Lecture 9: Conditional Probability

February 28, 2019

1 How to use math to talk your way out of a failed drug test (if you're old).

According to US government data from 2016, 3% of Americans aged 65 and above reported using marijuana in the past year.

Laboratory testing of an over-the-counter test strip for detecting marijuana use reports that 96% of those identified as users by the gold-standard laboratory test (gas chromatography/mass spectrometry) tested positive on the OTC test, while 97% of those testing negative on the gold standard also tested negative on the OTC test. In other words, assuming that the lab test is completely accurate, the over-the-counter test has a 3% false positive rate, and a 4% false negative rate.[1]

Let's imagine that we randomly sample 1000 people from the 65-and-above age group. According to the data cited, approximately 30 of them have used marijuana in the past year, and 970 have not. Of the 30 users, 96%, or about 29, test positive. Of the 970 non-users, 3%, also about 29, test positive. In all, half the positive results are false positives!

So, if 'it's legal in my state', or 'I have glaucoma' aren't working for you, try 'there's a very good chance that the test is wrong'.

(This line of argument is less persuasive in a different age group: In the 19-21 age range, 23% report marijuana use in the past *month*, and if you repeat the above calculation you'll find that the proportion of false positives far smaller.)

[1] Source: https://www.drugtestsuccess.com/drug-tests/specifications-thc-marijuana.

2 I have two children...

This is an old puzzle. An rather annoying math professor, asked whether he has any children, instead of giving a straight answer, says, 'I have two children, and one of them is a girl.' What, the puzzle asks, is the probability that the *other child* is a girl?

Here is a superficially similar problem, with a completely different answer: The professor says, 'I have two children, and the older child is a girl.' What is the probability that the other child is a girl?

Here are coin-tossing versions of both these puzzles. In the first, I toss two coins, and cover them up so that you cannot see the result (but I can). If they are both tails, I lift the cover to reveal this fact, and we move on to the next game. If, alternatively, at least one of the coins came up heads, I uncover a coin with heads to show you this fact, and invite you to bet if the other one is heads.

Alternatively, I ask *you* to lift the cover on one of the coins. If it is tails, we move on to the next game. If it is heads, you bet on whether the other one is heads.

In the second children puzzle, the professor is just asking you if the younger child is a girl; in the second coin puzzle, you are simply asked to guess whether the hidden coin is heads. The probability is $\frac{1}{2}$.

It is tempting to argue that the same is true for the first puzzle in each group, but it's more complicated than that: In the case of coins, there are four equally likely outcomes to the toss: HH,HT,TH,TT. We throw out the outcomes of the form TT and only play the game if one of the first three outcomes occurs. Of the three remaining outcomes, only one of them has both coins coming up heads. So the answer is $\frac{1}{3}$. The analysis is the same for the first children puzzle: We are limited to the three equally likely outcomes GG, GB, BG, and in only one of the three is the other child a girl.

3 Conditional Probability.

Let E and F be events in a probability space. $P(E|F)$ denotes the *conditional probability* of E conditioned on F. What this means, roughly, is what proportion of the times that F occurs, does E also occur? Formally, this is defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

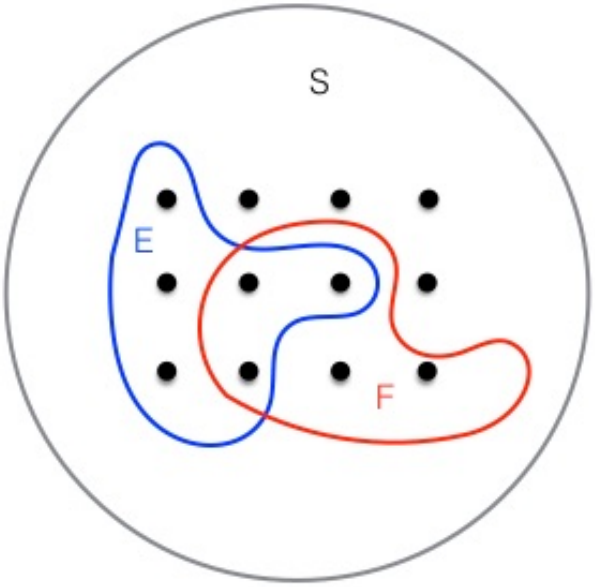

Figure 1: Illustrating conditional probability

4 Examples.

- Figure 1 graphically illustrates the definition. Imagine that the dots represent outcomes each with probability $\frac{1}{12}$. Then $P(E) = \frac{1}{2}$, $P(F) = \frac{5}{12}$, and $P(E \cap F) = \frac{3}{12}$. Thus $P(E|F) = \frac{3}{5}$, and $P(F|E) = \frac{1}{2}$.

- Consider the roll of a single fair die. Let F be the event 'the number showing is even', and E_1, E_2 the events 'the number showing is 1', and 'the number showing is 2', respectively. Here $F = \{2, 4, 6\}$, $E_1 = \{1\}$, $E_2 = \{2\}$, $E_1 \cap F = \emptyset$, and $E_2 \cap F = \{2\}$. Then $P(E_1|F) = 0$, while $P(E_2|F) = \frac{1}{3}$.

- Let's redo the first of the two coins puzzles. The sample space consists of the four equally likely outcomes HH,HT,TH,TT. We are asking, what is the probability that both coins are heads, given that at least one of them is heads? That is, we are asking for $P(E|F)$ where E is the event 'both coins come up heads', and F is the event 'at least one coin comes up heads'. As sets, $F = \{HH, HT, TH\}$ so $P(F) = \frac{3}{4}$. In this case, $E \cap F = E =$

$\{HH\}$, so $P(E) = \frac{1}{4}$. Thus

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{1}{4}/\frac{3}{4} = \frac{1}{3}.$$

We can also work the other coin puzzle this way. If F' is the event 'the first coin comes up heads', then $F' = \{HH, HT\}$, and when we repeat the calculation, we get $P(E|F') = \frac{1}{2}$.

- (Connection with independence.) If E, F are independent, then $P(E \cap F) = P(E) \cdot P(F)$, so it follows that $P(E|F) = P(E)$. Conversely, if $P(E|F) = P(E)$, it follows from the definition that $P(E \cap F) = P(E) \cdot P(F)$. So we can characterize independence this way in terms of conditional probability. Because of the symmetry in the problem, this also implies $P(F|E) = P(F)$.

- (Chain rule for conditional probability.) If we consider the intersection of three events, and apply the definition twice, we get

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3) &= P(E_1|E_2 \cap E_3) \cdot P(E_2 \cap E_3) \\ &= P(E_1|E_2 \cap E_3) \cdot P(E_2|E_3) \cdot P(E_3), \end{aligned}$$

and similarly for any number of events.

- Here is a somewhat more involved example. A wholesaler of light bulbs supplies bulbs from three different manufacturers: A, B, and C. 60% of the bulbs in stock are from A, 30% from B, and 10% from C. Approximately 4% of the light bulbs from manufacturer A are defective, in the sense that they burn out before the 'guaranteed lifetime'. Similarly, 2% of light bulbs from manufacturer B are defective, and 1% from manufacturer C are defective.

  You receive a random light bulb from the wholesaler. What is the probability that it is defective?

  The underlying experiment is the selection of a random light bulb supplied by the manufacturer. The associated events are

  E: The bulb is defective. A: The bulb comes from manufacturer A. B: The bulb comes from manufacturer B. C: The bulb comes from manufacturer C. The conditions given in the problem are

$$P(A) = 0.6, \quad P(B) = 0.3, \quad P(C) = 0.1.$$

$$P(E|A) = 0.04, \quad P(E|B) = 0.02, \quad P(E|C) = 0.01.$$

The problem is to determine $P(E)$. We can write the sample space as $S = A \cup B \cup C$, and this is a pairwise disjoint union. We then have

$$\begin{aligned} P(E) &= P(E \cap S) \\ &= P(E \cap (A \cup B \cup C) \\ &= P((E \cap A) \cup (E \cap B) \cup (E \cap C)) \\ &= P(E \cap A) + P(E \cap B) + P(E \cap C) \\ &= P(E|A) \cdot P(A) + P(E|B) \cdot P(B) + P(E|C) \cdot P(C) \\ &= 0.04 \cdot 0.6 + 0.02 \cdot 0.3 + 0.01 \cdot 0.1 \\ &= 0.031 \end{aligned}$$

Calculations of this kind are sometimes referred to as applications of the *Law of Total Probability*: We reconstruct the probability of E from the conditional probabilities of E relative to a collection of mutually exclusive conditions.

5 Bayes's Theorem

The definition implies

$$P(E|F) \cdot P(F) = P(E \cap F) = P(F|E) \cdot P(E),$$

which we can rewrite as

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

This is called *Bayes's Theorem.* ('Theorems' should have proofs, but that was the proof! It's just a trivial manipulation of the definition.)

Examples.

- Let's redo our drug test example. We imagine an experiment that consists of picking a member of the target population at random, and performing the drug test. We define events U:'the person is a user, N: the person is a nonuser, and $+$: 'the subject tests positive'. We have, from the data in the problem

$$P(U) = 0.03, P(N) = 0.97, P(+|U) = 0.96, P(+|N) = 0.03.$$

Bayes's Theorem gives

$$\begin{aligned} P(U|+) &= \frac{P(+|U) \cdot P(U)}{P(+)} \\ &= \frac{P(+|U) \cdot P(U)}{P(+|U) \cdot P(U) + P(+|N) \cdot N} \\ &= \frac{0.96 \times 0.03}{0.96 \times 0.03 + 0.03 \times 0.97} \\ &= 0.497. \end{aligned}$$

This is the probability that someone who tests positive is a user. Thus slightly more than half of the positive tests are false positives, as we found above.

- In the late 1990's, the distribution of colors in milk chocolate M&M's was 10% each blue,orange, green; 20% each red and yellow, and 30% brown. In 2008, the distribution was changed drastically to 24% blue, 20% orange, 16% green, 14% yellow, and 13% each red and brown. [2] You are employed by an eccentric billionaire who keeps a large number giant glass urns, each containing tens of thousands of M&M's, in the trophy room of his mansion. He knows that about 10% of these urns were filled with shipments of the vintage 1990's M&M's, while the remaining were filled from the more recent distribution. Your job is to determine which of the two classes a given urn belongs to. You begin by selecting an urn at random, reaching in, and pulling out 10 M&M's, of which 5 are brown, 4 are red, and one is orange. That looks more like the 1990's distribution, but it's a pretty small sample. What is the probability that this is a 1990's urn?

  We use Bayes's Theorem for this. We begin by defining the underlying events: V stands for 'vintage urn', M for 'modern urn', and E for the event of pulling out 5 brown, 4 red, and 1 orange. The problem is to determine $P(V|E)$. Our 'prior', to use Bayesian talk, is $P(V) = 0.1$, and consequently $P(M) = 0.9$. We want to know how this probability changes as the result of the outcome of the sampling experiment. $P(E|V)$ is given by the

[2] I did not make these numbers up! See https://blogs.sas.com/content/iml/2017/02/20/proportion-of-colors-mandms.html for more than you even wanted to know about this subject.

expression

$$\binom{10}{5,4,1} \cdot 0.3^5 \cdot 0.2^4 \cdot 0.1.$$

Recall that the factor at the start of the expression is the multinomial coefficient giving the number of sequences of 5 B's, 4 R's and 1 O. This is given by the formula

$$\binom{10}{5,4,1} = \frac{10!}{5!4!1!}.$$

Likewise, $P(E|M)$ is

$$\binom{10}{5,4,1} \cdot 0.13^5 \cdot 0.13^4 \cdot 0.2.$$

When we apply Bayes's Theorem, the multinomial coefficients conveniently cancel out, and we are left with

$$\begin{aligned} P(V|E) &= \frac{0.3^5 \cdot 0.2^4 \cdot 0.1 \cdot 0.1}{0.3^5 \cdot 0.2^4 \cdot 0.1 \cdot 0.1 + 0.13^5 \cdot 0.13^4 \cdot 0.2 \cdot 0.9} \\ &= 0.953, \end{aligned}$$

so it's a very good bet that this is one of the vintage urns. Observe that the prior probability is a big part of the calculation–if only one in one hundred of the urns were from the vintage class, then $P(V|E)$ would drop to 0.65.

6 Naïve Bayes Classifier

This slightly silly problem about the M&Ms given above forms the basis for an important tool in machine learning. We want to determine which of the two classes, V and M, a given urn belongs to. We sample some M&M's and get some result E of the sampling experiment. The task is to find out which of $P(V|E)$ and $P(M|E)$ is larger. We have

$$P(V|E) = \frac{P(E|V) \cdot P(V)}{P(E}, \quad P(M|E) = \frac{P(E|M) \cdot P(M)}{P(E}.$$

Since the denominators are the same, we only have to compare

$$P(E|V) \cdot P(V), \quad P(E|M) \cdot P(M).$$

The same multinomial coefficient appears in both $P(E|V)$ and $P(E|M)$, so we can get rid of this as well, and simply compare the values

$$(p_{V,\text{blue}})^{\ell} \cdot (p_{V,\text{yellow}})^{y} \cdot (p_{V,\text{green}})^{g} \cdot (p_{V,\text{orange}})^{o} \cdot (p_{V,\text{red}})^{r} \cdot (p_{V,\text{brown}})^{b} \cdot P(V)$$

and

$$(p_{M,\text{blue}})^{\ell} \cdot (p_{M,\text{yellow}})^{y} \cdot (p_{M,\text{green}})^{g} \cdot (p_{M,\text{orange}})^{o} \cdot (p_{M,\text{red}})^{r} \cdot (p_{M,\text{brown}})^{b} \cdot P(M),$$

where, for example $p_{M,\text{yellow}}$ denotes the probability of picking a yellow M& M in the modern distribution (14% in our example), and ℓ, y, g, o, r, b the numbers of blue,yellow, green,orange, red and brown M&Ms in the sample.

This same classification method can be used for certain tasks in natural language processing. Suppose we want to determine whether a certain message is spam or not. We first train our classifier on a large number messages that have been classified by hand, and find the distribution of words in a large collection of spam messages, and likewise the distribution of words in a large number of messages that are not spam. (For example, in a dataset of both spam and non-spam SMS messages, the word 'won' was 100 times more likely to occur in a spam message than in a legitimate message.) These word distributions are then treated exactly like the color distributions for M&M's. If we are given a fresh document, D, we view it simply as a collection of words, and compute two scores for it, one relative to the spam distribution found during training, and the other relative to the non-spam distribution, and choose the class associated with the higher score.

What is 'naïve' about this method is that it ignores things like the occurrences of key phrases, or anything else having to do with the order of words in the document D, and instead treats the generation of a document as simply pulling a bunch of words out of a bag of words. In fact, this is called the 'bag of words' model in the machine learning literature. Yet it is surprisingly effective.

There are a few wrinkles that have to be ironed out to make this work: It is certain that some test messages will contain words that did not appear in any of the training documents. In that case, the model will assign the probability 0 to such a word. (This would be like finding that one of the urns contained a *gold* M& M, and which would give us a score of 0 for both classes.) So we need a little trick to get around this problem. The method is described in much more detail in the problem assignment, where you get to apply this technique.