Week 5 Recap

In week 5, we began by continuing our probability lecture from last week, and then, as an exercise, tried to prove the correct answer to the Monty Hall problem using Bayes Rule.

Monty Redux

Here's that solution again. Remember our formula for Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Here, all the work is in figuring out what A and B are, and then plugging in. As a guide to figuring out A and B, remember that conceptually what we're trying to do is update our estimate of the likelihood that the car is behind the door we already picked, given our new evidence (that Monty opened a specific other door). For convenience, we can label the door we picked as Q, the door Monty opened as R, and the door that was neither picked nor opened as Z, and then what we're after is the conditional probability that the car is behind Q given that Monty opened R. Then it makes the most sense to formalize B as car is behind Q, and A as Monty opened R.

Let's start with the numerator of our fraction on the right side of the formula. What's $P(A|B)$? Well, if the car is behind Q, then Monty has a choice between opening door R or door S, and assuming he doesn't want to give away any more free information about where the car is, we have to assume that he picks one of the two at random. So $P(A|B) = \frac{1}{2}$. And $P(B)$ is just our prior on door Q having the car, which is $\frac{1}{3}$. (I'm using fractions here because they're all really simple fractions and it makes the arithmetic easier at the end, but these are still probabilities, not odds.)

As usual, the denominator takes a bit more work to figure out. What's the unconditional probability of Monty opening door R? Well, we need to use the law of total probability again. Since the only three states of the universe are car-behind-Q, car-behind-R, and car-behind-S, we need to calculate

$$P(R{-}open|car{-}Q)P(car{-}Q){+}P(R{-}open|car{-}R)P(car{-}R){+}P(R{-}open|car{-}S)P(car{-}S)$$

- We know that the second term in each of those multiplications is $\frac{1}{3}$ because the prior on each door having the car is a third.
- We have already concluded that $P(R - open|car - Q) = \frac{1}{2}$ — remember, that's just the evidence we have, we already used it in the numerator.

- We know that Monty can't open the door with the car behind it, so we know that $P(R - open|car - R) = 0$.

The third term is a little difficult to figure out—and this, I suspect, is where all the mathematicians who wrote angrily to Marilyn vos Savant went wrong. Here's a mistake you might make: since all Monty has to do is open a door without the car behind it, and neither Q nor S has the car behind it, $P(R-open|car-S) = \frac{1}{2}$. If you think that, you'll end up with a denominator of $\frac{1}{3}$ and an ultimate, incorrect but intuitive, calculation of $P(B|A) = \frac{1}{2}$. But this would be wrong. The thing you need to remember is that we've been writing out the problem in shorthand. Monty's decision as to which door to open was made not only with knowledge of where the car is, but also with knowledge of which door the player picked. So the longhand version of the calculation for our denominator is actually:

$$P(R-open|car-Q, picked-Q)P(car-Q)+P(R-open|car-R, picked-Q)P(car-R)+P(R-open|car-S, picke$$

And the other rule that we can't forget is that Monty also can't open a door that the player picked. So the player picked Q—if the car is behind S, the only door Monty can open is R. $P(R - open|car - S, picked - Q) = 1$.

This gives us the denominator:

$$\frac{1}{2}\frac{1}{3} + 0 + \frac{1}{3}$$

which we can easily simplify:

$$\frac{1}{6} + \frac{2}{6} = \frac{1}{2}$$

meaning that our overall Bayes Rule formula is:

$$\frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}}$$

simplify by canceling the $\frac{1}{2}$ -s and you get the correct answer.

We learned that most of the work in applying Bayes Rule is figuring out how to define the events for which we're trying to make probability judgments—how to formalize our idea of what we started out knowing, what we've learned (our evidence), and what we're trying to figure out.

Incidentally, for a discussion of how lawyers appear to be susceptible to screwing up Bayes Rule, an article called Miss Rate Neglect in Legal Evidence reports on a number of experiements with legally trained subjects where they do things like inappropriately flip around conditional probabilities.

Scavenger hunt

Our second major task this week was to do a Data Scavenger Hunt. Things started a little bumpy, because lots of people had trouble getting the data loaded, mainly due to the diversity of platforms in the class. (And also because I went around to help windows users, and didn't notice that many were actually using the Azure version, and hence didn't have their local downloads... sorry about that.)

We're going to finish this on Monday, and then I'll post some solutions after we're done.

I have also fixed the ergonomics of getting data. There's a private web server for this class at https://gobbledygook.herokuapp.com/. I'll give you a password for it in class (and will also post it on ICON). Whenever you need to load a dataset, you will just be able to use Pandas to load it directly from this server. Here's how. Assume the password is PASSWORD and the file you're trying to load is FILENAME.CSV. Then if you have Pandas imported as `pd`, the code
`df=pd.read_csv("https://gobbledygook.herokuapp.com/data?file=FILENAME.CSV&password=PASSWORD'`
will save the dataset as a Pandas dataframe to the variable `df`. No more mess with Windows file paths and local downloads vs. Azure notebooks or downloading files from ICON or any of that nonsense.

We'll be using a variety of datasets in the rest of this class, so get used to that command. Also, the server might start up slowly sometimes, so if it looks like it's hanging when you try to load a dataset, just wait 30 seconds or so and it'll be fine.

(Incidentally, this method of authenticating users by passing passwords in GET request parameters is a terrible idea, it's totally insecure. Don't do it for anything you ever build. I'm just doing it here because we don't need real security, just minimal privacy against casual misuse. And also because I'm going to tear down the server the moment class is over.)