Bayesian Statistics

Chapter 1. Introduction

Hojin Yang

Department of Statistics
Pusan National University

1.1. Introduction

- We often use probabilities informally to express our information and beliefs about unknown quantities
- In a mathematical sense, probabilities can numerically represent a set of rational beliefs
- There is a relationship between probability and information
- Bayes' rule provides a rational method for updating beliefs in light of new information
- The process of inductive learning via Bayes' rule is referred to as Bayesian inference

- More generally, Bayesian methods are data analysis tools that are derived from the principles of Bayesian inference
- In addition to their formal interpretation as a means of induction, Bayesian methods provide
  - 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
- Throughout this course, we will explore the broad uses of Bayesian methods for a variety of inferential and statistical tasks

Bayesian Learning

- Statistical induction is the process of learning about the general characteristics of a population from a subset of that population
- Numerical values of population characteristics are typically expressed in terms of a parameter θ and numerical descriptions of the dataset y
- The numerical values of both the population characteristics and the dataset are uncertain
- After a dataset y is obtained, the information can be used to decrease our uncertainty about the population characteristics
- Quantifying this change in uncertainty is the purpose of Bayesian inference

- The sample space $\mathcal{Y}$ is the set of all possible datasets, where a single dataset y denotes result
- The parameter space Θ is the set of possible parameter values
- Bayesian learning begins with a numerical formulation of joint beliefs about y and Θ, expressed in terms of probability distributions over $\mathcal{Y}$ and Θ

  1. For any $\theta \in \Theta$, our prior distribution $p(\theta)$ describes our belief that θ represents the true population characteristics
  2. For any $\theta \in \Theta$ and $y \in \mathcal{Y}$, our sampling model $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true

     Once we obtain the data y, the last step is to update our beliefs about θ
  3. For any $\theta \in \Theta$, our posterior distribution $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y

1.2. Examples

- Bayes theorem is often used in diagnostic tests for cancer
- A young person was diagnosed as having a type of cancer that occurs extremely rarely in young people
- Naturally, he was very upset
- A friend told him that it was probably a mistake
- His friend reasoned as follows
- No medical test is perfect
- There are always incidences of false positives and false negatives

- Let C stand for the event that he has cancer.
- Let $+$ stand for the event that an individual responds positively to the test
- Assume $P(C) = 1/1,000,000 = 10^{-6}$, $P(+|C) = .99$, and $P(+|C^c) = .01$
- So only one per million people in his age have the disease
- The test is extremely good relative to most medical tests, giving only 1% false positives and 1% false negatives
- Find the probability that he has cancer given that he has a positive response

- After you make this calculation, we will not be surprised to learn that he did not have cancer
- By Bayes' rule

$$
\begin{aligned}
P(C|+) &= \frac{P(+|C)P(C)}{P(+|C)P(C) + P(+|C^c)P(C^c)} \\
&= \frac{(.99)(10^{-6})}{(.99)(10^{-6}) + (.01)(.999999)} \\
&= \frac{.00000099}{.01000098} = .00009899
\end{aligned}
$$

- Deciding paternity
- Legal cases of disputed paternity in many countries are resolved using blood tests
- Laboratories make genetic determinations concerning the mother, child, and alleged father
- Some cases involve different types of evidence (for instance, the mother or the alleged father may not be available, but his brother is available, and so on)
- Most labs apply Bayes rule in communicating the testing results
- They calculate the probability that the alleged father is in fact the child's father given the genetic evidence

- For the sake of brevity, we will pare down the genetic evidence usually introduced and deal only with ABO blood type
- All the probabilities you need will be given
- Suppose you are on a jury considering a paternity suit brought by Suzy Smith's mother against Al Edged
- The following is part of the background information: Suzy's mother has blood type O and Al Edged is type AB
- All your probabilities are calculated conditional on this information

- You put all testimony concerning that Al Edged is real father in assessing $P(F)$, which is probability that Al is Suzy's father
- The evidence of interest is Suzy's blood type
- If it is O, then Al Edged is excluded from paternity-he is not her father, unless there has been a gene mutation or a laboratory error
- Suzy's blood type turns out to the B; call this event B

- According to Bayes rule

$$P(F|B) = \frac{P(B|F)P(F)}{P(B|F)P(F) + P(B|F^c)P(F^c)}$$

- According to Mendelian genetics $P(B|F) = \frac{1}{2}$
- You need to compute $P(B|F^c)$
- They calculate this as the proportion of B genes to the total number of ABO genes in blood bank
- It is known that a typical value among Caucasians is 9%

- Hence,

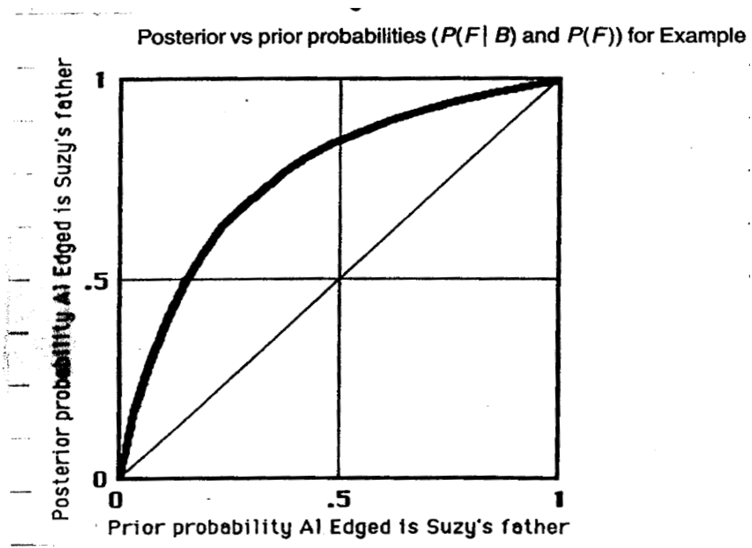
$$P(F|B) = \frac{(1/2)P(F)}{(1/2)P(F) + (0.09)P(F^c)}$$
$$= \frac{50 \cdot P(F)}{41 \cdot P(F) + 9}$$

- This is a substantial increase over $P(F)$
- For example, it is about 85% when $P(F) = \frac{1}{2}$
- The reason such a large increase is possible is that Suzy's paternal gene (B) is relatively rare
- The probability of paternity would increase for any male who has a B gene

- The relationship between our unconditional probability, $P(F)$, and our conditional probability, $P(F|B)$, can be shown using the following table

$P(F)$	0	.100	.250	.500	.750	.900	1
$P(F \mid B)$	0	.382	.649	.847	.943	.980	1

- Another way to show the same thing is to use a graph, such as the one in Figure



- The diagonal on this graph corresponds to evidence which contains no information about F
- Comparing this diagonal with the actual curve shows how much the evidence changes one's prior probability of paternity
- Tables and graphs are effective ways for juries and others to understand the strength of the evidence
- Blood banks and other laboratories that analyze genetic factors in paternity cases have a name for the Bayes factor in favor of F

$$\text{Paternity index} = PI = \frac{P(B|F)}{P(B|F^c)} = \frac{1/2}{.09} = 5.56$$

- The evidence (child has type B blood) is 5.56 times as likely if Al Edged is the father than if he is not

- The posterior probability of paternity (based on the equivalent version of Bayes' rule) is

$$\begin{aligned} P(F|B) &= \frac{1}{1 + \frac{P(B|F^c)P(F^c)}{P(B|F)P(F)}} \\ &= \frac{1}{1 + \frac{1 \cdot P(F^c)}{PI \cdot P(F)}} \\ &= \frac{PI}{PI + \frac{P(F^c)}{P(F)}} \end{aligned}$$

- Laboratories choose $P(F) = 1/2$ and report a probability (or likelihood) of paternity as though there is no prior probability involved

1.3. Where We Are Going

- The uses of Bayesian methods are quite broad
- We have seen how the Bayesian approach provides
  - models for rational, quantitative learning
  - estimators that work for small and large sample sizes
  - methods for generating statistical procedures in complicated problems
- We will review of probability in Chapter 2
- Learn the basics of Bayesian data analysis for one-parameter statistical models in Chapters 3 and 4
- Chapters 5, 6 and 7 discuss Bayesian inference with the normal and multivariate normal models
- Advanced topics are covered in Chapters 8, 9, and 10