DSL Seminar: MCMC (5)

Kyung-han Kim

Data Science Lab

August, 2023

목차: Bayesian Brainwashing

- Frequentist vs. Bayesian
    - C.I. (Confidence Interval and Credible Interval)
- Bayes' Theorem
- Prior, Likelihood, and Posterior
    - Conjugate Prior
- 베이즈통계에서 MCMC 사용하기 (6주차)

Frequentist vs. Bayesian

- 현대 통계학의 주류는 빈도론자(Frequentist)파이며,
  여러분이 알고 계시는 확률 또한 빈도론자의 정의 방식일 가능성이 큽니다.
- 빈도론자와 베이지안(Bayesian)은 확률을 다른 방식으로 정의합니다.
- Frequentist: 확률을 Long-run frequency로 정의함
  - 같은 사건을 매우 많이 반복했을 때, 해당 사건이 일어나는 비율
- Bayesian: 확률을 **Degree of Belief**로 정의함
  - 사건의 반복 횟수와 무관하게, 해당 사건이 발생할 것으로 생각되는 정도
- Degree of belief... 말이 너무 주관적인가요?
- 하지만 이젠 돌이킬 수 없습니다.
  베이지안에 입문한 여러분을 환영합니다.
  주관적인 것이 아니라, **자연스러운 것**입니다.

Self-Test: Are you Bayesian?

- 여러분은 빈도론자인지, 베이지안인지 검증해 봅시다.
- 동전을 던졌을 때, 앞면이 나올 확률과 뒷면이 나올 확률이 같다고 합시다.
- Q1] 지금 이 동전을 던져서 앞면이 나올 확률은?
- Q2] 제가 지금 동전을 던져 책상에 엎어 두었습니다.
  이 동전이 앞면일 확률은 얼마일까요?
- 여러분은 답이 뭐라고 생각하시나요? 정답이 있는 질문은 아닙니다.

Bayesian: 관점의 차이

- 베이지안은 **이미 고정된 것**에도 확률을 부여할 수 있습니다!
- 고정되어 있더라도 우리가 그 값이 얼마인지 정확히 모른다면, **믿음의 정도**에 따라 확률을 부여할 수 있다고 보는 것입니다.
- Q2에서 빈도론자는 0과 1 외의 확률을 부여할 수 없습니다. 수없이 반복하더라도 이미 고정되어 결과가 변하지 않기 때문입니다.
- 통계학에서 배우는 대표적인 고정된(Fixed) 값으로는 모수(Parameter)가 있습니다. - Unknown, but fixed constant.
- 즉, 베이즈통계에서는 모수가 가질 수 있는 값에 확률을 부여할 수 있습니다! 다르게 말하면 **모수를 확률변수 취급**합니다.
- 확률변수에는 확률분포가 있고, 그렇다면 우리는 모수가 따르는 확률분포를 정의해 줘야 합니다. (조금 뒤에 설명)

두 개의 C.I. - Confidence & Credible (1)

- 신뢰 구간 (Confidence Interval) 과 관련한 흔한 오해가 있습니다.
- 95% 신뢰 구간: 모수가 신뢰 구간 안에 위치할 확률이 95%? (X)
- 모수는 고정 (fixed) 이고, 신뢰 구간이 움직입니다!

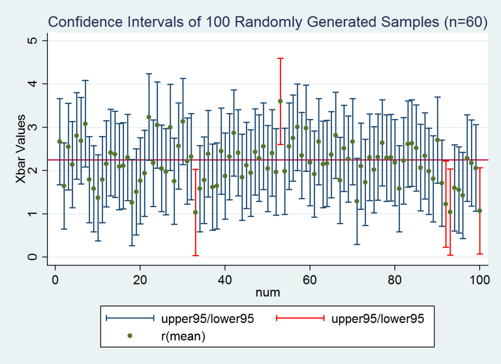

Figure 1: Confidence Interval

두 개의 C.I. - Confidence & Credible (2)

- 베이즈통계에서는 신뢰 구간이 아닌 Credible Interval을 사용합니다.
- 베이즈통계에서는 모수가 확률변수이기 때문에,
  모수도 값이 변할 수 있습니다.
- 95% Credible Interval을 구하면,
  해당 구간 안에 모수가 위치할 확률이 0.95라고 말할 수 있습니다!
- cf] 95% Confidence Interval의 올바른 해석:
  같은 방법으로 신뢰 구간을 여러 번 구하면,
  그 구간들 중 95%가 모수를 포함한다.

Bayes' Theorem

- 베이즈 정리 (Bayes' Theorem) 은 조건부확률과 관련된 정리이자, 베이즈 통계의 근간이 되는 중요한 정리입니다!

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

- 여기에서 A 를 X 로 바꾸고, B 를 θ 로 바꾸면?

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \propto P(\theta)P(X|\theta) = P(\theta)L(\theta|X)$$

- 위의 식을 꼭 기억하시기 바랍니다!
- 여기에서 A, B, X, θ 모두 확률변수를 나타냅니다.
- 특히 θ 는 모수입니다. **우리에게 모수는 확률변수입니다.**

필요 배경 지식: Likelihood (1)

- Likelihood(우도)는 수리통계학(2)에서 처음 소개되는 개념으로, pdf/pmf를 다른 관점에서 바라본 결과물로 생각할 수 있습니다.
- Ex] 지수분포의 pdf: $f(x) = f(x|\lambda) = \lambda e^{-\lambda x}$
- 특정 x에서 pdf 값을 알기 위해서는 모수(λ) 값을 알아야만 합니다.
- 즉, pdf 값은 모수 값이 주어져 있을 때 어떤 확률변수가 특정 값을 가질 가능성으로 해석할 수 있습니다. 또한 pdf는 x에 대한 함수입니다.

필요 배경 지식: Likelihood (2)

- 반면 Likelihood는 똑같은 식을 모수(λ)에 대한 함수로 보는 것입니다.
- Likelihood: 관측된 확률변수 값(=데이터)이 주어져 있을 때, 모수가 특정 값이었을 가능성!
- 우리가 겪는 실제 상황은 pdf보다는 likelihood에 더 적합합니다.
- 관측치가 여러 개일 경우 그 값들을 모두 곱한 것을 Likelihood function (우도함수)라고 합니다.
- Likelihood Function: $L(\lambda|x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$
- pdf/pmf는 x에 대한 함수, likelihood는 모수(λ)에 대한 함수!

베이즈통계의 기본 구조

- 앞서 베이즈정리에서 도출한 식을 다시 가져 옵시다.

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \propto P(\theta)P(X|\theta)$$

- 여기에서 우리는 각각의 항을 이렇게 부릅니다:

$$\underbrace{P(\theta|X)}_{\text{Posterior}} \propto \underbrace{P(\theta)}_{\text{Prior}}\underbrace{P(X|\theta)}_{\text{Likelihood}}$$

- 통계학의 기본 목표는 **모수를 추정**하는 것입니다.
- 베이즈통계에서는 모수가 확률분포를 가지므로,
  모수를 추정하기 위해서는 **모수가 따르는 확률분포**를 알아내야 합니다.
- 우리의 사전 지식과 데이터를 통해 얻은 정보를 혼합해 얻어낸
  모수가 따를 것으로 예상되는 확률분포가 바로 사후분포 (Posterior) 입니다.

Prior, Likelihood, Posterior (1): Prior

- $\underbrace{P(\theta|X)}_{\text{Posterior}} \propto \underbrace{P(\theta)}_{\text{Prior}} \underbrace{P(X|\theta)}_{\text{Likelihood}}$
  - Prior, Likelihood, Posterior 모두 θ에 대한 함수입니다!!!
- $P(\theta)$: Prior (distribution) - 사전 분포
- 분석을 본격적으로 시작하기 전에 설정하는,
  모수 θ가 따를 것으로 생각되는 확률분포의 pdf 혹은 pmf입니다.
- Remind! 베이즈통계에서 **모수는 확률변수 취급**하기 때문에
  자연스럽게 확률분포를 가져야만 합니다.
- Prior는 연구자가 임의로 설정할 수 있지만, 해당 모수의 성질에 맞게끔 설정하는 것이 좋습니다.
  - Ex] $\theta > 0$: $\theta \sim \Gamma(a, b)$ / $0 \leq \theta \leq 1$: $\theta \sim \text{Beta}(a, b)$
- Prior는 Likelihood와 결합해서 Posterior가 됩니다!
- Prior의 영향력을 최대한 줄이는 것이 베이지안의 관심사입니다.
  이를 Non-Informative Prior라고 합니다.

Prior, Likelihood, Posterior (2): Likelihood

- $\underbrace{P(\theta|X)}_{\text{Posterior}} \propto \underbrace{P(\theta)}_{\text{Prior}} \underbrace{P(X|\theta)}_{\text{Likelihood}}$
- $P(X|\theta)$: Likelihood (function) - 우도 함수
- 이 식은 기본적으로는 X **의 pdf/pmf**입니다.
  그리고 pdf/pmf는 x에 대한 함수입니다. (모수 값을 알아야 계산 가능)
- 하지만 실제 상황에서는 우리는 오히려 **모수 (θ) 값을 모르고**,
  해당 모수를 이용해 정의된 확률분포에서 얻어진 X를 알고 있습니다.
- 즉, X가 주어진 상태에서 θ를 알아내야 하는 상황이 되므로
  우리는 이 식을 θ **에 대한 함수**로 보기로 합니다.
- 그리고 그것이 Likelihood의 정의였습니다.
- 경우에 따라 Likelihood가 θ에 대한 함수임을 강조하기 위해 $P(X|\theta)$가
  아니라 $L(\theta|X)$라고 쓰기도 합니다. (저는 후자를 선호해요)

Prior, Likelihood, Posterior (3): Posterior

- $\underbrace{P(\theta|X)}_{\text{Posterior}} \propto \underbrace{P(\theta)}_{\text{Prior}}\underbrace{P(X|\theta)}_{\text{Likelihood}}$
- $P(\theta|X)$: Posterior (distribution): 사후 분포
- X, 즉 데이터가 주어져 있을 때 (given) 모수가 따르는 확률분포입니다.
- Prior에 Likelihood 값을 곱하면, 그 값이 Posterior에 비례합니다.
- 다시 말하면, 데이터 수집 전에 설정한 모수의 확률분포인 Prior가 데이터를 수집해 얻은 Likelihood 값을 통해 Update되면, 그것을 우리가 Posterior라고 부릅니다.
- 모수가 따르는 확률분포를 우리가 수집한 데이터로 수정해 얻은 새 분포!
- 사후분포는 사전분포의 영향도 받지만, 수집한 데이터 (likelihood) 의 영향도 받습니다.
- 데이터를 더 많이 수집할수록 사전분포의 영향력이 약해집니다.
- Posterior distribution을 알아내는 것이 최종 목표입니다!

Prior, Likelihood, Posterior (4): Normalizing Constant

- 원래 $P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$ 였습니다.
- 분모의 $P(X)$는 뭐길래 계속 무시했을까요?
- $P(X)$는 별다른 역할도 없으면서 구하기는 어려운 어떤 상수입니다.
- 유일한 역할은 전체 확률의 합을 1로 만들어주는 것입니다. 이러한 상수를 Normalizing constant라고 합니다.
- 베이즈통계에서는 주로 무시됩니다.

$$\text{결론: (Posterior)} = \frac{\text{(Prior)} \times \text{(Likelihood)}}{\text{(Normalizing Constant)}} \propto \text{(Prior)} \times \text{(Likelihood)}$$

cf] MCMC의 장점

- MCMC 외에도 복잡한 확률분포에서 표본을 생성하는 방법은 많습니다. 그 중에서 베이즈통계에서 MCMC가 널리 쓰이는 이유는 무엇일까요?
- MCMC는 Un-normalized density에도 사용이 가능합니다. (Why?)
- 다른 기법의 경우 반드시 Normalizing constant까지 알아야만 사용이 가능한 경우가 일반적입니다.
- 그리고 베이즈통계에서는 Normalizing constant를 구하는 것이 까다로워, MCMC와 궁합이 좋습니다.

How to choose Prior?

- $P(\theta|X) \propto P(\theta)P(X|\theta) \leftrightarrow$ (Posterior) $\propto$ (Prior) $\times$ (Likelihood)
- 주로 $\pi(\theta|X) \propto p(\theta)L(\theta|X)$ 로 씁니다.
- Prior도 확률분포, Likelihood도 확률분포이기 때문에 그 둘을 곱하면 분포가 일반적으로 복잡해집니다.
- 즉, 일반적인 경우에 Posterior는 어떤 잘 알려진 분포를 따른다고 말하기 어렵습니다.
- 하지만 몇몇 Prior-Likelihood 조합에서는 그 둘의 곱이 잘 알려진 분포를 바로 따르게 되어 분석이 편해집니다.
- 그런 경우를 의도적으로 만들기 위해 선정하는 Prior가 Conjugate Prior입니다.

Conjugate Prior

- Likelihood는 데이터의 성격을 나타내는 부분으로, 우리가 임의로 고를 수 없는 영역입니다. (fixed)
- 하지만 Prior는 우리가 기본적으로 마음대로 고를 수 있으므로, Likelihood에 잘 어울리는 형태의 Prior를 골라서 Posterior가 간단한 분포로 정리되게끔 만들 수 있습니다!

Likelihood	Prior	Posterior	Parameter
Binomial	Beta	Beta	$B(n, \mathbf{p})$
Poisson	Gamma	Gamma	$\mathrm{Pois}[\boldsymbol{\lambda}]$
Normal	Inverse-Gamma	Inverse-Gamma	$N(\mu, \boldsymbol{\theta})$
Normal	Normal	Normal	$N(\boldsymbol{\mu}, \theta)$

Table 1: Examples of conjugate prior for various likelihood functions

- 실제로 증명해 봅시다. (오늘의 과제입니다!)

Some Useful Distributions

1. Binomial: $X \sim B(N, p) :_N C_x p^x (1-p)^{N-x}$
2. Poisson: $X \sim Pois(\lambda) : \lambda e^{-\lambda x}$
3. Normal: $X \sim N(\mu, \sigma^2) : \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
4. Beta: $X \sim Beta[\alpha, \beta] : \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
5. Gamma: $X \sim \Gamma[\alpha, \beta] : \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ **(형태 주의)**
6. Inverse-Gamma: $X \sim IG(\alpha, \beta) : \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$

Wrap-up (1)

- 베이즈통계에서는 모수를 확률변수 취급한다!
  그러므로 각각의 모수 별로 pdf/pmf가 존재한다.
- 연구자는 모수가 따를 것으로 생각되는 확률분포를 먼저 가정한다.
  이것을 사전분포(Prior Distribution)라고 한다.
  사전분포는 데이터가 수집되면 수정된다.
- 수집된 데이터는 우도함수(Likelihood Function)에 반영된다.

Wrap-up (2)

- 우리의 목표는 처음에 우리가 가정한 사전분포를,
  수집된 데이터로 수정 (update) 한 결과물인 새로운 모수의 확률분포이다.
- 이 새로운 확률분포를 사후분포 (Posterior Distribution) 라고 한다.
- Posterior는 Prior와 Likelihood의 곱에 비례한다!
- Posterior는 두 함수의 곱이기 때문에 형태가 복잡한 경우가 많다.
  예외적으로 Conjugate prior를 사용하면 형태가 간단해진다.
- 형태가 복잡한 분포에서의 sampling?: **MCMC**!
- Posterior의 평균을 closed-form으로 구할 수 없기 때문에,
  Metropolis-Hastings algorithm을 이용해 Posterior mean을 추정한다.

다음주 예고

- GLM and Bayesian Regression
- Metropolis-Hastings algorithm and Bayesian Statistics