

Bayesian inference II

The key problem with (the good) Monte Carlo techniques is that they require a proposal distribution q that is good *everywhere*. What we'd like to do is to have a proposal distribution q that is good locally. Of course, the question is "locally to *what?*"

In Markov Chain Monte Carlo techniques, we maintain a "random walk" of parameters over parameter space. That is, instead of drawing a bunch of samples $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(R)}$ independently from a proposal distribution, we will draw $\theta^{(r+1)}$ *conditioned* on θ^r . This introduces a few new problems (samples are no longer independent), but will enable us to define a better proposal distribution.

Metropolis-Hastings

Let $q(\theta' | \theta)$ be a proposal distribution. MH works as follows:

1. Initialize $\theta^{(1)}$
2. For $r = 1 \dots R - 1$,
 - (a) Draw θ' according to $q(\theta' | \theta^{(r-1)})$
 - (b) Compute acceptance probability:

$$a = \min \left\{ 1, \frac{p(\theta')q(\theta^{(r)} | \theta')}{p(\theta^{(r)})q(\theta' | \theta^{(r)})} \right\}$$
 - (c) If accepted, set $\theta^{(r+1)}$ to θ' ; otherwise, set to $\theta^{(r)}$

The idea behind the acceptance probability is as follows. We want to move to θ' if it has high $p()$ probability; we want to stay in $\theta^{(r)}$ if it has high $p()$ probability. If q unnecessarily favors θ' , we don't want to move there; if it unnecessarily favors $\theta^{(r)}$, we want to move away.

Gibbs

Gibbs sampling is a bit different from other sampling algorithms we've talked about. First, it only works in very particular cases. Pretty much, if you've constrained yourself to conjugate distributions, then it works. It also *only* works for multivariate distributions.

Let's say our parameters are a vector $\theta = \langle \theta_1, \dots, \theta_D \rangle$. We'll write θ_{-d} for θ without position d ; namely, $\theta_{-d} = \langle \theta_1, \dots, \theta_{d-1}, \theta_{d+1}, \dots, \theta_D \rangle$.

Now, we have to assume that we can directly sample from the distribution $p(\theta_d | \theta_{-d})$ for all d . While this seems strong, it's actually not that unheard of. We'll shortly see an example.

The Gibbs sampler works as follows:

1. Initialize $\theta^{(1)}$
2. For $r = 2 \dots R$,
 - (a) Set $\theta^{(r)} = \theta^{(r-1)}$

- (b) For each dimension d ,
 - i. Sample $\theta_d^{(r)} \sim p(\theta_d^{(r)} \mid \theta_{-d}^{(r)})$

Latent Dirichlet Allocation

We'll now explore one particular model: LDA. LDA is a probabilistic model of text. It posits that a document is composed of a mixture of topics. Eg., we might have a sports topic, an entertainment topic, a tech topic and a science topic. Something about blue-ray might be a mix of tech and science.

The model is formally specified as follows. We have a vocabulary over V words. We have K "topics." For each topic k , there is a multinomial β_k over V . The β s have a symmetric Dirichlet prior with concentration η . The corpus is over D documents. Each word w_{dn} in document d is assigned a discrete latent variable z_{dn} that specifies which topic ($1 \dots K$) this word comes from. Documents themselves have a mixture parameter θ that is a K -dimensional multinomial with symmetric (global) Dirichlet prior with concentration α .

There generative model is as follows:

1. Choose $\alpha \sim \text{Uni}(0, 10)$
2. For each topic $k = 1 \dots K$,
 - (a) Choose $\beta_k \sim \text{Dir}(\eta, \dots, \eta)$
3. For each document $d = 1 \dots D$,
 - (a) Choose $\theta_d \sim \text{Dir}(\alpha, \dots, \alpha)$
 - (b) For each word $n = 1 \dots N$,
 - i. Choose topic $z_{dn} \sim \text{Mult}(\theta_d)$
 - ii. Choose word $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$

Or, written as a hierarchical model:

$$\begin{aligned}
 \alpha &| && \sim \text{Uni}(0, 10) \\
 \beta_k &| \eta && \sim \text{Dir}(\eta, \dots, \eta) \\
 \theta_d &| \alpha && \sim \text{Dir}(\alpha, \dots, \alpha) \\
 z_{dn} &| \theta_d && \sim \text{Mult}(\theta_d) \\
 w_{dn} &| z_{dn}, \beta && \sim \text{Mult}(\beta_{z_{dn}})
 \end{aligned} \tag{1}$$

It turns out we can construct a simple Gibbs sampler for this model. A useful fact for Gibbs samplers is that if we have a graphical model, then $p(\theta_d \mid \theta_{-d})$ only depends on the *Markov blanket* of d . The Markov blanket contains d 's parents, d 's children, and the parents of d 's children.

Using the Markov blanket, we can easily see that α depends only on the θ s, β_k depends only on η , the w s and the z s, etc. We get the following Gibbs updates:

$$\begin{aligned}
p(\alpha \mid - \alpha) &= \mathcal{Uni}(\alpha \mid 0, 10) \prod_d \mathcal{Dir}(\theta_d \mid \alpha) \\
p(\beta_k \mid - \beta_k) &= \mathcal{Dir}(\beta_k \mid \eta) \prod_d \prod_n \mathcal{Mult}(w_{dn} \mid \beta_k)^{\mathbf{1}_{z_{dn}=k}} \\
p(\theta_d \mid - \theta_d) &= \mathcal{Dir}(\theta_d \mid \alpha) \prod_n \mathcal{Mult}(z_{dn} \mid \theta_d) \\
p(z_{dn} \mid - z_{dn}) &= \mathcal{Mult}(z_{dn} \mid \theta_d) \mathcal{Mult}(w_{dn} \mid \beta_{z_{dn}})
\end{aligned}$$

Now, we can apply conjugacy to obtain expressions for most of these:

$$\begin{aligned}
p(\beta_k \mid - \beta_k) &= \mathcal{Dir}(\beta_k \mid \eta + \sum_d \sum_n \mathbf{1}_{z_{dn}=1}, \dots, \eta + \sum_d \sum_n \mathbf{1}_{z_{dn}=K}) \\
p(\theta_d \mid - \theta_d) &= \mathcal{Dir}(\theta_d \mid \alpha + \sum_n \mathbf{1}_{z_{nd}=1}, \dots, \alpha + \sum_n \mathbf{1}_{z_{nd}=K})
\end{aligned}$$

For handling z_{dn} , we can just compute probabilities for all possible values and sample (z is discrete). For α , we can use numeric optimization techniques.