DSL Seminar: MCMC (6)

Kyung-han Kim

Data Science Lab

February, 2023

목차

- 베이즈통계에서 MCMC 사용하기
- GLM (Generalized Linear Model)
- Bayesian Regression

베이즈통계에서 MCMC 사용하기

- MCMC는 원래 복잡한 분포에서 표본을 얻어내는 기법으로 베이즈통계에 국한되어 사용되는 기법은 아닙니다.
- 하지만 베이즈통계에서 필연적으로 다루게 되는 Posterior가 주로 복잡한 분포를 가지는만큼, MCMC는 베이즈통계와 궁합이 좋습니다!
- 그렇다면 구체적으로 어떻게 사용되는지 과제를 통해 확인합시다.
- 주의] 어려울 수 있습니다.
- 주의] 여지껏 들어본 적 없는 내용이 나올 수 있습니다.

Metropolis-Hastings algorithm with Bayesian Statistics

Bayesian Statistics HW2 Q2]
Now we have $x_1, ..., x_{1000}$ samples generated from the Cauchy distribution. These samples are stored as **x** in (hw02.Rdata). We will use the prior of $\theta \sim N(0, 100)$ and $\eta \sim U(0, 10)$ for the Bayesian inference. Our goal is to generate (θ, η) from the posterior distribution.

(a) Implement the Metropolis-Hastings algorithm to generate $10,000$ samples of (θ, η) from the posterior distribution.
You should tune the proposal distribution that have acceptance probability around $0.2 \sim 0.5$ for both θ and η. Report the trace plots, histograms, acceptance probabilities of your MCMC samples. (40 pts)

(b) Report the posterior means, 95% HPD intervals of θ and η. (10 pts)

Question 2-(a) (1): Likelihood and Prior

Since $X \sim \text{Cauchy}(\theta, \eta)$, $L(\theta, \eta|X) = \prod_{i=1}^{1000} \frac{1}{\theta\pi(1 + (\frac{x_i - \eta}{\theta})^2)}$.

Also, our prior of θ and η are $N(0, 100)$ and $U(0, 10)$ respectively.

So, $p(\theta) = \frac{1}{10\sqrt{2\pi}} \exp(-\frac{\theta^2}{200})$, and $p(\eta) = \frac{1}{10} I(0 \leq \eta \leq 10)$.

$$\therefore \pi(\theta, \eta|X) \propto (\prod_{i=1}^{1000} \frac{1}{\theta\pi(1 + (\frac{x_i - \eta}{\theta})^2)}) \frac{1}{10\sqrt{2\pi}} \exp(-\frac{\theta^2}{200}) \frac{1}{10} I(0 \leq \eta \leq 10)$$

$$\propto (\prod_{i=1}^{1000} \frac{1}{\theta(1 + (\frac{x_i - \eta}{\theta})^2)}) \exp(-\frac{\theta^2}{200}) I(0 \leq \eta \leq 10).$$

Now we can update θ and η with M-H algorithm!

Question 2-(a) (2): M-H algorithm - update θ

By using M-H algorithm, we can update multiple parameters at once. However, it may need more number of iterations since it accepts proposed sample only if all of them are accepted respectively.
Therefore, I'm going to update each parameter one at a time.

1) Update θ

1. If $t = 0$, set initial value $\theta^{(0)}$.
2. Generate $\theta' \sim N(\theta^{(t)}, \sigma_1^2)$
   σ_1^2 is a hyperparameter. Try some values and find the best σ^2!
   Be careful: θ' must be greater than 0. Do something!
3. Find the threshold $\log(\alpha) = \log(\pi(\theta', \eta^{(t)}|X)) - \log(\pi(\theta^{(t)}, \eta^{(t)}|X))$.
   Here, we can ignore $I(0 \leq \eta \leq 10)$ since it doesn't depend on θ.
4. Generate $U \sim U(0, 1)$ and compare $\log(U)$ and $\log(\alpha)$.
5. If $\log(U) \leq \log(\alpha)$, $\theta^{(t+1)} = \theta'$. Else, $\theta^{(t+1)} = \theta^{(t)}$.
6. Keep repeat until you get sufficiently large samples.

Question 2-(a) (3): M-H algorithm - update η

2) Update η

1. If $t = 0$, set initial value $\eta^{(0)}$.
2. Generate $\eta' \sim N(\eta^{(t)}, \sigma_2^2)$. Note that σ_1^2 and σ_2^2 can be different.
3. Find the threshold
   $\log(\alpha) = \log(\pi(\theta^{(t+1)}, \eta'|X)) - \log(\pi(\theta^{(t+1)}, \eta^{(t)}|X))$.
   Don't forget: $0 \leq \eta \leq 10$. What happens if η is out of range?
   Here, we can ignore $\exp(-\frac{\theta^2}{200})$ part since it doesn't depend on η.
   Also, we need to use $\theta^{(t+1)}$ since θ is already updated.
4. Generate $U \sim U(0, 1)$ and compare $\log(U)$ and $\log(\alpha)$.
5. If $\log(U) \leq \log(\alpha)$, $\eta^{(t+1)} = \eta'$. Else, $\eta^{(t+1)} = \eta^{(t)}$.
6. Keep repeat until you get sufficiently large samples.

Also, note that we need to update θ and η alternatively.
That is, the sequence should be $\theta_1 \to \eta_1 \to \theta_2 \to \eta_2 \to \cdots \to \theta_N \to \eta_N$
where N is total number of iterations. (In our question, $N = 10000$.)

단순선형회귀분석

- 회귀분석 (Regression Analysis): 하나 또는 여러 개의 독립변수가 종속변수에 어떻게 영향을 미치는지 분석하는 기법

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim^{iid} N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \text{ (matrix notation)}$$

- X: 독립 (Independent) 변수, 설명 (Explanatory) 변수
- Y: 종속 (Dependent) 변수, 반응 (Response) 변수
- 회귀분석의 목표는 회귀계수 (β_i) 를 알아내는 것입니다!
- 다양한 회귀분석 모형 가운데,
  - x와 β의 선형결합으로 y가 구해지고, (선형)
  - 독립변수가 1개인 (단순)

  모형을 단순선형회귀라고 합니다.

Bayesian Regression

- Bayesian regression의 핵심은, 우리가 알아내고자 하는 모수인 회귀계수 (β_i)들과 σ^2를 확률변수 취급하는 것입니다.

$$\{y_i\}_{i=1}^n | \beta_0, \beta_1, \sigma^2 \sim^{iid} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\beta_0 \sim N(\mu, \tau)$$

$$\beta_1 \sim N(\mu, \tau)$$

$$\sigma^2 \sim IG(a, b)$$

GLM (Generalized Linear Model)

- GLM은 선형회귀를 확장시킨 것입니다.
- 일반적인 선형회귀에서, $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ 이고 $\epsilon \sim N(0, \sigma^2)$ 이므로, 종속변수의 평균인 $E[\mathbf{Y}|\mathbf{X}] = \mu(\mathbf{X}) = \mathbf{X}\beta$ 입니다.
- GLM은 종속변수의 범위를 실수 전체에서 특정 범위로 제한한 모형이고, 이를 위해 **Link function**이라는 것을 사용합니다.
- 우리는 일반적으로 x 값이 정해질 때, y 값이 그에 따라 정해지는 형태의 함수가 익숙합니다.
  따라서 $g(\mathbf{X}\beta) = E[\mathbf{Y}|\mathbf{X}]$ 가 되면 적절할 것 같습니다.
- 하지만 실제로 Link function은 **반대로 정의**됩니다. 왜일까요?

GLM의 Link Function

- $\mathbf{X}\beta$ 쪽에 함수를 씌우면 더 이상 linear regression이 아니기 때문에, 똑같은 효과더라도 반드시 $E[\mathbf{Y}|\mathbf{X}]$ 쪽에 Link function을 씌웁니다.

$$g(E[\mathbf{Y}|\mathbf{X}]) = \mathbf{X}\beta$$

- 즉, Link function $g : S \to \mathbb{R}, S \subseteq \mathbb{R}$ 입니다.
- 베이즈통계 수업에서는 로지스틱 회귀 (Logistic Regression) 과 포아송 회귀 (Poisson Regression) 만 다룹니다.
- 로지스틱 회귀는 종속변수를 0 또는 1, 포아송 회귀는 종속변수를 음이 아닌 정수로 제한합니다.
- 다시 말해, 각각의 GLM은 아래와 같은 Link function을 필요로 합니다:
  - 로지스틱: $[0, 1] \to \mathbb{R}$
  - 포아송: $\mathbb{R}_{\geq 0} \to \mathbb{R}$

Logistic Regression

- 로지스틱 회귀는 종속변수가 Binary$(0,\ 1)$일 때 사용합니다.
  즉, 종속변수(Y)는 베르누이 분포를 따른다는 설정입니다.
- 이 때 Link function은 정의역이 $0 \sim 1$, 치역이 실수 전체인 함수여야 합니다. ($\mathbf{X}\beta$의 범위는 여전히 실수 전체)
- 이 조건을 만족시키는 대표적인 함수가 logit function입니다.

$$\text{logit}(p) = \log \frac{p}{1-p}$$

- $Y|X \sim \text{Bernoulli}(\mu(\mathbf{X}))$,
  $g(\mu(\mathbf{X})) = \text{logit}(\mu(\mathbf{X})) = \log \frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})} = \mathbf{X}\beta$,
  $\mu(\mathbf{X}) = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)}$ (로짓함수의 역함수 형태).
-
- cf) Probit Regression: $g(\cdot) = \Phi^{-1}(\cdot)$

Poisson Regression

- 포아송 회귀는 종속변수가 음이 아닌 정수 (0,1,2,...) 일 때 사용합니다. 즉, 종속변수 (Y) 는 포아송 분포를 따른다는 설정입니다.
- 이 때 Link function은 정의역이 양의 실수, 치역이 실수 전체인 함수여야 합니다.
- 이 조건을 만족시키는 대표적인 함수가 로그함수입니다.
- $Y|X \sim \text{Pois}(\mu(\mathbf{X}))$,
  $g(\mu(\mathbf{X})) = \log(\mu(\mathbf{X})) = \mathbf{X}\beta$,
  $\mu(\mathbf{X}) = \exp(\mathbf{X}\beta)$.

다음주 예고

- GLM 복습
- Gibbs Sampler
- Bayesian GLM with MCMC (M-H algorithm, Gibbs sampler)