More Programming and Statistics Boot Camps?

This week (only): PG Wed Office Hour 8 Feb at 3pm

Prob Set 1: due Mon night 13 Feb (no extensions …)

Note: Added part to problem 2 (later today [Tues])
Red (upper): The probability $1 - \frac{365!}{(365-n)!365^n}$ that at least two birthdays coincide within a group of $n$ people, as function of $n$.

Green (lower): The probability $1 - \left(\frac{364}{365}\right)^{n-1}$ of a birthday coinciding with yours within a group of $n$ people including you.
This class

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/19</td>
<td>2</td>
</tr>
<tr>
<td>01/27</td>
<td>3</td>
</tr>
<tr>
<td>02/20</td>
<td>3</td>
</tr>
<tr>
<td>03/30</td>
<td>2</td>
</tr>
<tr>
<td>04/01</td>
<td>2</td>
</tr>
<tr>
<td>04/04</td>
<td>3</td>
</tr>
<tr>
<td>05/08</td>
<td>2</td>
</tr>
<tr>
<td>05/16</td>
<td>2</td>
</tr>
<tr>
<td>05/31</td>
<td>2</td>
</tr>
<tr>
<td>06/02</td>
<td>2</td>
</tr>
<tr>
<td>06/11</td>
<td>2</td>
</tr>
<tr>
<td>06/15</td>
<td>2</td>
</tr>
<tr>
<td>06/25</td>
<td>2</td>
</tr>
<tr>
<td>07/01</td>
<td>2</td>
</tr>
<tr>
<td>07/24</td>
<td>2</td>
</tr>
<tr>
<td>08/13</td>
<td>2</td>
</tr>
<tr>
<td>08/17</td>
<td>2</td>
</tr>
<tr>
<td>10/05</td>
<td>3</td>
</tr>
<tr>
<td>10/17</td>
<td>2</td>
</tr>
<tr>
<td>11/07</td>
<td>2</td>
</tr>
<tr>
<td>11/27</td>
<td>2</td>
</tr>
<tr>
<td>12/10</td>
<td>2</td>
</tr>
<tr>
<td>12/27</td>
<td>2</td>
</tr>
<tr>
<td>12/28</td>
<td>2</td>
</tr>
</tbody>
</table>

(See PS1 #4)
Bayes’ Rule

A simple formula follows from the above definitions and symmetry of the joint probability:

\[ p(A|B)p(B) = p(A, B) = p(B, A) = p(B|A)p(A): \]

\[ p(A|B) = \frac{p(B|A)p(A)}{p(B)} \]

Called “Bayes’ theorem” or “Bayes’ rule” — connects inductive and deductive inference

(Rev. Thomas Bayes (1763), Pierre-Simon Laplace (1812), Sir Harold Jeffreys (1939))

For mutually disjoint sets \( A_i \) with \( \bigcup_{i=1}^{n} A_i = S \), Bayes’ rule takes the form

\[ p(A_i|B) = \frac{p(B|A_i)p(A_i)}{p(B|A_1)p(A_1) + \ldots + p(B|A_n)p(A_n)}. \]
**Example 1:** Consider a casino with loaded and unloaded dice.

For a loaded die \((L)\), probability of rolling a 6 is 50%:

\[
p(6|L) = 1/2, \text{ and } p(i|L) = 1/10 \ (i = 1, \ldots, 5)
\]

For a fair die \((\overline{L})\), the probabilities are \(p(i|\overline{L}) = 1/6 \ (i = 1, \ldots, 6)\).

Suppose there’s a 1% probability of choosing a loaded die:

\[
p(L) = 1/100.
\]

If we select a die at random and roll three consecutive 6’s with it, what is the posterior probability, \(P(L|6,6,6)\), that it was loaded?
The probability of the die being loaded, given 3 consecutive 6’s, is

\[
p(L|6, 6, 6) = \frac{p(6, 6, 6|L)p(L)}{p(6, 6, 6)} = \frac{p(6|L)^3p(L)}{p(6|L)^3p(L) + p(6|\bar{L})^3p(\bar{L})}
\]

\[
= \frac{(1/2)^3 \cdot (1/100)}{(1/2)^3 \cdot (1/100) + (1/6)^3 \cdot (99/100)}
\]

\[
= \frac{1}{1 + (1/3)^3 \cdot 99} = \frac{1}{1 + 11/3} = \frac{3}{14} \approx .21
\]

so only a roughly 21% chance that it was loaded.

(Note that the Bayesian “prior” in the above is \(p(L) = 1/100\), giving the expected probability before collecting the data from actual rolls, and significantly affects the inferred posterior probability.)
Example 2: Duchenne Muscular Dystrophy (DMD):

model as simple recessive sex-linked disease caused by a mutated X chromosome ($\tilde{X}$)

$\tilde{XY}$ male expresses the disease

$\tilde{XX}$ female is a carrier but does not express the disease

Neither of a woman’s parents expresses the disease, but her brother does.
The probability of the die being loaded, given 3 consecutive 6's, is

\[ p(L|6,6,6) = p(6,6,6|L) p(L) p(6|L) + p(6|L) \]

\[ = \frac{1}{2} \cdot \frac{1}{100} \cdot \frac{1}{2} \cdot \frac{1}{100} + \frac{1}{6} \cdot \frac{99}{100} \]

\[ = \frac{1}{1 + \left( \frac{1}{3} \right) \cdot \frac{99}{100}} \]

\[ = \frac{3}{14} \]

so only a roughly 21% chance that it was loaded.

(Note that the Bayesian "prior" in the above is \( p(L) = \frac{1}{100} \), giving the expected probability before collecting the data from actual rolls, and significantly affects the inferred posterior probability.)

Example 2: Duchenne Muscular Dystrophy (DMD): model as simple recessive sex-linked disease caused by a mutated X chromosome (\( e^X \)).

\( e^X Y \) male expresses the disease

\( e^X X \) female is a carrier but does not express the disease

Neither of a woman’s parents expresses the disease, but her brother does.

Then the woman’s mother must be a carrier, and the woman herself therefore has an \textit{a priori} 50/50 chance of being a carrier, \( p(C) = 1/2 \).

She gives birth to a healthy son (h.s.).

What now is her probability of being a carrier?
Her probability of being a carrier, given a healthy son, is

\[
p(C|h.s.) = \frac{p(h.s.|C)p(C)}{p(h.s.)}
\]

\[
= \frac{p(h.s.|C)p(C)}{p(h.s.|C)p(C) + p(h.s.|\overline{C})p(\overline{C})}
\]

\[
= \frac{(1/2) \cdot (1/2)}{(1/2) \cdot (1/2) + 1 \cdot (1/2)}
\]

\[
= \frac{1}{3}
\]

(where \( \overline{C} \) means “not carrier”).
Intuitively:

if she’s not a carrier, then there are two ways she could have a healthy son

(from either of her good X’s)

whereas: if she’s a carrier there’s only one way

So a total of three ways and the probability that she’s a carrier is 1/3, given the knowledge that she’s had exactly one healthy son.
Also worth noting:

the woman has a hidden state $C$ or $\overline{C}$

determined **once and for all**

she doesn’t make an independent coin flip each time she has a child as to whether or not she’s a carrier

Prior to generating data about her son or sons, she has a “Bayesian prior” of 1/2 to be a carrier.

Subsequent data permits a principled reassessment of that probability:
continuously decreasing for each successive healthy son
or jumping to 1 if she has a single diseased son
Example 3:

Suppose there’s a rare genetic disease that affects 1 out of a million people

\[ p(D) = 10^{-6} \]

Suppose a screening test for this disease is 100% sensitive

(i.e., is always correct if one has the disease)

and 99.99% specific

(i.e., has a .01% false positive rate)

Is it worthwhile to be screened for this disease?
Is it worthwhile to be screened for this disease? What is \(p(D|+)\)?

The above sensitivity and specificity imply:

\[ p(+|D) = 1 \quad \text{and} \quad p(+|\overline{D}) = 10^{-4} \]

so the probability of having the disease, given a positive test (+), is

\[
p(D|+) = \frac{p(+|D)p(D)}{p(+)} = \frac{p(+|D)p(D)}{p(+|D)p(D) + p(+|\overline{D})p(\overline{D})} = \frac{1 \cdot 10^{-6}}{1 \cdot 10^{-6} + 10^{-4}(1-10^{-6})} \approx 10^{-2}
\]

and there’s little point to being screened (only once).
Intuitively:

if one million people were screened,

we would expect roughly one to have the disease,

but the test will give roughly 100 false positives.

So a positive result would imply only roughly a 1 out of 100 chance for one of those positives to have the disease.

In this case the result is biased by the small [one in a million] Bayesian prior $p(D)$
More Joint Probabilities

\[ p(A, B \mid C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A \cap B \cap C)}{p(C)} \]

\[ p(A \mid B, C) = \frac{p(A, B, C)}{p(B, C)} = \frac{p(A \cap B \cap C)}{p(B \cap C)} \]

“Chain Rule”

\[ p(A, B, C) = p(A \mid B, C) p(B, C) = p(A \mid B, C) p(B \mid C) p(C) \]

\[ p(A_n, A_{n-1}, \ldots, A_1) = p(A_n \mid A_{n-1}, \ldots, A_1) p(A_{n-1}, \ldots, A_1) = \ldots \]

\[ = p(A_n \mid A_{n-1}, \ldots, A_1) p(A_{n-1} \mid A_{n-2}, \ldots, A_1) \cdots p(A_3 \mid A_2, A_1) p(A_2 \mid A_1) p(A_1) \]
Binary Classifiers:

Use a set of features to determine whether objects have binary (yes or no) properties.

Examples: whether or not a text is classified as medicine, or whether an email is classified as spam.

In those cases, the features of interest might be the words the text or email contains.
“Naive Bayes” methodology:

statistical method (making use of the word probability distribution)

as contrasted with a “rule-based” method
(where a set of heuristic rules is constructed and then has to be maintained over time)

Advantage of the statistical method:

features automatically selected and weighted properly,
no additional ad hoc methodology

easy to retrain as training set evolves over time, using same straightforward framework
Spam Filters

Spam filter = binary classifier where property is whether message is spam (\(S\)) or non-spam (\(\overline{S}\)).

Features = words of the message.

Assume we have a training set of messages tagged as spam or non-spam and use the document frequency of words in the two partitions as evidence regarding whether new messages are spam.

Example 1 (Rosen p. 422):

Suppose the word "Rolex" appears in 250 messages of a set of 2000 spam messages, and in 5 of 1000 non-spam messages.

Then we estimate 
\[
p(\text{"Rolex"} | S) = \frac{250}{2000} = .125
\]
\[
p(\text{"Rolex"} | \overline{S}) = \frac{5}{1000} = .005
\]

Assuming a "flat prior" (\(p(S) = p(\overline{S}) = \frac{1}{2}\)) in Bayes' law gives

\[
p(S | \text{"Rolex"}) = \frac{p(\text{"Rolex"} | S) p(S)}{p(\text{"Rolex"} | S) p(S) + p(\text{"Rolex"} | \overline{S}) p(\overline{S})}
\]

\[
= \frac{.125}{.125 + .005} = .962
\]

With a rejection threshold of .9, this would be rejected.
Example 1 (Rosen p. 422):

Suppose the word “Rolex” appears in 250 messages of a set of 2000 spam messages, and in 5 of 1000 non spam messages.

Then we estimate \( p(\text{“Rolex”}|S) = \frac{250}{2000} = .125 \)
and \( p(\text{“Rolex”}|\overline{S}) = \frac{5}{1000} = .005 \).

Assuming a “flat prior” \( p(S) = p(\overline{S}) = 1/2 \) in Bayes’ law gives

\[
p(S|\text{“Rolex”}) = \frac{p(\text{“Rolex”}|S)p(S)}{p(\text{“Rolex”}|S)p(S) + p(\text{“Rolex”}|\overline{S})p(\overline{S})} = \frac{.125}{.125 + .005} = \frac{.125}{.130} = .962 .
\]

With a rejection threshold of .9, this would be rejected.
Example 2 (two words, “stock” and “undervalued”):

Now suppose in a training set of 2000 spam messages and 1000 non-spam messages, the word “stock” appears in 400 spam messages and 60 non-spam, and the word “undervalued” appears in 200 spam and 25 non-spam messages.

Then we estimate

\[
p(\text{“stock”}|S) = 400/2000 = .2
\]

\[
p(\text{“stock”}|\overline{S}) = 60/1000 = .06
\]

\[
p(\text{“undervalued”}|S) = 200/2000 = .1
\]

\[
p(\text{“undervalued”}|\overline{S}) = 25/1000 = .025
\]
Accurate estimates of joint probability distributions \( p(w_1, w_2 | S) \) and \( p(w_1, w_2 | \overline{S}) \) would require too much data.

Key assumption: assume statistical independence to estimate as

\[
p(w_1, w_2 | S) = p(w_1 | S) \cdot p(w_2 | S)
\]

\[
p(w_1, w_2 | \overline{S}) = p(w_1 | \overline{S}) \cdot p(w_2 | \overline{S})
\]

(This assumption is not true in practice: words are not statistically independent. But we’re only interested in determining whether above or below some threshold, not trying to calculate an accurate \( p(S | \{w_1, w_2, \ldots, w_n\}) \))
Write $w_1 = \text{“stock”}$ and $w_2 = \text{“undervalued”}$, and recall:

\[
p(w_1|S) = \frac{400}{2000} = .2 \quad p(w_1|\bar{S}) = \frac{60}{1000} = .06, \\
p(w_2|S) = \frac{200}{2000} = .1 \quad p(w_2|\bar{S}) = \frac{25}{1000} = .025
\]

So assuming a flat prior ($p(S) = p(\bar{S}) = 1/2$), and independence of the features gives

\[
p(S|w_1, w_2) = \frac{p(w_1, w_2|S)p(S)}{p(w_1, w_2|S)p(S) + p(w_1, w_2|\bar{S})p(\bar{S})} \\
= \frac{p(w_1|S)p(w_2|S)p(S)}{p(w_1|S)p(w_2|S)p(S) + p(w_1|\bar{S})p(w_2|S)p(\bar{S})} = \frac{.2 \cdot .1}{.2 \cdot .1 + .06 \cdot .025} = .930
\]

at a .9 probability threshold a message containing those two words would be rejected as spam.
More generally, for $n$ features (words)

$$p(S|\{w_1, w_2, \ldots, w_n\}) = \frac{p(\{w_1, w_2, \ldots, w_n\}|S)p(S)}{p(\{w_1, w_2, \ldots, w_n\})}$$

= $\frac{p(\{w_1, w_2, \ldots, w_n\}|S)p(S)}{p(\{w_1, w_2, \ldots, w_n\}|S)p(S) + p(\{w_1, w_2, \ldots, w_n\}|\overline{S})p(\overline{S})}$

= $\frac{p(w_1|S)p(w_2|S)\cdots p(w_n|S)p(S)}{p(w_1|S)p(w_2|S)\cdots p(w_n|S)p(S) + p(w_1|\overline{S})p(w_2|\overline{S})\cdots p(w_n|\overline{S})p(\overline{S})}$

= $\frac{p(S)\prod_{i=1}^{n} p(w_i|S)}{p(S)\prod_{i=1}^{n} p(w_i|S) + p(\overline{S})\prod_{i=1}^{n} p(w_i|\overline{S})}$