

December 10, 2025

- 강의명: CS109A: 데이터 과학 입문
- 주차: Lecture 11
- 교수명: Pavlos Protopapas, Kevin Rader, Chris Gumb
- 목적: Lecture 11의 핵심 개념 학습

## Contents

1 베이지안 용어 정리 . . . . .	1
2 베이지안 추론이란 무엇인가? . . . . .	3
2.1 1. 빈도주의 (Frequentist) 관점 . . . . .	3
2.2 2. 베이지안 (Bayesian) 관점 . . . . .	3
3 베이즈 정리: 믿음을 업데이트하는 공식 . . . . .	5
4 사전분포(Prior) 선택하기 . . . . .	7
4.1 1. 정보적 사전분포 (Informative Prior) . . . . .	7
4.2 2. 비정보적 사전분포 (Uninformative Prior) . . . . .	7
4.3 3. 컬레 사전분포 (Conjugate Prior) . . . . .	7
5 핵심 예제: 정규-정규 모델 (Normal-Normal Model) . . . . .	9
6 베이지안 추정: 점과 구간 . . . . .	9
6.1 점 추정 (Point Estimation) . . . . .	9
6.2 구간 추정 (Interval Estimation) . . . . .	10
7 베이지안 선형 회귀 (Bayesian Linear Regression) . . . . .	11
8 심화: 릿지(Ridge)와 라쏘(Lasso)의 베이지안 해석 . . . . .	12
8.1 1. 릿지(Ridge) = MAP + 정규(Normal) 사전분포 . . . . .	12
8.2 2. 라쏘(Lasso) = MAP + 라플라스(Laplace) 사전분포 . . . . .	13
9 계산 방법: MCMC와 갑스 샘플링 . . . . .	14
9.1 문제: 사후 분포가 너무 복잡할 때 . . . . .	14
9.2 해결: 시뮬레이션 (MCMC) . . . . .	14
9.3 갑스 샘플링 (Gibbs Sampling) . . . . .	14

10 자주 묻는 질문 (FAQ) 및 점검 . . . . .	16
11 빠르게 훑어보기 (1-Page Summary) . . . . .	16

### ▣ 핵심 요약

이 문서는 베이지안(Bayesian) 모델링의 핵심 개념을 설명합니다.

베이지안 추론은 \*\*데이터(증거)를 바탕으로 기존의 믿음(사전 확률)을 업데이트\*\*하여, 더 합리적인 새로운 믿음(사후 확률)을 도출하는 과정입니다.

단 하나의 '정답'을 찾는 대신, \*\*파라미터의 '가능성'을 확률 분포로 표현\*\*합니다.

이 문서를 통해 베이즈 정리의 기본 원리를 이해하고, 이것이 어떻게 릿지(Ridge), 라쏘(Lasso)와 같은 머신러닝 회귀 모델과 연결되는지 학습합니다.

## 1 베이지안 용어 정리

베이지안 모델링을 처음 접할 때 가장 혼란스러운 것은 용어입니다. 핵심 용어들을 일상적인 언어로 먼저 정리합니다.

용어	쉬운 설명 (직관)	원어	비고
베이즈 정리	나의 '기존 믿음'과 '새로운 증거'를 합쳐 '최종 결론'을 내는 수학 공식	Bayes' Rule / Theorem	$P(A B) = \frac{P(B A)P(A)}{P(B)}$
사전 확률	데이터를 보기 전, 내가 이미 가지고 있던 믿음(가설)	Prior Probability	$P(A)$
가능도	나의 '믿음'이 맞다고 가정할 때, 이 '증거'가 나타날 확률	Likelihood	$P(B A)$
사후 확률	'증거'를 반영하여 새롭게 업데이트 된 '최종 믿음'(결론)	Posterior Probability	$P(A B)$
증거	그냥 '증거'가 관찰될 확률(전체 확률). 주로 정규화 상수로 쓰임.	Evidence / Marginal	$P(B)$
빈도주의	파라미터(예: 평균)는 '고정된 값'이며, 데이터가 무작위라고 보는 관점.	Frequentist	베이지안과 대비되는 통계학의 주류 관점.
베이지안	파라미터 자체를 '확률 변수'(믿음의 대상)로 보는 관점.	Bayesian	주관적 확률(belief)을 다룸.
초매개변수	'사전 확률'을 정의하기 위해 필요 한 또 다른 파라미터. (예: $\mu \sim N(\mu_0, \sigma_0^2)$ 에서 $\mu_0, \sigma_0^2$ )	Hyperparameter	Prior의 파라미터.
켤레 사전분포	계산을 편하게 해주는 '특수 조합'. (예: 이항-베타, 정규-정규)	Conjugate Prior	특정 가능도와 결합 시, 사후 분포가 사전 분포와 '같은 가족'이 됨.
MAP	사후 확률 분포에서 '가장 높은 지점'(최 빈값)을 추정치로 사용.	Max A Posteriori	사후 확률을 최대화하는 파라미터.
신용 구간	"파라미터가 이 구간 안에 있을 확률이 95%다." (매우 직관적)	Credible Interval	베이지안 베전의 구간 추정.
신뢰 구간	"같은 실험 100번 시, 95개의 구간이 '참 값'을 포함할 것이다."	Confidence Interval	빈도주의 베전의 구간 추정. (해석이 까다로움)

## 2 베이지안 추론이란 무엇인가?

통계학에는 세상을 바라보는 두 가지 큰 관점이 있습니다: 빈도주의(Frequentist)와 베이지안(Bayesian)입니다.

### 2.1 1. 빈도주의 (Frequentist) 관점

빈도주의는 우리가 흔히 배우는 전통적인 통계학입니다.

\* 핵심 가정: 우리가 찾으려는 파라미터(모수,  $\theta$ , 예: 전교생의 실제 평균 키)는 하나의 '고정된' 상수입니다.  
 \* 데이터: 이 '참 값'을 알기 위해 우리가 뽑는 표본(데이터,  $X$ )이 무작위(random)입니다.  
 \* 목표: 데이터를 많이 뽑아서 저 '고정된 참 값'( $\theta$ )을 정확히 추정(estimate)하는 것입니다.

#### □ 예제: 빈도주의 비유: 고정된 보물 찾기

어딘가에 하나의 '보물'( $\theta$ )이 숨겨져 있습니다.

우리는 이 보물의 위치를 모르지만, 보물에 대한 힌트( $X$ , 데이터)를 여러 번 얻을 수 있습니다.

빈도주의는 이 힌트(데이터)들을 조합하여 "보물은 ( $x, y$ ) 지점에 있을 것이다"라고 하나의 지점을 추정하는 방식입니다.

### 2.2 2. 베이지안 (Bayesian) 관점

베이지안 통계학은 '믿음(belief)'의 관점에서 접근합니다.

\* 핵심 가정: 우리가 가진 데이터( $X$ )는 '고정된' 관찰 값입니다. (일단 관찰했으므로)  
 \* 파라미터: 오히려 우리가 모르는 파라미터( $\theta$ )가 무작위(random)이며, 확률 분포를 가집니다.  
 \* 목표: 데이터( $X$ )를 관찰함으로써, 파라미터  $\theta$ 에 대한 우리의 믿음(belief)을 업데이트하는 것입니다.

#### □ 예제: 베이지안 비유: 확률적인 보물 지도

베이지안은 보물이 "어디쯤 있을지"에 대한 '믿음의 지도'(사전 확률)로 시작합니다. (예: "A 지역 60%, B 지역 40%")

이때 힌트( $X$ , 데이터)를 하나 얻습니다. (예: "보물은 강 근처에 있다.")

이 힌트를 바탕으로, "A 지역이 강 근처일 확률"과 "B 지역이 강 근처일 확률"을 계산하여, 원래의 믿음을 업데이트한 '새로운 지도'(사후 확률)를 만듭니다. (예: "A 지역 80%, B 지역 20%")

colback=mygray, colframe=darkgray, breakable, title=□ 빈도주의 vs. 베이지안 핵심 비교

- **빈도주의 (Frequentist):**

- 파라미터( $\theta$ ): 고정된 상수 (Unknown constant)
- 데이터( $X$ ): 랜덤 변수 (Random variable)
- 해석: ”무한히 반복하면, 95%의 신뢰 구간이 참 값을 포함한다.”

- **베이지안 (Bayesian):**

- 파라미터( $\theta$ ): 랜덤 변수 (Random variable)
- 데이터( $X$ ): 고정된 관찰 값 (Fixed data)
- 해석: ”우리의 데이터에 따르면, 95% 확률로 파라미터가 이 신용 구간 안에 있다.”

베이지안의 ’신용 구간’이 우리가 일상적으로 ”확률”을 이해하는 방식과 더 가깝습니다.

### 3 베이즈 정리: 믿음을 업데이트하는 공식

베이지안 추론은 베이즈 정리(Bayes' Rule)라는 하나의 공식에서 출발합니다.

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

이 공식의 각 항은 다음과 같은 의미를 가집니다.

- $P(\theta|X)$ : 사후 확률 (Posterior)
  - ”데이터  $X$ 를 관찰한 후에, 파라미터  $\theta$ 에 대한 우리의 최종 믿음”
  - 이것이 우리가 구하고자 하는 결과물입니다.
- $P(X|\theta)$ : 가능성 (Likelihood)
  - ”파라미터  $\theta$ 가 참이라고 가정할 때, 데이터  $X$ 가 관찰될 가능성”
  - 이것은 우리가 가진 데이터로부터 계산되는 증거의 힘입니다.
- $P(\theta)$ : 사전 확률 (Prior)
  - ”데이터  $X$ 를 관찰하기 전에, 파라미터  $\theta$ 에 대해 우리가 가졌던 초기 믿음”
  - 이것은 우리의 주관적인 지식이나 가정입니다.
- $P(X)$ : 증거 (Evidence)
  - ”그냥 데이터  $X$ 가 관찰될 전체 확률”
  - 실제 계산에서는  $\theta$ 와 관련이 없으므로,  $P(\theta|X)$ 의 합이 1이 되도록 만드는 정규화 상수(Normalizing Constant) 역할을 합니다.

따라서 베이즈 정리는 다음과 같이 요약할 수 있습니다.

#### ▣ 핵심 요약

사후 확률 (최종 믿음)  $\propto$  가능성 (증거의 힘)  $\times$  사전 확률 (초기 믿음)

$$f(\theta|X) \propto f(X|\theta)f(\theta)$$

#### ▣ 예제: 동전 뒤집기 예제 (이산 확률)

주머니 속에 3개의 동전이 있습니다:

- 동전 A: 앞면이 나올 확률  $p = 0.1$  (불량 동전)
- 동전 B: 앞면이 나올 확률  $p = 0.5$  (공정 동전)
- 동전 C: 앞면이 나올 확률  $p = 0.9$  (불량 동전)

1. **사전 확률 (Prior):** 내가 동전 하나를 무작위로 뽑았습니다. 이 동전이 A, B, C일 확률은 얼마일까요? 데이터가 없으므로, 각각  $1/3$ 입니다.

$$P(\theta = 0.1) = 1/3 \quad | \quad P(\theta = 0.5) = 1/3 \quad | \quad P(\theta = 0.9) = 1/3$$

2. **데이터 (Data):** ◎ 동전을 4번 던졌더니, 앞면이 3번, 뒷면이 1번 나왔습니다. ( $X = 3$ )

3. **가능도 (Likelihood):** 각 동전(가설)이 이 데이터를 만들어낼 가능성은 얼마일까요? (이항 분포 사용:  $\binom{4}{3} p^3 (1-p)^1$ )

- $P(X = 3|\theta = 0.1) = \binom{4}{3} (0.1)^3 (0.9)^1 = 4 \times 0.001 \times 0.9 = 0.0036$
- $P(X = 3|\theta = 0.5) = \binom{4}{3} (0.5)^3 (0.5)^1 = 4 \times 0.125 \times 0.5 = 0.2500$
- $P(X = 3|\theta = 0.9) = \binom{4}{3} (0.9)^3 (0.1)^1 = 4 \times 0.729 \times 0.1 = 0.2916$

데이터는 동전 C ( $p = 0.9$ ) 일 가능성을 가장 높게 시사합니다.

4. 사후 확률 (Posterior): 이제  $Posterior \propto Likelihood \times Prior$ 를 계산합니다.

- $P(\theta = 0.1|X = 3) \propto 0.0036 \times (1/3) \approx 0.0012$
- $P(\theta = 0.5|X = 3) \propto 0.2500 \times (1/3) \approx 0.0833$
- $P(\theta = 0.9|X = 3) \propto 0.2916 \times (1/3) \approx 0.0972$

5. 정규화 (Normalize): 사후 확률의 총합( $0.0012 + 0.0833 + 0.0972 = 0.1817$ )으로 나누어 합이 1이 되게 합니다.

- $P(\theta = 0.1|X = 3) = 0.0012/0.1817 \approx \mathbf{0.7\%}$
- $P(\theta = 0.5|X = 3) = 0.0833/0.1817 \approx \mathbf{45.8\%}$
- $P(\theta = 0.9|X = 3) = 0.0972/0.1817 \approx \mathbf{53.5\%}$

결론: 데이터를 보기 전 우리의 믿음은 (33%, 33%, 33%) 였지만, "4번 중 3번 앞면"이라는 데이터를 본 후, 우리의 믿음은 (0.7%, 45.8%, 53.5%)로 업데이트되었습니다. 우리는 이제 이 동전이  $p = 0.9$  동전(C)이라고 가장 강하게 믿게 되었습니다.

## 4 사전분포(Prior) 선택하기

베이지안 모델링의 가장 중요하고 주관적인 부분이 바로 사전분포(Prior,  $f(\theta)$ )를 정하는 것입니다. 사전분포는 ”내가 데이터를 보기 전에 파라미터에 대해 무엇을 알고 있는가?”를 확률 분포로 표현하는 것입니다. 사전분포를 선택하는 3가지 주요 접근 방식이 있습니다.

### 4.1 1. 정보적 사전분포 (Informative Prior)

이전 연구, 전문가의 의견, 또는 과거의 데이터를 바탕으로 ’정보가 있는’ 사전분포를 설정합니다.

\* 목적: 이미 알고 있는 지식을 모델에 적극적으로 반영합니다. \* 예시: \* 내일 정오의 기온( $\mu$ )을 예측하는 모델을 만든다고 가정합니다. \* 사전분포: ”어제 정오 기온이 20도였고, 최근 30일간 일교차 표준편차가 2도였다”는 정보를 바탕으로, \*  $\mu \sim N(\mu_0 = 20, \sigma_0^2 = 2^2)$  처럼 정규분포를 설정할 수 있습니다. \* 이는 ”내일 기온도 20도 근처일 것이고, 95% 확률로 16도에서 24도 사이일 것이다”라는 강한 믿음을 표현합니다.

### 4.2 2. 비정보적 사전분포 (Uninformative Prior)

파라미터에 대해 아는 것이 거의 없거나, 의도적으로 데이터의 영향력을 극대화하고 싶을 때 사용합니다. ”최소한의 정보”를 제공하는 사전분포입니다.

\* 목적: 사전 지식의 영향을 최소화하고, 데이터( $X$ )가 사후 분포를 거의 결정하도록 만듭니다. \* 예시: \* 새로운 치료법의 성공 확률  $p$  (0에서 1 사이)를 모델링 합니다. \* 사전분포:  $p$ 에 대해 아는 것이 전혀 없으므로, 0과 1 사이의 모든 값이 동일하게 가능하다고 가정합니다. \*  $p \sim Uniform(0, 1)$  (0과 1 사이의 균등 분포) \* 이는 ”모든 확률값이 공평하다”는 약한 믿음을 표현합니다.

### 4.3 3. 결례 사전분포 (Conjugate Prior)

수학적, 계산적 편의성을 위해 특정 ’궁합이 잘 맞는’ 사전분포-가능도 조합을 사용하는 것입니다.

\* 정의: 어떤 가능성도 함수  $f(X|\theta)$ 에 대해, 특정 사전분포  $f(\theta)$ 를 사용했더니, 그 결과물인 사후분포  $f(\theta|X)$ 가 사전분포와 동일한 분포 가족(family)이 되는 경우, 이 사전분포를 ’결례 사전분포’라고 부릅니다. \* 비유: ”파란색 물감(Prior) + 노란색 물감(Likelihood) = 초록색 물감(Posterior)” 결례 관계는 ’파란색 + 노란색 = 초록색’이라는 공식을 아는 것과 같습니다. 만약 결례가 아니면, ”파란색 + 분홍색 = ??” 처럼, 그 결과를 한눈에 알 수 없고 계산이 복잡해집니다. \* 왜 사용하는가? 복잡한 적분 계산 ( $P(X) = \int f(X|\theta)f(\theta)d\theta$ )을 피하고, 사후 분포의 파라미터를 간단한 공식으로 바로 유도할 수 있게 해줍니다.

주요 컬렉션 분포 조합				
가능도 (데이터)	파라미터	컬렉션 분포	→	사후 분포
Binomial (이항)	$p$ (성공 확률)	Beta (베타)	→	Beta (베타)
Poisson (푸아송)	$\lambda$ (비율)	Gamma (감마)	→	Gamma (감마)
Normal (정규)	$\mu$ (평균)	<b>Normal</b> (정규)	→	<b>Normal</b> (정규)
Normal (정규)	$1/\sigma^2$ (정밀도)	Gamma (감마)	→	Gamma (감마)
Exponential (지수)	$\lambda$ (비율)	Gamma (감마)	→	Gamma (감마)

## 5 핵심 예제: 정규-정규 모델 (Normal-Normal Model)

베이지안 추론의 가장 기본이 되는 ”정규-정규” 모델을 통해 사전-사후 분석이 어떻게 작동하는지 살펴봅니다. 이는 ”데이터(가능도)도 정규분포를 따르고, 파라미터(사전분포)도 정규분포를 따른다”는 의미입니다.

문제 설정:

- 가능도 (데이터):  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$  (우리가 관찰한 데이터는 평균이  $\mu$ 이고 분산이  $\sigma^2$ 인 정규분포에서 나왔다.)
- 가정:  $\sigma^2$  (데이터의 분산)는 이미 알고 있다고 가정합니다.
- 파라미터:  $\mu$  (데이터의 평균)는 모른다.
- 사전분포 (초기 믿음):  $\mu$  역시 정규분포를 따를 것이라고 가정합니다.  $\mu \sim N(\mu_0, \sigma_0^2)$  (여기서  $\mu_0$ 는 ’사전 평균’,  $\sigma_0^2$ 는 ’사전 분산’이며, 둘 다 초매개변수(Hyperparameter)입니다.)

결과 (사후 분포): 결례 관계에 의해, 사후 분포  $f(\mu|X)$  역시 정규분포가 됩니다!  $f(\mu|X) \sim N(\mu_n, \sigma_n^2)$  이때 업데이트된 사후 평균( $\mu_n$ )과 사후 분산( $\sigma_n^2$ )은 다음과 같습니다.

$$\mu_n = \frac{\sigma^2 \mu_0 + n \sigma_0^2 \bar{X}}{\sigma^2 + n \sigma_0^2} \quad | \quad \sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n \sigma_0^2}$$

### ▣ 핵심 요약

#### 정규-정규 모델의 직관적 해석

위 공식은 복잡해 보이지만, 매우 중요한 직관을 담고 있습니다.

1. 사후 평균( $\mu_n$ ) = ”사전 믿음과 데이터의 가중 평균” \*  $\mu_n$ 은 사전 평균( $\mu_0$ )과 데이터 평균( $\bar{X}$ )의 가중 평균입니다. \* 데이터가 많아지면 ( $n \rightarrow \infty$ )? 분자의  $n \sigma_0^2 \bar{X}$  항이 압도적으로 커집니다.  $\mu_n$ 은  $\bar{X}$  (데이터 평균)에 수렴합니다. 결론: 충분한 데이터는 결국 사전 믿음을 이깁니다. \* 사전 믿음이 매우 강하면 ( $\sigma_0^2 \rightarrow 0$ )? ”나의 초기 믿음  $\mu_0$ 는 절대 틀리지 않아!” (분산=0)  $\mu_n$ 은  $\mu_0$  (사전 평균)에 수렴합니다. 데이터( $\bar{X}$ )는 무시됩니다. \* 사전 믿음이 매우 약하면 ( $\sigma_0^2 \rightarrow \infty$ )? ”나는 아무것도 몰라!” (분산=무한대)  $\mu_n$ 은  $\bar{X}$  (데이터 평균)에 수렴합니다.
2. 사후 분산( $\sigma_n^2$ ) = ”데이터가 많을수록 확신이 커진다” \*  $\sigma_n^2$ 의 분모에  $n \sigma_0^2$  항이 있습니다. \* 데이터가 많아질수록 ( $n \rightarrow \infty$ ), 분모가 커져서  $\sigma_n^2$ 는 0에 수렴합니다. \* 결론: 데이터를 더 많이 관찰할수록, 우리의 사후 믿음(추정)은 더욱 확실해집니다. (분산 감소)

## 6 베이지안 추정: 점과 구간

사후 분포  $f(\theta|X)$ 는 파라미터  $\theta$ 에 대한 우리의 최종 믿음을 나타내는 ’분포’입니다. 하지만 종종 보고를 위해 하나의 ’값’(점 추정)이나 ’구간’(구간 추정)이 필요합니다.

### 6.1 점 추정 (Point Estimation)

사후 분포(PDF)를 대표하는 하나의 값을 뽑는 방법입니다.

- 사후 평균 (Posterior Mean):  $E[\theta|X]$ . 사후 분포의 기댓값(평균)을 사용합니다.
- 사후 최빈값 (Posterior Mode, MAP):  $\text{argmax}_\theta f(\theta|X)$ . 사후 분포에서 확률 밀도가 가장 높은 지점

(가장 높은 봉우리)을 사용합니다. 이를 **MAP (Maximum A Posteriori)** 추정이라고 부릅니다.

- 사후 중앙값 (**Posterior Median**): 사후 분포의 50% 지점을 사용합니다.

## 6.2 구간 추정 (Interval Estimation)

### 주의사항

#### 신용 구간 (Credible Interval) vs. 신뢰 구간 (Confidence Interval)

이 들은 매우 다르며, 베이지안의 '신용 구간'이 훨씬 직관적입니다.

- 베이지안 95% 신용 구간 (Credible Interval)
  - 해석: "관찰된 데이터를 바탕으로, 파라미터  $\theta$  가 이 구간 안에 있을 확률이 95%이다."
  - 계산: 사후 분포  $f(\theta|X)$ 의 양쪽 꼬리 2.5%를 잘라내고 남은 95%의 중앙 구간.
  - 직관: 우리가 원하는 확률적 해석과 일치합니다.
- 빈도주의 95% 신뢰 구간 (Confidence Interval)
  - 해석: "이 실험(표본 추출)을 100번 반복하면, 100개의 '신뢰 구간' 중 95개의 구간이 '고정된 참 값  $\theta$ '를 포함할 것이다."
  - 주의: 이미 계산된 하나의 신뢰 구간(예: [10, 20])을 보고 " $\theta$  가 [10, 20] 사이에 있을 확률이 95%다"라고 말하면 틀립니다. 빈도주의에서  $\theta$ 는 고정된 값이라 확률이 없으며, 확률은 '구간'에 걸려있습니다.

## 7 베이지안 선형 회귀 (Bayesian Linear Regression)

이러한 베이지안 원리를 선형 회귀 모델  $y = \beta_0 + \beta_1 x + \epsilon$ 에 적용할 수 있습니다.

**1. 가능성 (데이터 모델):** 선형 회귀의 기본 가정은 ”오차( $\epsilon$ )가 정규분포를 따른다”는 것입니다. 이는  $y$  역시  $x$ 에 따라 평균이 변하는 정규분포를 따른다는 의미입니다.

$y_i \sim N(\mu_i, \sigma^2)$ , 여기서  $\mu_i = \beta_0 + \beta_1 x_i$

**2. 파라미터:** 우리가 추정해야 할 파라미터는  $\beta_0$  (절편),  $\beta_1$  (기울기),  $\sigma^2$  (오차 분산)입니다.

**3. 사전분포 (초기 믿음):** 베이지안 접근법은 이 모든 파라미터에 사전분포를 할당합니다. (켤레 사전분포를 사용한다고 가정)

\*  $\beta_0 \sim N(\mu_0, \sigma_0^2)$  (절편에 대한 사전 믿음) \*  $\beta_1 \sim N(\mu_1, \sigma_1^2)$  (기울기에 대한 사전 믿음) \*  $1/\sigma^2 \sim Gamma(a_0, \lambda_0)$  (오차의 정밀도(분산의 역수)에 대한 사전 믿음)

**4. 사후분포 (결과):** 데이터( $X, y$ )를 관찰하고 베이즈 정리를 적용하면,  $\beta_0, \beta_1, \sigma^2$  각각에 대한 사후 분포를 얻게 됩니다.

### 핵심 요약

#### 베이지안 회귀의 의미

일반 선형 회귀(빈도주의)는  $\beta_1 = 5.0$ 처럼 하나의 값을 추정합니다.

베이지안 회귀는  $\beta_1$ 에 대한 하나의 확률분포를 제공합니다. (예: ” $\beta_1$ 의 사후 분포는 평균이 5.0이고 표준편차가 0.5인 정규분포와 유사하다.”)

이를 통해 우리는 다음과 같은 강력한 확률적 해석이 가능해집니다:

- ” $\beta_1$  (기울기)가 0보다 클 확률은 99.8%이다.”
- ” $\beta_1$ 의 95% 신용 구간은 [4.02, 5.98]이다.”

### 예제: 범주형 변수(더미 변수) 해석하기

한 강의의 수강생 그룹(4개)에 따라 숙제 시간을 예측하는 회귀 모델이 있습니다.

$$\hat{y} = 11.0 - 2.0x_{cs1090} + 3.5x_{cscie109} + 5.0x_{stat109}$$

- **기준 그룹 (Reference Group):** 모델 식에 없는 그룹, 즉  $ac_{209}$  가 기준 그룹(Baseline)입니다.
- **절편 (Intercept = 11.0) 해석:** 모든  $x$ 가 0일 때의  $\hat{y}$  값입니다. 즉, 기준 그룹( $ac_{209}$ ) 학생들의 평균 숙제 시간은 11.0 시간입니다.
- **계수 (Coefficient = 5.0) 해석:**  $x_{stat109}$ 의 계수 5.0은 기준 그룹( $ac_{209}$ )과의 평균 시간 차이'를 의미합니다. ” $stat_{109}$  그룹 학생들은  $ac_{209}$  그룹 학생들보다 평균 5.0시간 더 많이 숙제한다.” ( $stat_{109}$ 의 평균 시간 =  $11.0 + 5.0 = 16.0$  시간)
- **예측 (Prediction):**  $cscie109$  학생의 숙제 시간을 예측하려면,  $x_{cscie109} = 1$ 이고 나머지는 0을 대입 합니다.  $\hat{y} = 11.0 - 2.0(0) + 3.5(1) + 5.0(0) = 14.5$  시간

## 8 심화: 릿지(Ridge)와 라쏘(Lasso)의 베이지안 해석

베이지안 모델링은 릿지(Ridge)와 라쏘(Lasso) 회귀가 왜 그렇게 작동하는지에 대한 강력한 직관을 제공합니다.

### 배경: 손실 함수 (Loss Function)

\* **OLS (최소제곱법):** Loss =  $\sum(y_i - \hat{y}_i)^2$  (오차의 제곱합,  $L_2$ -loss) \* **Ridge (릿지):** Loss =  $\sum(y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2$  (오차의 제곱합 +  $L_2$ -Penalty) \* **Lasso (라쏘):** Loss =  $\sum(y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j|$  (오차의 제곱합 +  $L_1$ -Penalty)

### 핵심 연결 고리: MAP 추정

베이지안에서 MAP (사후 최빈값) 추정은 사후 확률  $f(\beta|X)$ 를 최대화하는  $\beta$ 를 찾는 것입니다.

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}} f(\beta|X) = \underset{\beta}{\operatorname{argmax}} [f(X|\beta)f(\beta)]$$

로그(log)를 써워도 최대화 지점은 동일합니다. (로그는 단조 증가 함수이므로)

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}} [\log(f(X|\beta)) + \log(f(\beta))]$$

이를 '최소화' 문제로 바꾸면, 음수(-)를 붙이면 됩니다.

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}} [-\log(f(X|\beta)) - \log(f(\beta))]$$

이제 이 식을 위 손실 함수들과 비교해 봅시다.

- $f(X|\beta)$ 는  $y \sim N(\beta X, \sigma^2)$  정규분포(가능도)입니다.  $-\log(f(X|\beta))$ 는  $\sum(y_i - \hat{y}_i)^2$  (오차 제곱합) 항에 비례합니다.
- $f(\beta)$ 는  $\beta$  계수들에 대한 사전분포(Prior)입니다.  $-\log(f(\beta))$ 는 폐널티(Penalty) 항에 비례합니다.

### 핵심 요약

손실 함수 최소화 ≡ 사후 확률 최대화 (MAP)

(OLS Loss + Penalty) ≡ (로그-가능도 + 로그-사전분포)

### 8.1 1. 릿지(Ridge) = MAP + 정규(Normal) 사전분포

릿지의  $L_2$ -폐널티  $\lambda \sum \beta_j^2$ 는 어떤 사전분포  $f(\beta)$ 에서 유래할까요?  $-\log(f(\beta)) \propto \beta^2$ 를 만족하는 분포를 찾으면 됩니다.

이는 평균이 0인 정규분포(Gaussian Prior)입니다.  $f(\beta) \propto \exp(-\frac{\beta^2}{2\sigma_p^2}) \implies -\log(f(\beta)) \propto \beta^2$

colback=myblue!5!white, colframe=myblue!75!black, title=□ 릿지(Ridge)의 베이지안 해석 릿지 회귀는  $\beta$  계수들에 대해 평균 0의 정규(Normal) 사전분포를 가정한 베이지안 회귀의 MAP 추정치와 같습니다.

직관: 이 사전분포는 ” $\beta$  계수들은 0 근처에 완만하게(bell-curve) 모여 있을 것이다”라고 믿습니다. 이 믿음이 계수들을 0에 가깝게 ’당기지만(shrinkage)’ 0으로 만들지는 않습니다.

## 8.2 2. 라쏘(Lasso) = MAP + 라플라스(Laplace) 사전분포

라쏘의  $L_1$ -페널티  $\lambda \sum |\beta_j|$ 는 어떤 사전분포  $f(\beta)$ 에서 유래할까요?  $-\log(f(\beta)) \propto |\beta|$ 를 만족하는 분포를 찾으면 됩니다.

이는 평균이 0인 라플라스(Laplace) 사전분포(이중 지수 분포)입니다.  $f(\beta) \propto \exp(-\frac{|\beta|}{b}) \implies -\log(f(\beta)) \propto |\beta|$

colback=mygreen!5!white, colframe=mygreen!75!black, title=□ 라쏘(Lasso)의 베이지안 해석 라쏘 회귀는  $\beta$  계수들에 대해 평균 0의 라플라스(Laplace) 사전분포를 가정한 베이지안 회귀의 MAP 추정치와 같습니다.

직관(Sparsity): 라플라스 분포는 정규분포와 달리  $\beta = 0$  지점에서 ’매우 뾰족한 첨탑’ 모양을 가집니다. 이 사전분포는 ”대부분의  $\beta$  계수들은 정확히 0일 것이다”라고 매우 강하게 믿습니다. 이 강한 믿음이 중요하지 않은 계수들을 0으로 만들어 변수 선택(feature selection)을 수행합니다.

colback=mygray, colframe=darkgray, breakable, title=시각적 비교: 정규 사전분포 vs. 라플라스 사전분포

- 정규분포(Normal, for Ridge):  $\beta = 0$  주변이 ’둥근 언덕’ 모양입니다. 0 근처의 값을 선호하지만, 정확히 0이어야 한다고 강하게 주장하지 않습니다.
  - 라플라스분포(Laplace, for Lasso):  $\beta = 0$  지점이 ’뾰족한 첨탑’ 모양입니다. 확률 질량이 0에 훨씬 많이 몰려있어,  $\beta$ 가 0이 될 확률을 훨씬 높게 부여합니다.
- (슬라이드 37의 두 그래프에 대한 텍스트 설명입니다.)

## 9 계산 방법: MCMC와 갑스 샘플링

### 9.1 문제: 사후 분포가 너무 복잡할 때

켤레 사전분포를 사용하면 사후 분포를 수학 공식으로 깔끔하게 유도할 수 있습니다. 하지만 모델이 복잡해지거나 켤레가 아닌 사전분포를 사용하면, 사후 분포  $f(\theta|X)$ 의 공식을 해석적으로(analytically) 풀 수 없습니다.

### 9.2 해결: 시뮬레이션 (MCMC)

공식을 직접 푸는 대신, 그 사후 분포에서 수많은 샘플을 추출(sampling)하여 분포의 모양을 근사합니다. 이를 MCMC (Markov Chain Monte Carlo) 방법이라고 부릅니다.

\* 비유 (케이크 레시피): 내가 모르는 '신비한 케이크'(사후 분포)가 있습니다.

해석적 방법 (Conjugate): 케이크의 '레시피'(수학 공식)를 알아내는 것입니다.

MCMC 방법 (Simulation): 레시피는 모르지만, 이 케이크를 수만 조각 '샘플링'하여 맛볼 수 있는 기계가 있습니다. 수만 조각을 맛본(샘플링) 후, 우리는 이 케이크의 평균 당도(사후 평균), 당도의 편차(사후 분산) 등을 경험적으로(empirically) 추정할 수 있습니다.

MCMC 샘플( $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$ )을 얻은 후, 추정은 매우 간단해집니다.

\* 사후 평균: 샘플들의 단순 평균  $\frac{1}{N} \sum \theta^{(i)}$  \* 95% 신용 구간: 샘플들을 정렬한 뒤, 2.5% 분위수와 97.5% 분위수를 찾습니다.

### 9.3 갑스 샘플링 (Gibbs Sampling)

MCMC의 대표적인 알고리즘 중 하나로, 다차원 파라미터를 다룰 때 유용합니다. (예:  $\theta_1, \theta_2, \theta_3$  세 개의 파라미터를 동시에 추정해야 할 때)

갑스 샘플링은 결합 사후 분포  $f(\theta_1, \theta_2, \theta_3|X)$ 를 직접 샘플링하기 어려울 때, 대신 '전체 조건부 분포 (Full Conditional Distributions)'를 이용해 번갈아 샘플링합니다.

\*  $f(\theta_1|\theta_2, \theta_3, X) * f(\theta_2|\theta_1, \theta_3, X) * f(\theta_3|\theta_1, \theta_2, X)$

알고리즘 (3개 파라미터 기준): 1. 초기 값 ( $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}$ )을 임의로 설정합니다. 2. 반복 (Iteration  $t = 1, 2, \dots$ ): a.  $\theta_1^{(t)}$ 를  $f(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, X)$ 에서 샘플링. b.  $\theta_2^{(t)}$ 를  $f(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, X)$ 에서 샘플링 (방금 뽑은 새 값 사용). c.  $\theta_3^{(t)}$ 를  $f(\theta_3|\theta_1^{(t)}, \theta_2^{(t)}, X)$ 에서 샘플링 (방금 뽑은 새 값을 사용). 3. 초기 수천 개의 샘플(예: 1 1000번째)은 초기 값의 영향을 받으므로 'Burn-in' 기간이라 부르며 폐기합니다. 4. Burn-in 이후의 샘플 ( $\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}$ )들이 우리가 원하는 결합 사후 분포의 샘플이 됩니다.

colback=mygray, colframe=darkgray, breakable, title=시각적 비유: 메트로폴리스-헤이스팅스 (Metropolis-Hastings) MCMC의 또 다른 유명한 알고리즘은 '산(사후 분포)을 오르는 맹인 등반가'로 비유할 수 있습니다.

1. 등반가는 현재 위치( $\theta$ )에서 임의의 다음 지점( $\theta^*$ )을 제안합니다. 2. **UPHILL (오르막):**  $f(\theta^*) > f(\theta)$  이면, 무조건 이동합니다. (더 가능성 높은 곳) 3. **DOWNHILL (내리막):**  $f(\theta^*) < f(\theta)$  이면, 확률적( $f(\theta^*)/f(\theta)$ )으로 이동합니다. (가파른 내리막(낭떠러지)은 거의 가지 않고, 완만한 내리막은 가끔 갑니다.) 4. 이 과정을 반복하면, 등반가는 결국 산의 정상(MAP) 주변에서 대부분의 시간을 보내게 됩니다.

(슬라이드 48의 그림에 대한 텍스트 설명입니다.)

## 10 자주 묻는 질문 (FAQ) 및 점검

### 주의사항

**Q:** 릿지(Ridge) 모델을 '훈련'할 때는 페널티 항( $\lambda \sum \beta^2$ )을 쓰는데, 왜 '검증(validation)' 할 때는 페널티 항을 제외한 MSE만 사용하나요?

**A:** '훈련'과 '검증'의 목적이 다르기 때문입니다.

- **훈련 (Training)의 목적:** 최적의  $\beta$  계수를 찾는 것입니다. 이때 페널티 항( $\lambda \sum \beta^2$ )은 계수가 너무 커지지 않도록 막는 '규제 장치(regularizer)'입니다. 이 장치는 모델을 만드는 과정(process)의 일부입니다.
- **검증 (Validation)의 목적:** 완성된 모델이 새로운 데이터를 얼마나 잘 예측하는지 평가하는 것입니다. 평가를 할 때는 "그래서 예측 값( $\hat{y}$ )이 실제 값( $y$ )과 얼마나 차이 나는가?"라는 '자연스러운' 예측 오차(natural error metric)만 측정해야 합니다. 훈련 과정의 도구였던 페널티 항을 평가에까지 포함시키면, 모델의 순수한 예측 성능을 왜곡하게 됩니다.

## 11 빠르게 훑어보기 (1-Page Summary)

### ▣ 핵심 요약

#### 1. 베이지안이란?

- 파라미터( $\theta$ )를 '고정된 값'이 아닌 '확률 분포(믿음)'로 본다.
- $P(\theta|X) \propto P(X|\theta)P(\theta)$  공식을 사용한다.
- (최종 믿음)  $\propto$  (증거의 힘)  $\times$  (초기 믿음)

#### 2. 4가지 핵심 요소

- 사전확률 (Prior  $P(\theta)$ ): 내 초기 믿음. (예:  $Uniform(0, 1)$ ,  $N(0, 10)$ )
- 가능성 (Likelihood  $P(X|\theta)$ ): 데이터가 말하는 증거. (예:  $Binomial(n, \theta)$ )
- 사후확률 (Posterior  $P(\theta|X)$ ): 데이터 반영 후 업데이트된 최종 믿음. (우리의 결과물)
- 증거 (Evidence  $P(X)$ ): 정규화 상수. (합을 1로 만듦)

#### 3. 사전분포의 종류

- Informative (정보적): 전문가 지식 반영.
- Uninformative (비정보적): 데이터가 말하게 함. (예:  $Uniform$ )
- Conjugate (켤레): 계산의 편의성. (예: 이항-베타, 정규-정규)

#### 4. 베이지안 vs. 빈도주의 구간

- 신용 구간 (Bayesian): " $\theta$  가 이 안에 있을 확률 95%" (직관적)
- 신뢰 구간 (Frequentist): "구간 100개 뽑으면 95개가  $\theta$ 를 포함" (해석 주의)

#### 5. 릿지/라쏘와의 연결 (MAP 추정)

- Ridge (릿지):  $\beta$ 에 정규(Normal) 사전분포를 가정한 것. (계수를 0 근처로 당김)
- Lasso (라쏘):  $\beta$ 에 라플라스(Laplace) 사전분포를 가정한 것. (뾰족한 분포  $\rightarrow$  계수를 0으로 만듦  
→ 변수 선택)

#### 6. 계산 (MCMC)

- 사후 분포의 공식을 풀기 어려울 때, 대신 수만 개의 샘플을 뽑아 분포를 근사한다.

- **Gibbs Sampling:** 조건부 분포를 이용해 파라미터를 하나씩 번갈아 샘플링.