Bayesian Statistics

Chapter 2. Beliefs and Probabilities

Hojin Yang

Department of Statistics
Pusan National University

Introduction

- We first discuss what properties a reasonable belief function should have, and show that probabilities have these properties
- We review the basic properties of discrete and continuous random variables and probability distributions
- Finally, we explore the link between independence and exchangeability

2.1. Belief Functions and Probabilities

- Let F, G, and H be three possibly overlapping statements about the world
- For example:

$$F = \{ \text{ a person graduates from college} \}$$

$$G = \{ \text{ a person's income is in the highest 10\%} \}$$

$$H = \{ \text{ a person lives in a large city} \}$$

- Let $Be(\cdot)$ be a belief function: assigns numbers to statements such that the larger the number, the higher the degree of belief

- Some philosophers have tried to relate beliefs to preferences over bets
- $Be(F) > Be(G)$: prefers to bet "F is true" than "G is true"
- We also want $Be(\cdot)$ to describe our beliefs under certain conditions
- $Be(F|H) > Be(G|H)$: prefers to bet that "F is also true" than bet "G is also true" if we knew that "H were true"
- $Be(F|G) > Be(F|H)$: if we were forced to bet on F, we would prefer to do it under the condition that "G is true" rather than "H is true"

Axioms of Beliefs

- Any function that is to numerically represent our beliefs should have the following properties:

  B1. $Be(\text{not } H|H) \leq Be(F|H) \leq Be(H|H)$

  B2. $Be(F \text{ or } G|H) \geq \max\{Be(F|H), Be(G|H)\}$

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

- How should we interpret these properties? Are they reasonable?

- B1 says that the number we assign to $Be(F|H)$, our conditional belief in F given H, is bounded below and above by the numbers we assign to complete disbelief ($Be(\text{not } H|H)$) and complete belief ($Be(H|H)$)
- B2 says that our belief that the truth lies in a given set of possibilities should not be smaller than any separate possibilities
- B3 says that if we have to decide whether or not F and G are true, knowing that H is true, we could do this by first deciding whether or not G is true given H, and if so, then deciding whether or not F is true given G and H

Axioms of Probability

- Now let's compare B1, B2 and B3 to the standard axioms of probability
- Suppose $F \cup G$ means F or G, $F \cap G$ means F and G an $\emptyset$ is the empty set
- A function, $P(\cdot)$ satisfying P1, P2 and P3, also satisfies B1, B2 and B3

  P1. $0 = P(\text{not } H|H) \leq P(F|H) \leq P(H|H) = 1$

  P2. $P(F \cup G|H) = P(F|H) + P(G|H)$ if $F \cap G = \emptyset$

  P3. $P(F \cap G|H) = P(G|H)P(F|G \cap H)$

- Therefore, if we use a probability function to describe our beliefs, we have satisfied the axioms of belief

2.2. Events, Partitions and Bayes' Rule

Definition: Partition

A collection of sets $\{H_1, \ldots, H_K\}$ is a partition of the set $\mathcal{H}$ if

1. the events are disjoint, which we write as $H_i \cap H_j = \emptyset \;\; \forall i \neq j$
2. the union of the sets is $\mathcal{H}, i.e., \cup_{j=1}^{K} H_j = \mathcal{H}$

- Examples
  - Let $\mathcal{H}$ be someone's religious orientation. Partitions include
    - {Protestant, Catholic, Jewish, other, none }
    - {Christian, non-Christian }
  - Let $\mathcal{H}$ be someone's number of children. Partitions include
    - { 0, 1, 2, 3 or more }
    - { 0, 1, 2, 3, 4, 5, 6, ... }

- Suppose $\{H_1, \ldots, H_K\}$ is a partition of $\mathcal{H}$, $P(\mathcal{H}) = 1$, and E is some specific event
- The axioms of probability imply the following:
- Rule of total probability

$$\sum_{k=1}^{K} P(H_k) = 1$$

- Rule of marginal probability

$$P(E) = \sum_{k=1}^{K} P(E \cap H_k) = \sum_{k=1}^{K} P(E|H_k)P(H_k)$$

- Bayes' rule

$$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{P(E)}$$
$$= \frac{P(E|H_j)P(H_j)}{\sum_{j=1}^{K} P(E|H_j)P(H_j)}$$

- We consider data on the education level and income for a sample of males over 30 years of age
    - Let $\{H_1, H_2, H_3, H_4\}$ be the lower 25th percentile, the second 25th percentile, the third 25th percentile and the upper 25th percentile in terms of income
    - So, $\{P(H_1), P(H_2), P(H_3), P(H_4)\} = \{0.25, 0.25, 0.25, 0.25\}$
    - $\{H_1, H_2, H_3, H_4\}$ is a partition and so these probabilities sum to 1

- Let E be the event that a randomly sampled person from the survey has a college education
- From the survey data

  $$\{P(E|H_1), P(E|H_2), P(E|H_3), P(E|H_4)\} = \{.11, .19, .31, .53\}$$

- These probabilities do not sum to 1, because they represent the proportions of people with college degrees in the four different income subpopulations H_1, H_2, H_3 and H_4
- Income distribution of the college-educated population:

  $$\{P(H_1|E), P(H_2|E), P(H_3|E), P(H_4|E)\} = \{.09, .17, .27, .47\}$$

- This distribution differs from $P(H_j) = 0.25$ and these probabilities do sum to 1

- In Bayesian inference $\{H_1, H_2, H_3, H_4\}$ often refer to disjoint hypotheses or states of nature and E refers to the outcome of a study

- To compare hypotheses post-experimentally, we often calculate the following ratio

$$
\begin{aligned}
\frac{P(H_i|E)}{P(H_j|E)} &= \frac{P(E|H_i)P(H_i)/P(E)}{P(E|H_j)P(H_j)/P(E)} \\
&= \frac{P(E|H_i)P(H_i)}{P(E|H_j)P(H_j)} \\
&= \frac{P(E|H_i)}{P(E|H_j)} \times \frac{P(H_i)}{P(H_j)} \\
&= \text{“Bayes factor”} \times \text{“prior beliefs”}
\end{aligned}
$$

- Bayes' rule tells us how our beliefs should change after seeing the data

Independence

Definition: Independence

Two events F and G are conditionally independent given H if

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

- How do we interpret conditional independence?
- By Axiom P3, $P(F \cap G|H) = P(G|H)P(F|H \cap G)$
- If F and G are conditionally independent given H, then

$$P(G|H)P(F|H \cap G) \stackrel{\text{always}}{=} P(F \cap G|H) \stackrel{\text{indep}}{=} P(F|H)P(G|H)$$
$$P(G|H)P(F|H \cap G) = P(F|H)P(G|H)$$
$$P(F|H \cap G) = P(F|H)$$

- Conditional independence therefore implies that $P(F|H \cap G) = P(F|H)$
- If we know H is true and F and G are conditionally independent given H, then knowing G does not change our belief about F

Random Variables

- Let Y be a random variable
- Let $\mathcal{Y}$ be the set of all possible values of Y
- Y is discrete if the set of possible outcomes is countable, meaning that Y can be expressed as $\mathcal{Y} = \{y_1, y_2, \ldots\}$
- The event that the outcome Y of our survey has the value y is expressed as $\{Y = y\}$
- For each $y \in \mathcal{Y}$, $P(Y = y)$ will be $p(y)$ and this function of y is called the probability density function of Y
    1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$
    2. $\sum_{y \in \mathcal{Y}} p(y) = 1$
- In general, $P(Y \in A) = \sum_{y \in A} p(y)$

- Let $\mathcal{Y}$ be R the set of all real numbers
- Probability distributions for Y define a cumulative distribution
$$F(y) = P(Y \leq y)$$
- Note that $F(\infty) = 1$, $F(-\infty) = 0$, and $F(b) \leq F(a)$ if $b < a$
    1. $P(Y > a) = 1 - F(a)$
    2. $P(a < Y \leq b) = F(b) - F(a)$
- If F is continuous, we say that Y is a continuous random variable

- For every continuous cdf F, there exists a positive function $p(y)$ such that

$$F(a) = \int_{-\infty}^{a} p(y)dy$$

- $p(y)$ is called the probability density function of Y

  1. $0 \leq p(y)$ for all $y \in \mathcal{Y}$
  2. $\int_{y \in R} p(y)dy = 1$

- In general, $P(Y \in A) = \int_{y \in A} p(y)dy$
- Unlike the discrete case, $p(y)$ is not the probability $Y = y$
- However, if $p(y_1) > p(y_2)$, we will informally say that y_1 has a higher probability than y_2

Descriptions of Distributions

- The mean or expectation of an unknown quantity Y

  $E[Y] = \sum_{y \in \mathcal{Y}} yp(y)$ if Y is discrete

  $E[Y] = \int_{y \in \mathcal{Y}} yp(y)dy$ if Y is continuous

- This is the center of mass of the distribution but it is not in general equal to either of

  mode: the most probable value of Y

  median: the value of Y in the middle of the distribution

- Measure of spread is the variance of a distribution

$$
\begin{aligned}
Var[Y] &= E[(Y - E(Y))^2] \\
&= E[Y^2] - E[Y]^2
\end{aligned}
$$

- Standard deviation is the square root of $Var[Y]$
- Alternative measures of spread are based on quantiles
- The α-quantile is the value y_α such that

$$F(y_\alpha) = P(Y \leq y_\alpha) = \alpha$$

- The interquartile range is the interval $(y_{0.25}, y_{0.75})$
- This range contains 50% of the mass of the distribution
- Similarly, the interval $(y_{0.025}, y_{0.975})$ contains 95% of the mass of the distribution

Joint Distributions

- Let

  $\mathcal{Y}_1$, $\mathcal{Y}_2$ be two countable sample spaces

  Y_1, Y_2 be two random variables, taking values in $\mathcal{Y}_1$, $\mathcal{Y}_2$ respectively.

- The joint pdf or joint density of Y_1 and Y_2 is defined as

$$p_{Y_1,Y_2}(y_1, y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \ \forall y_1 \in \mathcal{Y}_1, y_2 \in \mathcal{Y}_2$$

- Marginal density of Y_1 can be from the joint density

$$p_{Y_1}(y_1) = P(Y_1 = y_1) = \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1,Y_2}(y_1, y_2)$$

- Conditional density of Y_2 given $\{Y_1 = y_1\}$ can be as

$$p_{Y_2|Y_1}(y_2) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}$$

- We should convince that
  - $\{p_{Y_1}, p_{Y_2|Y_1}\}$ can be derived from p_{Y_1,Y_2}
  - $\{p_{Y_2}, p_{Y_1|Y_2}\}$ can be derived from p_{Y_1,Y_2}
  - p_{Y_1,Y_2} can be derived from $\{p_{Y_1}, p_{Y_2|Y_1}\}$
  - p_{Y_1,Y_2} can be derived from $\{p_{Y_2}, p_{Y_1|Y_2}\}$
  - but p_{Y_1,Y_2} cannot be derived from $\{p_{Y_1}, p_{Y_2}\}$
- The subscripts of density functions are often dropped, in which $p(y_1)$ refers to p_{Y_1}, $p(y_1, y_2)$ refers to $p_{Y_1,Y_2}(y_1, y_2)$, $p(y_1|y_2)$ refers to $p_{Y_1|Y_2}(y_1|y_2)$, etc

- If Y_1 and Y_2 are continuous, a cdf is given by

$$F_{Y_1,Y_2}(a,b) = P(\{Y_1 \leq a\} \cap \{Y_2 \leq b\})$$

- There is a function p_{Y_1,Y_2} such that

$$F_{Y_1,Y_2}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2$$

- The function p_{Y_1,Y_2} is the joint density of Y_1 and Y_2
    - $p_{Y_1}(y_1) = \int_{\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2) dy_2$
    - $p_{Y_2|Y_1}(y_2) = p_{Y_1,Y_2}(y_1, y_2)/p_{Y_1}(y_1)$
- Mixed continuous and discrete variables are also possible

Bayes' Rule and Parameter Estimation

- Let

  θ: parameter or a certain characteristic of the population

  Y: data from population who has the characteristic
- We might treat θ as continuous and Y as discrete
- Estimation of θ derives from the calculation of $p(\theta|y)$
- y is the observed value of Y
- This calculation first requires that we have a joint density $p(\theta, y)$ representing our beliefs about θ and the survey outcome Y

- It is natural to construct this joint density from
  - $p(\theta)$ beliefs about θ
  - $p(y|\theta)$ beliefs about Y for each value of θ
- Having observed $\{Y = y\}$, we need to compute our updated beliefs about θ

$$p(\theta|y) = p(\theta, y)/p(y) = p(\theta)p(y|\theta)/p(y)$$

- Posterior density of θ_a relative to θ_b , conditional on $Y = y$

$$\begin{aligned}\frac{p(\theta_a|y)}{p(\theta_b|y)} &= \frac{p(\theta_a)p(y|\theta_a)/p(y)}{p(\theta_b)p(y|\theta_b)/p(y)} \\ &= \frac{P(\theta_a)p(y|\theta_a)}{p(\theta_b)p(y|\theta_b)}\end{aligned}$$

- This means that we do not need to compute $p(y)$ in the relative posterior probabilities

- Another way to think about it is that, as a function of θ

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

- The constant of proportionality is $1/p(y)$, which could be computed from

$$p(y) = \int_\Theta p(y,\theta)d\theta = \int_\Theta p(y|\theta)p(\theta)d\theta$$

- Hence

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

- The numerator is the critical part

2.6. Independent Random Variables

- $Y_1, \ldots, Y_n$: r.v.s and θ: a parameter describing the population

- We say that $Y_1, \ldots, Y_n$ are conditionally independent given θ if for every collection of n set $\{A_1, \ldots, A_n\}$

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n|\theta) = P(Y_1 \in A_1|\theta) \times \cdots \times P(Y_n \in A_n|\theta)$$

- From our previous calculations, if independence holds,

$$P(Y_i \in A_i|\theta, Y_j \in A_j) = P(Y_i \in A_i|\theta)$$

- Conditional independence can be interpreted as meaning that Y_j gives no additional information about Y_i beyond that in knowing θ

- Under independence, the joint density is given by

$$P(y_1, \ldots, y_n|\theta) = P(y_1|\theta) \times \cdots \times P(y_n|\theta) = \prod_{i=1}^{n} P(y_i|\theta)$$

- For such a case, we say that $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed (i.i.d.) denoted by

$$Y_1, \ldots, Y_n|\theta \sim p(y|\theta)$$

2.7. Exchangeability

Definition: Exchangeability

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

- Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels convey no information about the outcomes.

- Independence versus dependence
  - $P(Y_{10} = 1) = a$
  - $P(Y_{10} = 1 | Y_1 = Y_2 = \cdots = Y_9) = b$
  - Should we have $a < b$, $a = b$, or $a > b$?
  - If $a \neq b$ then Y_{10} is NOT independent of $Y_1, \ldots, Y_9$

Claim

If $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.

Proof

If $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ. Then for any permutation π of $\{1, \ldots, n\}$ and any set of values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$

$$
\begin{aligned}
p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n|\theta)p(\theta)d\theta \ \text{ marginal probability} \\
&= \int \left\{ \prod_{i=1}^{n} P(y_i|\theta) \right\} p(\theta)d\theta \ \text{ conditionally i.i.d} \\
&= \int \left\{ \prod_{i=1}^{n} P(y_{\pi_i}|\theta) \right\} p(\theta)d\theta \ \text{ product not depend on order} \\
&= p(y_{\pi_1}, \ldots, y_{\pi_n}) \ \text{ marginal probability}
\end{aligned}
$$

2.8. de Finetti's Theorem

- We have seen that

$$\left.\begin{array}{l} Y_1, \ldots, Y_n|\theta \ \text{ i.i.d.} \\ \theta \sim p(\theta) \end{array}\right\} \Rightarrow Y_1, \ldots, Y_n \ \text{ are exchangeable}$$

- What about an arrow in the other direction?

Theorem: (de Finetti)

Let $y_i \in \mathcal{Y}$ for all $i \in \{1, 2, \ldots\}$. Suppose that, for any n, our belief model for $Y_1, \ldots, Y_n$ is exchangeable:

$$p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$$

for all permutations π. Then our model can be written as

$$p(y_1, \ldots, y_n) = \int \left\{ \prod_{i=1}^{n} P(y_i|\theta) \right\} p(\theta) d\theta$$

for some parameter θ, $p(y|\theta)$, $p(\theta)$

- The main ideas of this and the previous section can be summarized as follows

$$\left.\begin{array}{l} Y_1, \ldots, Y_n | \theta \ \text{ i.i.d.} \\ \theta \sim p(\theta) \end{array}\right\} \Leftrightarrow Y_1, \ldots, Y_n \text{ are exchangeable for all } n$$

- For this condition to hold, we must have exchangeability and repeatability

- Exchangeability will hold if the labels convey no info

- Repeatability is reasonable, including the following

    - $Y_1, \ldots, Y_n$ are outcomes of a repeatable experiment

    - $Y_1, \ldots, Y_n$ are sampled from a finite population with replacement

    - $Y_1, \ldots, Y_n$ are sampled from an infinite population without replacement