Chapter 2: Belief, probability, and exchangeability

Jesse Mu

September 8, 2016

Contents

This section will be brief, as it is just a review of probability. However, as exchangeability and de Finetti's theorem are important especially in Bayesian statistics, I'll go into more depth there.

Belief functions and properties

Let F, G, and H be events. A "belief" function $\text{Be}(\cdot)$ should correspond to certain intuitions about our beliefs about the likelihood of events. Probabilitiy functions have axioms that satisfy our notions of belief:

1. Contradictions and tautologies: $0 = P(\text{not } H \mid H) \leq P(F \mid H) \leq P(H \mid H) = 1$

2. Addition rule: $P(F \cup G \mid H) = P(F \mid H) + P(G \mid H)$ if $F \cap G = \emptyset$
3. Multiplication rule: $P(F \cap G \mid H) = P(G \mid H)P(F \mid G \cap H)$

Note that the axioms of probability and theorems discussed in this section are the same whether you subscribe to a Bayesian or frequentist interpretation of probability.

Events, partitions, and Bayes' rule

Consider a set $\mathcal{H}$, which is the "set of all possible truths." We can partition $\mathcal{H}$ into discrete subsets $\{H_1, \ldots, H_k\}$, where only one subset consists of the truth.

We can assign probabilities whether each of these sets contains the truth. First, some event in $\mathcal{H}$ is true, so $P(\mathcal{H}) = 1$. Let E be some observation (in this case related to the truth of one H_i). Then,

- Rule of total probability: $\sum_i P(H_i) = 1$
- Marginal probability: $P(E) = \sum_i P(E \cap H_i) = \sum_i P(E \mid H_i)P(H_i)$
  - The total probability of an event occurring is the sum of all of its probabilities under the possible partitions of truths
- Bayes' rule:

$$P(H_i \mid E) = \frac{\overbrace{P(E \mid H_i)}^{\text{likelihood}}\overbrace{P(H_i)}^{\text{prior}}}{P(E)} \tag{1}$$

Independence

Two events F and G are *independent* if $P(F \cap G) = P(F)P(G)$.

Two events F and G are *conditionally independent* given H if

$$P(F \cap G \mid H) = P(F \mid H)P(G \mid H) \tag{2}$$

This implies $P(F \mid H \cap G) = P(F \mid H)$. That is, if we know about H, and F and G are conditionally independent given H, then knowing G does not change our belief about H. This is a key property leveraged in Bayesian networks.

Random variables and their distributions

A random variable Y is *discrete* if the set of all its possible values $\mathcal{Y}$ is countable, i.e. they can be enumerated $\mathcal{Y} = \{y_1, y_2, \ldots\}$ Examples include the binomial and poisson distributions. Discrete random variables have a *probability mass function* (PMF) $f(y) = P(Y = y)$ which assigns a certain probability to every discrete point in its sample space. From this probability mass function, a *continuous distribution function* (CDF) is also defined:

$$F(y) = P(Y \leq y) = \sum_{y_i \leq y} f(y_i)$$

Y is *continuous* if $\mathcal{Y}$ can take *any* value in an interval. Therefore, the probability of Y taking a single value in the sample space is 0. So instead we describe such distributions with *probability density functions* (PDFs) $f(y)$, which must be integrated over an interval to obtain a probability: $P(a \leq y \leq b) = \int_a^b f(y)\, dy$. These variables also have CDFs:

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(x)\, dx \tag{3}$$

Examples include the normal and beta distributions.

Descriptions of distributions

In the same way that we use the mean, mode, and median to describe samples, we can use them to describe distributions. Notice that for many distributions, these quantities are *not* the same.

We also use variance and quantiles to measure the *spread* of distributions.

Joint distributions

Discrete

Let Y_1 and Y_2 be random variables with sample spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$. Then the joint pdf/density of Y_1 and Y_2 is defined as:

$$p(y_1, y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})$$

and the *marginal density* of Y_1 is obtained by summing over all possible values of Y_2:

$$p(y_1) = \sum_{y_2 \in \mathcal{Y}_2} p(y_1, y_2) \tag{4}$$

$$= \sum_{y_2 \in \mathcal{Y}_2} p(y_1 \mid y_2) p(y_2). \tag{5}$$

The *conditional density* is

$$p(y_2 \mid y_1) = \frac{p(y_1, y_2)}{p(y_1)} \tag{6}$$

Notice that given the joint density $p(y_1, y_2)$, we can calculate marginal and conditional densities $\{p(y_1), p(y_2), p(y_1 \mid y_2), p(y_2 \mid y_1)\}$ by simply summing up the relevant variables. Additionally, given $p(y_1)$ and $p(y_2 \mid y_1)$, (or the reverse), we can reconstruct the joint distribution. However, given only marginal densities $p(y_1)$ and $p(y_2)$, we can't reconstruct the joint distribution, since we don't know whether the events are independent.

Continuous

In the continuous case, the probability density function is a function of y_1 and y_2 such that the CDF is

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} p(y_1, y_2)\, dy_2\, dy_1.$$

Obtaining the marginal densities can be done by integrating out the irrelevant variable:

$$p(y_1) = \int_{-\infty}^{\infty} p(y_1, y_2)\, dy_2 \tag{7}$$

$$p(y_2) = \int_{-\infty}^{\infty} p(y_1, y_2)\, dy_1 \tag{8}$$

$$\tag{9}$$

With the marginal densities, you can compute the conditional densities $p(y_2 \mid y_1) = p(y_1, y_2)/p(y_1)$, etc.

Independent random variables

Let $Y_1, \ldots, Y_n$ be random variables dependent on a common parameter θ. Then $Y_1 \ldots, Y_n$ are conditionally independent given θ if

$$p(y_1, \ldots, y_n \mid \theta) = p(y_1 \mid \theta) \times \cdots \times p(y_n \mid \theta). \tag{10}$$

Note this extends naturally from the definition of independent of two random variables, $P(A \cap B) = P(A)P(B)$. Thus, knowing about any Y_i does not give any information about the other Y_j. Lastly, the joint density of these variables can be defined as

$$p(y_1, \ldots, y_n \mid \theta) = \prod_i p(y_i \mid \theta).$$

We say that $Y_1, \ldots, Y_n$ are conditionallly independent and identically distributed (i.i.d.):

$$Y_1, \ldots, Y_n \mid \theta \sim \text{i.i.d. } p(y \mid \theta).$$

Exchangeability

In many situations with several random variables, we would intuit that the specific order of observation of these random variables aren't important. For example, consider a random sample of 3 participants from an infinite population which may or may not have a property (1 or 0) It makes sense that

$$p(0, 0, 1) = p(1, 0, 0) = p(0, 1, 0).$$

since the likelihood of a person having the property or not is θ, regardless of the sample. This property is called exchangeability.

Exchangability. Let $Y_1, \ldots, Y_n$ be random variables. 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 exchangable.

If $Y_1, \ldots Y_n$ are i.i.d., then they are exchangeable:

$$
\begin{aligned}
p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n \mid \theta)\, d\theta && (11)\\
&= \int \left(\prod_i p(y_i \mid \theta) \right) p(\theta)\, d\theta \qquad \text{(i.i.d.)} && (12)\\
&= \int \left(\prod_i p(y_{\pi_i} \mid \theta) \right) p(\theta)\, d\theta \qquad \text{(order of product doesn't matter)} && (13)\\
&= p(y_{\pi_1}, \ldots, y_{\pi_n}). && (14)
\end{aligned}
$$

Classical assumption of Bernoulli variables $X_1, X_2, \ldots, X_n$ as outcomes of the same experiment (e.g. a coin flip): *independence*. But continuing to observe X_js should result in a change of opinion about the distribution of coin flip outcomes (e.g. gradually learning coin bias). So Bayesian statisticians should assume *exchangeability*, a weaker condition than *independence*.

de Finetti's theorem

de Finetti's theorem. Let $Y_1, \ldots, Y_n$ be a finite subset of an infinitely exchangeable but **not necessarily i.i.d.** random variables:

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

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

$$p(y_1, \ldots, y_n) = \int \left(\prod_1^n p(y_i \mid \theta) \right) p(\theta)\, d\theta.$$

So, in general,

$$Y_1, \ldots, Y_n \mid \theta \text{ are i.i.d.} \Leftrightarrow Y_1, \ldots, Y_n \text{ are exchangable for all } n$$

Importantly, if we sample from a sufficiently large population, then we can model the sample variables as being approximately conditionally i.i.d.

Proof for {0, 1} random variables

From Heath & Sudderth (1976). De Finetti's Theorem on Exchangeable Variables. Also see this neat StackExchange answer.

Alternative statement of theorem

For every infinite sequence of exchangeable random variables (X_n) having values in $\{0, 1\}$ and a finite subsequence $X_1, \ldots, X_n$, there is a probability distribution F such that

$$P(X_1 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0) = \tag{15}$$

$$P\left(\underbrace{1, \ldots, 1}_{k \text{ times}}, \underbrace{0, \ldots, 0}_{n-k \text{ times}}\right) = \int_{0,1} \theta^k (1-\theta)^{n-k} F(d\theta) \tag{16}$$

Where F is the prior over Θ, i.e. the values that θ can take.

Proof

Lemma. Let $p_{k,n} = P(X_1 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0)$, the quantity on the left side of the above equation, and suppose $X_1, \ldots, X_m$ are exchangable Boolean random variables as before.

Let $q_r = P\left(\sum_{j=1}^m X_j = r\right)$, that is, q_i is the probability that exactly i of the m variables are 1.

Also denote $(x)_k = \prod_{j=0}^{k-1}(x-j) = \frac{x!}{(x-k)!}$.

Then,

$$p_{k,n} = \sum_{r=0}^{m} \frac{(r)_k (m-r)_{n-k}}{(m)_n} q_r \text{ for } 0 \leq k \leq n \leq m \tag{17}$$

where $\frac{(r)_k (m-r)_{n-k}}{(m)_n}$ is some complicated combinatorics term for the number of ways to that r ones and $m-r$ zeros can be placed in a sequence of length m.

With the sustitution $\frac{r}{m} = \theta$, we can re-package the summation of discrete r as an integral over a stepwise function F_m with domain $[0, 1]$ that jumps q_r at every interval $0 \leq r/m \leq 1$:

$$p_{k,n} = \int_0^1 \frac{(\theta m)_k ((1-\theta)m)_{n-k}}{(m)_n} F_m(d\theta).$$

As m goes to infinity, the steps of the stepwise function grow infinitely small, such that the function converges to the continuous probability distribution F. By Helly's theorem and some hand-waving in this explanation, the components of the integrand converge to the exponential terms in the theorem: $\theta^k (1-\theta)^{n-k}$.

Exercises

2.1

a

The marginal probability distribution of a father's occupation is found by summing over the rows of the son's occupation. Formally, let F be the father's occupation and S be the son's. Then $P(F = f) = \sum P(F = f, S = s)$.

- Farm: 0.11
- Operatives: 0.279
- etc. . .

b

This time sum over columns.

c

In the row of the table where the father's occupation is "farm", normalize the values of the son's occupation such that they sum to one. This can be done by dividing each value by the sum of the row. Formally, this is calculating

$$P(S = s \mid F = \text{farm}) = \frac{P(S = s, F = \text{farm})}{P(F = \text{farm})} \tag{18}$$

$$= \frac{P(S = s, F = \text{farm})}{\sum_{s' \in S} P(S = s', F = \text{farm})}. \tag{19}$$

d

Normalize the column where the son's occupation is "farm" like above.

2.2

a

Since expectation is linear,

$$\mathbb{E}(a_1 Y_1 + a_2 Y_2) = a_1 \mathbb{E}(Y_1) + a_2 \mathbb{E}(Y_2) \tag{20}$$

$$= a_1 \mu_1 + a_2 \mu_2 \tag{21}$$

Since Y_1 and Y_2 are independent, $\text{Cov}(Y_1, Y_2) = 0$. When adding variances, it's necessary to combine terms like below:

$$\text{Var}(a_1 Y_1 + a_2 Y_2) = a^2 \text{Var}(Y_1) + b^2 \text{Var}(Y_2) + 2ab\text{Cov}(Y_1, Y_2) \tag{22}$$

$$= a^2 \sigma_1^2 + b^2 \sigma_2^2 \tag{23}$$

b

$$\mathbb{E}(a_1 Y_1 - a_2 Y_2) = a_1 \mathbb{E}(Y_1) - a_2 \mathbb{E}(Y_2) \tag{24}$$

$$= a_1 \mu_1 - a_2 \mu_2 \tag{25}$$

$$\text{Var}(a_1 Y_1 - a_2 Y_2) = a^2 \text{Var}(Y_1) + b^2 \text{Var}(Y_2) - 2ab\text{Cov}(Y_1, Y_2) \tag{26}$$

$$= a^2 \sigma_1^2 + b^2 \sigma_2^2 \tag{27}$$

$$= \text{Var}(a_1 Y_1 + a_2 Y_2) \tag{28}$$

2.3

Let X, Y, Z be random variables with joint density $p(x, y, z) \propto f(x, z) g(y, z) h(z)$.

a

$$p(x \mid y, z) = \frac{p(x, y, z)}{p(y, z)} \tag{29}$$
$$= \frac{p(x, y, z)}{\int p(x, y, z)\ dx} \tag{30}$$
$$\propto \frac{f(x, z)g(y, z)h(z)}{\int f(x, z)g(y, z)h(z)\ dx} \tag{31}$$
$$\propto \frac{f(x, z)g(y, z)h(z)}{g(y, z)h(z) \int f(x, z)\ dx} \tag{32}$$
$$\propto \frac{f(x, z)}{\int f(x, z)\ dx} \tag{33}$$

b

$$p(y \mid x, z) = \frac{p(x, y, z)}{p(x, z)} \tag{34}$$
$$\propto \frac{f(x, z)g(y, z)h(z)}{\int p(x, y, z)\ dy} \tag{35}$$
$$\propto \frac{f(x, z)g(y, z)h(z)}{\int f(x, z)g(y, z)h(z)\ dy} \tag{36}$$
$$\propto \frac{f(x, z)g(y, z)h(z)}{f(x, z)h(z) \int g(y, z)\ dy} \qquad \propto \frac{g(y, z)}{\int g(y, z)\ dy} \tag{37}$$

c

It is sufficient to show that $p(x \mid y, z) = p(x \mid z)$:

From (a), $p(x \mid y, z) \propto \frac{f(x,z)}{\int f(x,z)\ dx}$. Now

$$p(x \mid z) = \frac{p(x, z)}{p(z)} \tag{38}$$
$$\propto \frac{\int p(x, y, z)\ dy}{\int \int p(x, y, z)\ dy\ dx} \tag{39}$$
$$\propto \frac{\int f(x, z)g(y, z)h(z)\ dy}{\int \int f(x, z)g(y, z)h(z)\ dy\ dx} \tag{40}$$
$$\propto \frac{f(x, z)h(z) \int g(y, z)\ dy}{h(z) \left(\int g(y, z)\ dy\right) \left(\int f(x, z)\ dx\right)} \tag{41}$$
$$\propto \frac{f(x, z)}{\int f(x, z)\ dx} \tag{42}$$
$$\propto p(x \mid y, z) \tag{43}$$

Now, since $p(x \mid z) \propto p(x \mid y, z)$ but additionally $\int p(x \mid z) = \int p(x \mid y, z) = 1$, $p(x \mid z) = p(x \mid y, z)$, so x and y are conditionally independent given z.

2.4

a

Use **P3** and condition on an always true event:

$$P(H_j \cap E \mid 1 = 1) = P(H_j \mid 1 = 1)P(E \mid H_j \cap 1 = 1) = P(H_j)P(E \mid H_j) \tag{44}$$
$$P(H_j \cap E \mid 1 = 1) = P(E \mid 1 = 1)P(H_j \mid E \cap 1 = 1) = p(E)p(H_j \mid E) \tag{45}$$
$$= P(H_j)P(E \mid H_j) \tag{46}$$

b

If we can assume that $p(\mathcal{H}) = 1$, then

$$P(E) = P(E \cap H) \tag{47}$$
$$= P(E \cap (H_1 \cup H_2 \cup \cdots \cup H_k)) \tag{48}$$
$$= P((E \cap (H_1)) \cup (E \cap (H_2 \cup \cdots \cup H_k))) \quad \text{Distributing} \tag{49}$$
$$= P(E \cap H_1) + P(E \cap (H_2 \cup \cdots \cup H_k)) \quad \text{P2; Implied condition on a true event} \tag{50}$$

c

Proceed inductively from b.

d

$$P(H_j)P(E \mid H_j) = P(E)P(H_j \mid E) \tag{51}$$
$$\implies P(H_j \mid E) = \frac{P(H_j)P(E \mid H_j)}{P(E)} \tag{52}$$
$$\implies P(H_j \mid E) = \frac{P(H_j)P(E \mid H_j)}{\sum_k P(E \cap H_k)} \quad \text{From c} \tag{53}$$
$$\implies P(H_j \mid E) = \frac{P(H_j)P(E \mid H_j)}{\sum_k P(E \mid H_k)P(H_k)} \quad \text{P3} \tag{54}$$
$$\tag{55}$$

2.5

a

```
x = rbind(
  # Y = 0 Y = 1
  c(.5 * .4, .5 * .6), # X = 0
  c(.5 * .6, .5 * .4) # X = 1
)
rownames(x) = c("X = 0", "X = 1")
kable(x, col.names = c("Y = 0", "Y = 1"))
```

	Y = 0	Y = 1
X = 0	0.2	0.3
X = 1	0.3	0.2

b

$$\mathbb{E}(Y) = 0.5$$

. Probability the ball is green is 0.5.

c

$$\begin{aligned}
\text{Var}(Y \mid X = 0) &= \mathbb{E}((Y \mid X = 0)^2) - \mathbb{E}(Y \mid X = 0)^2 && (56)\\
&= 1^2 p(Y = 1 \mid X = 0) + 0^2 p(Y = 0 \mid X = 0) - && (57)\\
&\quad (1p(Y = 1 \mid X = 0) + 0p(Y = 0 \mid X = 0))^2 && (58)\\
&= .6 - (.6)^2 && (59)\\
&= 0.24 && (60)\\
&= \text{Var}(Y \mid X = 1) \qquad \text{Not going to do this one} && (61)
\end{aligned}$$

$$\begin{aligned}
\text{Var}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 && (62)\\
&= 1^2 p(Y = 1) + 0^2 p(Y = 0) - (0.5)^2 && (63)\\
&= 0.5 - (0.5)^2 = 0.25 && (64)
\end{aligned}$$

$\text{Var}(Y)$ is slightly larger since we are more uncertain about the value of Y when we have not yet flipped the coin. Knowing which urn we will draw from clarifies our probabilites of obtaining a green or red ball.

d

$$P(X = 0 \mid Y = 1) = .6 \tag{65}$$

2.6

If $A \perp B \mid C$ then

$$\begin{aligned}
P(A, B \mid C) &= P(A \mid C)P(B \mid C) && (66)\\
&= (1 - P(A^c \mid C))P(B \mid C) && (67)\\
&= P(B \mid C) - P(B \mid C)P(A^c \mid C) && (68)
\end{aligned}$$

so

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

.

Also notice

$$P(B \mid C) = P(A, B \mid C) + P(A^c, B \mid C) \quad \text{LTP} \tag{69}$$
$$\implies P(A^c, B \mid C) = P(B \mid C) - P(A, B \mid C) \tag{70}$$

Equating the two equations above, we have

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

We do this similarly for the other events.

For a case where $A \perp B \mid C$ holds but $A \perp B \mid C^c$ does not, consider a Bayesian network where knowing C results in 100% belief in the values of A and B (e.g. $P(A, B \mid C) = 1$), but absent of C, values A and B have some probability of manifesting that affect each other.

2.7

a

No matter how likely the event, the maximum money you will obtain is \$1. Thus no rational person would give more than \$1 for the possibility of obtaining \$1, as such an action would always result in lost money.

b

Since either E or E^c will definitively happen, we would want someone who is betting on the occurrences of either E or E^c to even out.

2.8

a

i

A frequentist idea: if I were to pick at random many, many x sampled from the census roll, how many of them would be Hindu?

A subjectivist interpretation: how strongly do I believe that a random person from this census roll is Hindu?

(Should be 0.15)

ii

If I were to pick at random many, many x from the census roll, in the long run average, what proportion of them would be 6452859?

How strongly do I believe that this x I am going to sample is 6452859?

iii

An interesting one

Assume person 6452589 is a random person sampled from Sri Lanka that is either Hindu or not. In the long run, if I observed many persons 6452589, how many of them would be Hindu?

Person 6452589 is *one* person who may or may not be Hindu. How strongly do I believe that he/she is Hindu?

b

i

ii

iii

Skipping

c