Solutions Chapter 6 - The devil is in the denominator

alexandrahotti

October 2019

6.1 Too many coin flips

Suppose we flip two coins. Each coin i is either fair ($Pr(H) = \theta_i = 0.5$) or biased towards heads ($Pr(H) = \theta_i = 0.9$) however, we cannot visibly detect the coin's nature. Suppose we flip both coins twice and record each result.

Problem 6.1.1.

Suppose that we specify a discrete uniform prior on both θ_1 and θ_2. Find the joint distribution of the data and the coins' identity.

Answer:
The joint probability can be rewritten using the conditional probability rule:

$$Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) = Pr(X_1, X_2 | Y_1, Y_2, \theta_1, \theta_2) Pr(Y_1, Y_2, \theta_1, \theta_2) \tag{1}$$

Now, we use that X_1, X_2 are independent of Y_1, Y_2, θ_2 given θ_1 and again the conditional probability rule.

$$Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) = Pr(X_1, X_2 | \theta_1) Pr(Y_1, Y_2 | \theta_1, \theta_2) Pr(\theta_1, \theta_2)$$

Lastly, we use that θ_1 and θ_2 are independent and that Y_1, Y_2 are independent of θ_1.

$$Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) = Pr(X_1, X_2 | \theta_1) Pr(Y_1, Y_2 | \theta_2) Pr(\theta_1) Pr(\theta_2) \tag{2}$$

Now, since we are modeling coin flips with one trial for each coin a suitable model seems to be the Bernoulli distribution. Defined as:

$$Pr(X = k | \theta) = \theta^k (1 - \theta)^{1-k} \tag{3}$$

Using this in (2) gives us:

$$Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) = \theta_1^{X_1} \theta_1^{X_2} (1 - \theta_1)^{1 - X_1} (1 - \theta_1)^{1 - X_2} \cdot \theta_2^{Y_1} \theta_2^{Y_2} (1 - \theta_2)^{1 - Y_1} (1 - \theta_2)^{1 - Y_2} \cdot 0.5 \cdot 0.5$$

$$Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) = \theta_1^{X_1 + X_2} (1 - \theta_1)^{2 - X_1 - X_2} \cdot \theta_2^{Y_1 + Y_2} (1 - \theta_2)^{2 - Y_1 - Y_2} \cdot 0.5^2$$

Problem 6.1.2.

Show that the above distribution is a valid probability distribution.

Answer:

- Since the two θ values are in the range [0,1] all values will be non-negative.
- The distribution should sum to 1. This seems a bit tedious though since we have so many parameters.

Problem 6.1.3.

We flip each coin twice and obtain for coin 1 {HH} and coin 2 {HT}. Assuming that the result of each coin flip is independent of the previous result write down a likelihood function.

Answer:

$$Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2) = Pr(\{H,T\}|\theta_2)Pr(\{H,H\}|\theta_1)$$

Using that heads is given by: $X = 1$ and tails: $X = 0$.

$$Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2) = \theta_1^1\theta_1^1 \cdot \theta_2(1-\theta_2)^{1-0} = \theta_1^2\theta_2(1-\theta_2)$$

Problem 6.1.4.

What are the maximum likelihood estimators of each parameter?

Answer: We know from the assignment description that the θ parameter can be either 0.5 or 0.9. The likelihood gets as big as possible if:
$\hat{\theta_1}$: has the value 0.9. Since $0.9^2 = 0.81 > 0.5^2 = 0.25$.
$\hat{\theta_2}$: has the value 0.5. Since $0.5(1-0.5) = 0.25 > 0.9(1-0.9) = 0.09$.

Problem 6.1.5.

Calculate the marginal likelihood of the data (that is, the denominator of Bayes' rule).

Answer:
The marginal likelihood of the data, using the law of total probability, is:

$$Pr(\{H,H\},\{H,T\}) = \sum_{\theta_1\in 0.5,0.9}\sum_{\theta_2\in 0.5,0.9} Pr(\{H,H\},\{H,T\},\theta_1,\theta_2) = \sum_{\theta_1\in 0.5,0.9}\sum_{\theta_2\in 0.5,0.9} Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2)Pr(\theta_1,\theta_2) =$$

$$\sum_{\theta_1\in 0.5,0.9}\sum_{\theta_2\in 0.5,0.9} Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2)Pr(\theta_1)Pr(\theta_2) = \sum_{\theta_1\in 0.5,0.9} Pr(\{H,H\}|\theta_1)Pr(\theta_1)\sum_{\theta_2\in 0.5,0.9} Pr(\{H,T\}|\theta_2)Pr(\theta_2) =$$

$$(Pr(\{H,H\}|\theta_1 = 0.5)Pr(\theta_1 = 0.5) + Pr(\{H,H\}|\theta_1 = 0.9)Pr(\theta_1 = 0.9) \cdot (Pr(\{H,T\}|\theta_2 = 0.5)Pr(\theta_2 = 0.5)+$$

$$Pr(\{H,T\}|\theta_2 = 0.9)Pr(\theta_2 = 0.9))$$

Assuming a uniform prior for two independent coin throws gives us

$$Pr(\{H,H\},\{H,T\}) = \sum_{\theta_1\in 0.5,0.9} \theta_1^2 Pr(\theta_1)\sum_{\theta_2\in 0.5,0.9} \theta_2(1-\theta_2)Pr(\theta_2) =$$

$$(0.5^2\cdot 0.5^2 + 0.9^2\cdot 0.5^2)(0.5(1-0.5)\cdot 0.5^2 + 0.9(1-0.9)\cdot 0.5^2) = 0.5^2(0.5^2+0.9^2)(0.5(1-0.5)+0.9(1-0.9)) =$$

$$0.25(0.25+0.81)(0.25+0.09) = 0.0901$$

Problem 6.1.6.

Hence calculate the posterior distribution, and demonstrate that this is a valid probability distribution.

Answer:

$$Pr(\theta_1,\theta_2|\{H,H\},\{H,T\}) = \frac{Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2)Pr(\theta_1,\theta_2)}{Pr(\{H,H\},\{H,T\})} = \frac{Pr(\{H,H\},\{H,T\}|\theta_1,\theta_2)Pr(\theta_1)Pr(\theta_2)}{Pr(\{H,H\},\{H,T\})} \quad (4)$$

Now, we have everything we need to calculate (5). Again assuming a uniform prior for two independent coin throws.

$$Pr(\theta_1, \theta_2|\{H,H\},\{H,T\}) = \frac{\theta_1^2\theta_2(1-\theta_2)\cdot\frac{1}{4}}{0.0901} \tag{5}$$

We can show that this is a valid probability distribution by summing over the joint distribution for all values of θ_1 and θ_2:

$$\sum_{\theta_1\in 0.5,0.9}\sum_{\theta_2\in 0.5,0.9}\frac{\theta_1^2\theta_2(1-\theta_2)\cdot\frac{1}{4}}{0.0901} = \frac{2500}{901}\sum_{\theta_1\in 0.5,0.9}\theta_1^2\sum_{\theta_2\in 0.5,0.9}\theta_2(1-\theta_2) = \frac{2500}{901}(0.25+0.81)\cdot(0.25+0.09) = 1$$

Problem 6.1.7.

Find the posterior mean of θ_1. What does this signify.

Answer:

$$\mathbb{E}[\theta_1|\{H,H\}] = \sum_{\theta_1\in 0.5,0.9}\theta_1\cdot\frac{\theta_1^2}{0.5^2+0.9^2} \approx 0.81$$

Therefore, there is a greater mass towards $\theta_1 = 0.9$ meaning the biased coin.

Problem 6.1.7.

Problem 6.1.8. Now suppose that away from our view a third coin is flipped, and denote Z = 1 for a heads. The result of this coin affects the bias of the other two coins that are flipped subsequently so that,

$$Pr(\theta_i = 0.5|Z) = 0.8^Z 0.1^{1-Z} \tag{6}$$

Suppose we again obtain for coin 1 HH and coin 2 HT. Find the maximum likelihood estimators (θ_1, θ_2, Z). How do the inferred biases of coin 1 and coin 2 compare to the previous estimates?

Answer: First, note that as the bias of a fair coin is affected by the outcome of the third coin via (6), the effect on a coin biased towards head, i.e with $\theta = 0.9$ should be:

$$Pr(\theta_i = 0.9|Z) = 1 - Pr(\theta_i = 0.5|Z) = 1 - 0.8^Z 0.1^{1-Z} \tag{7}$$

Next, we rewrite the likelihood in terms that we can compute.

$$Pr(\{H,H\}_1,\{H,T\}_2,\theta_1,\theta_2|Z) = Pr(\{H,H\}_1,\theta_1|Z)Pr(\{H,T\}_2,\theta_2|Z) =$$

$$Pr(\{H,H\}_1|Z,\theta_1)Pr(\theta_1|Z)Pr(\{H,T\}_1|Z,\theta_2)Pr(\theta_2|Z)$$

Given θ the data is conditionally independent of Z. Thus, we finally get the likelihood as:

$$Pr(\{H,H\}_1,\{H,T\}_2,\theta_1,\theta_2|Z) = Pr(\{H,H\}_1|\theta_1)Pr(\theta_1|Z)Pr(\{H,T\}_1|\theta_2)Pr(\theta_2|Z)$$

Now we have 3 parameters with 2 possible values, thus we have $2^3 = 8$ possible parameter combinations. The likelihood for each combination can be found in Table 1.

θ_1	θ_2	Z	$\mathcal{L}$
0.5	0.5	0	$0.5^2\cdot 0.5^2\cdot 0.1^2 = 0.000625$
0.5	0.5	1	$0.5^2\cdot 0.5^2\cdot 0.8^2 = 0.04$
0.5	0.9	0	$0.5^2\cdot 0.9\cdot 0.1\cdot 0.1\cdot 0.9 = 0.002025$
0.5	0.9	1	$0.5^2\cdot 0.9\cdot 0.1\cdot 0.8\cdot 0.2 = 0.0036$
0.9	0.5	0	$0.9^2\cdot 0.5^2\cdot 0.1\cdot 0.9 = 0.018225$
0.9	0.5	1	0.0325
0.9	0.9	0	0.059049
0.9	0.9	1	0.002916

Table 1: Likelihood for different parameter combinations for θ_1, θ_2 and Z.

Problem 6.1.9.

Calculate the marginal likelihood for the coin if we suppose that we specify a discrete uniform prior on Z, i.e. $Pr(Z = 1) = 0.5$.

Answer: The marginal likelihood (an invalid probability distribution) is given by: $p(data) = p(\{H, H\}_1, \{H, T\}_2)$. We can obtain this by marginalizing out (summing out) all the parameters that the data is dependent on with the following general formula:

$$Pr(data) = \sum_{All\theta} = Pr(data|\theta)Pr(\theta) \tag{8}$$

Note that as $Pr(Z = 1) = 0.5$ the probability of heads becomes $Pr(Z = 0) = 0.5$. Therefore, we get the discrete marginal probability by multiplying the likelihood column in table 1 by 0.5 and then adding up these values. This gives us 0.0794.

Problem 6.1.10.

Suppose we believe that the independent coin flip model (where there is no third coin) and the dependent coin flip model (where the outcome of the third coin affects the biases of the two coins) are equally likely a priori. Which of the two models do we prefer?

Answer: According to Murphy [1, p. 165] model selection is equivalent to picking the model with the highest marginal likelihood. We select a model by computing the Bayes factor as the ratio of the marginal likelihoods:

$$BF_{1,0} \triangleq \frac{p(D|M_1)}{p(D|M_0)} / \frac{p(M_1)}{p(M_0)} = \frac{p(D|M_1)}{p(D|M_0)} / 1 = \frac{p(D|M_1)}{p(D|M_0)} = \frac{0.0901}{0.0794} \approx 1.13 \tag{9}$$

Since the Bayes factor is larger than 1, but greater than 3, the Jeffrey's scale tells us that this is weak evidence for M_1 [1, p. 165].

6.2 Coins combined

Suppose that we flip two coins, each of which has $Pr(H) = \theta_i$ where $i \in \{1, 2\}$, which is unknown. If their outcome is both the same then we regard this as a success; otherwise a failure. We repeatedly flip both coins (a single trial) and record whether the outcome is a success or failure. We do not record the result of flipping each coin. Suppose we model the number of failures, X, we have to undergo to attain n successes.

Problem 6.2.1.

Stating any assumptions that you make specify a suitable probability model here.

Answer:

- Since θ is unknown and we record the outcomes of the coin flips these events are dependent. Meaning that when we do not know θ, knowing something about one outcome tells us something about another outcome.
- We are dealing with counts of discrete events, i.e. counts of successes and failures.

These two assumptions fit with the negative binomial distribution. Which has the following pmf:

$$Pr(X = y|n, \theta) = \binom{n + y - 1}{n - 1} (1 - \theta)^y \theta^n \tag{10}$$

Since we are flipping two coins we get a success if both coins land on either heads or tails. The probability of success is thus given by one of these two events:

$$p_{\text{success}} = \theta_1\theta_2 + (1 - \theta_1)(1 - \theta_2) \tag{11}$$

Where the first term corresponds to both coins displaying heads and the second to both showing tails. Using this in (12) gives us:

$$Pr(X = y|n, \theta_1, \theta_2) = \binom{n+y-1}{n-1} (1 - (\theta_1\theta_2 + (1-\theta_1)(1-\theta_2)))^y \cdot (\theta_1\theta_2 + (1-\theta_1)(1-\theta_2))^n \tag{12}$$

Note that even though the outcomes of the model are dependent (successes and failures), the coin flips of coin 1 and 2 are independent.

Problem 6.2.2.

We obtain the data in *denominator_NBCoins.csv* for the number of failures to wait before 5 successes occur. Suppose that we specify the following priors $\theta_1 \sim U(0,1)$ and $\theta_2 \sim U(0,1)$. Calculate the denominator of Bayes' rule. (Hint: use a numerical integration routine.)

Answer: The denominator is obtained by marginalizing out the parameters in the likelihood (12). By computing:

$$\int_0^1 \int_0^1 \binom{n+y-1}{n-1} (1 - (\theta_1\theta_2 + (1-\theta_1)(1-\theta_2)))^y \cdot (\theta_1\theta_2 + (1-\theta_1)(1-\theta_2))^n d\theta_1 d\theta_2 \approx 2.48731 \cdot 10^{-170} \tag{13}$$

Problem 6.2.3.

Draw a contour plot of the posterior. Why does the posterior have this shape?

Answer: Since we are only interested in the shape of the posterior we could disregard its numerator, since the numerator only affects the height and not the shape of the posterior(p117). Also, since we have a uniform prior the shape is completely determined by the likelihood(p93). The prior divided by the denominator simply becomes a constant that we multiply with our likelihood.

In the code the log-likelihood and log-posterior is used to avoid underflow.

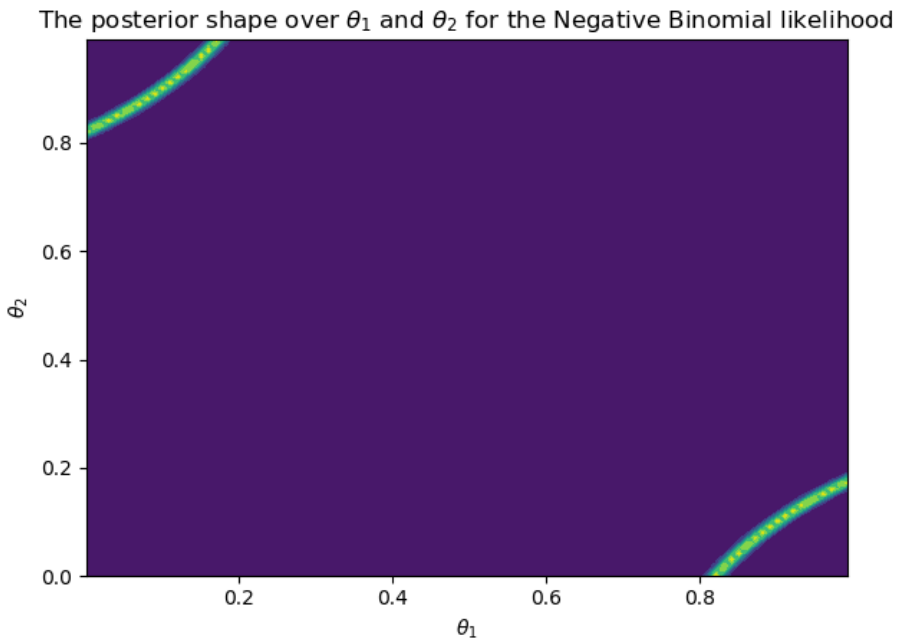

Figure 1: The log-likelihood near the MLE estimate.

The contour plot of the posterior is shown in Figure 1. The probability mass consists of thin band and are associated with the functions $\theta_1\theta_2$ and $(1-\theta_1)(1-\theta_2)$. The values in the data are quite large, meaning that there are relatively many failures before we achieve 5 successes. Therefore, it is more likely to get different values for the two coins. This is seen in the plot as the probability mass mainly is located in the corners where one parameter has a value close to 1 and the other close to 0.

Problem 6.2.4.

Comment on any issues with parameter identification for this model and how this might be rectified.

Answer: One issue is that we are using a uniform prior. An alternative prior could have been one that has a high weight for the lower and higher theta values. Another alternative could be to use this posterior as a prior and then collect more data.

References

[1] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.