STAT 535: Chapter 2: Using Bayes' Rule

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

An Illustrative Example

- Suppose we wish to categorize an online news item as "fake news" or "real news".
- When we encounter such an article, we can't observe with certainty whether it is "fake".
- But we can observe some important characteristics about the article.
- In addition, we may have some prior knowledge about how often online news items are "fake news".
- In the `fake_news` data set in the `bayesrules` package, there are 150 articles posted on Facebook, and experts have identified 60 of them as "fake news" and the rest as real.
- Assuming this is a representative sample, this could inform our prior knowledge about how likely an article is to be fake news.

Using Prior and Data Information on the Example

- In addition, we can observe that 16 of the 60 identified fake news items have exclamation points in the headline, and only 2 of 90 of the real news items have exclamation points in the headline.
- This is a piece of information we can observe when we encounter a new article, so we can consider it *data information* since it is observable.
- When we encounter a news item, we could *update* our prior information (about 2/5 of Facebook news items are "fake news") with the data information (whether or not the encountered article has an exclamation point in the headline, which is associated with being fake) to get a better estimated probability that the item is fake.
- The prior and data information combine to yield **posterior** information about the parameter of interest.

A Formal Prior Model

- Let B denote the event that a random news item is fake news.
- Based on our prior knowledge, we can set $P(B) = 0.4$, which means the probability that the item is real news is $P(B^c) = 0.6$.
- This is a coherent prior, since the events encompass all possible outcomes and the probabilities sum to 1.

Incorporating the Data into the Model

- Now consider our observable data: Let A denote the event that the news item's title has an exclamation point.
- Based on our data, we can state that $P(A|B) \approx 16/60 = 0.2667$ and $P(A|B^c) \approx 2/90 = 0.0222$.
- These are conditional probabilities; $P(A|B)$ is the probability that the item's headline has an exclamation point **given that** the item is fake news.
- If the conditional probability $P(A|B)$ equals the unconditional probability $P(A)$, then events A and B are **independent**.

The Likelihood Function

- If A is an observed event that is **known** to have occurred, then we can use our knowledge that A occurred to determine the *likelihood* that event B also occurs.
- We denote the likelihood with L, and in discrete/categorical situations, $L(B|A) = P(A|B)$, while $L(B^c|A) = P(A|B^c)$.
- Note that the likelihood function is **not** a valid probability function: In our example,
  $L(B|A) + L(B^c|A) = 0.2667 + 0.0222 = 0.2889$, not 1.
- But the likelihood allows us to answer the question: How compatible is the data that we observed with some hypothetical "state of nature"?

Joint and Marginal Probability

- Recall that the likelihood function is not a valid probability function.
- We will see that the marginal probability, $P(A)$, can be used as a **normalizing constant** that will create a valid probability distribution.
- The **joint probability** $P(A \cap B)$ is the probability of observing both events A and B.
- In the news example,
  $P(A \cap B) = P(A|B)P(B) = (0.2667)(0.4) = 0.1067$.
- In addition,
  $P(A \cap B^c) = P(A|B^c)P(B^c) = (0.0222)(0.6) = 0.0133$.

The Law of Total Probability

- The total probability that an article title has an exclamation point is the sum of:
  - the probability that an article is fake news and its title has an exclamation point, and
  - the probability that an article is real news and its title has an exclamation point.
- So $P(A) = P(A \cap B) + P(A \cap B^c) = 0.1067 + 0.0133 = 0.12$.
- This last formula is a special case of the *Law of Total Probability* (LTP).

Bayes' Rule and Posterior Probability

- What we *really* want to know is: Given that we observe an article with an exclamation point in its title, what is the probability that it is "fake news"?
- This is $P(B|A)$.
- Recalling that $L(B|A) = P(A|B)$ and $L(B^c|A) = P(A|B^c)$, Bayes' Rule for events states:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)L(B|A)}{P(B)L(B|A) + P(B^c)L(B^c|A)}$$

- In words, this states that the posterior = (prior)$\times$(likelihood)/(normalizing constant).

Bayes' Rule and Posterior Probability with the News Example

- News example:

$$P(B|A) = \frac{(0.4)(0.2667)}{0.12} = 0.889.$$

- Given that we observe an article with an exclamation point in its title, the **posterior probability** that it is "fake news" is 0.889.
- Recall that before we observed the data (the exclamation point), the **prior probability** that an article is "fake news" was 0.4.
- Observing the data has **updated** our probability estimate.

Another Bayes' Rule Example

- **Example**: (1975 British national referendum on whether the UK should remain part of the European Economic Community)
- Suppose 52% of voters supported the Labour Party and 48% the Conservative Party. Suppose 55% of Labour voters wanted the UK to remain part of the EEC and 85% of Conservative voters wanted this.
- What is the probability that a person voting "Yes" to remaining in EEC is a Labour voter?

$$P(L|Y) = \frac{P(Y|L)P(L)}{P(Y)}$$

Another Bayes' Rule Example

Note
$P(Y) = P(Y \cap L) + P(Y \cap L^c) = P(Y|L)P(L) + P(Y|L^c)P(L^c)$.
So

$$P(L|Y) = \frac{P(Y|L)P(L)}{P(Y|L)P(L) + P(Y|L^c)P(L^c)}$$
$$= \frac{(.55)(.52)}{(.55)(.52) + (.85)(.48)} = 0.41.$$

Bayes' Law with Multiple Events

Let **D** represent some observed data and let A, B, and C be mutually exclusive (and exhaustive) events conditional on **D**. Note that

$$\begin{aligned} P(\mathbf{D}) &= P(A \cap \mathbf{D}) + P(B \cap \mathbf{D}) + P(C \cap \mathbf{D}) \\ &= P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C). \end{aligned}$$

By Bayes' Rule,

$$P(A|\mathbf{D}) = \frac{P(\mathbf{D}|A)P(A)}{P(\mathbf{D})}$$

$$\Rightarrow P(A|\mathbf{D}) = \frac{P(\mathbf{D}|A)P(A)}{P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C)}.$$

Bayes' Rule with Multiple Events

- Denoting A, B, C by θ_1, θ_2, θ_3, we can write this more generally as

$$P(\theta_i|\mathbf{D}) = \frac{P(\theta_i)P(\mathbf{D}|\theta_i)}{\sum_{j=1}^{3} P(\theta_j)P(\mathbf{D}|\theta_j)}.$$

- If there are k distinct discrete outcomes $\theta_1, \ldots, \theta_k$, we have, for any $i \in \{1, \ldots, k\}$:

$$P(\theta_i|\mathbf{D}) = \frac{P(\theta_i)P(\mathbf{D}|\theta_i)}{\sum_{j=1}^{k} P(\theta_j)P(\mathbf{D}|\theta_j)},$$

- The denominator equals $P(\mathbf{D})$, the **marginal** distribution of the data.
- Note if the values of θ are portions of the continuous real line, the sum may be replaced by an integral.

Bayes' Rule Example (4 Classes)

Example: In the 1996 General Social Survey, for males (age 30+):

- 11% of those in the lowest income quartile were college graduates.
- 19% of those in the second-lowest income quartile were college graduates.
- 31% of those in the third-lowest income quartile were college graduates.
- 53% of those in the highest income quartile were college graduates.

What is the probability that a college graduate falls in the lowest income quartile?

Bayes' Rule Example (4 Classes)

$$P(Q_1|G) = \frac{P(G|Q_1)P(Q_1)}{\sum_{j=1}^{4} P(G|Q_j)P(Q_j)}$$

$$= \frac{(.11)(.25)}{(.11)(.25) + (.19)(.25) + (.31)(.25) + (.53)(.25)} = 0.09.$$

Exercise: Find $P(Q_2|G)$, $P(Q_3|G)$, $P(Q_4|G)$ also. How does this **conditional** distribution differ from the **unconditional** distribution $\{P(Q_1), P(Q_2), P(Q_3), P(Q_4)\}$?

Statistics Using Bayes' Rule

- We now consider inference about parameters, based on data.
- Generically denote an unobserved parameter of interest as θ.
- Generically denote our data as $\mathbf{D}$.
- Our probability model for the data, given a value of θ, is denoted $p(\mathbf{D}|\theta)$.
- Our model for our prior knowledge about θ is denoted $p(\theta)$.
- This could be highly specific or quite vague, depending how uncertain we are about θ.

Statistics Using Bayes' Rule

- We seek to make probability statements about θ, **given** some observed data: $p(\theta|\mathbf{D})$.
- By Bayes' Rule,

$$p(\theta|\mathbf{D}) = \frac{p(\theta)p(\mathbf{D}|\theta)}{p(\mathbf{D})}.$$

- Note $p(\mathbf{D})$ **does not** depend on θ and thus carries no information about θ.
- It is simply a **normalizing constant** which makes $p(\theta|\mathbf{D})$ sum (or integrate) to 1.

Statistics Using Bayes' Rule

- For inference about θ, it is just as good to write

$$p(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta)$$

- The LHS is called the **posterior distribution** of θ and represents a compromise between the **prior** information about θ in $p(\theta)$ and the information from the sample about θ in $p(\mathbf{D}|\theta)$.
- Some useful **summaries** of the posterior are the **posterior mean**

$$E[\theta|\mathbf{D}] = \int \theta p(\theta|\mathbf{D})\, d\theta$$

and the **posterior variance**

$$
\begin{aligned}
var[\theta|\mathbf{D}] &= E\left\{(\theta - E[\theta|\mathbf{D}])^2|\mathbf{D}\right\} \\
&= \int (\theta - E[\theta|\mathbf{D}])^2 p(\theta|\mathbf{D})\, d\theta \\
&= \int \theta^2 p(\theta|\mathbf{D})\, d\theta - 2E[\theta|\mathbf{D}] \int \theta p(\theta|\mathbf{D})\, d\theta \\
&\quad + \left(E[\theta|\mathbf{D}]\right)^2 \int p(\theta|\mathbf{D})\, d\theta \\
&= E[\theta^2|\mathbf{D}] - \left(E[\theta|\mathbf{D}]\right)^2
\end{aligned}
$$

- If the values of θ are discrete, sums would replace the integrals.

An Example: A Binomial Probability

- In 1997, World Chess Champion Garry Kasparov faced chess computer Deep Blue in a 6-game contest.
- Let π represent the unknown probability of Kasparov winning any particular game, and assume the number of game wins Kasparov obtains in the 6 games is a Binomial$(6, \pi)$ random variable.
- The *Bayes Rules!* book gives an analysis of this setup which is interesting but rather artificial: It assumes a prior distribution that puts positive probability on only three specific values of π, as if these three were the only possible values of π.
- A more realistic analysis would spread the prior probability distribution for π over the whole interval from 0 to 1.
- We will explore such models in the next chapter.