Day1 : Introduction to Bayesian Approach

[FastCampus] AI센터 베이지안 통계과정

강사: 전인수 (isjeon@vision.snu.ac.kr)

JUN 14, 2019

목차

- Review of Probability
- Probability Model
  - Gaussian Distribution
  - Maximum Likelihood Estimation (MLE)
- Basic Bayesian Theory
  - 동전던지기 실험
  - Conjugate Prior
  - MAP estimation
  - Bayes Update
- Conjugate Relation
  - Gaussian-Gaussian
  - Gaussian-Gamma
  - Beta-Bernoulli

Ch 1. Review of Probability

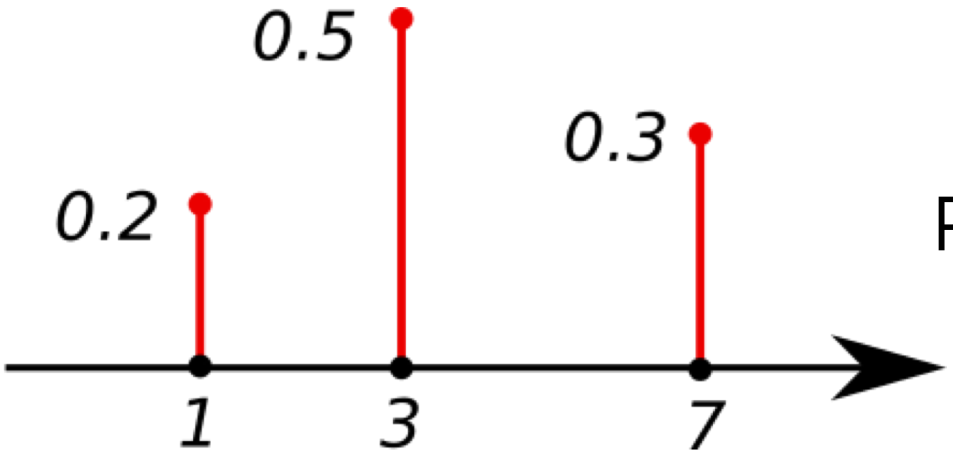
기초적인 확률 이론을 리뷰한다

Probability Distribution

- 확률분포(Probability Dsitribution)란? 표본공간의 부분집합(혹은 특정 사건)에 발생 확률을 수치적으로 부여해 주는 함수로 크게
  - (Discrete) 이산확률변수의 경우 : 확률질량함수 (pmf, probability mass function)
  - (Continuous) 연속확률변수의 경우 : 확률밀도함수 (pdf, probability density function

로 구분 한다.

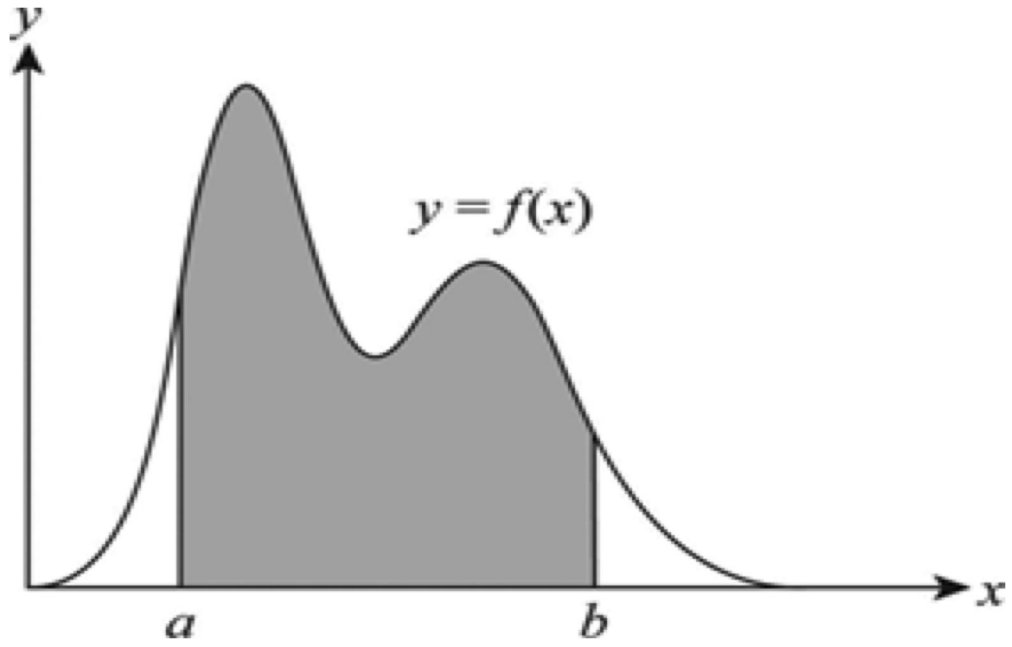
Discrete : Probability Mass Function (PMF)



$$P(x) = \begin{cases} 0.2 & X = 1 \\ 0.5 & X = 3 \\ 0.3 & X = 7 \\ 0 & otherwise \end{cases}$$

1. $0 \leq p_X(x) \leq 1$
2. $\sum_x p_X(x) = 1$
3. $P(X \in B) = \sum_{x \in B} p_X(x)$

Continuous: Probability Density Function (PMF)



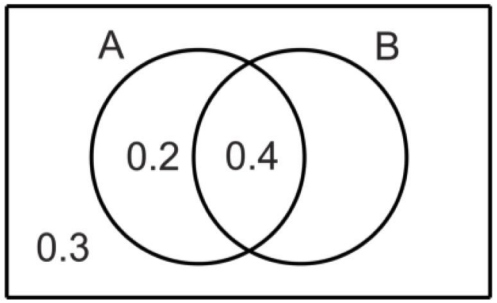
$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

1. $f_X(x) > 1$ is possible.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. $P(X \in B) = \int_{x \in B} f_X(x)dx$

Rule of Probability

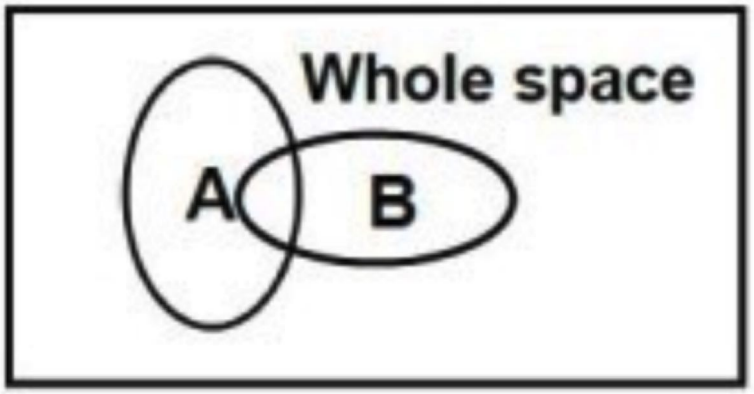
The addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
The multiplication rule: $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$
Independent events: $P(A \cap B) = P(A)P(B)$ *or* $P(A \mid B) = P(A)$
The complement rule: $P(A^c) = 1 - P(A)$

예) Given $P(A \cap B)=0.4$, $P(A \cap B')=0.2$ and $P(A' \cap B')=0.3$. Find $P(B)$ and $P(A|B)$.



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.4}{0.5} = \frac{4}{5} = 0.8.$$

Rule of Probability with Picture



$$P(A) = \frac{\text{A}}{\text{Whole space}} \qquad P(B) = \frac{\text{B}}{\text{Whole space}}$$

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

$$P(A \cap B) = \frac{\text{A}}{\text{Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{A}}{\text{Whole space}} \times \frac{\text{B}}{\text{A}} = \frac{\text{A}}{\text{Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{B}}{\text{Whole space}} \times \frac{\text{A}}{\text{B}} = \frac{\text{A}}{\text{Whole space}} = P(A \cap B)$$

Other Probability Rules

Chain Rule

$$P(X,Y) = P(X|Y)P(Y)$$
$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$
$$P(X_1,\dots,X_n) = \prod_{i=1}^{N} P(X_i|X_1,\dots,X_{i-1})$$

Marginalization

$$P(Y) = \int_x P(X,Y)\,dx \quad \text{(continuous)}$$
$$P(Y) = \sum_x P(X,Y) \quad \text{(discrete)}$$

Probability (with Contingency Table)

- 1973 년 UC-Berkeley Admission Dataset

UCBAdmissions

	남성	여성	total
합격	1198	557	1755
불합격	1493	1278	2771
total	2691	1835	4526

Joint Probability

p(남성, 합격) = 1198/4526 ≈ 0.265
P(남성, 불합격) = 1493/4526 ≈ 0.330
p(여성, 합격) = 557/4526 ≈ 0.123
P(여성, 불합격) = 1278/4526 ≈ 0.282

Marginal Probability

p(합격) = 1755/4526 ≈ 0.388
p(불합격) = 2771/4526 ≈ 0.612

p(남성) = 2691/4526 ≈ 0.595
p(여성) = 1835/4526 ≈ 0.405

Conditional Probability

p(합격|남성) = 1198/2691 ≈ 0.445
p(불합격|남성) = 1493/2691 ≈ 0.555

p(합격|여성) = 557/1835 ≈ 0.304
p(불합격|여성) = 1278/1835 ≈ 0.696

Probability (with Contingency Table)

- 1973 년 UC-Berkeley Admission Dataset

	Dept	A	B	C	D	E	F
합격	남성	512	353	120	138	53	22
	여성	89	17	202	131	94	24
불합격	남성	313	207	205	279	138	351
	여성	19	8	391	244	299	317

	Dept	A	B	C	D	E	F
합격	남성	0.62	0.63	0.37	0.33	0.28	0.06
	여성	0.82	0.68	0.34	0.35	0.24	0.07
불합격	남성	0.38	0.37	0.63	0.67	0.72	0.94
	여성	0.18	0.32	0.66	0.65	0.76	0.93

Conditional Probability

p(합격|남성,A)
= 512/(512+313) ≈ 0.620

p(불합격|남성,A)
= 313/(512+313) ≈ 0.380

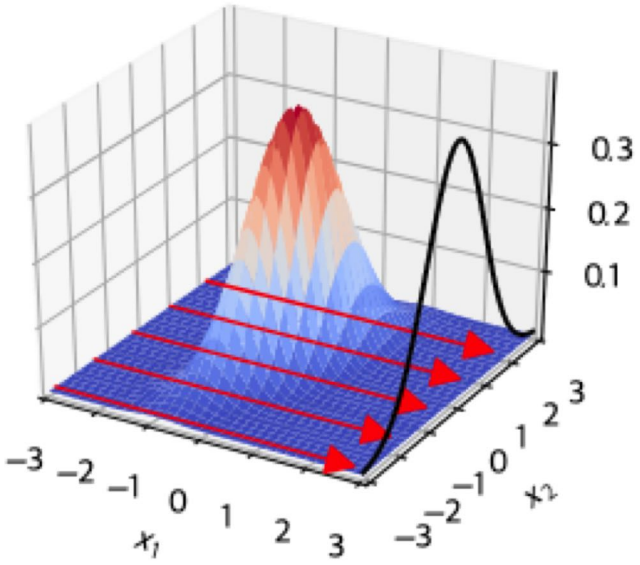
p(합격|여성,A)
= 89/(89+19) ≈ 0.824

p(불합격|여성,A)
= 1278/1835 ≈ 0.175

…

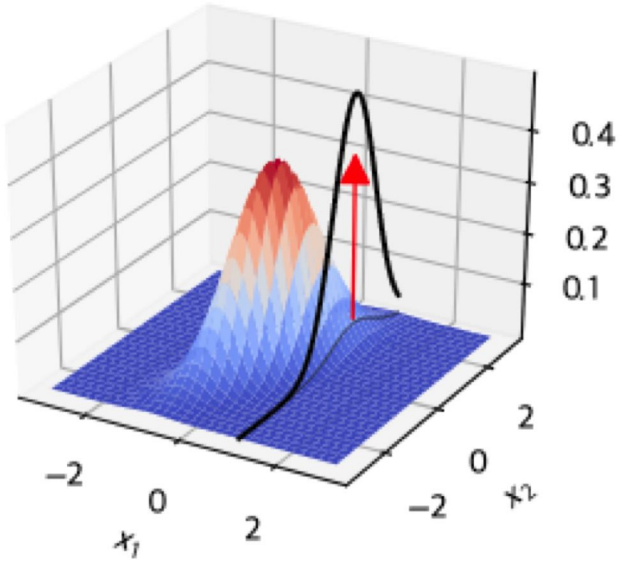
새로운 변수의 도입이 Conditional Probability를 변화 시킬 수 있음

(Continuous) Marginal & Conditional Probability



Marginalization

$$P(X_2) = \int_{x_1} P(X_1, X_2)\, dx_1$$



Conditional

$$P(X_2|X_1 = 1)$$

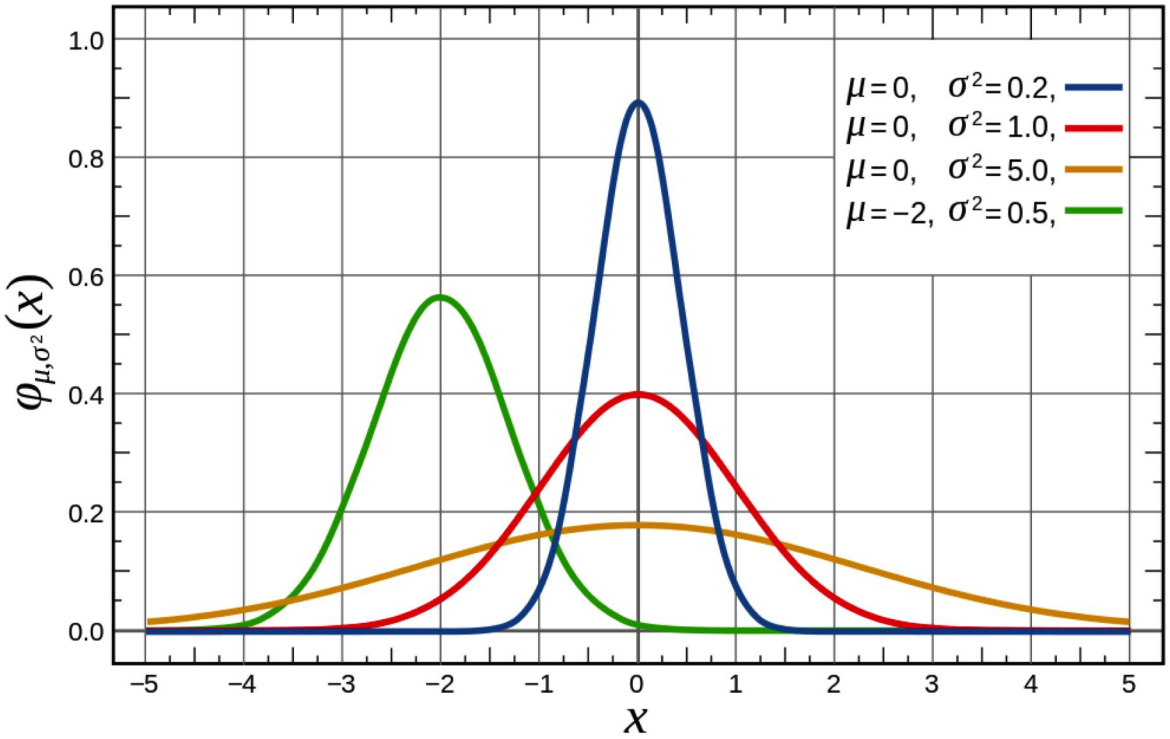
Ch 2. Probability Model & Maximum Likelihood Estimation (MLE)

Gaussian Distribution통해 확률 모형의 개념을 학습한다

확률 모형(Probability Model)이란?

- 확률 모형(Probability Model)이란 수집 및 관측된 데이터의 발생 확률(또는 분포)를 잘 근사하는 "모형"으로 일반적으로 $p(x|\theta)$ 로 표기한다.
  - 확률 모형 (Probability Model) ≈ 통계 모형 (Statistical Model) ≈ 확률 분포 (Probability Distribution).
- 이때, θ는 확률 "모형"을 정의하는 데 중요한 역할을 하는 값으로 모수(parameter, 패라미터) 또는 요약 통계량 (descriptive measure) 부른다.
- 확률 모형은 상황에 따라 $p(x;\theta)$, $p_\theta(x)$, 혹은 $p(x)$와 같이 모수 (또는 parameter) θ 를 생략하고 표기하기도 한다.

Gaussian Distribution



Gaussian 분포란?

기초적인 확률 모형 (또는 확률 분포)로 관찰된 전체 데이터 집합이 평균을 중심으로 하여 뭉쳐져 있는 형태를 표현 하는데 적합하다

$$p(x) \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x; \mu, \sigma^2) \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

모수(parameter) θ 의 추정

- 데이터들이 어떤 확률 분포 $p(x|\theta^*)$에 따라 sampling 되어 구해졌다고 생각해 보자

$$\mathrm{X} = \{x_1, x_2, x_3, \ldots, x_n\},\ x_i \sim p(x|\theta^*)$$

- 모수 θ 추정의 목적은 관측된 데이터의 실제 확률 분포 $p(X|\theta^*)$를 최대한 잘 근사 하는 수학적 모형을 찾는 것이다.

- 사실, 실제의 데이터 확률 분포 $p(x|\cdot)$ 또는 parameter θ^* 등을 정확히 알 수는 없다.

- 따라서, 임의의 확률 모형 $p(x|\cdot)$를 가정한 뒤, 적어도 그 모형이 데이터를 가장 잘 설명하는 parameter θ $(\approx \theta^*)$ 를 찾는 과정을 모수 추정이라고 한다. (예 $p(x|\theta) = \mathcal{N}(x|\mu, \sigma^2), \theta = [\mu, \sigma]$)

MLE (Maximum Likelihood Estimation)

- 모수추정을 위해서 사용하는 기초적인 방법에는 MLE (Maximum Likelihood Estimation)가 있다.
- (MLE) 관측된 데이터 $\mathrm{X} = \{x_1, x_2, x_3, \dots, x_n\}$ 를 토대로 우리가 상정한 확률 모형이 데이터를 가장 잘 설명하도록 θ를 찾는 방법:

$$\hat{\theta} = arg\max_{\theta} L(\theta)$$
$$= \arg\max_{\theta} p(X|\theta)$$
$$= \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)$$

- 이때, $L(\theta)$는 θ에 대한 함수로서 가능도 함수(likelihood function)라 부른다 (또는 우도 함수).
  - (주의) 가능도 함수는 관측된 데이터 $\mathrm{X} = \{x_1, x_2, x_3, \dots, x_n\}$ 를 토대로 정의된 것으로 데이터에 관한 함수가 아니다. 상황에 따라 $L(\theta|X)$ 와 같이 표기하기도 한다.

MLE (Maximum Likelihood Estimation)

$$\hat{\theta}_{ML} = \text{argmax}_\theta \; p(x \mid \theta)$$

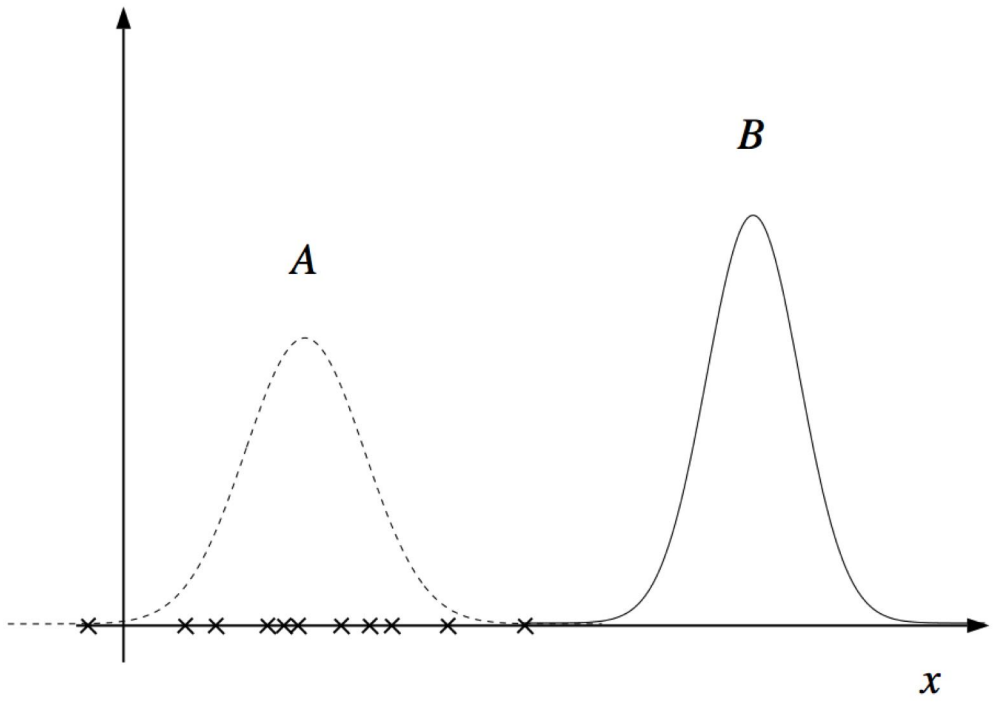


Figure 5.1: A univariate density estimation problem. (See Section 5.2.1 for a discussion of density estimation). The data $\{x_1, x_2, \ldots, x_N\}$ are given as X's along the abscissa. The parameter vector θ is the mean μ and variance σ^2 of a Gaussian density. Two candidate densities, involving different values of θ, are shown in the figure. Density A assigns higher probability to the observed data than density B, and thus would be preferred according to the principle of maximum likelihood.

Log-likelihood function $l(\theta)$

- MLE 식을 최적화하기 위해서 실제로는 log를 취한 로그 우도 log-likelihood $l(\theta)$의 형태를 대신 사용하며 다음과 같이 표기한다.

$$\hat{\theta} = arg\max_{\theta} L(\theta)$$
$$= arg\max_{\theta} l(\theta)$$
$$= \arg\max_{\theta} \log \prod_{i=1}^{n} p(x_i|\theta)$$
$$= \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)$$
$$= \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta)$$

- 마지막 식을 (데이터셋 $\{x_i\}_{i=1}^{n}$으로 부터 정의된) empirical expectation 이라고 부르고 다음과 같이 표현하기도 한다.

$$\mathrm{E}_{\{x \sim p(x|\theta^*)\}} [\log p(x_i|\theta)] \approx \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta)$$

MLE with Gaussian

- 예를들어 확률 분포를 Gaussian이라고 가정하고 (예 $p(x|\theta) = \mathcal{N}(x|\mu, \sigma^2)$, $\theta = [\mu, \sigma]$) 모수 추정값을 계산해보자

$$\hat{\theta} = arg\max_{\theta} l(\theta)$$

$$l(\mu, \sigma) = \sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu, \sigma^2)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

MLE with Gaussian

- $l(\mu, \sigma)$을 최대로 만드는 μ를 구하기 위해 각 μ 로 미분을 취한 후 0 이 되는 식에서 μ를 유도해보면…

$$\frac{dl(\mu,\sigma)}{d\mu} = \frac{d}{d\mu}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \right)$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2x_i + 2\mu)$$

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

- 즉 , $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ 임을 알 수 있다.

MLE with Gaussian

- 마찬가지로 $l(\mu,\sigma)$을 최대로 만드는 σ를 구하기 위해 미분을 취한 후 0이 되는 식에서 σ 를 유도해보면…

$$\frac{dl(\mu,\sigma)}{d\sigma} = \frac{d}{d\,\sigma}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$
$$= -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\frac{d}{d\,\sigma}\left(\frac{1}{\sigma^2}\right)$$
$$= -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\times 2\frac{1}{\sigma}\left(-\frac{1}{\sigma^2}\right)$$

$$-\frac{\mathrm{n}}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

- 즉 , $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ 임을 알 수 있다.

MLE (Maximum Likelihood Estimation)

$$\hat{\theta}_{ML} = \text{argmax}_\theta \; p(x \mid \theta)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

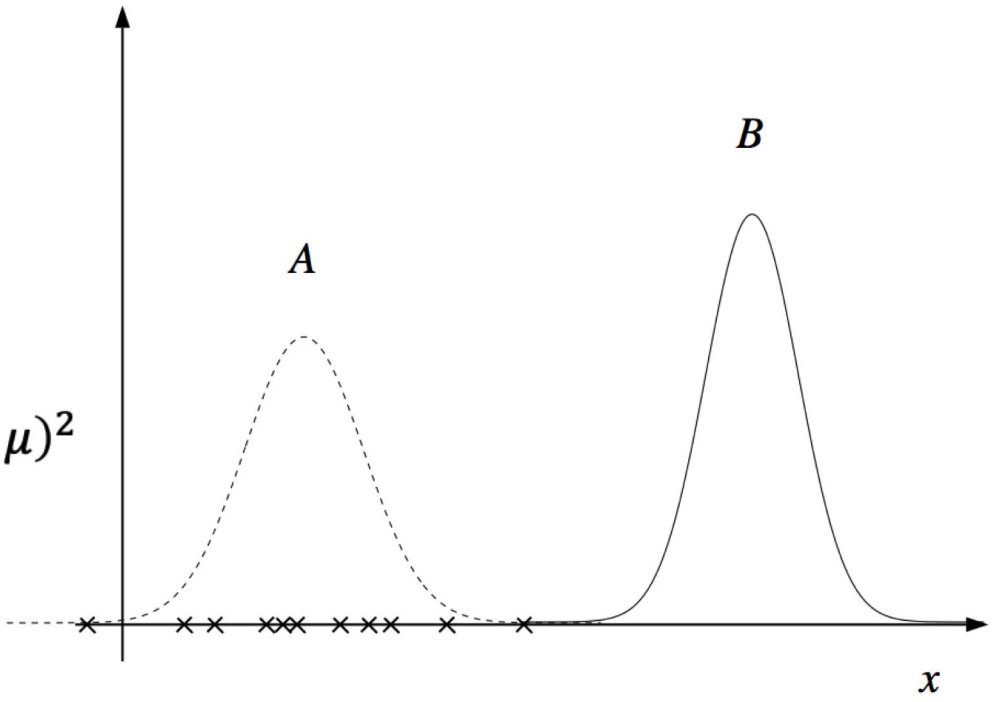


Figure 5.1: A univariate density estimation problem. (See Section 5.2.1 for a discussion of density estimation). The data $\{x_1, x_2, \ldots, x_N\}$ are given as X's along the abscissa. The parameter vector θ is the mean μ and variance σ^2 of a Gaussian density. Two candidate densities, involving different values of θ, are shown in the figure. Density A assigns higher probability to the observed data than density B, and thus would be preferred according to the principle of maximum likelihood.

Ch 2. Summary

- 확률 모형(Probability Model)이란 수집 및 관측된 데이터의 발생 확률(또는 분포)를 잘 근사하는 "모형"으로 일반적으로 $p(x|\theta)$ 로 표기한다.

- (MLE) 관측된 데이터 $\mathrm{X} = \{x_1, x_2, x_3, \ldots, x_n\}$ 를 토대로 우리가 상정한 확률 모형이 데이터를 가장 잘 설명하도록 θ를 찾는 방법이다.

- MLE에서 $L(\theta)$= $p(X|\theta)$ 는 θ에 대한 함수로 해석하고 우도 함수(likelihood function)라 부른다.

- MLE 방식로 모수 추정을 하기 위해서는 수식에 log를 붙인 로스 우도 함수 (log-likelihood function) $l(\theta)$의 형태를 사용한다.

Ch 3. Bayesian Theory

동전던지기 실험을 통해 Bayesian Paradigm을 학습해 보자

Bernoulli Distribution

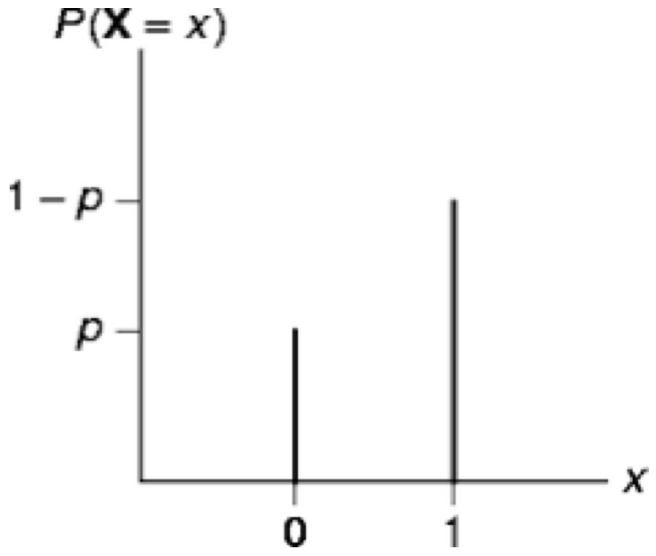
- 베르누이 분포란?
  - 확률 이론 및 통계학에서 자주 사용 되는 분포로서 동전 던지기의 앞면 뒷면, 입시의 합격과 불합격, 사업의 성공과 실패, 수술 후 환자의 치유 여부 등, 어떤 실험이 두 가지 가능한 결과만을 가질 경우 이를 표현하는 확률 모형이다.

Bernoulli RV

$$X \in \{0,1\}$$

$$P(X=1) = p$$

$$f(x) = p^x(1-p)^{1-x}$$



동전던지기 실험

- 동전 던지기 실험을 통해 다음과 같은 데이터를 얻었다.

Head Tail Tail Tail Tail

- 관측된 데이터로 부터, 같은 동전을 던질때 앞면이 나타날 확률은 Likelihood를 이용해서 계산한다 (학습 or 예측).

$$\underbrace{P(x|\theta)}_{\text{Likelihood}} = \prod_{i=1}^{5} P(x_i|\theta) \quad \text{(동전던지기는 독립적으로 시행)}$$

$$= \underbrace{\theta}_{\text{Parameter}} \cdot (1-\theta) \cdot (1-\theta) \cdot (1-\theta) \cdot (1-\theta)$$

Parameter: 각 시행에서 던진 동전의 앞면이 나타날 확률

Likelihood 계산

- Bernoulli Distribution 이하 각 데이터의 발생 확률을 최대로 하는 Parameter θ를 찾기 위해 Likelihood를 계산해보자

Likelihood of parameter θ_A: $L(\theta_A) = P(x|\theta_A) = \prod_{i=1}^{N} P(x_i|\theta_A)$

Likelihood of parameter θ_B: $L(\theta_B) = P(x|\theta_B) = \prod_{i=1}^{N} P(x_i|\theta_B)$

어떤 parameter θ가 관측된 데이터 x_i들을 잘 설명하는가?

가장 그럴 듯한 θ는 무엇인가?
어떻게 결정하나?
가장 그럴듯한 θ? $\theta = 1/5$

Maximum Likelihood Estimation (MLE)

- MLE를 통해 Log-likelihood를 최대화 하는 θ를 계산해보면..

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\ l(\theta)$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{5} \log P(x_i|\theta)$$

$$= \frac{\partial}{\partial \theta} \{\log \theta + 4 \log(1-\theta)\}$$

$$= \frac{1}{\theta} - \frac{4}{1-\theta}$$

Maximization condition

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \quad \Longleftrightarrow \quad \theta_{ML} = \frac{1}{5}$$

MLE의 가정

R. Fisher
(1890-1962)

- 데이터가 N개 관측 될수록, 예측 오차는 $1/N$ 오더씩 울어든다
- MLE 는 asymptotically unbiased 하다.

만약 무한대의 관측 데이터가 주어질 경우 MLE로 예측한 Parameter는 실제 Parameter로 수렴한다

그런데 만약..

우리에게 제한된 수량의 관측 데이터만 주어진다면..

MLE의 문제점

- 과연 5번의 시도 만으로 앞면의 확률이 1/5라고 단정지을 수 있을까?

Head Tail Tail Tail Tail

- MLE는 초기 관측에 쉽게 overfitting한다..
  - 극단적으로 동전던지기를 1번 해서 앞면이면.. 앞면의 확률이 100%?

이후 5번을 더 던졌더니…

Head Head Head Head Tail

Bayesian approach

경험적으로 확률이 반반인 동전이 많았다.
Parameter θ에 대해 우리의 경험을 바탕으로 한 확률적인 가정을 더하자.

$$\theta = \frac{1}{2}$$

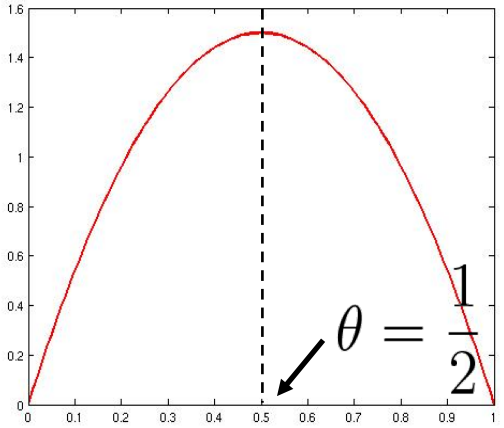
T. Bayes
(1702-1761)

Likelihood Prior

Posterior

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int d\theta P(x|\theta)P(\theta)}$$

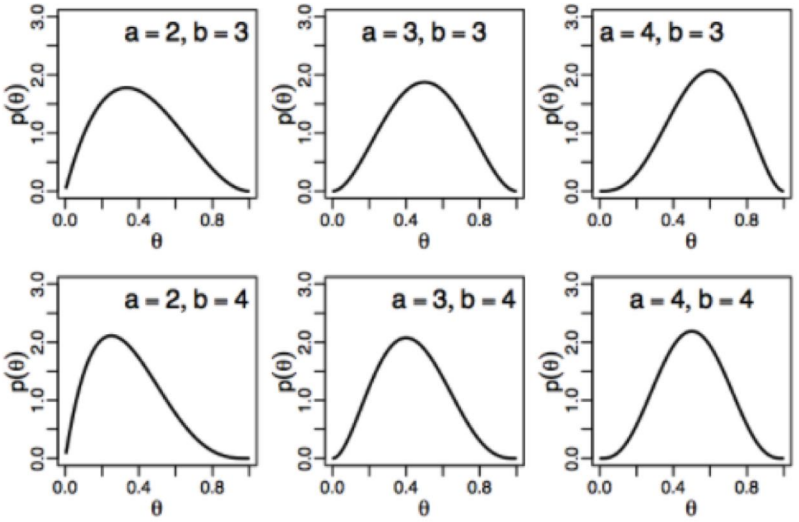
사후 정보 = 데이터로 얻어진 정보 + 사전 정보

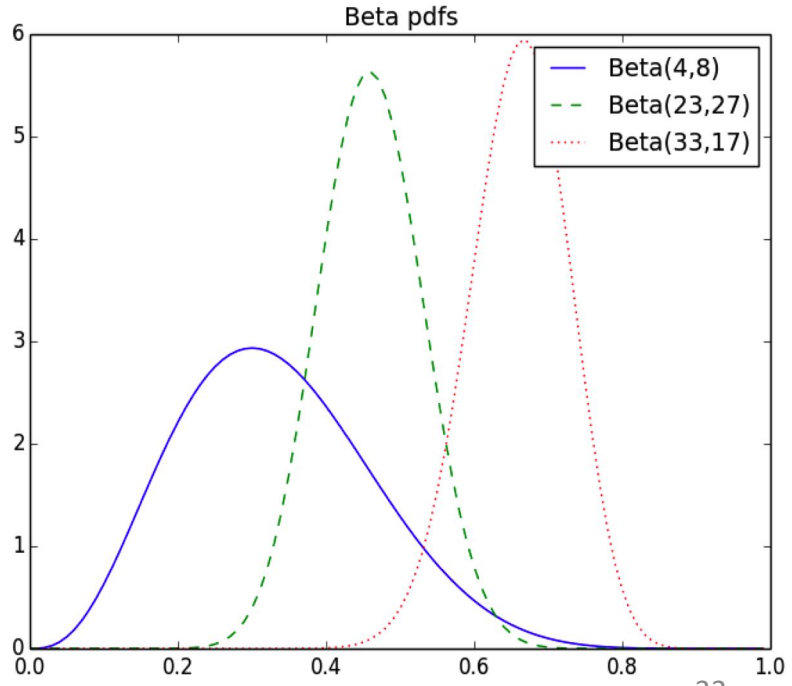


P(θ)는? Beta Distribution

- Bernoulli Likelihood에서 θ는 확률 동전앞면이 나올 확률이었다.
- Prior $P(\theta)$는 확률에 대한 확률 분포로서 **Beta Distribution**을 사용한다.

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$





Conjugate Prior란?

- Conjugate prior는 Bayes Rule에 의해 식을 유도하였을 때, posterior가 prior와 같은 distribution 형태를 갖게 하는 prior이다.

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

Bayes Theorem

$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Bernoulli Likelihood

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

Beta Prior

$$P(\theta|x) = \frac{\theta^x(1-\theta)^{1-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(x)B(\alpha,\beta)} \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+(1-x)-1} = Beta(\hat{\alpha},\hat{\beta})$$

Posterior

Beta with updated parameter

Conjugate Relation 예시

Form of the Likelihood Function as a Function of the Parameter of Interest	Parameter	Conjugate prior	Posterior distribution of the parameter $f_{\theta\|\mathbf{x}}(\theta\|\underline{x})$
1. Bernoulli likelihood: $f_{\mathbf{x}\|\theta}(\mathbf{x}\|p) \propto p^{n\bar{x}}(1-p)^{n(1-\bar{x})}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = p$	$\theta \sim beta(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$ $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta\|\mathbf{x} \sim beta(\alpha', \beta')$, where $\alpha' = \alpha + n\bar{x}$ $\beta' = \beta + n(1-\bar{x})$
2. Poisson likelihood: $f_{\mathbf{x}\|\theta}(\mathbf{x}\|\mu) \propto \begin{cases} \mu^{n\bar{x}} e^{-n\mu} & \begin{pmatrix}\text{if } x_i = 0,1,\ldots,\\ \text{for } i = 1,2,\ldots,n\end{pmatrix} \\ 0 & \text{otherwise,} \end{cases}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = \mu$	$\theta \sim G(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$ $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}} & \text{if } \theta > 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta\|\mathbf{x} \sim G(\alpha', \beta')$, where $\alpha' = \alpha + n\bar{x}$ $\beta' = \frac{\beta}{1+n}$
3. Negative binomial likelihood: $f_{\mathbf{x}\|\theta}(\mathbf{x}\|p) \propto \begin{cases} p^{nr\bar{x}}(1-p)^{n(\bar{x}-r)} & \begin{pmatrix}\text{if } x_i \geq r\\ \text{for } i = 1,2,\ldots,n\end{pmatrix} \\ 0 & \text{otherwise,} \end{cases}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\theta = p$	$\theta \sim beta(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$ $g(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta\|\mathbf{x} \sim beta(\alpha', \beta')$, where $\alpha' = \alpha + nr$ $\beta' = \beta + n(\bar{x}-r)$
4. Uniform likelihood: $f_{\mathbf{x}\|\theta}(\mathbf{x}\|L) \propto \begin{cases} \frac{1}{L^n} & \text{if } L > \max\{x_i\} \\ 0 & \text{otherwise} \end{cases}$	$\theta = L$	$\theta \sim Pareto(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$ $g(\theta) = \begin{cases} \frac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}} & \text{if } 0 > \beta \\ 0 & \text{otherwise} \end{cases}$	$\theta\|\mathbf{x} \sim Pareto(\alpha', \beta')$, where $\alpha' = \alpha + n$ $\beta' = \max\{x_1,\ldots,x_n,\beta\}$
5. Pareto likelihood: $\mathbf{x}\|b \sim Pareto(a,b)$ $f_{\mathbf{x}\|\theta}(\mathbf{x}\|b) \propto \begin{cases} b^{na} & \text{if } \min\{x_1,\ldots,x_n\} > b \\ 0 & \text{otherwise} \end{cases}$	$\theta = b$	$\theta \sim Pareto(\alpha,\beta)$ with $\alpha > 0$ and $\beta > 0$ $g(\theta) = \begin{cases} \frac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}} & \text{if } 0 > \beta \\ 0 & \text{otherwise} \end{cases}$	$\theta\|\mathbf{x} \sim Pareto(\alpha', \beta')$, where $\alpha' = \alpha - an$, $\beta' = \beta$ with $\alpha > na$

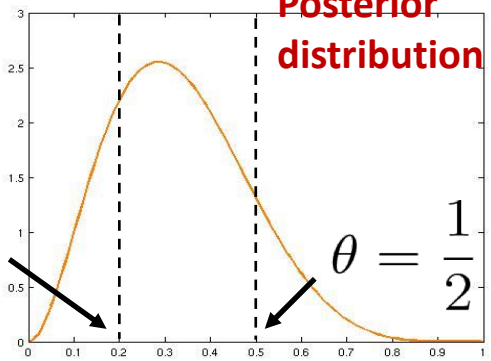
Posterior Distribution

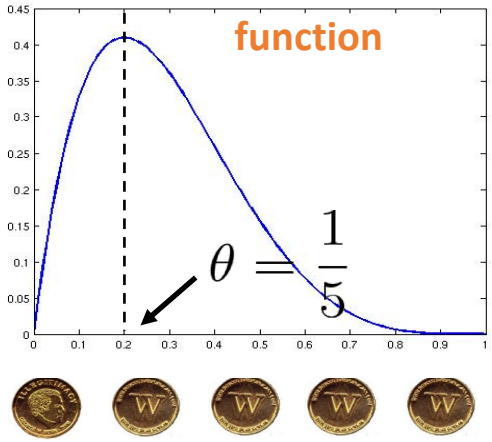
Posterior

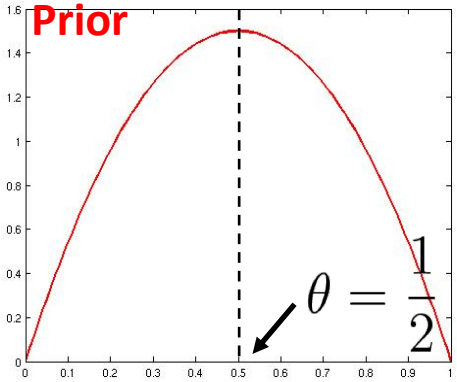
Likelihood Prior

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int d\theta P(x|\theta)P(\theta)}$$

$\theta = \frac{1}{5}$







Posterior

$$P(\theta|x) = \frac{\theta^x(1-\theta)^{1-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(x)B(\alpha,\beta)}$$

$$\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+(1-x)-1} = Beta(\hat{\alpha},\hat{\beta})$$

updated Prior with observed data

Maximum A Posterior Estimation (MAP)

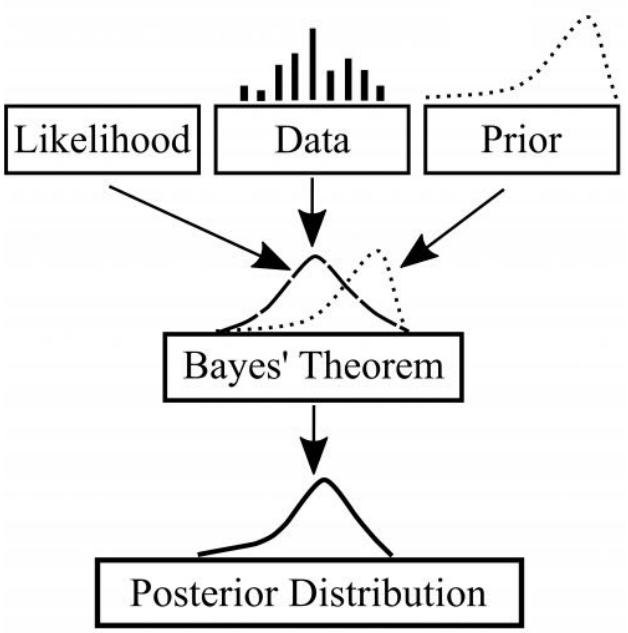
- Likelihood에 추가적으로 Prior를 함께 고려하여, Posterior 분포를 최대화 하는 모수 θ를 추정하는 방법이다.

$$\hat{\theta}_{MAP} = arg\max_{\theta} p(\theta|X)$$
$$= \arg\max_{\theta} p(X|\theta)p(\theta)/p(X)$$
$$= \arg\max_{\theta} p(X|\theta)p(\theta)$$
$$= \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)\, p(\theta)$$

- 일반적으로 log를 취한 후 Optimization을 통해 계산한다.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(X|\theta)\, p(\theta)$$
$$= \arg\max_{\theta} \log p(X|\theta) + \log p(\theta)$$
$$= \arg\max_{\theta} \log \prod_{i=1}^{n} p(x_i|\theta) + \log p(\theta)$$
$$= \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta) + \log p(\theta)$$

Full Bayesian Approach (= Posterior 추정)

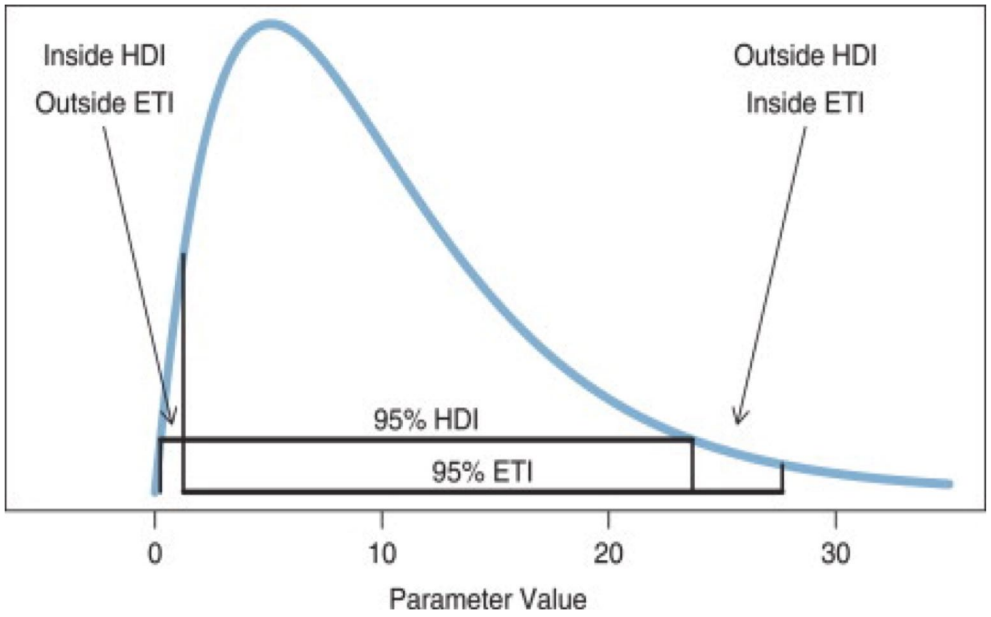


$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

- MAP estimation은 Optimal한 θ에 대한 point estimation
- Full Bayesian Approach는 Posterior Distribution 자체를 구하는 것이 issue
- Prior와 likelihood가 conjugate 관계의 경우 Posterior를 쉽게 구할 수 있음
- Posterior Distribution == 데이터를 관측한 후 θ에 대한 믿음

신용 구간 (Credibility Interval)

- Posterior Distribution를 참조하여 Parameter에 대한 신용 구간 (Credibility Interval)을 구할 수 있다



High-Density Interval (HDI)

Equal-Tailed Interval (ETI)

- 베이즈 신용 구간은 사전 분포로부터 문제에 특화된 문맥 정보와 혼합되지만 빈도주의 신뢰 구간은 오로지 데이터에만 기초한다.

Predictive Distribution

- Posterior 분포 $p(\theta|X)$를 이용해서, posterior predictive distribution을 다음과 같이 계산 할 수 있다.

$$\mathrm{p}(\tilde{x}|X) = \int_{\theta} \mathrm{p}(\tilde{x}|\theta, X)p(\theta|X)\, d\theta$$

- 우리가 가지고 있는 데이터 X에 기반하여, 관측하지 않은 새로운 데이터 $\tilde{x}$에 대한 확률을 모델링 한 것.

- 직관적으로 Ensemble (Boostrap aggregation, Baggina)기법을 생각해 보자..

베이즈 갱신 (Bayes Update)

- Posterior는 likelihood를 통해서 Data를 관측하고 정보가 업데이트된 Prior다.

$P(z)$

$$P(z|x_1) = \frac{P(x_1|z)\,P(z)}{P(x_1)}$$

$$P(z|\{x_1, x_2\}) = \frac{P(x_2|z\,)\,P(z|x_1)}{P(x_2)} \quad (= \frac{P(x_2|z, x_1\,)\,P(z|x_1)}{P(x_2|x_1)})$$

$$P(z|\{x_1, x_2, x_3\}) = \frac{P(x_3|z\,)\,P(z|\{x_1, x_2\})}{P(x_3)}$$

…

$P(z|\boldsymbol{X})$ -> 데이터가 많이 쌓일 수록 Posterior는 정확해진다.

Ch 3. Summary

- 관측된 데이터를 가장 잘 설명하는 Parameter θ (분포)를 찾기 위해 MLE 사용

- Bayesian Approach는 Parameter에 확률 분포를 추가적으로 가정한 후 Posterior를 통해서 Optimal Parameter θ를 추정 (MAP)

- MAP estimation의 Analytical (Closed Form) Solution을 얻기 위해서는 conjugate 관계가 필수적
  - Conjugate 관계가 성립하지 않을 시? EM, MCMC, Variational Inference 등 활용

- Posterior는 likelihood를 통해서 Data를 관측하고 정보가 업데이트된 Prior (Prior로 재활용 가능)

Ch 4. Conjugate Relation

Gaussian-Gaussian, Gamma-Gaussian, Beta-Bernoulli

Gaussian-Gaussian Conjugacy

$$P(X|\theta) = \mathcal{N}(X|\theta, \sigma^2)$$

$$\mathcal{A}(v) = \mathcal{N}(\theta|a, b^2)$$

$\mathcal{N}(X|\theta, \sigma^2)$ $\mathcal{N}(\theta|m, s^2)$

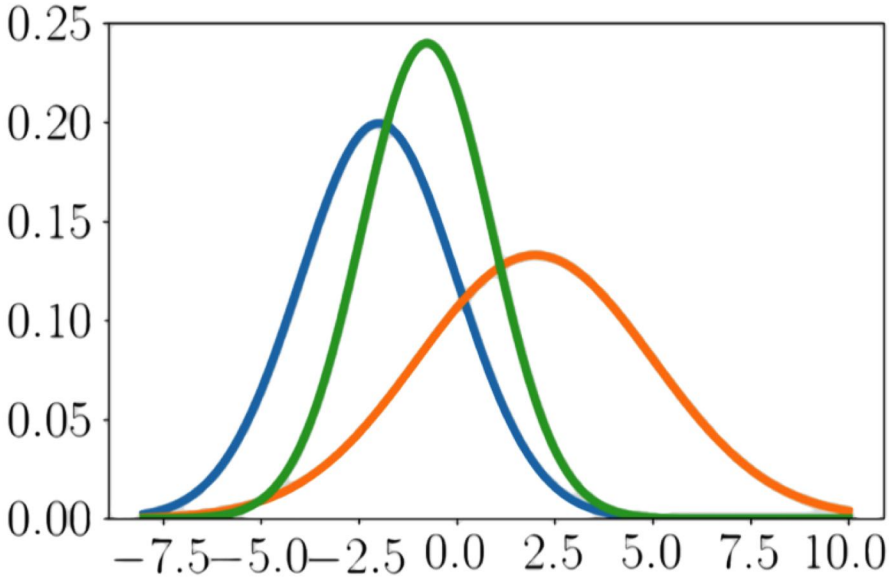
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$\mathcal{N}(\theta|a, b^2)$

Gaussian-Gaussian Conjugacy

$$P(X_1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad P(X_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \text{const} \cdot e^{-\text{parabola}}$$



Gaussian-Gaussian Conjugacy

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

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

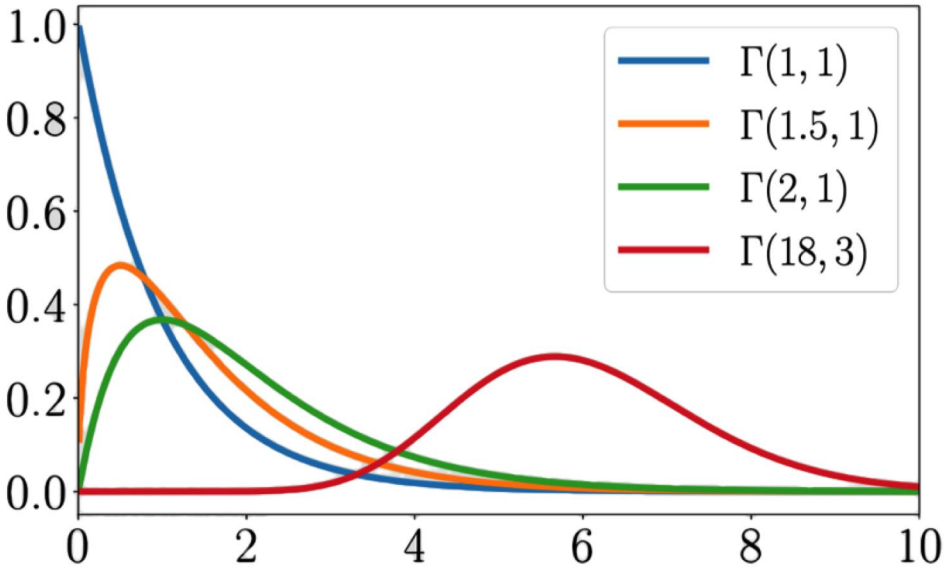
$$p(\theta|x) \propto e^{-(\theta-\frac{x}{2})^2}$$

$$p(\theta|x) = \mathcal{N}(\theta|\frac{x}{2},\frac{1}{2})$$

Gamma Distribution

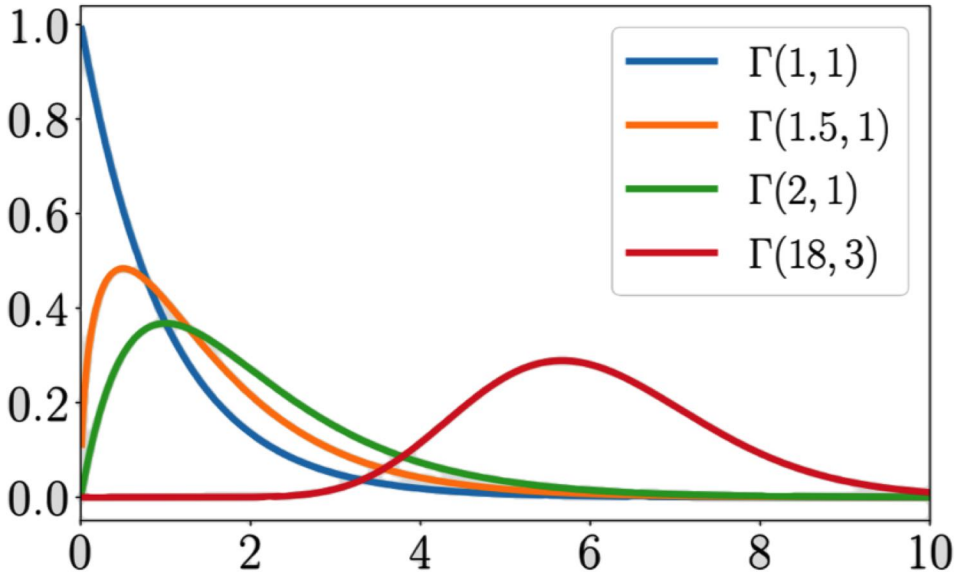
$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$$\gamma, a, b > 0$$



Gamma Distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

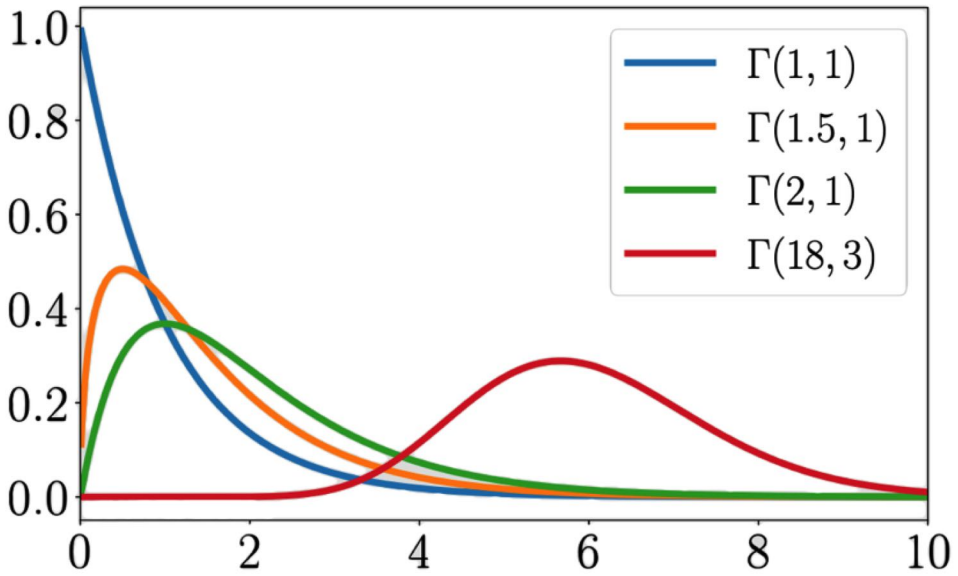


Gamma Distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$\Gamma(n) = (n-1)!$

$\Gamma(5) = 24$



Gamma Distribution – Statistics

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$$\mathbb{E}[\gamma] = a/b$$

$$\mathrm{Mode}[\gamma] = \frac{a-1}{b}$$

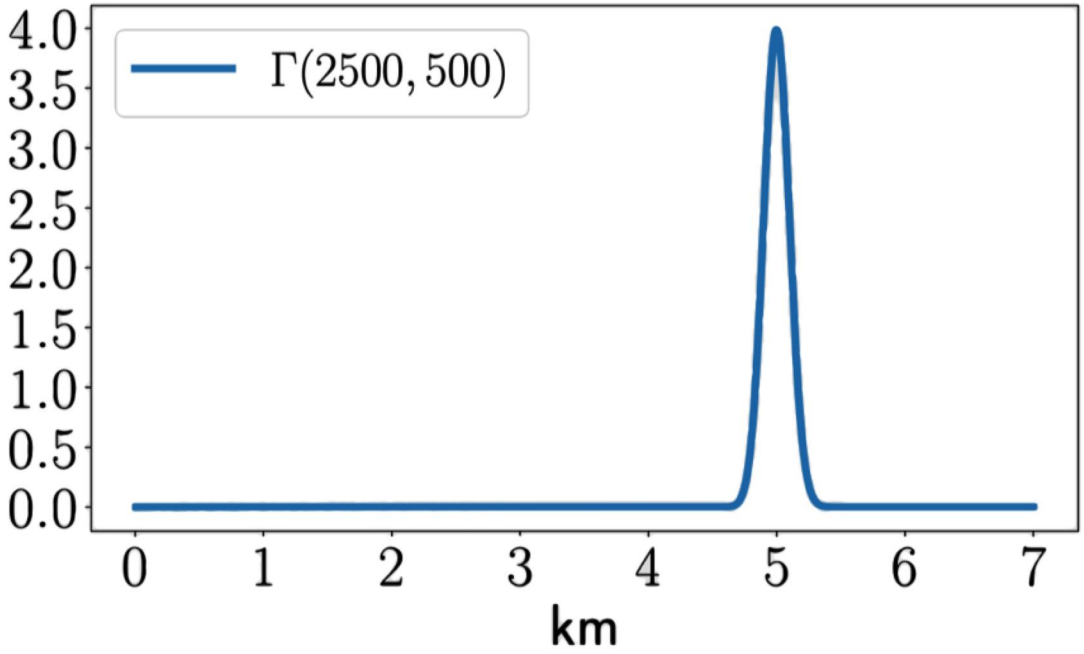
$$\mathrm{Var}[\gamma] = a/b^2$$

Gamma Distribution - Example

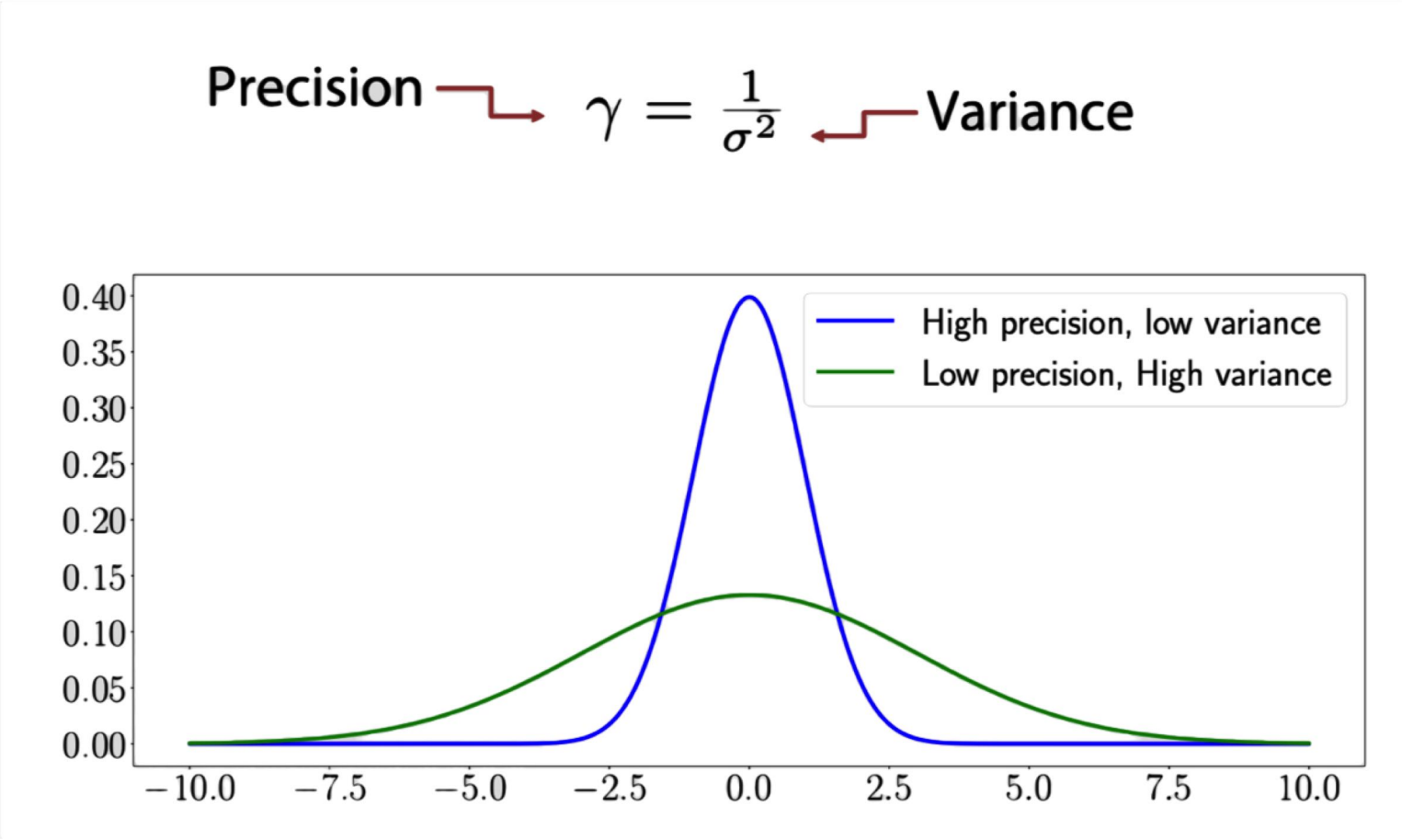
You run 5km $\pm$ 100m a day

$$\mathbb{E}[x] = {}^{a}/_{b} = 5,\ Var[x] = {}^{a}/_{b^2} = 0.1^2$$

$$\Rightarrow a = 2500, \mathrm{b} = 500$$



Precision


Precision
$\gamma = \frac{1}{\sigma^2}$
Variance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
High precision, low variance
Low precision, High variance

Gaussian (Variance vs Precision)

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathcal{N}(x|\mu,\gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}} e^{-\gamma\frac{(x-\mu)^2}{2}}$$

Gamma-Gaussian Conjugacy

$$p(\gamma) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma}$$

$$p(\gamma|x) \propto p(x|\gamma)p(\gamma)$$

$$p(\gamma|x) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(x-\mu)^2}{2}}\right)\cdot\left(\gamma^{a-1}e^{-b\gamma}\right)$$

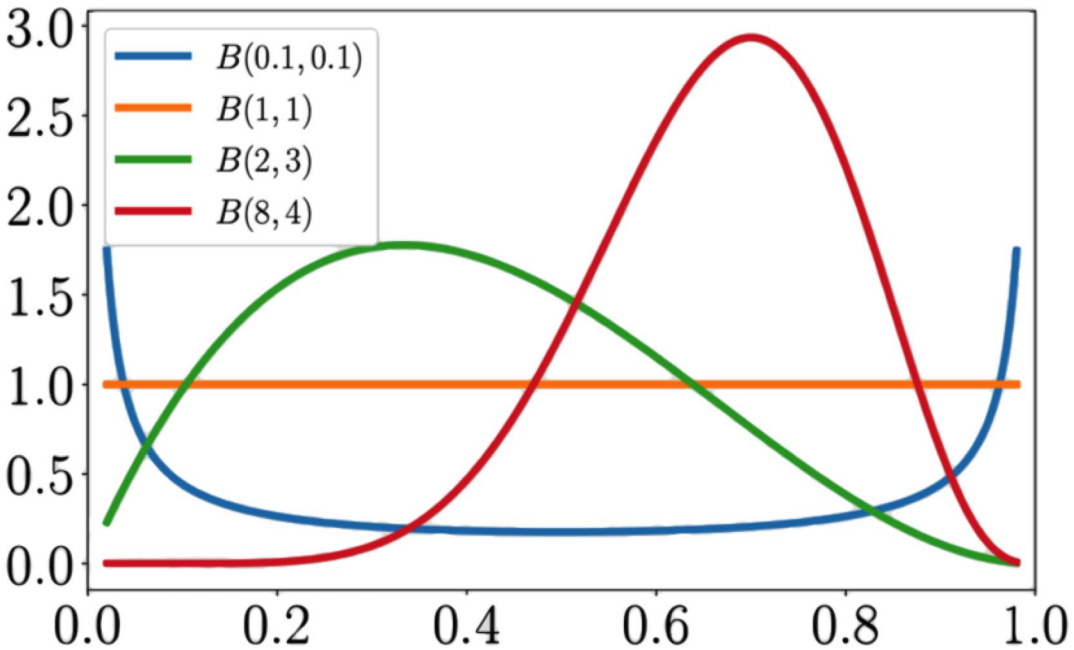
$$p(\gamma|x) \propto \gamma^{\frac{1}{2}+a-1}e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

$$p(\gamma|x) = \Gamma(a+\frac{1}{2}, b+\frac{(x-\mu)^2}{2})$$

Beta Distribution

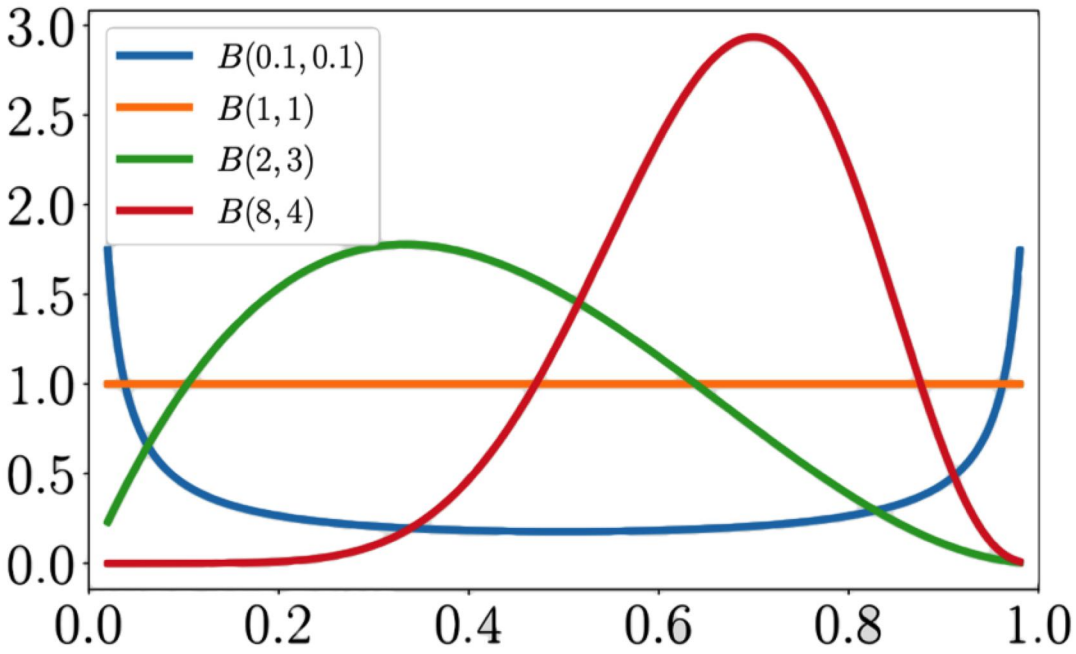
$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$x \in [0,1], \quad a,b > 0$$



Beta Distribution

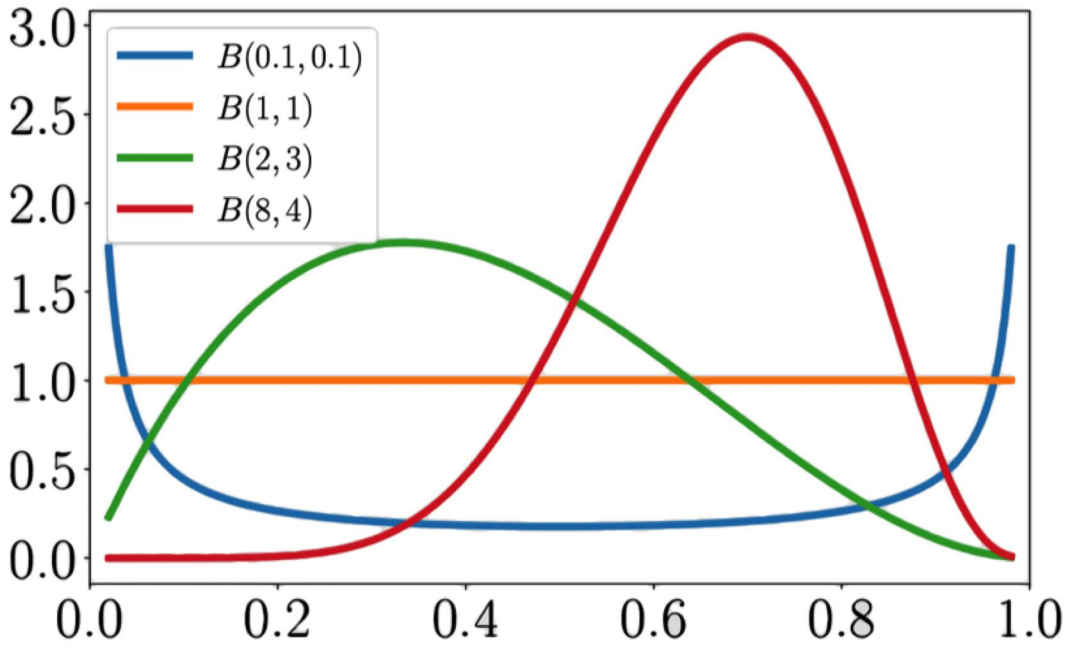
$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$



Beta Distribution

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$



Beta Distribution - Statistics

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$\mathbb{E}x = \frac{a}{a+b}$$

$$\text{Mode}[x] = \frac{a-1}{a+b-2}$$

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)}$$

Beta Distribution - Example

Movie rank is 0.8 ± 0.1

1 — best movie

0 — Batman & Robin

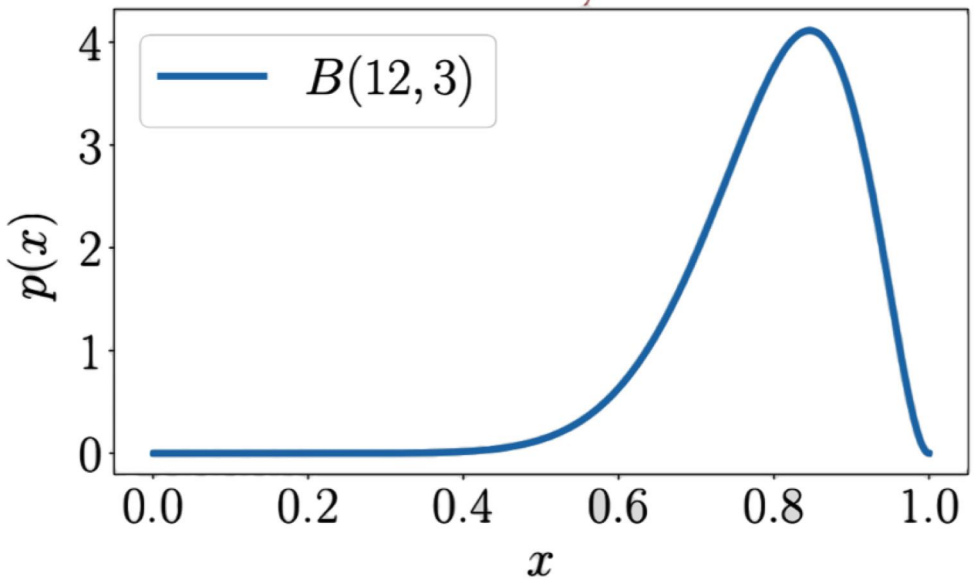
Beta Distribution - Example

Movie rank is 0.8 ± 0.1

$$\mathbb{E}x = \frac{a}{a+b} = 0.8$$

$$\mathrm{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)} = 0.1^2$$

$$\Rightarrow a = 12, b = 3$$



Beta-Bernoulli conjugacy

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

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

$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$p(\theta|X) = B(N_1+a, N_0+b)$$

수고하셨습니다!