베이즈 통계와 계산

이광민

Seoul National University

my1989@snu.ac.kr

November 28, 2018

Overview

1. 베이즈 추론의 이해

2. 사후분포의 계산
   - 몬테카를로 방법
   - 깁스추출법
   - 메트로폴리스-헤이스팅스 알고리듬
   - MCMC 진단

베이즈 추론의 이해

통계적 추론의 문제

- 모수 (θ) : 자연의 법칙을 나타내는 미지의 값
- 관측치 (x) : 확률분포 $p(x|\theta)$를 따르는 확률변수
- 목표 : x를 근거로 모수 θ에 대한 추론을 하고자 한다.

예시

압정을 던졌을 때, λ모양이 나올 확률을 구하고자 한다.
$\Longrightarrow$ λ모양이 나올 확률을 모수 θ로 설정한다.

예시

관측치 혹은 자료는 다음과 같이 압정을 10번 던졌을때 관측된 값이다.

베이지안 추론

베이지안 추론은 다음의 세 가지 요소로 이루어져 있다.

- 사전분포 : θ의 분포로서 자료를 보기 전의 분석자의 모수(θ)에 관한 불확실성(믿음의 정도)을 나타낸다. $\pi(\theta)$로 나타낸다.
- 자료분포 : $x \sim p(x|\theta)$.
- 사후분포 : 자료가 주어졌을 때 θ의 확률분포로 자료를 본 후의 분석자의 θ에 관한 불확실성을 나타낸다. $\pi(\theta|x)$로 나타낸다.

예시

압정 예시)

- 모수 θ : λ가 나올 확률
- 관측치 x : 압정을 10번 던졌을 때 λ가 나온 횟수

예시

위 문제에 대한 베이즈 추론은 다음과 같이 수행한다.

- 사전분포 : $\theta \sim Uniform(0, 1)$.
  자료 x를 보기 전에, 모든 θ값이 동일한 가능성을 갖고 있다고 생각하여 균등분포를 부여한다.
- 가능도 : $x|\theta \sim Binomial(n, \theta)$.
- 사후분포 : 사전분포와 가능도에 대해 베이즈 정리 적용.

베이즈 법칙 : 사후분포의 계산방법

베이즈 정리는 사후분포를 계산하는 수학적 방법이다.

$$p(\theta, x) = \pi(\theta)p(x|\theta).$$

x가 주어졌을 때 θ의 사후분포는

$$\pi(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{\pi(\theta)p(x|\theta)}{p(x)},$$

여기서

$$p(x) = \int p(\theta, x)d\theta = \int \pi(\theta)p(x|\theta)d\theta.$$

예시

사전분포가 $\pi \sim U(0,1)$이고,
관측치에 대한 분포가 $x|\theta \sim Binomial(n,\theta)$일때,

- 사후분포

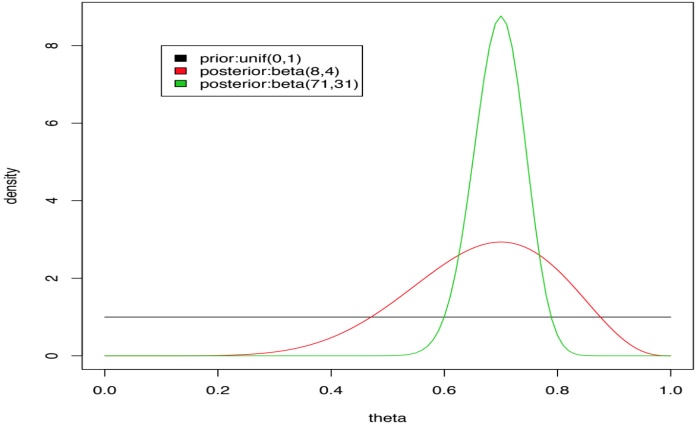
$$\begin{aligned} \pi(\theta|x) &\propto \pi(\theta) \times p(x|\theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^x (1-\theta)^{n-x} \end{aligned}$$

$$\theta|x \sim Beta(x+1, n-x+1).$$

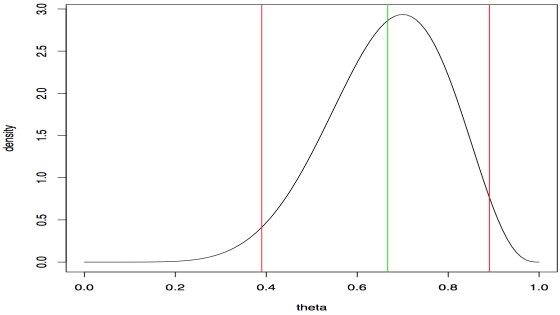
- 압정의 예시

$$\theta|x = 7 \sim Beta(8,4)$$

예시


prior:unif(0,1)
posterior:beta(8,4)
posterior:beta(71,31)
density
theta
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

베이즈 추정량 및 신뢰구간



- 베이즈 추정량은 사후분포의 한점 요약으로서 사후분포의 평균이다.
- 신뢰구간은 다음을 만족하는 (L, U) 이다.

$$P(L < \theta < U|x) = 1 - \alpha$$

베이즈 추정의 요약 및 문제

- 사전분포($\pi(\theta)$)와 관측치에 대한 분포($p(x|\theta)$)를 설정한다.
- 사전분포와 자료분포에 베이즈 정리를 적용하여 사후분포($\pi(\theta|x)$)를 구한다.
- 사후분포에 대한 요약으로서 베이즈 추정량과 신뢰구간을 다음과 같이 계산한다.

$$\int \theta\pi(\theta|x)d\theta.$$

$$\{t : \int_{-\infty}^{t} \pi(\theta|x)d\theta = \alpha\}.$$

베이즈 추정의 요약 및 문제

위 과정을 실제로 적용할 때 두가지 문제점이 있다.

- $\pi(\theta|x)$를 계산하기 어려움.
- $\pi(\theta|x)$를 포함하는 적분을 수행하기 어려움.

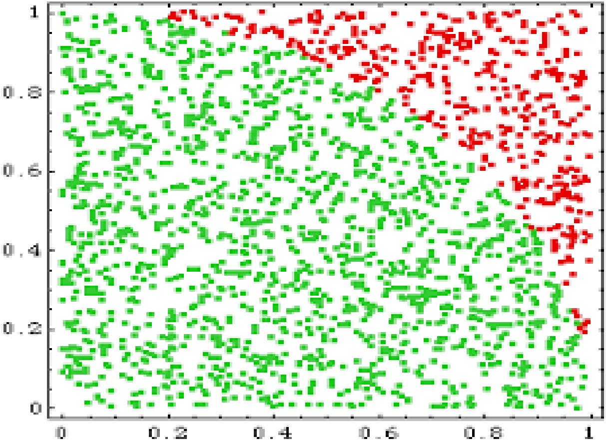
이를 수치적 방법으로 해결함.

사후분포의 계산

- **몬테카를로 방법**
- 깁스추출법
- 메트로폴리스-헤이스팅스 방법
- MCMC 진단

몬테카를로 방법

- 목적
  $X \sim g(x)$일 때, 적분값
  $$I = E[f(X)] = \int f(x)g(x)dx$$
  을 구하고자 한다.
- 알고리듬
  $X_1, \ldots, X_n \overset{iid}{\sim} g$이라 하자. 다음과 같이
  $$\hat{I} \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$
  를 계산하고, $\hat{I}$를 I를 추정하는데 사용한다.


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

몬테카를로 방법의 정당화

- 근거
  강대수의 법칙에 의해
$$\hat{I} \longrightarrow I, \text{ a.s.}$$
- 추정오차
$$SE(\hat{I}) = \sqrt{\frac{v}{n}},$$
  여기서,
$$v = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(X_i) - \frac{1}{n}\sum_{j=1}^{n} f(X_j)\right)^2.$$
- 오차비율 $n^{-1/2}$는 적분의 차원에 의존하지 않는다. 이는 다른 수치 방법들에 비한 장점이다.

몬테카를로 방법

- 문제
  $X|\theta \sim N(\theta, 1)$이고, 사전분포 $\theta \sim Cauchy(0, 1)$일 때, 사후분포의 평균을 구해보자.
- 풀이
  사후분포의 밀도함수는

$$\pi(\theta|x) \propto \frac{1}{1+\theta^2} e^{-\frac{1}{2}(x-\theta)^2}$$

이다. 따라서, 사후분포의 평균은

$$E(\theta|x) = \frac{\int \frac{\theta}{1+\theta^2} e^{-\frac{1}{2}(\theta-x)^2} d\theta}{\int \frac{1}{1+\theta^2} e^{-\frac{1}{2}(\theta-x)^2} d\theta}.$$

위 적분의 결과는 수식으로 주어지지 않는다.

몬테카를로 방법

- 알고리듬
  $\theta_1, \ldots, \theta_m \overset{i.i.d.}{\sim} N(x, 1)$를 추출한다. 그리고 분자와 분포의 적분을 각각 몬테카를로 방법으로 추정한다. 최종추정량은

$$\hat{\theta}^m \approx \frac{\sum_{i=1}^{m} \frac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^{m} \frac{1}{1+\theta_i^2}}$$

와 같이 주어진다.

중요도추출

문제
$X \sim g(x)$이고,

$$I = \int f(x)g(x)dx$$

를 구하려고 한다. 그런데 $g(x)$에서 표본을 추출하는 것은 어렵고, $\pi(x)$에서 표본을 추출하는 것은 쉽다고 하자.

중요도추출

1. $X_1, \ldots, X_n \overset{iid}{\sim} \pi(x)$를 추출한다.
2. 가중치 $w_i = \frac{g(x_i)}{\pi(x_i)}$, $i = 1, \ldots, n$ 를 계산한다.
3. I를 다음의 값으로 추정한다.

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} w_i f(x_i).$$

알고리듬의 근거-불편추정량

$$
\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n} w_i f(X_i)\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{g(X)}{\pi(X)} f(X)\right) \\
&= \int \frac{g(x)}{\pi(x)} f(x)\pi(x)dx \\
&= \int g(x)f(x)dx
\end{aligned}
$$

중요도추출

$$
\begin{aligned}
Var(\hat{I}) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}f(X_i)\Big) \\
&= \frac{1}{n}Var\Big(\frac{g(X_i)}{\pi(X_i)}f(X_i)\Big) \\
&\leq \frac{1}{n}E\Big(\frac{g(X)}{\pi(X)}f(X)\Big)^2 \\
&= \frac{1}{n}\int \frac{g^(x)}{\pi^2(x)}f^2(x)\pi(x)dx \\
&= \frac{1}{n}\int \frac{g^2(x)}{\pi(x)}f(x)dx
\end{aligned}
$$

이에 따라 n이 커짐에 따라 중요도 추출에 의한 적분 추정량의 분산이 0 으로 수렴하고, $\hat{I} \longrightarrow I$임을 알수 있다.

- 몬테카를로 방법
- **깁스추출법**
- 메트로폴리스-헤이스팅스 방법
- MCMC 진단

깁스추출법

$$I = \int f(x)g(x)dx, x \in \mathbb{R}^d$$

- $g(x)$가 표본추출하기 어려운 고차원 확률분포인 경우.
- $g(x)$에서 직접 표본추출 하는 것 대신, 깁스추출법을 이용하여 난수를 발생하여 사용할 수 있다.

깁스추출법 알고리듬

$(x_1, \ldots, x_p)$의 밀도함수가 $g(x_1, \ldots, x_p)$이라 하자.
$x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$라 표시하자.

체계적 순차 (systematic sweep) 깁스추출 알고리듬

초기값 $(X_1^{(0)}, \ldots, X_p^{(0)})$을 선정한다. $t = 1, 2, \ldots$에 대하여

1. $X_1^{(t)} \sim g_{X_1|X_{-1}}(\cdot|X_2^{(t-1)}, \ldots, X_p^{(t-1)})$을 추출한다.
2. $X_2^{(t)} \sim g_{X_2|X_{-2}}(\cdot|X_1^{(t)}, X_3^{(t-1)}, \ldots, X_p^{(t-1)})$을 추출한다.

   $\vdots$

3. $X_j^{(t)} \sim g_{X_j|X_{-j}}(\cdot|X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$을 추출한다.

   $\vdots$

4. $X_p^{(t)} \sim g_{X_p|X_{-p}}(\cdot|X_1^{(t)}, \ldots, X_{p-1}^{(t)})$을 추출한다.

깁스추출법의 예

- 이변량 정규분포 이변량 정규분포

$$\pi = N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_2^2 \end{bmatrix} \right)$$

  를 정상분포를 갖는 깁스추출표본을 구하자.

- 이변량 정규분포의 조건부 분포

$$X_1|X_2 = x_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2))$$

$$X_2|X_1 = x_1 \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$$

  가 된다.

깁스추출법의 예

초기값 $(x_1^{(0)}, x_2^{(0)})$을 정한다. $t = 1, 2, \ldots,$ 에 대해서,

1. $x_1^{(t)} \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2^{(t-1)} - \mu_2), \sigma_1^2(1-\rho^2))$를 발생한다.
2. $x_2^{(t)} \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1^{(t)} - \mu_1), \sigma_2^2(1-\rho^2))$를 발생한다.

$(x_1^{(t)}, x_2^{(t)}), t \geq 0$를 이용해 정규분포의 적률과 분위수를 구한다.

$$\int h(x_1, x_2)\hat{\pi}(dx_1, dx_2) = \frac{1}{n}\sum_{t=1}^{n} h(x_1^{(t)}, x_2^{(t)})$$

$$P(X_1 \geq \hat{0}, X_2 \geq 0) = \frac{1}{n}\sum_{t=1}^{n} I(x_1^{(t)} \geq 0, x_2^{(t)} \geq 0)$$

깁스추출법 정당화

- 확률분포 g에 대해 깁스추출법으로 추출된 확률변수의 수열 $(X^{(0)}, X^{(1)}, \ldots)$을 g분포에서 샘플된 난수로 취급할수 있는가?

- 마코프 체인 이론에 의해
충분히 큰 m에 대해, X_m은 g에서 샘플링 한것과 같다.

- 따라서, 몬테카를로 방법과 같이
$(X^{(0)}, X^{(1)}, \ldots)$의 표본적률과 표본분위수는 $f(x_1, \ldots, x_p)$의 적률과 분위수로 수렴한다.

- 몬테카를로 방법
- 깁스추출법
- **메트로폴리스-헤이스팅스 방법**
- MCMC 진단

메트로폴리스-헤이스팅스 알고리듬

- 동기
  깁스추출법을 하기 위해 필요한 Full conditional 분포를 얻기 힘들 경우
- 목표
  주어진 목표 분포 $g(x)$를 극한분포로 하는 일반적인 순차적 샘플링 방법을 제시한다.

메트로폴리스-헤이스팅스 알고리듬

1. (초기화) $x^{(0)}$를 정한다.
2. (메트로폴리스-헤이스팅스 반복)
   $t = 1, 2, \ldots, m$에 대해서 다음을 수행한다.
   - (후보값과 균등확률변수 추출) 서로 독립이 되도록 x, u를 추출한다.
     $x \sim g(\cdot|x^{(t-1)})$
     $u \sim Unif(0, 1)$.
   - (합격확률의 계산)

$$\alpha(x^{(t-1)}, x) = \min\{1, \frac{\pi(x)g(x^{(t-1)}|x)}{\pi(x^{(t-1)})g(x|x^{(t-1)})}\}$$

     계산
   - $x^{(t)}$값의 결정

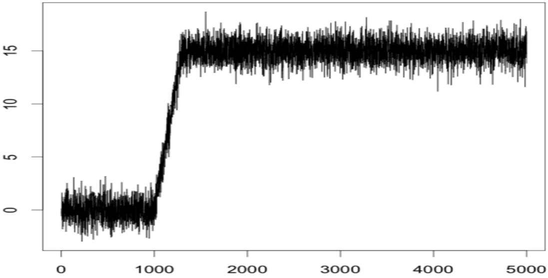
$$x^{(t)} = \begin{cases} x & \text{if } u \leq \alpha(x^{(t-1)}, x) \\ x^{(t-1)} & \text{if } u > \alpha(x^{(t-1)}, x). \end{cases} \tag{1}$$

- 몬테카를로 방법
- 깁스추출법
- 메트로폴리스-헤이스팅스 방법
- **MCMC 진단**

- Remind.

순차적으로 샘플된 값의 분포는 목표분포로 **수렴**한다.

순차적 샘플 값이 언제부터 수렴했는지 trace plot으로 확인.



앞의 1500개의 표본을 버릴 필요가 있다. 이를 번인(Burn-in)이라 함.

실효표본크기와 분산

깁스추출법과 메트로폴리스-헤이스팅스 방법으로 추출한 샘플은 시계열 자료라 할 수 있다.

- 정상시계열 y_t가 있을 때,

$$H_m = \frac{1}{m}\sum_{t=1}^{m} h(y_t) \longrightarrow Eh(y)$$

$$Var(H_m) \longrightarrow \sigma^2(1 + 2\sum_{k=1}^{\infty} \rho_k),$$

  이때 σ^2은 $h(y_t)$의 주변분포에 대한 분산이고, ρ_k는 시차 k에 대한 자기상관계수이다.

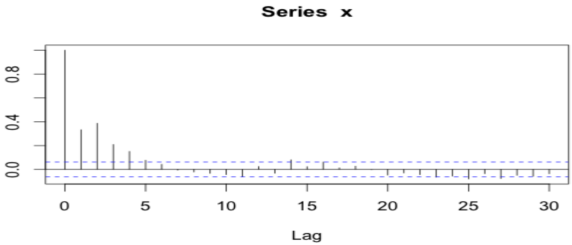
- 실효표본크기

$$M = \frac{m}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

  m개의 MCMC표본으로 만들어진 표본평균이 M개의 독립인 표본으로 만들어진 표본평균의 분산과 같다는 의미

가늘게하기

- 문제점 : MCMC 샘플의 자기 상관계수가 크다면, 실효표본크기가 작아진다.
- 가늘게하기(thinning) : 모든 표본을 다 사용하지 않고 r번째 표본만 사용하여 자기 상관계수를 줄인다.



The End