Introduction to Bayesian Inference : Bayes' rule, Examples, and Difficulties

Lee, Dohyoung

Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}}$$

θ: parameter which describe population characteristics
Θ: parameter space
y: sampled data
p(θ): "Our" prior distribution. "Our" belief of θ representing the true population characteristics.
p($y|\theta$): "Our" sampling model. "Our" belief that y would be the outcome if we knew θ to be true.
p($\theta|y$): "Our" posterior distribution. "Our" belief that θ is the true value, having observed data y

Bayes' rule does not tell us what our beliefs should be,
It tells us how they should change after seeing new information

Example 1: Estimating the probability of rare event

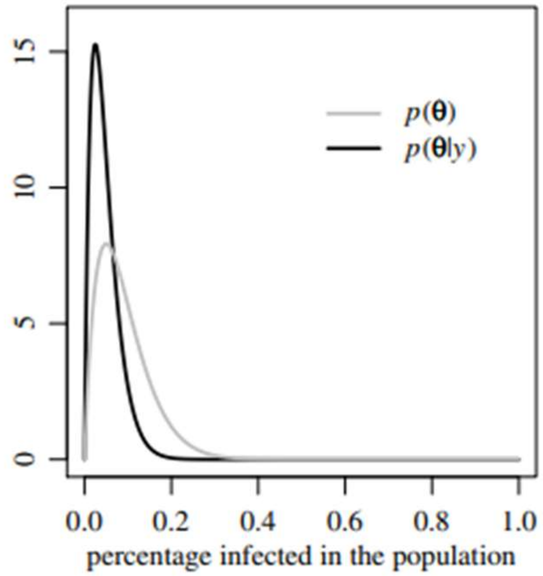
θ: The fraction of infected individuals in the city
y: The number of infected from the sample of 20 individuals

Let Y | θ ~ binomial(20, θ)

Let prior θ ~ beta(2,20) mean: 0.09 mode: 0.05

Observed Y=0

Posterior θ | {Y=0} ~ beta(2,40) (Why? Next week!) mean: 0.048 mode: 0.025



Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}}$$

Prior Information about a parameter	+	Information from data	$\Rightarrow$	Posterior Information about a parameter
Prior Distribution $\mathrm{p}(\theta)$	+	Likelihood function $\mathrm{p}(y|\theta)$	$\Rightarrow$	Posterior distribution $\mathrm{p}(\theta|y)$

Note 1. If y is conditionally iid given θ, Likelihood $\mathrm{L}(\theta|y)$=$\mathrm{p}(y|\theta)$= $\mathrm{p}(y_1|\theta)\cdot\cdots\cdot\mathrm{p}(y_n|\theta)$=pdf$(y_1|\theta)\cdot\cdots\cdot$pdf$(y_n|\theta)$

Note 2. $\int \mathrm{p}(\theta)\,d\,\theta = 1$, $\int \mathrm{p}(\theta|y)d\,\theta = 1$ (distribution of θ) but $\int \mathrm{p}(y|\theta)d\,\theta \neq 1$ (function of θ)

Note 3. $\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}$: normalizing constant to make $p(\theta|y)$ a distribution

Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p\left(y\middle|\tilde{\theta}\right)p(\tilde{\theta})\ d\tilde{\theta}}$$

The process of inductive learning via Bayes' rule is referred to as *Bayesian inference.*

Bayesian methods are data analysis tools that are derived from the principles of Bayesian inference.

Bayesian methods provide
-formal interpretation as a means of induction
-parameter estimates with good statistical properties
-parsimonious descriptions of observed data
-predictions for missing data and forecasts of future data
-a computational framework for model estimation, selection and validation

Example 1: Estimating the probability of rare event

Prior θ ~ beta(a,b)

Observed Y=y

Posterior θ | {Y=y} ~ beta(a+y, b+20-y) (Why? Next week!)

Posterior mean E[θ | Y=y] = $\frac{a+y}{a+b+20}$

$$= \frac{20}{a+b+20}\frac{y}{20} + \frac{a+b}{a+b+2}\frac{a}{a+b}$$

$$= \frac{n}{w+n}(sample\ mean) + \frac{w}{w+n}(prior\ mean)$$

n: data #
w= a+b (strength of prior belief)

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

Example 1: Estimating the probability of rare event

Comparison to non-Bayesian methods

Wald interval $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$

$\hat{\theta} = \frac{y}{n}$

If y=0? $\hat{\theta}$=0, Wald interval : 0

If $\hat{\theta}$ small, n small, then $\hat{\theta} - 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n} < 0$

Wald interval is asymptotically correct when n is large

cf. Bayesian estimator $\hat{\theta} = \frac{n}{w+n}(sample\ mean) + \frac{w}{w+n}(prior\ mean)$

Example 2: Building a predictive linear model

$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{64} x_{i,64} + \sigma\epsilon_i$

Y_i: diabetes progression of subject i
x_i: 64 explanatory variables

Train data: 342
Test data: 100

Our belief: most of the 64 explanatory variable have little to no effect on diabetes progression

Prior: p($\beta_i = 0) = 0.5$
Posterior: given train data,

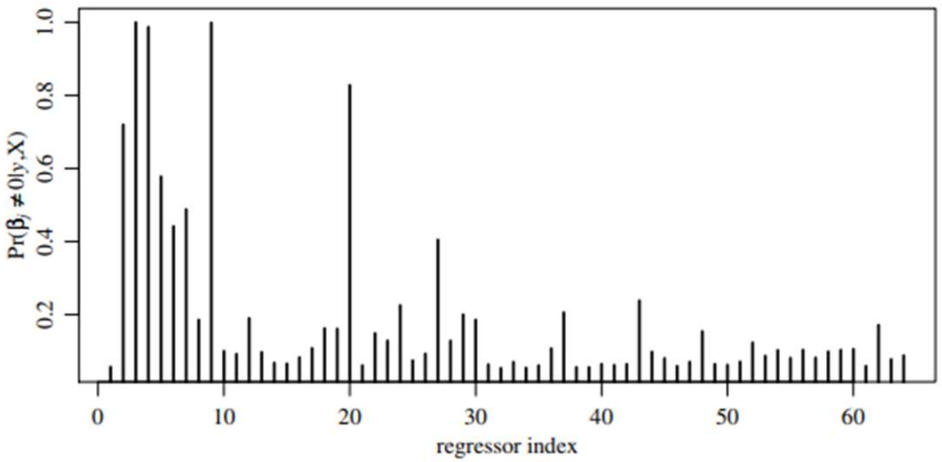


Fig. 1.3. Posterior probabilities that each coefficient is non-zero.

Example 2: Building a predictive linear model

Comparison to non-Bayesian methods

Let $\widehat{\boldsymbol{\beta}}_{Bayes} = E[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}]$: posterior mean given train data
$\widehat{\boldsymbol{y}}_{test} = \boldsymbol{X}_{test}\ \widehat{\boldsymbol{\beta}}_{Bayes}$
MSE : 0.45

$\widehat{\boldsymbol{y}}_{test} = \boldsymbol{X}_{test}\ \widehat{\boldsymbol{\beta}}_{OLS}$
MSE : 0.67

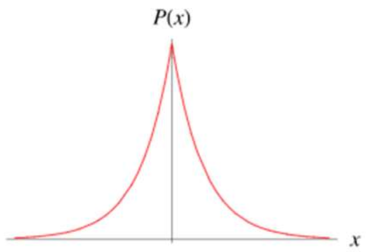
In this case, Bayesian prediction is better!

Example 2: Building a predictive linear model

OLS estimate poor when small sample size
->Standard remedy: fit a sparse regression model(Set some or many β_i to 0).

1. Bayesian approach
2. Lasso regression. Minimize $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$

In fact, the lasso estimate is equal to the posterior mode of $\boldsymbol{\beta}$ in which the prior distribution of each β_j is a double exponential distribution, whose peak is at 0.



double exponential distribution, a.k.a. Laplace distribution

Difficulties in Bayesian Inference

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}}$$

- Computation of normalizing constant

-conjugate prior : analytically integrable
-numerical integration: numerical approximation
-MCMC : avoiding integration

- Determination of prior

-Hard to precisely, mathematically represent our belief, and actually it's wrong
"all models are wrong, but some are useful"(Box and Draper, 1987, pg. 424)
-Which prior is useful?: sensitivity check, mixed prior, noninformative prior, reference prior,…
-Hierarchical model, empirical bayes,…

Reference

- Hoff, P. D. (2009). *First Course in Bayesian Statistical Methods*. Springer. : Chapter 1
- Kruschke, J. K. (2014). *Doing Bayesian Data Analysis*. Academic Press. : Chapter 5

CH2.

김두은

2-1. Belief

Example

Let F, G, H be three possibly overlapping statements about the world.

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 }

Let Be() be a belief function. ≡ 우리가 명제를 믿는 정도를 숫자로 나타내 주는 함수.

$Axioms\ of\ beliefs$

B1. $Be(\sim H|H) \leq Be(F|H) \leq Be(H|H)$ $0 = P(\sim H|H) \leq P(F|H) \leq P(H|H) = 1$

B2. $Be(F\ or\ G|H) \geq \max\{ Be(F|H), Be(G|H) \}$ $P(F \cup G|H) = P(F|H) + P(G|H)$ if $F \cap G = \emptyset$

B3. $Be(F\ and\ G|H)$ can be derived from $Be(G|H)$ and $Be(F|G\ and\ H)$

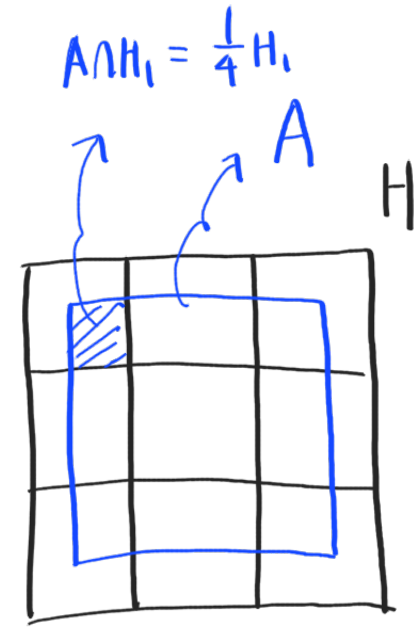
모두 확률의 공리와 일치. $P(F \cap G|H) = P(G|H) \cdot P(F|G \cap H)$

2-2. 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: $P(H_j|\mathrm{A}) = \frac{P(A|H_j)P(H_j)}{\sum_k (A|H_k)} = \frac{P(A \cap H_j)}{P(A)}$

2-3 ~ 2-6 / 짧은 요약

Conditional independence : parameter θ에 대한 조건부 환경에서도 일반적인 독립의 정의가 성립함.

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

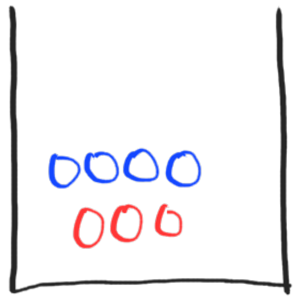
$$\rightarrow \mathrm{P}(Y_i \in A_i|\theta, Y_j \in A_j) = \mathrm{P}(Y_i \in A_i|\theta)$$

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

2-7. Exchangeability

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

Example 1. 항아리 문제



$P(\text{파}, \text{빨}) = \frac{4}{7} \times \frac{3}{6}$

$P(\text{빨}, \text{파}) = \frac{3}{7} \times \frac{4}{6}$

cf) $P(\text{빨}) \times P(\text{파}) = \frac{3}{7} \times \frac{4}{7}$

Example 2. 세 가지 정규분포

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

2-8. de Finetti's theorem

1) Conditional IID ⟹ exchangeability (곱셈)
2) Conditional IID ⟸ infinite exchangeability : de finetti's therem

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

⟹ decomposition of the model

Example. 스팸 메일 필터

prior) 반반

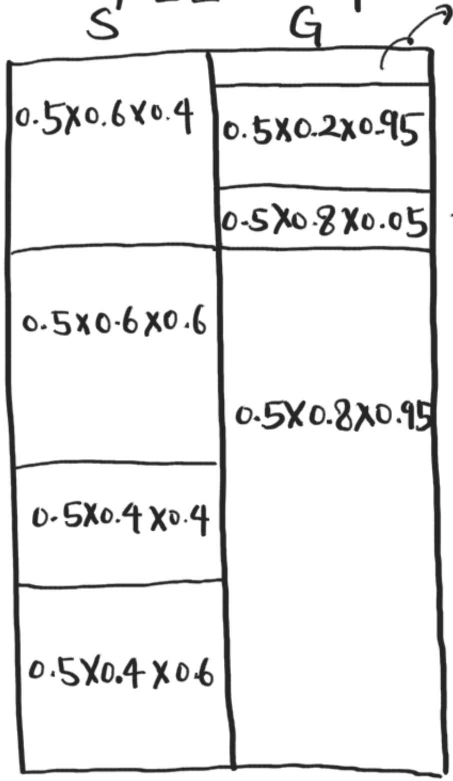
정보1)

	쌔	쌔 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

바삭, '만남' 0일때

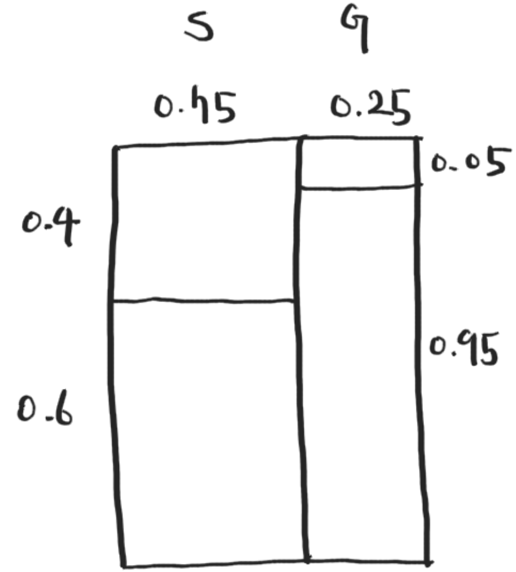
사후확률의 비

$= 0.5\times0.6\times0.4 : 0.5\times0.2\times0.05$

$= 24 : 1$

$= 96\% : 4\%$

Method 2) 순차처리. 바삭이 있을 때 사후확률을 전제로,



'만남' 0일때

사후확률의 비

$= 0.75\times0.4 : 0.25\times0.25$

$= 24 : 1$

$= 96\% : 4\%$

HW

1. 주 교재 226p / 연습문제 2.5, 2.6번

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.

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

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

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

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.

2. Suppose you have an urn containing R_0 red balls and W_0 white balls. Let $c \geq 0$ be a fixed integer. Draw a ball, note the color, replace the ball and put an additional c balls of that color in the urn as well. Rinse and repeat.

Define $X_i = \begin{cases} 1 & \text{if the ith ball is red} \\ 0 & \text{otherwise} \end{cases}$

Show that the random variables in the infinite sequence $X_1, X_2, \ldots$ are exchangeable.

Pólya's urn