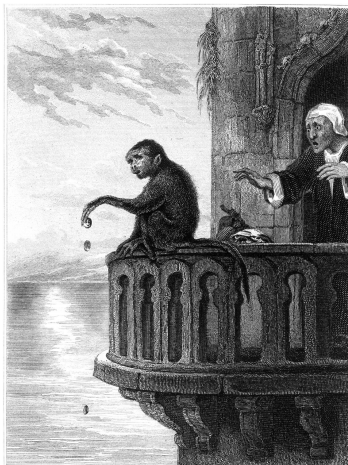# Four Acts on Bayesian Inference

Justin Domke

Fall 2017

1 Act 1: Magical Monkeys

Two monkeys, **Alfred** (a) and **Betty** (b) live in a parallel universe with two kinds of blocks, **green** (g) and **yellow** (y). Alfred likes green blocks, and Betty prefers the yellow blocks. One day, a Wizard decides to give one of the monkeys the magical power to send one block over to our universe each day.

The Wizard chooses the magical monkey by flipping a fair four-sided die. He casts a spell on Alfred if the outcome is 1, and Betty if the outcome is $2-4$. That is, the Wizard chooses Alfred with probability .25 and Betty with probability .75. Both Alfred and Betty send their favorite colored block with probability .8

After the Wizard has chosen, we see the first magical block, and it is **green**. Our problem is: <u>What is the probability that Alfred is the magical monkey?</u>

Intuitively speaking, we have two somewhat contradictory pieces of information. We know that the Wizard chooses Betty more often. But green is Alfred's preferred color. Given probabilities above, is Alfred or Betty more probable?

First, we can write down all the probabilities. These are

$$Pr(M = a) = .25$$
$$Pr(M = b) = .75$$
$$Pr(B = g|M = a) = .8$$
$$Pr(B = y|M = a) = .2$$
$$Pr(B = g|M = b) = .2$$
$$Pr(B = y|M = b) = .8.$$

Now, the quantity we are interested in is $Pr(M = a|B = g)$. We can mechanically apply Bayes' rule to obtain

$$Pr(M = a|B = g) = \frac{1}{Pr(B = g)}Pr(M = a)Pr(B = g|M = a)$$

$$= \frac{1}{Pr(B = g)}.25 \times .8.$$

Similarly, we can calculate that

$$Pr(M = b|B = g) = \frac{1}{Pr(B = g)}.75 \times .2.$$

But, of course, we know that one of the monkeys is magical, so these quantities must sum to one. Thus (since they both involve the quantity $Pr(B = g)$ that we haven't explicitly calculated) we can normalize to obtain that

$$Pr(M = a|B = g) = \frac{.25 \times .8}{.25 \times .8 + .75 \times .2} = \frac{4}{7}.$$

2   Act 2: Magical Monkeys on Multiple Days

We wait around two more days, and two more blocks appear, both of which are yellow. Assume that the monkeys make **an independent** choice each day to send their favorite or less favorite block. Now, what is the probability that Alfred is the magical monkey, given that $B = gyy$?

Intuitively speaking, what do we expect? Betty is more likely to be chosen, and 2/3 of the blocks we've seen are suggestive of Betty (since she prefers yellow).

Now, we can mechanically calculate the probabilities. This is just like before, except we use the fact that the blocks seen on each day are conditionally independent given a particular monkey.

$$Pr(M = a|B = gyy) = \frac{1}{Pr(B = gyy)}Pr(M = a)Pr(B = gyy|M = a)$$

$$= \frac{1}{Pr(B = gyy)}Pr(M = a)Pr(B = g|M = a)Pr(B = y|M = a)Pr(B = y|M = a)$$

$$= \frac{1}{Pr(B = gyy)}.25 \times .8 \times .2 \times .2$$

$$Pr(M = b|B = gyy) = \frac{1}{Pr(B = gyy)}Pr(M = b)Pr(B = g|M = b)Pr(B = y|M = b)Pr(B = y|M = b)$$

$$= \frac{1}{Pr(B = gyy)}.75 \times .2 \times .8 \times .8$$

Again, we know that these two probabilities sum to one. Thus, we can normalize and get that

$$Pr(M = a|B = gyy) = \frac{.25 \times .8 \times .2 \times .2}{.25 \times .8 \times .2 \times .2 + .75 \times .2 \times .8 \times .8} = \frac{1}{13}.$$

So now, Betty looks more likely by far.

3   Act 3: Magical Monkeys and the Weather

Now, suppose the Wizard gives us some additional information. The magical monkey (whilst relaxing over in the other universe) is able to perceive our weather, which is either **clear** (c) or **rainy** (r). Both the monkeys prefer clear weather. When the weather is clear, they send their preferred block with

probability .8, while if the weather is rainy, they angrily send their preferred block with probability .2. That is, we have

$$Pr(B = g|M = a, W = c) = .8$$
$$Pr(B = g|M = a, W = r) = .2$$
$$Pr(B = g|M = b, W = c) = .2$$
$$Pr(B = g|M = b, W = r) = .8.$$

Along with seeing the previous sequence of blocks $B = gyy$, we observed the weather sequence $W = rrc$. Now, what is the probability that Alfred is the magical monkey?

We ask the Wizard what the distribution over rainy and clear weather is. The wizard haughtily responds that this is **irrelevant**, but does confirm that the weather is independent of which monkey was made magical.

Can we do anything without knowing the distribution over the weather? Can we calculate the probability that Alfred is the magical monkey?

Let's give it a try. We can apply Bayes' equation to get that

$$Pr(M = a|B = gyy, W = rrc) = \frac{1}{Pr(B = gyy, W = rrc)} Pr(M = a) Pr(B = gyy, W = rrc|M = a).$$

Now, since the weather is independent of the monkey, we know that

$$Pr(B = gyy, W = rrc|M = a) = Pr(W = rrc|M = a) Pr(B = gyy|M = a, W = rrc)$$
$$= Pr(W = rrc) Pr(B = gyy|M = a, W = rrc).$$

Thus, we have that

$$Pr(M = a|B = gyy, W = rrc) = \frac{Pr(W = rrc)}{Pr(B = gyy, W = rrc)} Pr(M = a) Pr(B = gyy|M = a, W = rrc)$$
$$= \frac{Pr(W = rrc)}{Pr(B = gyy, W = rrc)} Pr(M = a) Pr(B = g|M = a, W = r)$$
$$\times Pr(B = y|M = a, W = r) Pr(B = y|M = a, W = c)$$
$$= \frac{Pr(W = rrc)}{Pr(B = gyy, W = rrc)} .25 \times .2 \times .8 \times .2.$$

Through the same logic we can calculate that

$$Pr(M = b|B = gyy, W = rrc) = \frac{Pr(W = rrc)}{Pr(B = gyy, W = rrc)} .75 \times .8 \times .2 \times .8.$$

Again, we know that these probabilities need to sum to one. Since the factor $Pr(W = rrc)/Pr(B = gyy, W = rrc)$ is constant between the two, we can normalize it out. Thus, we get that

$$Pr(M = a|B = gyy, W = rrc) = \frac{.25 \times .2 \times .8 \times .2.}{.25 \times .2 \times .8 \times .2. + .75 \times .8 \times .2 \times .8.} = \frac{1}{13}.$$

Again, it looks like Alfred was the more likely monkey. And– oh wait– we somehow got away with not knowing the distribution over the weather...

4   Act 4: Predicting the next block

Now, suppose that after seeing the previous sequences (namely $B = gyy$ and $W = rrc$) , on the fourth day, we find that it has rained. What is the probability that we will get a green block on the fourth day? Mathematically, we want to calculate

$$Pr(B' = g|W' = r, B = gyy, W = rrc).$$

How can we go about this? Well, **if** we knew that the magical monkey was Alfred (which we do not!), it would be easy to calculate. We would just have

$$Pr(B' = g|W' = r, M = a),$$

and similarly if the magical monkey were Betty. Now, we don't know which monkey is magical, but we know the **probabilities** that each monkeys is magical given the avilable information– we just calculated them in the previous section! So, we can factorize the distribution of interest as

$$Pr(B' = g|W' = r, B = ggy, W = rrc) = Pr(B' = g|W' = r, M = a)Pr(M = a|B = gyy, W = rrc)$$
$$+ Pr(B' = g|W' = r, M = b)Pr(M = b|B = gyy, W = rrc)$$
$$= .2 \times \frac{1}{13} + .8 \times \frac{12}{13}$$
$$= \frac{2}{130} + \frac{96}{130}$$
$$= \frac{3}{4}$$

So the next block is likely to be green, given that it's a rainy day. (This makes sense, since the monkey is probably Betty, and Betty most likely sends a green block on a rainy day.)

5 Epilogue: That's Bayesian Inference

This is how Bayesian inference works (in the simplest possible setting). In general Bayesian inference, you have:

- A set of possible inputs (for us these are weather patterns)

- A set of possible outputs (for us these are colored blocks)

- A set of possible models, that map a given input to a distribution over the outputs (for us these are monkeys)

- A "prior" distribution over which models are more likely (for us, this is the wizard and the four-sided die)

- A dataset of inputs and outputs (for us, this is $W = rrc$ and $B = gyy$)

- A new input (for us this is $W' = r$)

- An unknown output that we'd like to predict (for us this is B')

Bayesian inference essentially proceeds in two steps:

- In the first step, one uses the prior over which models are more likely along with the dataset of inputs and outputs to build a "posterior" distribution over which models are more likely. We did this by using

$$Pr(M|W, B) \propto Pr(M) \prod_i Pr(B_i|W_i, M),$$

  where i indexes the different days of observation.

- In the second step, one "integrates over the posterior". That is, each model gives us a distribution over B' given the new input W'. So, we simply sum over the different models

$$Pr(B'|W') = \sum_m Pr(M = m|W, B)Pr(B'|W', M = m).$$

In the general case, the set of possible models is much larger (typically infinite) and more complex computational methods need to be used to integrate over it. (Commonly Markov chain Monte Carlo or variational methods). Also, of course, we don't typically have a Wizard telling us exactly what the prior distribution over the models is, meaning one must make assumptions, or try to "learn" the prior as in, say, empirical Bayesian methods.