# Introduction to Bayesian Inference

## 1. Bayes' rule

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int_{\Theta} p(y \mid \tilde{\theta})\, p(\tilde{\theta})\, d\tilde{\theta}} \propto p(y \mid \theta)\, p(\theta)$$

**새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도(확률)를 갱신해 나가는 방법!**

### → 뭘로? Likelihood로!

확률적인 배경 지식을 가지고 특별한 추가 정보 없이 샘플을 분류하는 예시를 생각해보자.

- 아무 사람이나 데리고 와서 남자인지 여자인지 분류하라고 하면 어떻게 분류할까? → 세상에 절반은 남자, 절반은 여자이므로 50% 확률로 어림짐작 → $p$(성별 = 남자) 혹은 $p$(성별 = 여자)
- haha ha의 삼색이를 데리고 와서 성별을 확인해본다고 하자. → 일반적으로 삼색 고양이는 성염색체 관련 이유로 대부분이 암컷으로 알려져 있음. → 매우 높은 확률로 삼색이는 암컷일 것임!

$\theta$ : 모수

$\Theta$ : 모수 공간

$y$ : 표본 데이터

$p(\theta)$ : 사전 확률. 현재 가지고 있는 정보를 기초로 하여 정한 초기 확률. 확률 시행 전에 이미 가지고 있는 지식을 통해 부여한 확률.

$p(y \mid \theta)$ : 가능도. 어떤 값이 관측되었을 때 이 값이 어떤 확률분포에서 왔는지를 나타내는 확률. 추가되는 특정 정보!

$p(\theta \mid y)$ : 사후 확률. 사건 발생 후에 어떤 원인으로부터 일어난 것이라고 생각되어지는 확률. 추가된 정보로부터 사전 정보를 새롭게 수정한 확률(수정확률) → 사전 지식만을 가지고 구할 수 있다!

$p(y) = \int_{\Theta} p(y \mid \tilde{\theta})\, p(\tilde{\theta})\, d\tilde{\theta}$ : Evidence. Normalizing Constant.

$$\int p(\theta)\, d\theta \;=\; 1, \quad \int p(\theta \mid y)\, d\theta \;=\; 1, \quad \int p(y \mid \theta)\, d\theta \;\neq\; 1$$

### 기존에 알고 있던 사전 지식에 추가 정보를 얹어주는 방식으로 '판단 근거 (사후 확률)'를 찾는 것!

- 판단 근거 = 사전 지식 $\times$ 추가 정보
- 남자라고 판단할 근거 : $p$(성별 = 남자) $\times$ $p$(키 = 175 | 성별 = 남자)
- 여자라고 판단할 근거 : $p$(성별 = 여자) $\times$ $p$(키 = 175 | 성별 = 여자)

→ 사후확률은 가능도에 비례!

## 2. Example

① 사전 확률이 베타 분포(Beta Distribution)를 따르고, 가능도가 이항 분포를 따른다고 가정!

$$\theta \;\sim\; Beta(\alpha, \beta) \;\longrightarrow\; p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

### Prior

$$\theta \;\sim\; Beta(2, 20) \;\to\; p(\theta) = \frac{\Gamma(22)}{\Gamma(2)\Gamma(20)}\theta\,(1-\theta)^{19}$$

### Likelihood

$$Y|\theta \sim B(20,\theta) \rightarrow p(y\,|\,\theta) = \binom{20}{y}\theta^{y}\,(1-\theta)^{20-y}$$

이항 분포에서 데이터가 관측되는 확률은 (0, 1) 사이의 값을 가지고, 베타 분포가 (0, 1) 사이의 값을 support로 갖기 때문에, 사전 확률은 reasonable하다고 할 수 있다!

$$p(y\,|\,\theta)\,p(\theta) = \binom{20}{y}\theta^{y}\,(1-\theta)^{20-y}\,\frac{\Gamma(22)}{\Gamma(2)\Gamma(20)}\theta\,(1-\theta)^{19}$$

$$= \frac{\Gamma(20+1)}{\Gamma(y+1)\,\Gamma(20-y+1)}\,\frac{\Gamma(22)}{\Gamma(2)\Gamma(20)}\theta^{2+y-1}\,(1-\theta)^{20+20-y-1}$$

$$\left(Note: \binom{n}{y} = \frac{n!}{y!\,(n-y)!} = \frac{\Gamma(n+1)}{\Gamma(y+1)\,\Gamma(n-y+1)}\right)$$

$$p(y) = \int_0^1 p(y\,|\,\theta)\,p(\theta)\,d\theta$$

$$= \int_0^1 \binom{20}{y}\theta^{y}\,(1-\theta)^{20-y}\,\frac{\Gamma(22)}{\Gamma(2)\Gamma(20)}\theta\,(1-\theta)^{19}\,d\theta$$

$$= \frac{\Gamma(20+1)}{\Gamma(y+1)\,\Gamma(20-y+1)}\,\frac{\Gamma(22)}{\Gamma(2)\Gamma(20)}\int_0^1 \theta^{2+y-1}\,(1-\theta)^{20+20-y-1}\,d\theta$$

$$= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}$$

$$= \frac{\Gamma(20+1)\Gamma(2+20)\Gamma(y+2)\Gamma(20-y+20)}{\Gamma(y+1)\Gamma(20-y+1)\Gamma(2)\Gamma(20)\Gamma(2+20+20)}$$

$$p(\theta\,|\,y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\theta^{y+\alpha-1}\,(1-\theta)^{\beta+n-y-1}$$

$$p(\theta\,|\,y) = \frac{\Gamma(20+2+20)}{\Gamma(y+2)\Gamma(20-y+20)}\theta^{y+2-1}\,(1-\theta)^{20+20-y-1}$$

$$= \frac{\Gamma(2+40)}{\Gamma(y+2)\Gamma(40-y)}\theta^{y+2-1}\,(1-\theta)^{40-y-1}$$

$$\theta\,|\,Y=0\;\sim\;Beta(2+y,\,40-y)$$

$$p(\theta)=Beta(\theta\,|\,\alpha,\beta)\;\longrightarrow\;p(\theta\,|\,y)=Beta(\theta\,|\,\alpha+y,\,\beta+n-y)$$

prior과 posterior가 같은 분포 형태를 가질 때, “prior is conjugate to the likelihood”라고 한다!

$$\mathbb{E}(\theta)\;=\;\frac{\alpha}{\alpha+\beta},\quad \mathbb{E}(\theta\,|\,y)\;=\;\frac{\alpha+y}{\alpha+\beta+n}$$

$$\mathbb{E}(\theta\,|\,y)=\frac{\alpha+y}{\alpha+\beta+n}\;=\;\frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{(\alpha+\beta+n)}+\frac{y}{n}\frac{n}{(\alpha+\beta+n)}$$

$$(w=\alpha+\beta)$$

$$=\frac{n}{w+n}\times\,\hat{\theta}_{ML}\;(sample\ mean)\;+\;\frac{w}{w+n}\times\,\mathbb{E}(\theta)\;(prior\ mean)$$

how posterior information is affected by prior mean and strength of prior belief

사후확률의 평균은 사전확률의 평균과 $MLE$ 값의 Weighted sum으로 표현될 수 있다!

$n$이 커짐에 따라 prior mean은 posterior mean 에 영향을 끼치지 않게 된다! 즉, 우리가 사전에 설정한 prior knowledge가 옳지 않았다 하더라도 그 영향이 줄어들게 된다는 것으로서, 좋은 성질이라 할 수 있다!

- Wald Test

$$\hat{\theta}\,\pm\,1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\quad \hat{\theta}=\frac{y}{n}$$

Wald interval is asymptotically correct when $n$ is large

- Determination of prior

Hard to precisely, mathematically represent our belief, and actually it's wrong

$$all\ models\ are\ wrong,\ but\ some\ are\ useful.$$

## 3. Belief

$Be(\,\cdot\,) \equiv$ 우리가 명제를 믿는 정도를 숫자로 나타내 주는 함수

$F = \{\, a\ person\ votes\ for\ a\ left-of-center\ candidate\}$
$G = \{\, a\ person's\ income\ is\ in\ the\ lowest\ 10\%\ of\ the\ population\}$
$H = \{\, a\ person\ lives\ in\ a\ large\ city\}$

믿음의 공리

① $Be(\sim H\,|\,H)\ \le\ Be(F\,|\,H)\ \le\ Be(H\,|\,H)$

$$0\ =\ P(\sim H\,|\,H)\ \le\ P(F\,|\,H)\ \le\ P(H\,|\,H)\ =\ 1$$

② $Be(\,F\ or\ G\,|\,H\,)\ \ge\ \max\{\,Be(F\,|\,H),\,Be(G\,|\,H)\,\}$

$$P(F\cup G\,|\,H)\ =\ P(F\,|\,H)\ +\ P(G\,|\,H)\qquad if\ \ F\cap G\ =\ \emptyset$$

③
$Be(\,F\ and\ G\,|\,H\,)\ can\ be\ derived\ from\ Be(G\,|\,H)\ and\ Be(F\,|\,G\ and\ H)$

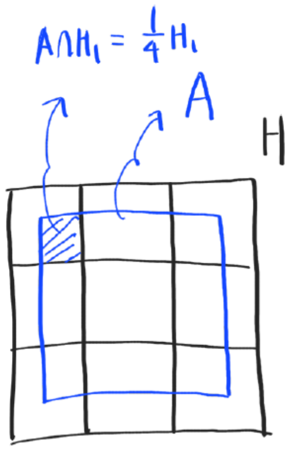
$$P(F\cap G\,|\,H)\ =\ P(G\,|\,H)\,\cdot\,P(F\,|\,G\cap H)$$

## 4. Events, Partitions and Bayes' rule

A collection of sets $\{H_1, \dots, H_k\}$ is a partition of another set $H$ if

1. the events are disjoint, which we write as $H_i \cap H_j = \emptyset$ for $i \neq j$
2. the union of the sets is $H$, which we write as $\cup_k H_k = H$.

From above, we get three rules from axioms of probability.




Rule of total probability: $\sum_k P(H_k) = 1$

Rule of marginal probability: $P(A) = \sum_k P(A \cap H_k) = \sum_k P(A|H_k)P(H_k)$

Bayes' rule: $P(H_j|\mathrm{A}) = \frac{P(A|H_j)P(H_j)}{\sum_k (A|H_k)} = \frac{P(A \cap H_j)}{P(A)}$

Conditional independence : parameter $\theta$에 대한 조건부 환경에서도 일반적인 독립의 정의가 성립함.

$$\mathrm{P}(Y_1 \in A_1, \dots, Y_n \in A_n|\theta) = \mathrm{P}(Y_1 \in A_1|\theta) \times \cdots \times \mathrm{P}(Y_n \in A_n|\theta)$$

$$\rightarrow \mathrm{P}(Y_i \in A_i|\theta, Y_j \in A_j) = \mathrm{P}(Y_i \in A_i|\theta)$$

$$\rightarrow p(y_1, \dots, y_n|\theta) = \prod_i p(y_i|\theta)$$

# 5. Exchangeability

$Let\ p(y_1, \dots, y_n)\ be\ the\ joint\ density\ of\ Y_1, \dots, Y_n.$
$If\ p(y_1, \dots, y_n) = p(y_{\pi_1}, \dots, y_{\pi_n})$
$for\ all\ permutations\ \pi\ of\ \{1, \dots, n\}, then\ Y_1, \dots, Y_n\ are\ exchangeable.$

Example 2. 세 가지 정규분포

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) \Big/ \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}\right) \Big/ \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 5 \\ 2 & 1 & 4 \\ 5 & 4 & 1 \end{bmatrix}\right)$$

**독립의 약 조건!**

# 6. de Finetti's theorem

베이즈 추론의 정당성을 부여하는 정리!

1) Conditional IID $\Rightarrow$ exchangeability (곱셈)
2) Conditional IID $\Leftarrow$ infinite exchangeability : de finetti's therem

$$p(y_1, \dots, y_n) = \int \left\{ \prod_i p(y_i|\theta) \right\} p(\theta)\, d\theta$$

$\Rightarrow$ decomposition of the model

교환 가능 확률 변수족 - 위키백과, 우리 모두의 백과사전

확률론과 통계학에서, 교환 가능 확률 변수족(交換可能確率變數族, exchangeable family of random variables)은 유한 개를 재배열하여도 결합 확률 분포가 변하지 않는 확률 변수 집합이다. 교환 가능 시그마 대수(交換可能σ代數, exchangeable sigma-algebra)는 유한 개의 확률 변수를 재배열하여도 발생 여부가 바뀌지 않는 사건들로 구성된 시

W https://ko.wikipedia.org/wiki/%EA%B5%90%ED%99%98_%EA%B0%80%EB%8A%A5_%ED%99%95%EB%A5%A0_%EB%B3%80%EC%88%98%EC%A1%B1

Prior) 반반

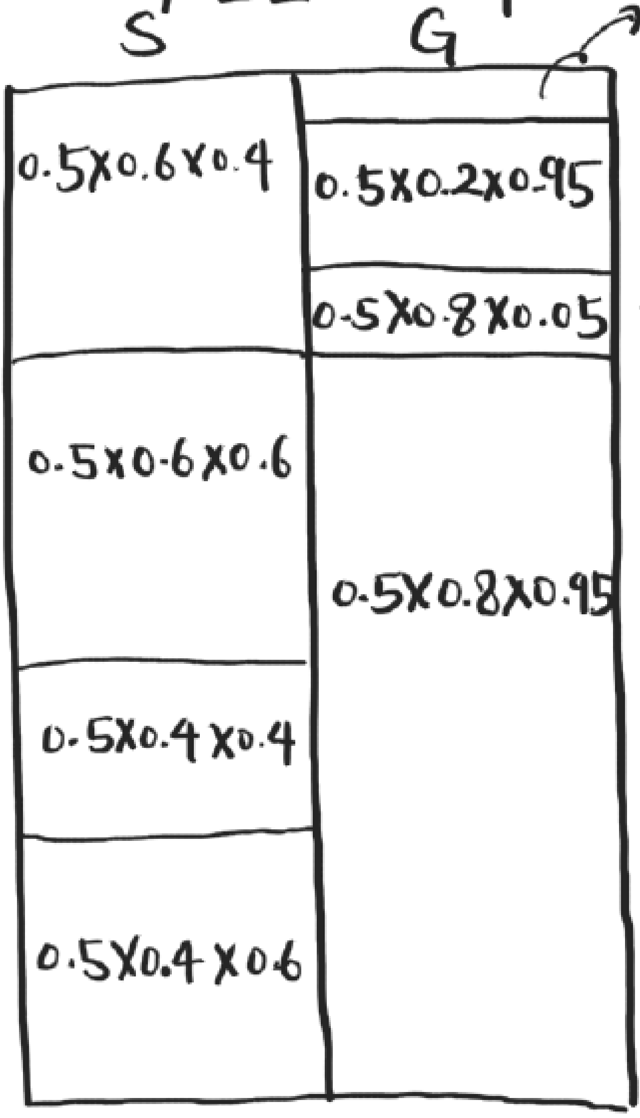
정보1)

| | 배 | 배 X |
|---|---|---|
| S | 0.6 | 0.4 |
| G | 0.2 | 0.8 |

정보2)

| | '만남' | '만남' X |
|---|---|---|
| S | 0.4 | 0.6 |
| G | 0.05 | 0.95 |

Method 1) 한번에 처리 .




0.5×0.2×0.05

비1, 맞았 0일때

사후확률의 비

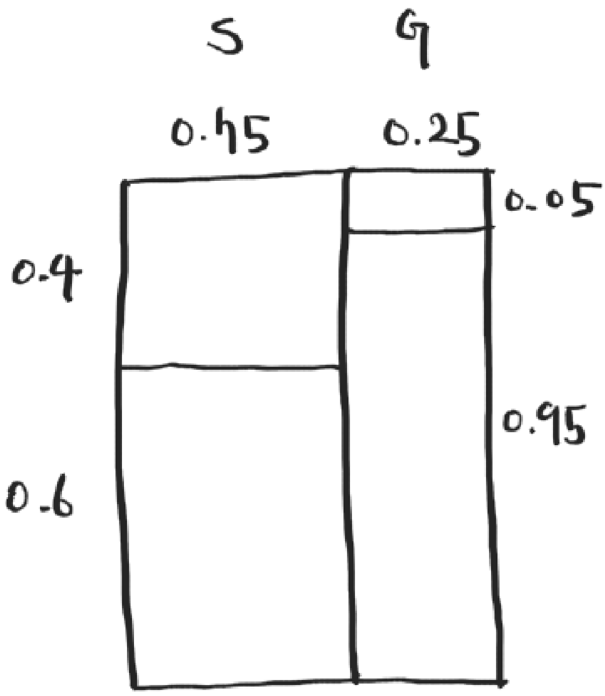
= 0.5×0.6×0.4 :

0.5×0.2×0.05

= 24 : 1

= 96% : 4%

Method 2) 순차처리. 사건이 있을 때 사후확률을 전제로,




'촨난' 이었을때

사후확률의 비

$= 0.75 \times 0.4 : 0.25 \times 0.25$

$= 24 : 1$

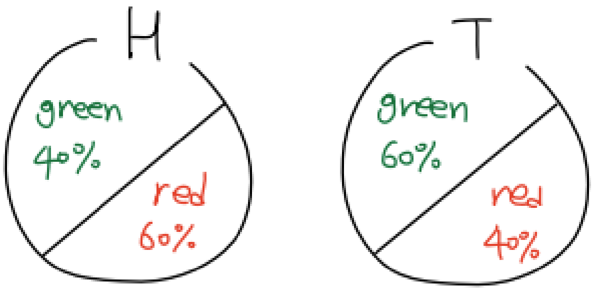
$= 96\% : 4\%$

# HW

## 2.5

Urns: Suppose urn $H$ is filled with 40% green balls and 60% red balls, and urn $T$ is filled with 60% green balls and 40% red balls. Someone will flip a coin and then select a ball from urn $H$ or urn $T$ depending on whether the coin lands heads or tails, respectively. Let $X$ be 1 or 0 if the coin lands heads or tails, and let $Y$ be 1 or 0 if the ball is green or red.

a) Write out the joint distribution of $X$ and $Y$ in a table.




동전 앞면 → urn H의 공 선택

동전 뒷면 → urn T의 공 선택

$$X = \begin{cases} 1, & \text{앞면 (head)} \\ 0, & \text{뒷면 (tail)} \end{cases} \qquad Y = \begin{cases} 1, & \text{녹색 (green)} \\ 0, & \text{빨간색 (red)} \end{cases}$$

| Y \ X | 1 | 0 |
|---|---|---|
| 1 | $\frac{1}{2} \times \frac{40}{100} = 0.2$ | $\frac{1}{2} \times \frac{60}{100} = 0.3$ |
| 0 | $\frac{1}{2} \times \frac{60}{100} = 0.3$ | $\frac{1}{2} \times \frac{40}{100} = 0.2$ |

b) Find $E[Y]$. What is the probability that the ball is green?

$$E[Y] = E[E[Y|X]]$$

$$= E[Y|X=1] \cdot P(X=1) + E[Y|X=0] \cdot P(X=0)$$

$$= \frac{40}{100} \times \frac{1}{2} + \frac{60}{100} \times \frac{1}{2} = 0.5$$

$$P(Y=1) = P(Y=1, X=1) + P(Y=1, X=0)$$

$$= 0.2 + 0.3 = 0.5$$

c) FInd $Var[Y \mid X = 0]$, $Var[Y \mid X = 1]$ and $Var[Y]$. Thinking of variance as measuring uncertainty, explain intuitively why one of these variances is larger than the others.

$$Var[Y|X=0] = E[Y^2|X=0] - \{E[Y|X=0]\}^2$$

$$= \frac{60}{100} - \left(\frac{60}{100}\right)^2 = 0.6 - 0.36 = 0.24$$

$$Var[Y|X=1] = E[Y^2|X=1] - \{E[Y|X=1]\}^2$$

$$= \frac{40}{100} - \left(\frac{40}{100}\right)^2 = 0.4 - 0.16 = 0.24$$

$$Var[Y] = E[Var[Y|X]] + Var[E[Y|X]]$$

$$= \left(\frac{1}{2} \times 0.24 + \frac{1}{2} \times 0.24\right) + \frac{1}{2}\{E[Y|X=1] - E[Y]\}^2$$

$$+ \frac{1}{2}\{E[Y|X=0] - E[Y]\}^2$$

$$= 0.24 + \left(\frac{10}{100}\right)^2 = 0.25$$

d) Suppose you see that the ball is green. What is the probability that the coin turned up tails?

posterior

Likelihood

prior

$$P(T|Y=1) = \frac{P(T \cap Y=1)}{P(Y=1)} = \frac{P(Y=1|T) \cdot P(T)}{0.5}$$

Evidence

$$= \frac{\frac{3}{5} \times \frac{1}{2}}{\frac{1}{2}} = \frac{3}{5}$$

## 2.6

Conditional independence: Suppose events $A$ and $B$ are conditionally independent given $C$, which is written $A \perp B \,|\, C$. Show that this implies that $A^c \perp B \,|\, C$, $A \perp B^c \,|\, C$, and $A^c \perp B^c \,|\, C$, where $A^c$ means "not $A$." Find an example where $A \perp B \,|\, C$ holds but $A \perp B \,|\, C^c$ does not hold.

de Finetti's theorem 에 따라

$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$ 이므로,

①

$$P(A^c \cap B \mid C) = P(B \mid C) - P(A \cap B \mid C)$$

$$= P(B \mid C) - P(A \mid C) \cdot P(B \mid C)$$

$$= P(B \mid C)\{1 - P(A \mid C)\} = P(A^c \mid C) \cdot P(B \mid C)$$

②

$$P(A \cap B^c \mid C) = P(A \mid C) - P(A \cap B \mid C)$$

$$= P(A \mid C) - P(A \mid C) \cdot P(B \mid C)$$

$$= P(A \mid C)\{1 - P(B \mid C)\} = P(A \mid C) \cdot P(B^c \mid C)$$

③

$$P(A^c \cap B^c \mid C) = P((A \cup B)^c \mid C) = 1 - P(A \cup B \mid C)$$

$$= 1 - P(A \mid C) - P(B \mid C) + P(A \cap B \mid C)$$

$$= 1 - P(A \mid C) - P(B \mid C)\{1 - P(A \mid C)\}$$

$$= \{1 - P(A \mid C)\}\{1 - P(B \mid C)\} = P(A^c \mid C) \cdot P(B^c \mid C)$$

Pólya urn model - Wikipedia

In statistics, a Pólya urn model (also known as a Pólya urn scheme or simply as Pólya's urn), named after George Pólya, is a type of statistical model used as an idealized mental exercise framework, unifying many treatments. In an urn model, objects of real interest (such as atoms, people, cars, etc.)

W https://en.wikipedia.org/wiki/P%C3%B3lya_urn_model