COS 513: Gibbs Sampling

Matthew Salesi

December 6, 2010

1 Overview

Concluding the coverage of Markov chain Monte Carlo (MCMC) sampling methods, we look today at *Gibbs sampling*. Gibbs sampling is a simple and widely used method for generating random samples from a joint distribution over several variables when this distribution is not known and/or is difficult to calculate. Instead, Gibbs sampling draws from the conditional distributions of the variables in a manner that approximates the joint distribution over time. This makes Gibbs sampling particularly useful for approximate inference in Bayesian networks.

2 Sampling Procedure

Consider a vector $\boldsymbol{x}$ of K random variables

$$\boldsymbol{x} = \langle x_1, x_2, \ldots, x_K \rangle$$

and a set of observed data $\mathcal{D}$. Gibbs sampling operates iteratively in the following manner: For each iteration, sample from the conditional distribution of x_k given $\boldsymbol{x}_{-k}$ (the vector of all variables except x_k) for $k = 1, 2, \ldots K$.

$$x_k \sim P(x_k \mid \boldsymbol{x}_{-k}, \mathcal{D}) \text{ for } k = 1 \ldots K$$

It can be shown that this method of sampling defines a Markov chain on $\boldsymbol{x}$ whose stationary distribution is $P(\boldsymbol{x} \mid \mathcal{D})$. Thus, as we gather more samples, we tend towards the joint distribution $P(\boldsymbol{x} \mid \mathcal{D})$.

3 Relation to Metropolis-Hastings

Gibbs sampling is in fact a specific case of the more general Metropolis-Hastings algorithm (previously discussed in class). Consider running Metropolis-Hastings with the transition probabilities $B_k = P(x_k \mid \boldsymbol{x}_{-k})$, i.e., the univariate conditional probabilities of each variable given all other variables. This implies that the selection factor A_k is

$$A_k(\boldsymbol{x}') = \min\left(1, \frac{P(\boldsymbol{x}')B_k(\boldsymbol{x}', \boldsymbol{x})}{P(\boldsymbol{x})B_k(\boldsymbol{x}, \boldsymbol{x}')}\right)$$

We may decompose the B_k's in the second argument of the min above as follows:

$$\frac{P(\boldsymbol{x}')B_k(\boldsymbol{x}', \boldsymbol{x})}{P(\boldsymbol{x})B_k(\boldsymbol{x}, \boldsymbol{x}')} = \frac{P(\boldsymbol{x}'_{-k})P(x'_k \mid \boldsymbol{x}'_{-k})P(x_k \mid \boldsymbol{x}_{-k})}{P(\boldsymbol{x}_{-k})P(x_k \mid \boldsymbol{x}_{-k})P(x'_k \mid \boldsymbol{x}_{-k})}$$

Since the probabilities involving x_k and x'_k are the same for all k, the terms in the numerator and denominator directly cancel; therefore, the entire expression is equal to 1. Gibbs sampling is thus equivalent to the Metropolis-Hastings algorithm where the quantity A_k is always equal to 1, a significant simplification.

4 Example

Consider the Bayesian mixture model represented by the graphical model in Fig. 1:

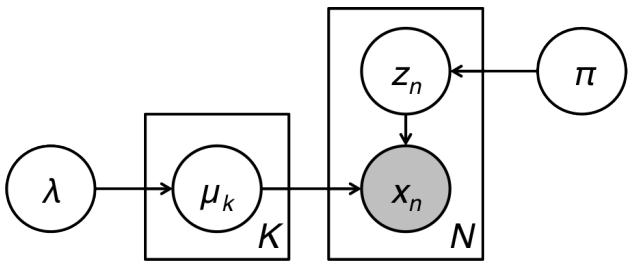

Figure 1: A Bayesian mixture model

In this model, we suppose that our data points $\boldsymbol{x}_{1:N}$ are drawn from the mixture of distributions defined by the means $\boldsymbol{\mu}_{1:K}$. In the Bayesian tradition, we also place a prior distribution over the means $\boldsymbol{\mu}_{1:K}$ with the

hyperparameter λ: $\mu_k \sim N(0, \lambda)$. We define $\boldsymbol{z}_{1:N}$ as the K-dimensional vectors of indicators that represent the mixture component from which each data point is drawn; that is, x_n is drawn from μ_k when $(z_n)_k = 1$; all other $(z_n)_i$ where $i \neq k$ will be 0. Thus, we model our data by a two-step random process: first, a distribution is selected from the mixture (parameterized by π), and then the data points are sampled from this distribution:

$$\begin{aligned} z_n &\sim \text{Multinomial}(\pi) \\ x_n \mid z_n, \boldsymbol{\mu}_{1:K} &\sim N(\mu_{z_n}, \sigma^2) \end{aligned}$$

An "expanded" example of the model can be seen in Fig. 2.

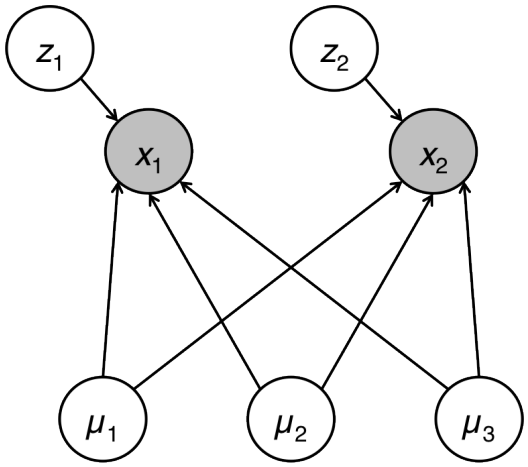

Figure 2: An expanded example of the model

In this example, we might be interested in computing approximations of various distributions, such as the conditional distribution of the mixture component location given the example data, $P(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{x}_{1:N})$. While exact methods are difficult for this type of problem, approximate inference sufficiently simplifies the process.

Suppose that we have sampled $S^{(m)} = \langle \boldsymbol{\mu}_{1:K}, \boldsymbol{z}_{1:N} \rangle$ from the distribution $P(\boldsymbol{\mu}_{1:K}, \boldsymbol{z}_{1:N} \mid \boldsymbol{x}_{1:N})$. We may use these samples for approximate inference:

$$\frac{1}{M} \sum_{m=1}^{M} \delta_{S^{(m)}} \boldsymbol{\mu}_{1:K}$$

For generating samples of this nature, we can use the Gibbs sampler with the appropriate univariate conditional probabilities:

$$(1) \qquad P(\mu_i \mid \boldsymbol{\mu}_{-i}, \boldsymbol{z}, \boldsymbol{x}) \propto P(\mu_i) \prod_{n:\, z_n = i} P(x_n \mid \mu_i)$$

$$(2) \qquad P(z_n \mid \boldsymbol{z}_{-n}, \boldsymbol{\mu}, \boldsymbol{x}) \propto P(z_n) P(x_n \mid \mu_{z_n})$$

For (1), this approximation is computable because we have chosen the prior $P(\mu_i)$ to be conjugate to the likelihood $P(x_n \mid \mu_i)$. For (2), we have $P(z_n)P(x_n \mid \mu_{z_n}) = P(z_n \mid x_n, \boldsymbol{\mu}_{1:K})$, a K-way multinomial. By iteratively sampling from (1) and (2), we may obtain the necessary estimates for calculating probabilities of interest.

5 Conclusions

Many issues and possible extensions may arise when using Gibbs sampling. The process of *collapsing* allows one to integrate out hidden variables in the conditional distributions and improve computation efficiency, as in the following example:

$$\begin{aligned} P(z_n \mid \boldsymbol{z}_{-n}, \boldsymbol{x}) &\propto \int_{\mu} P(z_n, \boldsymbol{\mu}_{1:K} \mid \boldsymbol{z}_{-n}, \boldsymbol{x}) \\ &= \int_{\mu} P(z_n \mid \boldsymbol{\mu}_{1:K}, \boldsymbol{z}_{-n}, x) P(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{z}_{-n}, \boldsymbol{x}) \\ &= P(z_n) \int_{\mu} P(x_n \mid \mu_{z_n}) P(\mu_{z_n} \mid \boldsymbol{x}_{m:\, z_m = z_n}) \end{aligned}$$

Collapsed Gibbs sampling is an example of a more general technique known as *Rao-Blackwellizaiton*, by which general statistical estimators of an unknown quantity can be improved using a sufficient statistic for that quantity.

Diagnostics for convergence are important to consider, since the sampling process must be close to converging in order to obtain reasonably accurate results. It is possible that better choices for the Metropolis-Hastings transition probabilities than the univariate conditionals used in Gibbs sampling may lend themselves to faster convergence; however, Gibbs sampling is useful at least for a "first try."

For more information on Gibbs sampling and other MCMC methods, see the book *Introducting Monte Carlo Methods with R* (Robert & Casella, 2004).

As an additional note, there exists software for laying out graphical models that can determine the Gibbs sampler for you—see BUGS, JAGS, and HBC.