Lecture Notes 14: Markov Chain Mento Carlo Algorithm

Professor: Zhihua Zhang *Scribe:*

14 MCMC

In statistics, Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample of the desired distribution.

14.1 Metropolis-Hastings Algorithm

Suppose we want to sample $f(x)$ and given a proposal distribution g. The steps of the algorithm is:

1. Initialize $X^{(0)} = x^{(0)}$
2. for $t = 1, 2, \cdots$
3. Give $X^{(t)} = x^{(t)}$, sample a candidate value x^* from a proposal distribution $g(x|x^{(t)})$
4. Compute the M-H ratio $R(x^{(t)}|x^*)$ where

$$R(u, v) = \frac{f(v)g(u|v)}{f(u)g(v|u)}$$

5. Sample $x^{(t+1)}$ according to following:

$$x^{(t+1)} = \begin{cases} x^* & \text{with probality } \min(R(x^{(t)}|x^*), 1) \\ x^{(t)} & otherwise \end{cases}$$

MetropolisHastings algorithm resides in designing a Markov process (by constructing transition probabilities). Suppose X_1, X_2 are two neighbouring states. The derivation of the algorithm starts with the condition of detailed balance:

$$f(x_2)g(x_1|x_2) = f(x_1)g(x_2|x_1)$$

If the condition of detailed balance establishes, then $f(x)$ is the stationary distribution. To make this equation established, we introduce an acceptance coefficient R, namely,

$$f(x_2)g(x_1|x_2)R(x_2, x_1) = f(x_1)g(x_2|x_1)R(x_1, x_2)$$

Let $R(x_1, x_2) = \min(\frac{f(x_2)g(x_1|x_2)}{f(x_1)g(x_2|x_1)}, 1)$.
if $f(x_2)g(x_1|x_2) \geq f(x_1)g(x_2|x_1)$, $R(x_1|x_2) = 1$, $R(x_2, x_1) = \frac{f(x_1)g(x_2|x_1)}{f(x_2)g(x_1|x_2)}$, so, $f(x_2)g(x_1|x_2)R(x_2, x_1) = f(x_1)g(x_2|x_1)R(x_1, x_2)$.
if $f(x_2)g(x_1|x_2) \leq f(x_1)g(x_2|x_1)$, the result is same. So the detailed balance condition is satisfied.

14.2 Gibbs Sampling

Gibbs sampling, in its basic incarnation, is a special case of the Metropolis-Hastings algorithm. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Suppose we want to obtain k samples of $\mathbf{X} = (x_1, \ldots, x_n)$ from a joint distribution $p(x_1, \ldots, x_n)$. Denote the ith sample by $\mathbf{X}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$. We proceed as follows:

1. We begin with some initial value $\mathbf{X}^{(0)}$.

2. For each sample $i \in \{1 \ldots k\}$, sample each variable $x_j^{(i)}$ from the conditional distribution $p(x_j | x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_n^{(i-1)})$. That is, sample each variable from the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled.

The samples then approximate the joint distribution of all variables. Furthermore, the marginal distribution of any subset of variables can be approximated by simply examining the samples for that subset of variables, ignoring the rest. In addition, the expected value of any variable can be approximated by averaging over all the samples.

14.3 Bayes Linear Models

Consider a standard linear regression problem, in which for $i = 1, ..., n$ we specify the conditional distribution of y_i given a predictor x_i:

$$y_i = x_i^{\mathrm{T}} b + \varepsilon_i$$

where ε_i are independent and identical normally distributed random variables:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Let $p(b, \sigma^2) = p(\sigma^2)\rho(b|\sigma^2)$, where $\rho(\sigma^2)$ is an inverse-gamma distribution.

$$\begin{aligned} p(b, \sigma^2) &= p(\sigma^2)\rho(b|\sigma^2) \\ &= N(\mu, \sigma^2 v) IG(\alpha, \beta) \\ &= \frac{\beta^\alpha \sigma^{2(-\alpha+\frac{p}{2}+1)}}{(2\pi)^{\frac{p}{2}} |v|^{\frac{1}{2}} \Gamma(\alpha)} \exp\left(-\frac{(b-\mu)^T V^{-1}(b-\mu) + 2\beta}{2\sigma^2}\right) \end{aligned}$$

Let $D = \{(x_i, y_i)\}_{i=1}^n$, thus

$$\begin{aligned} p(b, \sigma^2 | D) &= \frac{p(D|b, \sigma^2) p(b, \sigma)}{p(D)} \\ &\propto \prod_{i=1}^{n} p(\varepsilon_i | b, \sigma^2) p(b, \sigma) \end{aligned}$$

14.4 Bayes Classification

Let $y \in (0,1)$, suppose $p(y|b) = (\mu(b))^y(1-\mu(b))^{1-y}$, where $\mu(b) = h(x^T b)$.
h is a function range from 0 to 1, such as sigmoid function($h(x) = \frac{\exp(x)}{1+exp(x)}$) or Gaussian CDF($h(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})dt$).(probit model)

If give b a Gaussian prior, namely, $b \sim N(\mu, v)$. We have

$$\begin{aligned} p(b|D) &\propto p(D|b)p(b) \\ &\propto \prod_{i=1}^{n} (\mu_i(b))^{y_i}(1-\mu_i(b))^{1-y_i} p(b) \end{aligned}$$

It is possible to motivate the probit model as a latent variable model. Suppose there exists an auxiliary random variable. Let $z = x^T b + \varepsilon$, where $\varepsilon \sim N(0,1)$. So,

$$p(y=1|z,b) = \begin{cases} 1 & z > 0 \\ 0 & otherwise \end{cases}$$

and $p(y=0|z,b) = 1 - p(y=1|z,b)$. Hence,

$$\begin{aligned} p(y=1|b) &= p(y=1|z>0,b)p(z>0|b) + p(y=1|z\le 0,b)p(z\le 0|b) \\ &= p(z>0|b) \\ &= p(\varepsilon > -x^T b|b) \\ &= \int_{-x^T b}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})dt \\ &= \int_{-\infty}^{x^T b} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})dt \\ &= \Phi(x^T b) \end{aligned}$$