Day 6 - Intro to Bayesian Statistics

Your Name

9/14/2021

Agenda

0. Survey + Motivation
   - source: https://www.youtube.com/watch?v=lG4VkPoG3ko (https://www.youtube.com/watch?v=lG4VkPoG3ko)
1. Rice, Bayesian Statistics (a very brief introduction)
   - source: https://faculty.washington.edu/kenrice/BayesIntroClassEpi2018.pdf (https://faculty.washington.edu/kenrice/BayesIntroClassEpi2018.pdf)
2. Intuition for Bayes Theorem
3. the Bayes Factor
4. Bayesian estimation

2. Intuition for Bayes Theorem

Let’s return to our cancer screening example. For reference, here it is again:

> Suppose a 50-year old woman comes in for a routine mammogram. She gets a positive test result, and asks you what her chances of breast cancer are. You tell her:

	+ Result	- Result
Has Cancer	9	1
Cancer Free	89	901

recall that:

Sensitivity: how often does the test correctly identify patients who have cancer?

in this case, we look at the first row (patients who have cancer): the test is accurate 9 times out of 10, or 90%.

Specificity: how often does the test correctly identify patients who are cancer-free?

the second row, cancer free, correctly gives a negative result 901 out of (901 + 89) times, or ~91%.

so: let’s ask the question, if a woman tests positive, what are the odds she has cancer?

in the positive test column, we see that the probability of cancer **given a positive test** is then 9 out of (9 + 89), or 1/11, or ~9%.

---

tests **update** your odds of illness, they do not **predict** your odds of illness!!!

---

let’s go ahead and take a look again at Bayes’ Theorem:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

this is the classical formulation, and these species have specific names:

H: the hypothesis
E: the evidence
P(H): the **prior**
P(E): the marginal probability
P(E|H): the **likelihood**
P(H|E): the **posterior**

more on these later. let’s recast this expression:

$$\begin{aligned} P(H|E) &= \frac{P(E|H) \cdot P(H)}{P(E)} \\ &= \frac{P(H) \cdot P(E|H)}{P(E|H) \cdot P(H) \ + \ P(E|\neg H) \cdot P(\neg H)} \end{aligned}$$

we can do some algebraic manipulation to get this into odds form, which is considerably more intuitive (at least to me):

$$P(H|E) = O(H) \cdot \frac{P(E|H)}{P(E|\neg H)}$$

the second term, $\frac{P(E|H)}{P(E|\neg H)}$, is called the **Bayes Factor** (BF), and the odds of H is the **prior odds**.

this allows us to eyeball problems like the medical one above very quickly and very easily. it takes 3 steps:

1. express the prior in terms of odds
2. compute the Bayes Factor
3. multiply to get the posterior

repeating our above analysis one last time,
1. express the prior in terms of odds:
* 1% chance of having cancer: odds = 1:99
2. compute the Bayes Factor
* $\frac{P(+|\text{Cancer})}{P(+|\neg\text{Cancer})} = \frac{9/(9+1)}{89/(89+901)} = \frac{.9}{.08989} \approx 10$
3. multiply
* posterior odds: 10:99, ~1:10
4. optional: recast as probability
* 10 / (99+10) = 10/109 = ~9%

---

your turn!:

> Let’s “do our own research” and take a look at the clinical data for the InBios International, Inc. SCov-2 Ag ‘Detect’ Rapid Test for COVID (if you’ve had a test through UChicago, this is the one you got).

	+ Result	- Result

	+ Result	- Result
Has COVID	39	0
COVID Free	6	257

data source: https://www.fda.gov/media/148353/download (https://www.fda.gov/media/148353/download)

a. calculate the sensitivity

b. calculate the specificity

c. calculate the Bayes Factor

d. the current Chicago positive test rate is 3.7%. what is the probability you have COVID, given a positive test result, assuming you're currently in Chicago?

e. the current UChicago positive test rate among students is .22%. now what is the probability you have COVID, given a positive test result?

#Answers below

a. Sensitivity: 39 / (39 + 6) = 86.67%

b. Specificity: 257 / (257 + 0) = 100%

c. BF = 86.67% / 13.33% = 6.5

d. posterior ~= 3.7:96.3 * 6.5 = 24.05:96.3 = 19.98%

e. .22:99.78 * 6.5 = 1.43:99.78 = 1.4%

3. the Bayes Factor

> There are two subspecies of African Elephant: savannah and forest elephants, which differ in their genetic makeup. Interpol have seized an illegally-smuggled elephant tusk, and they want to know which subspecies of elephant the tusk came from. To try to answer this they collect DNA from the tusk and measure it at a number of locations ("markers" in genetics jargon) along the elephant genome. At each marker the DNA can be one of two types ("alleles" in genetics jargon), which for simplicity we will label 0 and 1. So the available data on the tusk might look something like this:

Marker	Allele	fS	fF
1	1	.4	.8
2	0	.12	.2
3	1	.21	.11

Marker	Allele	fS	fF
4	0	.12	.17
5	0	.02	.23
6	1	.32	.25

> Interpol also have information on the frequency of each allele in each of the two subspecies - this was obtained by measuring the DNA of a large number of savanna elephants and a large number of forest elephants. We will use fSj and fFj to denote the frequency of “1” allele at marker j in savanna and forest elephants respectively (and since there are only two alleles, the frequency of the 0 allele is 1−fSj and 1−fFj).

> The question before us is: Which subspecies of elephant did the tusk come from, and how confident should we be in this conclusion?

To get some intuition, let us examine the data at the first few markers. At marker 1 our tusk has the allele 1. This allele is less common in savanna elephants than forest elephants (40% of savanna elephants carry this allele, vs 80% of forest elephants), so this observation seems to support the sample coming from forest. However, at the same time 40% of savanna samples carry this allele, so it remains plausible that the sample came from savanna.

Moving to marker 2, our tusk has the allele 0, which is more common in savanna (88%) than forest elephants (80%). And at marker 3 the tusk has allele 1 which is also more common in savanna (21% vs 11%). So, in contrast to marker 1, the data at these two markers are more consistent with the tusk coming from a savanna elephant than a forest elephant.

Solution

We can phrase this problem as a “model comparison” problem. We have data $X = x$ from our tusk, and we have two different models for how those data might have arisen: it could have been sampled from a savanna elephant, or it could have been sampled from a forest elephant. We will use M_S and M_F as shorthand for these two models. A key point is that these two models imply different probability distributions for X: some values of X are more common under M_S and others are more common under M_F.

the probability distributions, then, are

$$p(x|M_S) = \prod_j f_{S_j}^{x_j}\ (1 - f_{S_j})^{1-x_j}$$

$$p(x|M_F) = \prod_j f_{F_j}^{x_j}\ (1 - f_{F_j})^{1-x_j}$$

the likelihood ratio (very similar to, and in this case, identical to, the Bayes Factor), is just the ratio of the likelihood of model 1 to the likelihood of model 2, conditioned on the data:

$$LR(M_1, M_0) := L(M_1)/L(M_0)$$

large values for the likelihood support M1, and small values (near 0) support M0. Values near 1 are inconclusive.

from the elephant data, we then see that:

```
x = c(1,0,1,0,0,1)
fS = c(0.40, 0.12,0.21,0.12,0.02,0.32)
fF = c(0.8,0.2,0.11,0.17,0.23,0.25)
L = function(f,x){
  prod(f^x*(1-f)^(1-x))
}
LR = L(fS,x)/L(fF,x)
print(LR)
```

```
## [1] 1.81359
```

Let’s end this section with an exercise:

> You are playing a game with a friend. The friend has two six-sided dice, one blue and one green. The sides on the blue dice are numbered 1,2,3,3,3, and 4. The sides on the green dice are labelled 1,2,2,3,4,4. He picks one of the dice without telling you, and rolls it 10 times, obtaining the results 3,3,2,3,1,2,3,3,4,3. Looking at these results, does intuition say that they support the green dice or the blue dice? Strongly or weakly? Phrase the problem as a model comparison problem. State your modelling assumptions, and compute a likelihood ratio. Does it support your intuition?

Now let’s expand this, to comparing multiple models to one another:

```
x = c(1,0,1,0,0,1)
ref_freqs = rbind(
  c(0.39, 0.14,0.22,0.12,0.03,0.38),
  c(0.41, 0.10,0.18, 0.12,0.02,0.28),
  c(0.40, 0.11,0.22, 0.11,0.01,0.3),
  c(0.75,0.25,0.11,0.18,0.25,0.25),
  c(0.85,0.15,0.11,0.16,0.21,0.26)
)

# define functions for computing posterior from Likelihood vector and pi vector
normalize = function(x){
  return(x/sum(x))
}
posterior_prob = function(L_vec, pi_vec){
  return(normalize(L_vec*pi_vec))
}

# define likelihood function
L = function(f,x){
  prod(f^x*(1-f)^(1-x))
}

# compute the likelihoods for each model by applying L to rows of ref_freqs
L_vec=apply(ref_freqs,1,L,x=x)
print(L_vec)
```

```
## [1] 0.023934466 0.016038570 0.020702326 0.009513281 0.013712299
```

```
posterior_prob(L_vec, c(0.2,0.2,0.2,0.2,0.2))
```

```
## [1] 0.2852705 0.1911608 0.2467472 0.1133871 0.1634344
```

4. Bayesian Parameter Estimation

We'll now take a Bayesian approach to estimating θ in Example 4.1. We treat the unknown parameter θ as a random variable and wish to find its posterior distribution after observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples. We will start with a very simplified, unrealistic prior distribution that assumes only five possible, equally likely values for θ 0.1, 0.3, 0.5, 0.7, 0.9.

```
# prior
theta = seq(0.1, 0.9, 0.2)
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta,
                     prior,
                     likelihood,
                     product,
                     posterior)

knitr::kable(bayes_table, digits = 4, align = 'r')
```

let’s go ahead and plot these:

```
# plots
plot(theta-0.01, prior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="orange", xlab='theta'
, ylab='', main="Flat Prior")
lines(theta+0.01, likelihood/sum(likelihood), type='h', xlim=c(0, 1), ylim=c(0, 1), col=
"skyblue")
lines(theta, posterior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="seagreen")
legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange",
"skyblue", "seagreen"))
```

Now consider a prior distribution which places probability 1/9, 2/9, 3/9, 2/9, 1/9 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. What does this prior distribution say about θ? Redo the previous analysis How does the posterior distribution change?

```
# prior
theta = seq(0.1, 0.9, 0.2)
prior = c(1,2,3,2,1)
prior = prior / 9.

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta, prior, likelihood, product, posterior)

knitr::kable(bayes_table, digits = 4, align = 'r')

# plots
plot(theta-0.01, prior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="orange", xlab='theta'
, ylab='', main="Scaled Prior 1")
lines(theta+0.01, likelihood/sum(likelihood), type='h', xlim=c(0, 1), ylim=c(0, 1), col=
"skyblue")
lines(theta, posterior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="seagreen")
legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange",
"skyblue", "seagreen"))
```

Now consider a prior distribution which places probability 5/15, 4/15, 3/15, 2/15, 1/15 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. What does this prior distribution say about θ? Redo the previous analysis. How does the posterior distribution change?

```
# prior
theta = seq(0.1, 0.9, 0.2)
prior = c(5,4,3,2,1)
prior = prior / 15.

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta, prior, likelihood, product, posterior)

knitr::kable(bayes_table, digits = 4, align = 'r')

# plots
plot(theta-0.01, prior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="orange", xlab='theta'
, ylab='', main="Scaled Prior 2")
lines(theta+0.01, likelihood/sum(likelihood), type='h', xlim=c(0, 1), ylim=c(0, 1), col=
"skyblue")
lines(theta, posterior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="seagreen")
legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange",
"skyblue", "seagreen"))
```