베이즈데이터분석 / 이재용 교수

06강

몬테카를로 방법

한국방송통신대학교

목차

목차

몬테카를로 방법

목적

$x \sim g(x)$일 때, 적분값

$$B = \mathbb{E}[f(x)] = \int f(x)g(x)dx,$$

을 구하고자 한다.

- (적분을 하기 힘든) 분포함수에서 정보를 추출하는 방법이다.
- 정보는 분포의 기대값과 분위수를 말한다.
- 분포함수에서 난수를 뽑는 알고리듬이 필요하다.

몬테카를로 방법

근거

강한 큰수의 법칙에 의해

$$\hat{B} \to B, \ a.s.$$

장점

- 컴퓨터로 난수를 생성하기 쉽다면 값 싸게 추정오차를 줄일 수 있다.
- 추정오차가 적분의 차원에 의존하지 않는다.

알고리듬

$x_1, \dots, x_n \overset{iid}{\sim} g$이라 하자. 다음과 같이

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

를 계산하고,
$\hat{B}$를 B를 추정하는데
사용한다.

추정오차

$$\hat{B}\text{의 추정오차} = \widehat{SE(\hat{B})} = \sqrt{\frac{v}{n}}.$$

여기서,

$$V = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i) - \frac{1}{n}\sum_{j=1}^{n} f(x_j)\right)^2$$

이다.

몬테카를로 방법의 예: π의 추정

π의 추정

$u \sim U(0,1), g(u) = 4\sqrt{1-u^2}$이라하면,

$$\pi = \int_0^1 g(u)du = Eg(u)$$

이다.

실습

몬테카를로 방법의 예: π의 추정

알고리듬

$u_1, u_2, \cdots, u_n \overset{i.i.d.}{\sim} U(0,1)$ 이라 하고,

$y_i = g(u_i), i = 1, \cdots, n$, 이라 하자.

π의 추정량으로

$$\hat{\pi} = \bar{y}$$

을 쓸 수 있다. 또한, π의 95% 신뢰구간은

$$\hat{\pi} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

이다. 여기서 s는 $y_1, \cdots, y_n$, 의 표본표준편차이다.

예. 정규모형과 코시 사전분포

문제

$x|\theta \sim N(\theta, 1)$이고,

사전분포가 $\theta \sim Ca(0, 1)$일 때,

사후분포의 평균을 구해보자.

예. 정규모형과 코시 사전분포. 풀이

사후분포의 밀도함수는

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

이다. 따라서, 사후분포의 평균은

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

이다. 위 적분의 결과는

수식으로 주어지지 않는다.

알고리듬

$\theta_1, \theta_2, \cdots, \theta_m \overset{i.i.d.}{\sim} N(x, 1)$를 추출한다.
그리고 분자와 분모의 적분을 각각 추정해서
사후분포의 평균을 구한다. 추정량은

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

와 같이 주어진다.

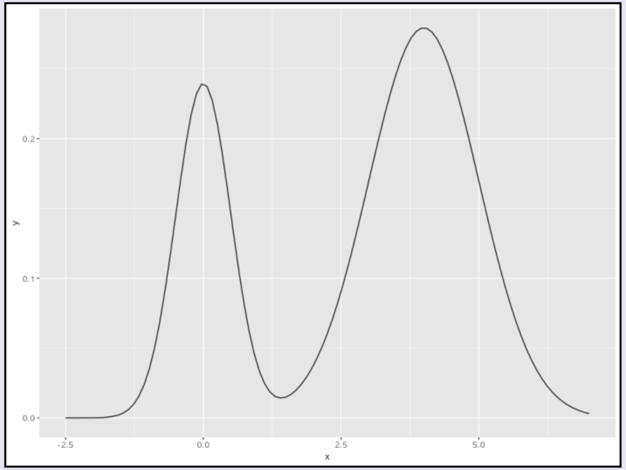
목차

몬테카를로 방법을 이용한 베이즈 추론

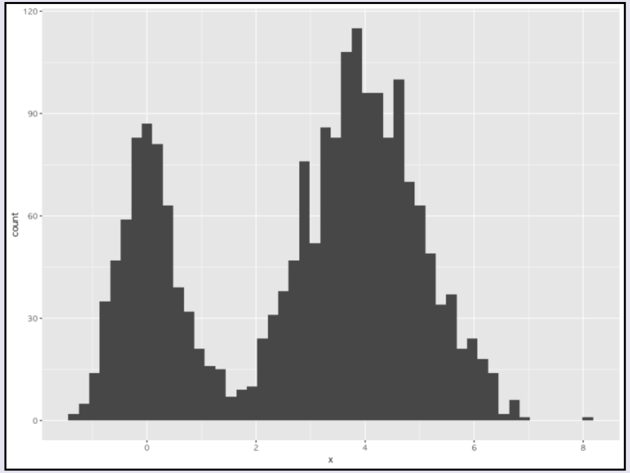
- $\pi(\theta)$: 사후분포
- $\theta_1, \theta_2, \ldots, \theta_m \sim \pi(\theta)$

$$\pi(\theta) \approx$$

θ	θ_1	θ_2	…	θ_m
확률	$\frac{1}{m}$	$\frac{1}{m}$	…	$\frac{1}{m}$



$\approx$



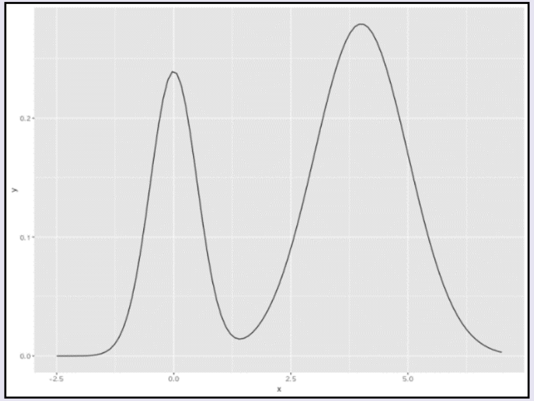
$$\int \theta \pi(\theta) d\theta \approx \frac{1}{m} \sum_{i=1}^{m} \theta_i$$

몬테카를로 방법을 이용한 베이즈 추론

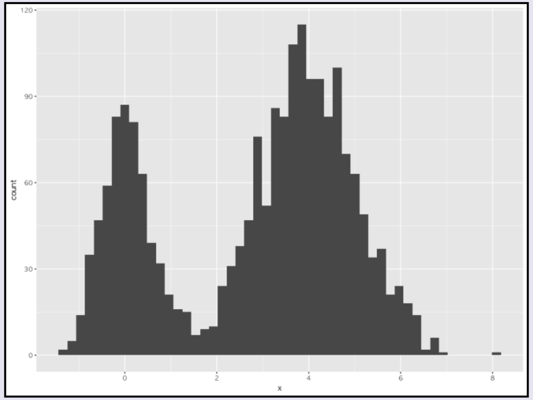
- $\pi(\theta)$: 사후분포
- $\theta_1, \theta_2, \ldots, \theta_m \sim \pi(\theta)$

$$\pi(\theta) \approx$$

θ	θ_1	θ_2	…	θ_m
확률	$\frac{1}{m}$	$\frac{1}{m}$	…	$\frac{1}{m}$



$\approx$



$$\int \theta \pi(\theta) d\theta \approx \frac{1}{m} \sum_{i=1}^{m} \theta_i$$

한국방송통신대학교

문제. 압정의 예

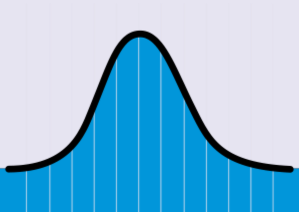
- 사전분포: $\theta \sim U(0,1)$.
- 가능도: $x|\theta \sim Bin(n,\theta)$.
- 사후분포

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

- θ의 베이즈 추정량을 구하시오.
- θ의 95% 신용구간을 구하시오.

사후표본의 표본평균은 사후분포의 기대값을 근사하고,
사후표본의 표본분위수는 사후분포의 분위수를 근사한다.

목차



중요도 추출(importance sampling)

문제

$x \sim g(x)$이고,

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

를 구하려고 한다.

그런데, $g(x)$에서 표본을 추출하는 것은 어렵고,

$\pi(x)$에서 표본을 추출하는 것은 쉽다고 하자.

중요도 추출(importance sampling)

알고리듬

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. B를 다음의 두 값으로 추정한다.

$$\hat{B}_1 = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i)$$

$$\hat{B}_2 = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i}.$$

중요도추출

1. $\hat{B}_1$은 B의 불편추정량이지만, $\hat{B}_2$은 불편추정량은 아니다.
   하지만 편향은 크지 않다.

2. $\hat{B}_1$를 구하기 위해서는 가중치를 정확히 알아야 한다.
   반면에 $\hat{B}_2$를 구하기 위해서는 가중치가 모르는 상수의 곱으로 나타나도 상관없다. 즉, $\hat{B}_1$을 구하기 위해서는 $g(x)$를 정확하게 알아야 하지만, $\hat{B}_2$를 구하기 위해서는

   $$g(x) = \text{모르는 상수} \times \text{아는 함수형태}$$

   이어도 상관없다.

3. 중요도추출은 $g(x)$에서 표본을 추출하기 어려울 때 사용할 수 있다.

예. 삼각분포의 평균

$$f(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 \le x \le 2 \end{cases}$$

라 할 때,

$$B = \int_0^2 x f(x) dx$$

를 중요도추출을 구하는 방법에 대해 알아보자.

예. 삼각분포의 평균. 풀이

$\pi(x) = \frac{1}{2} I(0 \le x \le 2)$를 $U(0,2)$의 밀도함수로 정의하자.

$$I = \int_0^2 x f(x) dx = \int_0^2 x \frac{f(x)}{\pi(x)} \pi(x) dx$$

임을 이용하여 다음의 알고리듬을 구성할 수 있다.

예. 삼각분포의 평균. 풀이

알고리듬

단계 1. $x_1, \dots, x_n \overset{i.i.d.}{\sim} U(0, 2)$를 생성한다.

단계 2. 가중치 $w_i = \dfrac{f(x_i)}{\pi(x_i)} = 2f(x_i), i = 1, 2, \dots, n,$ 를 계산한다.

단계 3. 다음 두 가지 중 하나의 식을 이용해 추정값을 계산한다.

단계 1..1 $\hat{I}_1 = \dfrac{1}{n}\sum_{i=1}^{n} x_i w_i$ 를 계산한다.

단계 2..2 $\hat{I}_2 = \dfrac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$ 를 계산한다.

다음시간

07강

정규모형