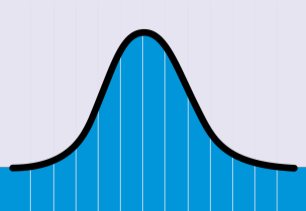
베이즈데이터분석 / 이재용 교수

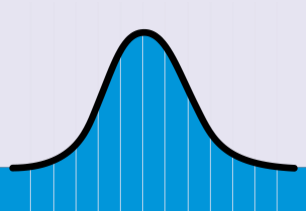
11강

해밀턴 몬테 카를로와 스탠

목차



목차



해밀톤 몬테 카를로

- 목표 밀도함수가 $\pi(\theta) \propto e^{-U(\theta)}$일 때,
  해밀턴(Hamiltonian) 몬테 카를로 혹은 해밀턴 MCMC는

$$\pi(\theta,\eta) \propto e^{-H(\theta,\eta)} = e^{-U(\theta)-K(\eta)}, K(\eta) = \frac{1}{2}\sum_i \frac{\eta_i^2}{m_i^2}$$

  를 정상분포로 갖는 마르코프 체인을 생성한다.

- (η는 잠재변수)

$$\int \pi(\theta,\eta)d\eta = \pi(\theta).$$

- MCMC 표본 $(\theta^{(t)},\eta^{(t)})_{t=1}^m$ 이 있을 때

$$(\theta^{(t)})_{t=1}^m \approx \pi(\theta).$$

해밀톤 동역학

- 해밀톤 동역학(Hamiltonian dynamics)는
  k –차원의 위치($\theta = (\theta_i, 1 \leq i \leq k)$)와
  운동량($\eta = (\eta_i, 1 \leq i \leq k)$)으로 물체의 운동을 표현하는 방정식이다.

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \eta_i}$$

$$\frac{d\eta_i}{dt} = -\frac{\partial H}{\partial \theta_i}, i = 1, 2, \ldots, k.$$

- $H(\theta, \eta)$는 해밀토니안(Hamiltonian)이라 하고,
  시스템의 총에너지를 나타낸다. 식으로는

$$H = U + K$$

와 같다. U는 위치에너지이고, K는 운동에너지를 나타낸다.

해밀톤 동역학

- 해밀톤 동역학은 해밀토니안 즉 에너지를 보존한다.

  즉, 해밀톤 동역학으로 이동한 값은

  밀도함수 $\pi(\theta, \eta) \propto e^{-H(\theta,\eta)}$의 값을 보존한다.

해밀톤 MCMC: 알고리듬

단계 1 (초기화) $\theta^{(0)}$와 $\eta^{(0)}$를 정한다.

단계 2 (HMC 반복)

$t = 1, 2, \ldots, m$에 대해서 다음을 수행한다.

(i) (운동량 변수 추출) $\eta^{+} \sim e^{-K(\eta)}$에서 추출한다.

(ii) (해밀톤 동역학을 이용한 위치와 운동량 변수 추출: 등넘기(leapfrog) 알고리듬)

- (등넘기 알고리듬으로 후보 추출)
  현재의 θ와 $\eta = \eta^{+}$값을 $\theta(t)$와 $\eta(t)$라 놓고,
  후보값 $\theta^{*} = \theta(t + \epsilon)$과 $\eta^{*} = \eta(t + \epsilon)$를
  다음과 같이 구한다.

해밀톤 MCMC: 알고리듬

$$\eta_i\left(t+\frac{\epsilon}{2}\right) = \eta_i(t) - \frac{\epsilon}{2}\frac{dU(\theta(t))}{d\theta_i}$$

$$\theta_i(t+\epsilon) = \theta_i(t) + \epsilon\frac{\eta_i(t+\epsilon/2)}{m_i^2}$$

$$\eta_i(t+\epsilon) = \eta_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{dU(\theta(t+\frac{\epsilon}{2}))}{d\eta_i}.$$

후보추출을 보통 L번 반복한다.

- (합격-불합격 결정)

$$\alpha = \min\{1, e^{-U(\theta^*)+U(\theta^{(t-1)})-K(\eta^*)+K(\eta^+)}\}$$

의 확률로 $(\theta^{(t)},\eta^{(t)}) = (\theta^*,\eta^*)$로 놓고,

그 외에는 $(\theta^{(t)},\eta^{(t)}) = (\theta^{(t-1)},\eta^+)$로 놓는다.

해밀톤 MCMC: 알고리듬

결과.

HMC 알고리듬으로 생성된

마르코프 체인 $(\theta^{(t)})$는 $\pi(\theta)$를 정상분포로 갖는다.

$(\theta^{(t)})$의 표본평균은 $\pi(\theta)$의 기대값을 근사하고,

$(\theta^{(t)})$의 표본분위수는 $\pi(\theta)$의 분위수를 근사한다.

해밀톤 MCMC: 특징들

- 임의보행 MH 알고리듬에 비해, 점프가 커서 마르코프 체인이 수렴상태로 빠르게 도달한다.
- 이산형 변수가 있는 분포에는 적용할 수 없다.
- L과 ϵ의 튜닝
  - (NUTS, No-U-Turn-Sampler) NUTS는 HMC의 변형으로 등넘기단계의 L을 정할 필요가 없게 해준다.
  - ϵ의 튜닝은 적응 메트로폴리스의 아이디어를 쓴다. 즉, 본격적인 변수들의 추출을 하기 전에 미리 샘플링을 해서 자기상관계수를 적당한 값으로 맞추도록 ϵ을 정한다.

해밀톤 MCMC: 특징들

결과.

HMC 알고리듬으로 생성된 마르코프 체인 $(\theta^{(t)})$는 $\pi(\theta)$를 정상분포로 갖는다.

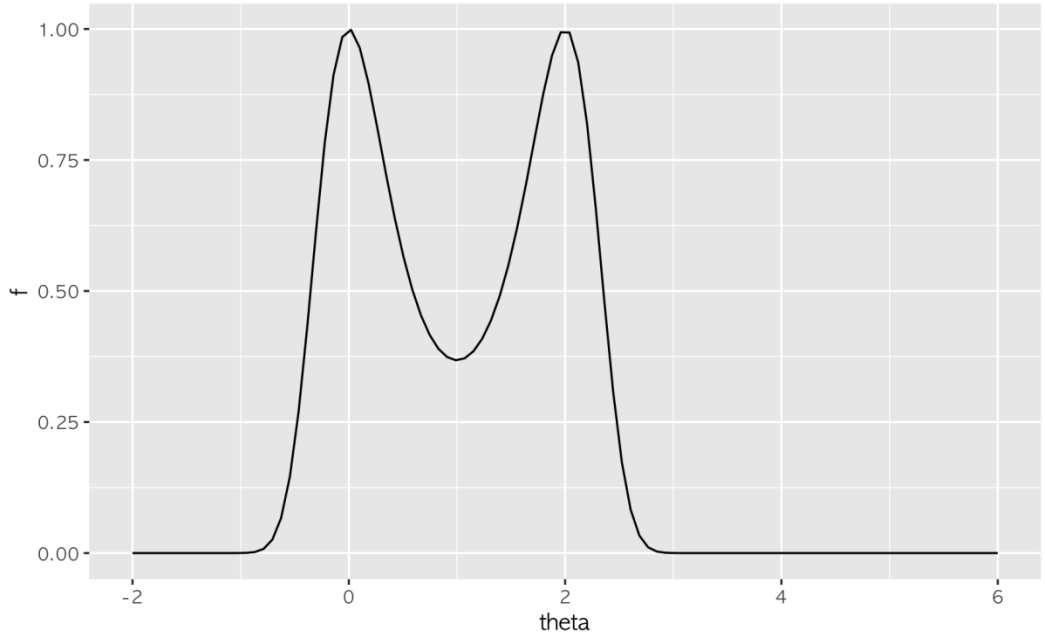
$(\theta^{(t)})$의 표본평균은 $\pi(\theta)$의 기대값을 근사하고,

$(\theta^{(t)})$의 표본분위수는 $\pi(\theta)$의 분위수를 근사한다.

예. 봉이 2개인 분포

목표 밀도함수

$$\pi(\theta) = e^{-\theta^2 (\theta - d)^2}, \theta \in \mathbb{R}, d > 0$$



예. 봉이 2개인 분포

해밀토니안

$$H(\theta,\eta) = U(\theta) + K(\eta)$$

$$U(\theta) = \theta^2(\theta - d)^2$$

$$K(\eta) = \frac{\eta^2}{2m}$$

$$\pi(\theta,\eta) = e^{-H(\theta,\eta)} = e^{-\theta^2(\theta-d)^2 - \frac{\eta^2}{2m}}$$

예. 봉이 2개인 분포

해밀토니안의 도함수

$$\frac{dU(\theta)}{d\theta} = \frac{d}{d\theta}\left(\theta^4 - 2d\theta^3 + d^2\theta^2\right)$$

$$= 4\theta^3 - 6d\theta^2 + 2d^2\theta$$

$$\frac{dK(\eta)}{d\eta} = \frac{d}{d\eta}\frac{\eta^2}{2m}$$

$$= \frac{\eta}{m}$$

해밀톤 MCMC 알고리듬

단계 1 (초기화) $\theta^{(0)}, \eta^{(0)}$, $\epsilon > 0$, L을 정한다.

단계 2 (HMC 반복)

$t = 1, 2, \ldots, m$에 대해서 다음을 수행한다.

(i) (운동량 변수 추출) $\eta^{+} \sim N(0, m)$

(ii) (해밀톤 동역학을 이용한 위치와 운동량 변수 추출:
등넘기(leapfrog) 알고리듬)

해밀톤 MCMC 알고리듬

- (등넘기 알고리듬으로 후보 추출) $\theta^* = \theta^{(t-1)}, \eta^* = \eta^+$이라 놓고,

  $j = 1, 2, \ldots, L$에 대해 다음을 수행한다.

$$\eta^* = \eta^* - \frac{\epsilon}{2}(4(\theta^*)^3 - 6d(\theta^*)^2 + 2d^2\theta^*)$$

$$\theta^* = \theta^* + \epsilon\frac{\eta^*}{m}$$

$$\eta^* = \eta^* - \frac{\epsilon}{2}(4(\theta^*)^3 - 6d(\theta^*)^2 + 2d^2\theta^*).$$

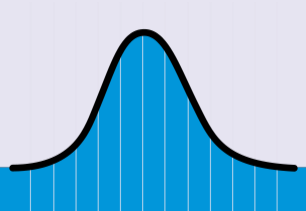
- (합격-불합격 결정)

$$\alpha = \min\{1, e^{-U(\theta^*)+U(\theta^{(t-1)})-K(\eta^*)+K(\eta^{(+)})}\}$$

의 확률로 $(\theta^{(t)}, \eta^{(t)}) = (\theta^*, \eta^*)$로 놓고,

그 외에는 $(\theta^{(t)}, \eta^{(t)}) = (\theta^{(t-1)}, \eta^+)$로 놓는다.

목차



STAN의 설치

- RStudio에서 아래의 코드를 수행한다.
- STAN을 따로 설치할 필요는 없다.

스탠을 설치한다.

```
remove.packages("rstan")
if (file.exists(".RData")) file.remove(".RData")
install.packages("rstan", repos = "https://cloud.r-project.org/",
dependencies = TRUE)
```

압정자료 분석

- 패키지 로딩

```
library(ggplot2) # 그림
library(GGally) # ggpairs
library(dplyr) # 자료 변형
library(rstan)
library(psych) # describe
library(reshape2) # 자료 변형 melt
```

- 병렬처리와 컴파일된 코드를 하드드라이브에 저장하기 위해.

```
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
```

압정자료 분석

자료 준비

```
x = 7
n = 10
data = list(x=x, n=n)
```

스탠 수행

```
pin.fit = stan(model_code=pin.code, data=data,
seed=1234567, chains=4, iter=2000, thin=1)
```

압정자료 분석

스탠 코드

```
pin.code ="
data {
    // data
    int<lower=0> x;
    int<lower=0> n;
}

parameters {
    real<lower=0, upper=1> theta;
}

model {
    x ~ binomial(n, theta);
    theta ~ uniform(0,1);
}
"
```

사후분석

```
print(pin.fit)


plot(pin.fit, plotfun="plot")
plot(pin.fit, plotfun="dens")
plot(pin.fit, plotfun="hist")
plot(pin.fit, plotfun="trace")
plot(pin.fit, plotfun="ac")
```

Stan의 기본 문법

코멘트

- 한 줄에서 ”//” 뒤는 코멘트로 처리된다.
- 혹은 ”/* .... */” 와 같이 쓸 수 있다.

Stan의 기본 문법

데이터 타입

- 기본형(primitive types): real과 int
- 벡터와 행렬: vector(열벡터),row vector(행벡터), matrix
- 어레이(array)

```
real x;
int a;
vector[10] x;
matrix[3,4] y;

array[10] real x;
array[6,7] matrix[3,3] m; // 각 원소가 3x3 행렬인 6x7 어레이
```

변수 범위

- 범위는 “<lower=a, upper=b>”의 형태로 표시된다.
  lower, upper 하나만 써도 된다.

```
int<lower = 1> N;

real<upper = 0> log_p;

vector<lower = -1, upper = 1>[3] rho;
```

Stan의 기본 문법

확률 분포들

```
log(beta) ~ normal(mu, sigma);
target += normal_lpdf(y | mu,sigma);

log(beta) ~ normal(mu, sigma) T[-0.5, 2.1];
// [-0.5, 2.1]에 제한된 정규분포

y ~ exponential(beta);
y ~ gamma(alpha, beta);

y ~ bernoulli(theta);
n ~ binomial(N, theta);
y ~ poisson(3.7);
```

자세한 내용은
stan reference manual과
stan function manual을
참조한다.

다음시간

12강

모형선택