Bayes’ Theorem

- The fundamental equation in Bayesian inference is Bayes’ Theorem, discovered by an English cleric, Thomas Bayes, and published posthumously. It was rediscovered and systematically exploited later by Laplace.

Bayes’ Theorem

- Bayes’ Theorem is a trivial result of the definition of conditional probability: when $P(D|H) \neq 0$,

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

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

$$\propto P(D|A \& H)P(A|H)$$

- Note that the denominator $P(D|H)$ is nothing but a normalization constant required to make the total probability on the left sum to 1

- Often we can dispense with the denominator, leaving its calculation until last, or even leave it out altogether!

Bayes’ Theorem

- Bayes’ theorem is a model for learning. Thus, suppose we have an initial or prior belief about the truth of $A$. Suppose we observe some data $D$. Then we can calculate our revised or posterior belief about the truth of $A$, in the light of the new data $D$, using Bayes’ theorem

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

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

$$\propto P(D|A \& H)P(A|H)$$

- The Bayesian mantra: posterior $\propto$ prior $\times$ likelihood

Bayesian Inference 9/2/04

Bayesian Inference 9/2/04

Bayesian Inference 9/2/04

Bayesian Inference 9/2/04
Bayes’ Theorem

- In the special case that there are only two states of nature, \( A_1 \) and \( A_2=\neg A_1 \), we can bypass the calculation of the marginal likelihood by using the *odds* ratio, the ratio of the probabilities of the two hypotheses:

\[
\text{Prior odds} = \frac{P(A_1 \mid H)}{P(A_2 \mid H)}
\]

\[
\text{Posterior odds} = \frac{P(D \mid A_1 \& H)}{P(D \mid A_2 \& H)} \times \frac{P(A_1 \mid H)}{P(A_2 \mid H)} = \text{Likelihood ratio} \times \text{Prior odds}
\]

- The marginal probability of the data, \( P(D \mid H) \), is the same in each case and cancels out

- The likelihood ratio is also known as the *Bayes factor*.

Bayes’ Theorem

- That’s it! In Bayesian inference there is one uniform way of approaching every possible problem in inference
- There’s not a collection of arbitrary, disparate “tests” or “methods”—everything is handled in the same way
- So, once you have internalized the basic idea, you can address problems of great complexity by using the same uniform approach
- Of course, this means that there are no black boxes. One has to *think* about the problem you have…establish the model, think carefully about priors, decide what summaries of the results are appropriate. It also requires clear thinking about what answers you really want so you know what questions to ask.

\[
\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}}
\]

\[
\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}
\]
Bayes’ Theorem

- The hardest practical problem of Bayesian inference is actually doing the integrals. Often these integrals are over high-dimensional spaces.
- Although some exact results can be given (and the readings have a number of them, the most important being for normally distributed data), in many (most?) practical problems we must resort to simulation to do the integrals. In the past 15 years, a powerful technique, Markov Chain Monte Carlo (MCMC) has been developed to get practical results.

Examples

- Consider two extreme cases. The states of nature are $A_1$ and $A_2$.
- We observe data $D$
- Suppose $P(D|A_1,H) = P(D|A_2,H)$. What have we learned?

Examples

- Consider two extreme cases. The states of nature are $A_1$ and $A_2$.
- We observe data $D$
- Suppose $P(D|A_1,H) = P(D|A_2,H)$. What have we learned?

Examples

- Consider two extreme cases. The states of nature are $A_1$ and $A_2$.
- We observe data $D$
- Suppose $P(D|A_1,H) = 1$, $P(D|A_2,H) = 0$. What have we learned?

Examples

- Consider two extreme cases. The states of nature are $A_1$, $A_2$ and $A_3$, and two possible data $D_1$ and $D_2$. Suppose the likelihood is given by the following table:

|        | $P(D|A)$ | Sum |
|--------|----------|-----|
|        | $D_1$    | $D_2$|    |
| $A_1$  | 0.0      | 1.0  | 1.0 |
| $A_2$  | 0.7      | 0.3  | 1.0 |
| $A_3$  | 0.2      | 0.8  | 1.0 |

- What happens to our belief about the three states of nature if we observe $D_1$? $D_2$?
Examples

Here’s a nice way to arrange the calculation (for these simple cases):

<table>
<thead>
<tr>
<th>A</th>
<th>Prior</th>
<th>D₁</th>
<th>D₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0.3</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>A₂</td>
<td>0.5</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>A₃</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Suppose we observe D₁. Then D₂ is irrelevant (we didn’t observe it) and we calculate the posterior:

| A   | Prior | D₁ | D₂ | Joint | P(A₁|D₁) | P(A₂|D₁) | P(A₃|D₁) |
|-----|-------|----|----|-------|--------|----------|----------|
| A₁  | 0.3   | 0.0| 1.0| 0.00  | 0.00   | 0.00     | 0.00     |
| A₂  | 0.5   | 0.7| 0.3| 0.35  | 0.90   | 0.10     | 0.10     |
| A₃  | 0.2   | 0.2| 0.8| 0.04  | 0.10   | 0.39     | 1.00     |

Examples

• Suppose we observe D₂. How do we calculate the posterior?

| A   | Prior | D₁ | D₂ | P(A₁|D₂) | P(A₂|D₂) | P(A₃|D₂) |
|-----|-------|----|----|--------|----------|----------|
| A₁  | 0.3   | 0.0| 1.0| 0.39   | 1.00     | 0.00     |
| A₂  | 0.5   | 0.7| 0.3| 0.90   | 0.39     | 1.00     |
| A₃  | 0.2   | 0.2| 0.8| 1.00   | 0.90     | 0.39     |

Note that in all of these examples, if we were to multiply the likelihood by a constant, the results would be unchanged since the constant would cancel out when we divide by the marginal probability of the data or when we compute the Bayes factor.

This means that we don’t need to worry about normalizing the likelihood (it isn’t normalized as a function of the states of nature anyway). This is a considerable simplification in practical calculations.
Examples

- The hemoccult test for colorectal cancer is a good example. Let $D$ be the condition that the patient has the disease, $+$ be the data that the patient tests positive for the condition, and $-$ the data that the patient tests negative.
- The test is not perfect. Sigmoidoscopy or colonoscopy are much more accurate, but much more expensive, too expensive to use for screening tests. In the general population, only 0.3% have undiagnosed colorectal cancer. We are interested in the proportion of false negatives and false positives that would occur if we used the test to screen the general population.
- The hemoccult test will be positive 50% of the time if the patient has the disease, and will be positive 3% of the time if the patient does not have the disease.

<table>
<thead>
<tr>
<th></th>
<th>Likelihood</th>
<th>Joint</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>0.003</td>
<td>0.50</td>
<td>0.0015</td>
</tr>
<tr>
<td>$\neg D$</td>
<td>0.997</td>
<td>0.03</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Marginal</td>
<td>0.0314</td>
<td>0.9686</td>
</tr>
</tbody>
</table>

From this table we see that if a person in the general population tests positive, there is still less than a 5% chance that he has the condition. There are a lot of false positives. This test is commonly used as a screening test, but it is not accurate and a positive test must be followed up by colonoscopy (the "gold standard").
- There are few false negatives; a negative test is good news.

Examples (Natural Frequencies)

- Many doctors and most patients do not understand the real meaning of a test like this, and it is sometimes difficult to get the idea across.
- One way is to use natural frequencies, which involves considering a particular size population and computing the expected number of each category in the population.
- This is a good way for both doctors and patients to understand the real meaning of the test results. It is also a good way for a professional statistician to communicate the meaning of any statistical situation to a statistically naive client.

Here, for example, we could consider a population of 10,000 patients screened. In that population:
- 0.3%, or 30 have the condition
  - Of these, 50%, or 15, test positive and 15 test negative
- The remaining 9,970 do not have the condition
  - Of these, 3%, or 299, test positive and 9,471 test negative
- Bottom line: less than 5% of the positives actually have the condition, and 0.16% of the negatives have it
- Thus the test is good for ruling out the condition, but not so good for detecting it (95% false positive rate).
Let’s Make a Deal (Formal Solution)

• We can set up the problem in the following table. You have chosen door 1, so the host cannot open that door. Supposes he opens door 2. If the prize is behind door 1, the host has a choice; if behind door 3, he does not.

<table>
<thead>
<tr>
<th></th>
<th>Prior</th>
<th>Likelihood</th>
<th>Joint</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1/3</td>
<td>1/2</td>
<td>1/6</td>
<td>1/3</td>
</tr>
<tr>
<td>D2</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1/3</td>
<td>1</td>
<td>1/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>

Marginal 1/2

• We see that it is twice as likely that the prize is behind door 3, so that it is advantageous to switch

• Exercise: Explain this problem to a friend using natural frequencies

Example: Mice Again

• We have a male and a female mouse, black coat.
• The female’s mother had a brown coat, so the female must be Bb.
• We don’t know about the male. We wish to determine the male’s genetic type
• Prior: Can set P(BB)=1/3, P(Bb)=2/3 (see problem in previous chart set)
• Suppose the male and female have a litter with 5 pups, all with black coat. What is the probability that the male is BB?

Bayesian Jurisprudence

• The prosecutor’s fallacy involves confusing the two inequivalent conditional probabilities $P(A|B)$ and $P(B|A)$. An example of this is the following argument that the accused is guilty:
  • The probability that the accused’s DNA would match the DNA found at the scene of the crime if he is innocent is only one in a million. Therefore, the probability that the accused is innocent of the crime is only one in a million
  • Confuses $P(\text{match} | \text{innocent})=10^{-6}$ with $P(\text{innocent} | \text{match})=10^{-6}$ ????

Bayesian Jurisprudence

• To do this correctly we must take the prior probabilities into account. Suppose that the crime takes place in a city of 10 million people, and suppose that this is the only other piece of evidence we have. Then a reasonable prior might be $P(\text{guilty})=10^{-7}$, $P(\text{innocent})=1-10^{-7}$

• Using natural frequencies (if you served on a jury, this would be a good way to explain the situation to your fellow jurors) we see that it is likely that there are 10 innocent people in a city of ten million whose DNA would match. And there is one guilty person, for a total of 11 matches. Thus on this data alone and using $P(\text{match} | \text{guilty})=1$
  • $P(\text{innocent} | \text{match})=10/11$

• Do a formal Bayesian analysis to confirm this result!
In a Bayesian approach to jurisprudence, we would have to assess the effect of each piece of evidence on the guilt or innocence of the accused, taking into account any dependence or independence. For example, in the DNA example we just cited, if we knew that the accused had an identical twin brother living in the city, we would expect an additional DNA match over and above the 11 expected by the naïve calculation, making $P(\text{innocence}) = \frac{11}{12}$ instead of $\frac{10}{11}$ (here, since we know about the twin, the DNA data isn’t independent across all in the city).

Depending on the kind of test done, if the accused had close relatives living in the town (who might also match), we might have to add them to the pool of potential matches, further increasing the probability of innocence.

Comment: Although it is common for expert witnesses to give very small DNA match false positive rates, in practice the real probabilities are much larger. Typical rates from commercial labs come in at the level of 0.5%-1%. The lab used in the OJ Simpson case tested at 1 erroneous match in 200. This can be due to many causes:

- Laboratory errors
- Coincidental match
- DNA from accused placed at crime scene either unintentionally or (as claimed by the defense in the OJ Simpson case) intentionally
- DNA from accused left at crime scene before or after the crime

We might consider also whether the accused had a motive. Motive is often considered an important component of any prosecution, because it is much more likely that a person would commit a crime if he/she had a motive than if not.

Thus for example, if a murder involved someone who had a lot of enemies or rivals who would benefit from his demise, there may be a lot more people with motive than for someone who was liked by nearly all. This would decrease the prior probability of guilt.

We might approach it this way: if the number of people in the city is $N_{\text{city}}$ then the prior probability of guilt is $\frac{1}{N_{\text{city}}}$ and the prior odds of guilt are

$$O(G) = \frac{P(G)}{1-P(G)} = \frac{1}{1-\frac{1}{N_{\text{city}}}}$$

$$= \frac{N_{\text{city}}-1}{N_{\text{city}}-1}$$

If the number of people in the city with a motive is $N_{\text{motive}}$, then the posterior odds of guilt would be

$$O(\text{motive} \mid G) \cdot O(G) = \frac{1}{N_{\text{motive}}-1} \cdot \frac{1}{N_{\text{city}}-1}$$

$$= \frac{1}{N_{\text{motive}}-1}$$
Bayesian Jurisprudence

- This calculation assumed independence. But if we use DNA evidence to narrow down the pool of potential murderers in determining our prior for the motive data, and if the suspect had a motive, then relatives of the suspect might also have a motive and the probabilities cannot be simply multiplied since they are no longer independent. Some care is required!

Bayesian Jurisprudence–Combining Data

- In general, when we consider multiple pieces of evidence, a correct Bayesian analysis will condition as follows:

\[
P(H \mid D_1, D_2) = \frac{P(D_2 \mid H, D_1) P(D_1 \mid H) P(H)}{P(D_2 \mid \overline{H}, D_1) P(D_1 \mid \overline{H}) P(\overline{H})}
\]

Thus we use the posterior after observing \(D_1\) as the prior for \(D_2\). We can “chain” as long as we wish, as long as we condition carefully and correctly.

- We can multiply independent probabilities iff the data are independent:

\[
\frac{P(H \mid D_1, D_2)}{P(\overline{H} \mid D_1, D_2)} = \frac{P(D_2 \mid H) P(D_1 \mid H) P(H)}{P(D_2 \mid \overline{H}) P(D_1 \mid \overline{H}) P(\overline{H})}
\]

OJ Simpson Case

- During the OJ Simpson case, Simpson’s lawyer Alan Dershowitz stated that “fewer than 1 in 2000 of batterers go on to murder their wives [in any given year].” He intended this information to be exculpatory, that is, to tend to exonerate his client.
- The prosecutor’s fallacy involved confusing two inequivalent conditional probabilities, usually \(P(A \mid B)\) for \(P(B \mid A)\). Here the fallacy is a little different: the failure to condition on all background information (remember my warning about this early on?)
- The actual effect of this data is to incriminate his client, as the following Bayesian argument shows [I.J. Good, Nature, 381, 481, 1996]

OJ Simpson Case

- Let \(G\) stand for “the batterer is guilty of the crime”.
- Let \(B\) stand for “the wife was battered by the batterer during the year”.
- Let \(M\) stand for “the wife was murdered (by someone) during the year”.
- Dershowitz’s statement implies that \(P(GB) = 1/2000\) (say)
- Also, \(P(\overline{G}B)\) is very close to 1, call it 1
- Also, \(P(MG\&B) = P(MG) = 1\). Surely if the batterer is guilty of murdering his wife, the wife was murdered.
- In this notation, the particular fallacy is in confusing \(P(GB)\) with \(P(GB\&M)\), which turn out to be very different.
OJ Simpson Case

- We can estimate \( P(M \sim G \& B) \) as follows… There are about 25,000 murders in the US per year, out of a population of 250,000,000, or a rate of 1/10,000. Of these, about a quarter are women (rough approximation), so that the probability of being murdered if you are a woman is half this, 1/20,000. Most of these are just random murders, for which the batterer is not guilty, so we can approximate

\[
P(M \sim G \& B) = P(M \sim G) = 1/20,000
\]

OJ Simpson Case

- Now we can estimate the posterior odds that the batterer is guilty of the murder, as follows:

\[
\frac{P(G \mid M \& B)}{P(\neg G \mid M \& B)} = \frac{P(M \mid G \& B) \times P(G \mid B)}{P(M \mid \neg G \& B) \times P(\neg G \mid B)}
\]

\[
= \frac{1/20,000 \times 1/2000}{1/20,000 \times 1} = 10
\]

OJ Simpson Case (Natural Frequencies)

- Out of every 100,000 battered women, about 5 will die each year due to having been murdered by a stranger (this is 100,000/20,000 where the 1/20,000 factor is from the previous chart)

- But according to Dershowitz, out of every 100,000 battered women, 50 will die each year due to having been murdered by their batterer.

- Thus, looking at the population of women who were battered and murdered in a given year, the ratio is 10:1. This is the change in odds in favor of the hypothesis that OJ murdered his wife, and not some random stranger, when we learn that OJ’s wife was both battered and murdered.

OJ Simpson Case (Natural Frequencies)

- We can look at this in tree form:
### Three Similar But Different Problems

- **Factory:** A machine has good and bad days. 90% of the time it is “good”, and 95% of the parts are “good”. 10% of the time it is “bad” and 70% of the parts are “good”.
- On a particular day, the first twelve parts are sampled. 9 are good, 3 are bad (that is our data \(D\)). Is it a good or a bad day?

### Three Similar But Different Problems

- In this example, note that we calculate the probability of the particular sequence:

\[
D = \{g, b, g, g, g, g, g, b, g, b\} = \{d_1, d_2, \ldots, d_{12}\}
\]

- If we considered only the count without regard for the sequence, there would be an additional factor of the binomial coefficient “12 Choose 9”:

\[
C^{12}_9 = \binom{12}{9} = \frac{12!}{9!(12 - 9)!}
\]

- However, each posterior probability gets the same additional factor, so it cancels (either in the Bayes’ factor or in the posterior probability).

### Three Similar But Different Problems

- It is crucial for this problem that the samples be independent, that is, the fact that we sampled a good (or bad) part gives us no information about the other samples.
  - It’s certainly possible that the samples might not be independent; e.g., when the machine is in its “Bad” state, we have \(P(b_n \mid b_{n-1}, \text{Bad}) \neq P(b_n \mid g_{n-1}, \text{Bad})\).
  - The archetypical example of such sampling is “sampling with replacement”. For example, suppose we have an urn with two colors of balls in it. We draw a ball at random, note the color, and replace it. This means that when we draw a sample from the urn, we do not affect the probabilities of the subsequent samples because we restore the *status quo ante*, so the samples are independent.

### Three Similar But Different Problems

- A town has 100 voters. We sample 10 voters to see whether they will vote yes or no on a proposition. We get 6 “yes”, 4 “no”. What can we infer about the probable result \(R\) of the election?
  - Guess \(R=100 \times 6/10\), but this is a frequentist guess. We want a Bayesian posterior probability on the result \(R\).
Three Similar But Different Problems

- Let $Y_i$ be the “yes” votes polled and $N_i$ the “no” votes
- $P(Y_1 | R) = R/100$
  $P(Y_2 | R \& Y_1) = (R-1)/99$
  $P(Y_3 | R \& Y_1 \& Y_2) = (R-2)/98$
  ...
  $P(Y_6 | R \& Y_1 \& Y_2 \& \ldots \& Y_5) = (R–5)/95$
- $P(N_1 | R \& Y_1 \& Y_2 \& \ldots \& Y_6) = (100–R)/94$
- $P(N_2 | R \& N_1 \& Y_1 \& Y_2 \& \ldots \& Y_6) = (99–R)/93$
- $P(N_4 | R \& N_1 \& \ldots \& N_3 \& Y_1 \& Y_2 \& \ldots \& Y_6) = (97–R)/91$
- Note that the likelihood is 0 if $R \leq 5$ or $R \geq 97$, as it must be since we know for sure that at the time of the poll 6 voters support the proposition and 4 oppose it
- To get the posterior distribution on $R$ we need a prior. We don’t know anything, so a conventional prior might be flat, $P(R)=$constant

Three Similar But Different Problems

- Then the posterior probability of $R$, assuming a flat prior, is given by
  $$P(R | D) \propto P(D | R)P(R) \propto P(D | R)$$

- The posterior distribution of course has to be normalized by dividing by the sum of $P(D|R)$ over all $R$.
- Are there any other assumptions that we should make explicit here?
Three Similar But Different Problems

- In this example, we have a lake with an unknown number \( N \) of identical fish. We catch \( n \) of them, tag them, and return them to the lake. At a later time (when we presume that the tagged fish have swum around and thoroughly mixed with the untagged fish) we catch \( k \) fish, and observe the number tagged.
- For example, \( n=60 \), \( k=100 \), of which 10 are tagged. What is the total number of fish in the lake?
- [This is another archetypical problem, the “catch-and-release” problem.]

Three Similar But Different Problems

- Likelihood in this case is similar to the voting problem, with a total population \( N \) (but this time \( N \) is unknown):
  \[
P(D \mid N) = \frac{60 \times 59 \times \cdots \times 51 \times (N - 60)(N - 61)\cdots(N - 149)}{N(N - 1)\cdots(N - 99)}
  \]
- Again for illustration take a flat prior (but this is unrealistic since we have knowledge that the lake cannot hold an infinite number of fish. Nonetheless…)
  \[
P(N \mid D) \propto P(D \mid N) P(N) \propto P(D \mid N)
  \]
- The prior is improper (sums to infinity) since there is no bound on \( N \). This will not cause problems as long as the posterior is proper (sums to a finite result)
- The posterior says that \( N \geq 150 \), known from the data

Three Similar But Different Problems

- Here is the posterior distribution under these assumptions
Three Similar But Different Problems

- The examples show the Bayesian style:
  - List all states of nature
  - Assign a prior probability to each state
  - Determine the likelihood (probability of obtaining the data actually observed as a function of state of nature)
  - Multiply prior times likelihood to obtain an unnormalized posterior distribution
  - If needed, normalize the posterior
- One has to make assumptions about the things that go into the inference. Bayesian analysis forces you to make the assumptions explicit. There is no black magic or black boxes.