical CDF (cumulative distribution function) for the Galaxy velocity data (as given in the file GalaxyData.dat).

   (b) Use the stick-breaking construction to simulate draws of P from the *prior* of $P \sim DP(\alpha P_0)$. Plot the CDF for some of the draws in a graph to illustrate the prior variation. Plot also the prior mean of P. Try this for $\alpha = 1$, $\alpha = 10$ and $\alpha = 100$.

   (c) Use the stick-breaking construction to simulate draws of P from the *posterior* of P based on the Galaxy data. Plot the CDF for some of the posterior draws in a graph to illustrate the posterior variation. Plot also the posterior mean of P. Again, try this for $\alpha = 1$, $\alpha = 10$ and $\alpha = 100$. Compare with the prior.

2. **Dirichlet process mixture**. Consider again the Galaxy velocity data, but now modelled by an infinite DP mixture model with a normal/Gaussian density kernel

$$\mathcal{K}(y_i|\theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - \mu_i)^2\right)$$

   where $\theta_i = (\mu_i, \sigma_i^2)$ are the parameters for the ith observation y_i. Let $(\mu_i, \sigma_i^2) \sim P$ where $P \sim DP(\alpha P_0)$ and P_0 is the conjugate prior

$$\mu_i|\sigma_i^2 \sim N(\mu_0, \sigma_i^2/\kappa_0) \text{ and } \sigma_i^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2).$$

   Set $\mu_0 = \sigma_0^2 = 20$ (I am cheating a bit here and using that the sample mean and sample variance are both close to 20). Choose your own values for $\kappa_0 > 0$ and $\nu_0 > 0$.

(a) Implement the Blocked Gibbs sampler on page 552 in BDA3. The results from Page 67-69 from the BDA3 book will be useful (scanned pages will be posted on the course web page).

(b) Analyze the Galaxy data using the Blocked Gibbs sampler in 2a) above. Plot the posterior distribution of the number of non-empty components. Plot a regular (non-Bayesian) histogram of the data and overlay the fitted DPM density. Investigate the effect of α by performing the analysis with $\alpha = 1$, $\alpha = 10$ and $\alpha = 100$.

(c) Now treat α as unknown with prior $\alpha \sim Gamma(a_\alpha, b_\alpha)$. Add an updating step for α in the Blocked Gibbs sampler (see page 553 in BDA3). Analyze the Galaxy data again using some suitable values for a_α and b_α. Plot the prior and posterior of α. Note that BDA3 uses the so called *rate parametrization* of the Gamma density where if $X \sim Gamma(a_\alpha, b_\alpha)$ then the pdf is of the form

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Good luck! All problems have solutions with probability one!