DS 102 Discussion 5

Wednesday, Sep 30, 2020

The past few lectures have looked at how to perform Bayesian inference using Markov Chain Monte Carlo sampling methods, such as Gibbs sampling. The goal of Bayesian inference is to get the posterior distribution of the parameters given the data, $\mathbb{P}(\theta \mid X)$. Often this is difficult to derive in closed form, so instead we'll try to sample from it.

1. **Rejection Sampling**

   The rejection sampling method generates sampling values from a target distribution X with arbitrary probability density function $f(x)$ by using a proposal distribution Y with probability density $g(x)$. The idea is that one can generate a sample value from X by instead sampling from Y and accepting the sample from Y with probability $f(x)/(Mg(x))$, repeating the draws from Y until a value is accepted. M here is a constant, finite bound on the likelihood ratio $f(x)/g(x)$, satisfying $1 < M < \infty$ over the support of X; in other words, M must satisfy $f(x) \leq Mg(x)$ for all values of x. Note that this requires that the support of Y must include the support of X—in other words, $g(x) > 0$ whenever $f(x) > 0$.

   The validation of this method is the envelope principle: when simulating the pair $(x, v = u \cdot Mg(x))$, one produces a uniform simulation over the subgraph of $Mg(x)$. Accepting only pairs such that $u < f(x)/(Mg(x))$ then produces pairs (x, v) uniformly distributed over the subgraph of $f(x)$ and thus, marginally, a simulation from $f(x)$.

   Under this scheme, what is the probability that we accept a sample, i.e. the probability $\mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right)$? What is the largest probability we can get by changing M?

   Solution:

   The unconditional acceptance probability is the proportion of proposed samples which are accepted, which is

$$\begin{aligned}
\mathbb{P}\left(U \le \frac{f(Y)}{Mg(Y)}\right) &= \mathrm{E}\,\mathbf{1}\left[U \le \frac{f(Y)}{Mg(Y)}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}\left[U \le \frac{f(Y)}{Mg(Y)}\right] | Y\right]\right] && \text{(tower property)} \\
&= \mathrm{E}\left[\mathbb{P}\left(U \le \frac{f(Y)}{Mg(Y)}\middle| Y\right)\right] \\
&= E\left[\frac{f(Y)}{Mg(Y)}\right] && \text{(U is uniform on $(0,1)$)} \\
&= \int\limits_{y:g(y)>0} \frac{f(y)}{Mg(y)} g(y)\,dy \\
&= \frac{1}{M} \int\limits_{y:g(y)>0} f(y)\,dy \\
&= \frac{1}{M} && \text{(since support of Y includes support of X)}
\end{aligned}$$

Since M must satisfy $f(x) \le Mg(x)$, the largest probability we can get is $\inf_x \frac{g(x)}{f(x)}$.

2. **Gibbs sampling for Gamma-Poisson model.**

We practice deriving one iteration of Gibbs sampling for the Gamma-Poisson model. When the dimension of the parameters is large, sampling from the posterior over *all* the parameters θ is also often difficult. The main insight behind Gibbs sampling is that it can be much easier to sample the posterior over just a *single* parameter, $\mathbb{P}(\theta_i \mid X, \theta_{-i})$ (where we use the index $-i$ to mean all indices except for i). Gibbs sampling then iterates through each parameter θ_i and samples from $\mathbb{P}(\theta_i \mid X, \theta_{-i})$. This loop is repeated, each time conditioning on the newly sampled values. Iterating through each parameter θ_i and sampling from $\mathbb{P}(\theta_i \mid X, \theta_{-i})$ is not the same thing as sampling from $\mathbb{P}(\theta \mid X)$. However, the good news is that given enough iterations, the former converges to the latter.

Consider the hierarchical Bayes model where

$$\begin{aligned}
\beta &\sim \mathrm{Gamma}(m, \alpha) \\
\theta_i \mid \beta &\sim \mathrm{Gamma}(k, \beta), \quad i = 1, \ldots, n \\
X_i \mid \theta_i &\sim \mathrm{Pois}(\theta_i), \quad i = 1, \ldots, n,
\end{aligned}$$

where the θ_i are independent of each other and the X_i are independent of each other. The β and θ_i are unknown parameters, and m, α, and k are fixed and known.

We'd like to infer the parameters β and the θ from the data X. That is, we'd like to sample from the posterior distribution $\mathbb{P}(\beta, \theta \mid X)$ using Gibbs sampling. This entails deriving the posterior of each parameter, conditioned on the data and all the other parameters.

(a) We'll start with β. Derive $\mathbb{P}(\beta \mid \theta_1, \dots, \theta_n, X_1, \dots, X_n)$.

Solution:

We have

$$\begin{aligned}
\mathbb{P}(\beta \mid \theta_1, \dots, \theta_n, X_1, \dots, X_n) &= \mathbb{P}(\beta \mid \theta_1, \dots, \theta_n) \\
&= \frac{\mathbb{P}(\beta) \prod_{i=1}^n \mathbb{P}(\theta_i \mid \beta)}{\int_0^\infty \mathbb{P}(b) \prod_{i=1}^n \mathbb{P}(\theta_i \mid b)\, db} \\
&\propto_\beta \beta^{m-1} e^{-\alpha\beta} \prod_{i=1}^n \beta^k e^{-\beta\theta_i} \\
&\propto_\beta \beta^{nk+m-1} e^{-(\alpha + \sum_{i=1}^n \theta_i)\beta} \\
&\propto_\beta \text{Gamma}(nk + m, \alpha + \sum_{i=1}^n \theta_i).
\end{aligned}$$

(b) Next, we'll look at each θ_i. Derive $\mathbb{P}(\theta_i \mid \beta, \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n, X_1, \dots, X_n)$

Solution:

$$\begin{aligned}
\mathbb{P}(\theta_i \mid \beta, \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n, X_1, \dots, X_n) &= \mathbb{P}(\theta_i \mid \beta, X_i) \\
&= \frac{\mathbb{P}(\theta_i, \beta, X_i)}{\mathbb{P}(\beta, X_i)} \\
&= \frac{\mathbb{P}(X_i \mid \theta_i, \beta)\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(\beta)}{\mathbb{P}(\beta, X_i)} \\
&= \frac{\mathbb{P}(\beta)\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(X_i|\beta, \theta_i)}{\mathbb{P}(\beta) \int_0^\infty \mathbb{P}(u|\beta)\mathbb{P}(X_i \mid \beta, u)\, du} \\
&= \frac{\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(X_i|\beta, \theta_i)}{\int_0^\infty \mathbb{P}(u|\beta)\mathbb{P}(X_i \mid \beta, u)\, du} \\
&\propto_{\theta_i} \theta_i^{k-1} e^{-\beta\theta_i} \theta_i^{X_i} e^{-\theta_i} \\
&\propto_{\theta_i} \theta_i^{X_i + k - 1} e^{-(\beta+1)\theta_i} \\
&\propto_{\theta_i} \text{Gamma}(X_i + k, \beta + 1).
\end{aligned}$$

(c) Using the posteriors you derived in the last two parts, write out the algorithm for the Gibbs sampler.

Solution:
Initialize $\beta^{(0)} \sim \text{Gamma}(m, \alpha)$ and $\theta_i^{(0)} \mid \beta = \beta^{(0)} \sim \text{Gamma}(k, \beta^{(0)})$ for all i.
For $t = 1, \ldots, T$ for some large stopping time T:

1. Start with $(\beta^{(t-1)}, \theta_1^{(t-1)}, \ldots, \theta_n^{(t-1)})$ from the previous iteration.
2. Sample $\beta^{(t)}$ according to

$$\beta^{(t)} \sim \mathbb{P}(\beta \mid \theta_1 = \theta_1^{(t-1)}, \ldots, \theta_n = \theta_n^{(t-1)}) = \text{Gamma}(nk + m, \alpha + \sum_{i=1}^{n} \theta_i^{(t-1)})$$

3. Sample the $\theta_i^{(t)}$ in parallel according to

$$\theta_i^{(t)} \sim \mathbb{P}(\theta_i \mid \beta = \beta^{(t)}, X_i = x_i) = \text{Gamma}(x_i + k, \beta^{(t)} + 1)$$

3. **Metropolis-Hastings (optional)**

This problem proves properties of the **Metropolis-Hastings Algorithm**.

Recall that the goal of MH was to draw samples from a distribution $p(x)$. The algorithm assumes we can compute $p(x)$ up to a normalizing constant via $f(x)$, and that we have a proposal distribution $g(x, \cdot)$. The steps are:

- Propose the next state y according to the distribution $g(x, \cdot)$.
- Accept the proposal with probability

$$A(x, y) = \min(1, \frac{f(y)}{f(x)} \frac{g(y, x)}{g(x, y)}).$$

- If the proposal is accepted, then move the chain to y; otherwise, stay at x.

The key to showing why Metropolis-Hastings works is to look at the **detailed balance equations**. Suppose we have a finite irreducible Markov chain on a state space $\mathcal{X}$ with transition matrix P. If there exists a distribution π on $\mathcal{X}$ such that for all $x, y \in \mathcal{X}$,

$$\pi(x)P(x, y) = \pi(y)P(y, x),$$

then π is a stationary distribution of the chain (i.e. $\pi P = \pi$).

(a) For the Metropolis-Hastings chain, what is $P(x, y)$ in this case? For simplicity, assume $x \neq y$.

Solution: $P(x, y)$ is the probability that we propose y with g and then accept it. Thus it is

$$P(x, y) = g(x, y)A(x, y) = g(x, y) \min(1, \frac{f(y)}{f(x)} \frac{g(y, x)}{g(x, y)})$$

(b) Show $p(x)$, our target distribution, satisfies the detailed balance equations with $P(x, y)$, and therefore is the stationary distribution of the chain.

Solution: We will check that detailed balance holds with $p(x)$ for a pair of states (x, y). Without loss of generality, assume that $f(y)g(y, x) \leq f(x)g(x, y)$, i.e. $A(x, y) = \frac{f(y)g(y,x)}{f(x)g(x,y)}$ (if this were not true, we could swap x and y). This means that $A(y, x) = 1$ and therefore that $P(y, x) = g(y, x)$. Then,

$$\begin{aligned}
p(x)P(x, y) &= p(x)g(x, y)A(x, y) \\
&= p(x)g(x, y)\frac{f(y)g(y, x)}{f(x)g(x, y)} \\
&= p(x)\frac{f(y)}{f(x)}g(y, x) \\
&= p(y)g(y, x) \\
&= p(y)P(y, x)
\end{aligned}$$

The second to last line comes from the fact that $p(x)$ and $f(x)$ are directly proportional. i.e. $f(x) = kp(x)$.

$$p(x)\frac{f(y)}{f(x)} = p(x)\frac{kp(y)}{kp(x)} = p(y)$$