EE 178/278A Probabilistic Systems Analysis

Spring 2014 Tse/Hussami Lecture 3

Conditional Probability[1]

A pharmaceutical company is marketing a new test for a certain medical disorder. According to clinical trials, the test has the following properties:

1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% (these are called "false negatives").
2. When applied to a healthy person, the test comes up negative in 80% of cases, and positive in 20% (these are called "false positives").

Suppose that the incidence of the disorder in the US population is 5%. When a random person is tested and the test comes up positive, what is the probability that the person actually has the disorder? (Note that this is presumably *not* the same as the simple probability that a random person has the disorder, which is just $\frac{1}{20}$.)

More specifically, the two questions we would like to answer in order to assess the reliability of the test are:

i- Given that the test is positive, what is the chance that the patient is healthy? (False positive rate)

ii- Given that the test is negative, what is the chance that the patient is affected? (False negative rate)

These are examples of *conditional probabilities*: for example, we are interested in the probability that a person has the disorder (event A) *given that* he/she tests positive (event B). Let's write this as $\mathbf{P}(A|B)$.

How should we define $\mathbf{P}(A|B)$? Well, since event B is guaranteed to happen, we should look not at the whole sample space Ω, but at the smaller sample space consisting only of the outcomes in B. What should the conditional probabilities of these outcomes be? If they all simply inherit their probabilities from Ω, then the sum of these probabilities will be $\sum_{\omega\in B}\mathbf{P}(\omega) = \mathbf{P}(B)$, which in general is less than 1. So we should *normalize* the probability of each outcome by $\frac{1}{\mathbf{P}(B)}$. In other words, for each outcome $\omega \in B$, the new probability becomes

$$\mathbf{P}(\omega|B) = \frac{\mathbf{P}(\omega)}{\mathbf{P}(B)}.$$

Now it is clear how to define $\mathbf{P}(A|B)$: namely, we just sum up these normalized probabilities over all outcomes that lie in both A and B:

$$\mathbf{P}(A|B) := \sum_{\omega\in A\cap B}\mathbf{P}(\omega|B) = \sum_{\omega\in A\cap B}\frac{\mathbf{P}(\omega)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}.$$

Definition 3.1 (conditional probability): For events A, B in the same probability space, such that $\mathbf{P}(B) > 0$, the *conditional probability of A given B* is

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}.$$

[1] Part of this note is adapted from the notes of EECS 70 at Berkeley.

Let's go back to our medical testing example. One can model the experiment as having 4 outcomes: the randomly chosen patient can be affected and test positive (outcome (affected, positive)), healthy and test positive (outcome (healthy, positive)), outcome (healthy, negative), and outcome (affected, negative)). The data for the problem do not directly specify the probabilities of these outcomes but instead are given in terms of conditional probabilities. Recall that A is the event that the patient is affected and B is the event that the test is positive. The data for the problem can be expressed in terms of conditional probabilities: $\mathbf{P}(B|A) = 0.9$ and $\mathbf{P}(B|A^c) = 0.2$. We also know that $\mathbf{P}(A) = 0.05$. From these the probabilities of the 4 outcomes can be calculated. For example:

$$\mathbf{P}((\text{affected}, \text{positive})) = \mathbf{P}(A \cap B) = \mathbf{P}(B|A)\mathbf{P}(A).$$

See table below.

Table 1: Outcome probabilities

	Positive Test	Negative Test
Affected	$0.9 \times 0.005 = 0.045$	$0.05 \times 0.1 = 0.005$
Healthy	$0.95 \times 0.2 = 0.19$	$0.95 \times 0.8 = 0.76$

From these, we can now compute the false positive and false negative rates:

$$\text{False positive rate} = \mathbf{P}(A^c|B) = \frac{\mathbf{P}(A^c \cap B)}{\mathbf{P}(B)} = \frac{0.19}{0.19 + 0.045} = 0.809$$

$$\text{False negative rate} = \mathbf{P}(A|B^c) = \frac{\mathbf{P}(A \cap B^c)}{\mathbf{P}(B^c)} = (\frac{0.05}{0.76 + 0.005} = 0.0065$$

The false positive rate is high because of the sheer number of people that are healthy (only 5% of the population is affected). This seems bad: if a person tests positive, there's only about a 19% chance that he/she actually has the disorder! This sounds worse than the original claims made by the pharmaceutical company, but in fact it's just another view of the data.

[Incidentally, note that $\mathbf{P}(B|A) = \frac{9}{10}$; so $\mathbf{P}(A|B)$ and $\mathbf{P}(B|A)$ can be very different. Of course, $\mathbf{P}(B|A)$ is just the probability that a person tests positive given that he/she has the disorder, which we knew from the start was 90%.]

To complete the picture, what's the (unconditional) probability that the test gives a correct result (positive or negative) when applied to a random person? Call this event C. Then

$$\mathbf{P}(C) = \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B^c) = 0.045 + 0.76 = 0.805$$

So the test is about 80% effective overall, a more impressive statistic.

But how impressive is it? Suppose we ignore the test and just pronounce everybody to be healthy. Then we would be correct on 95% of the population (the healthy ones), and wrong on the affected 5%. In other words, this trivial test is 95% effective! So we might ask if it is worth running the test at all. What do you think?

Here are a couple more examples of conditional probabilities, based on some of our sample spaces from the previous lecture note.

Balls and bins.

Suppose we toss $m = 3$ (labelled) balls into $n = 3$ bins; this is a uniform sample space with $3^3 = 27$ points. We already know that the probability the first bin is empty is $(1 - \frac{1}{3})^3 = (\frac{2}{3})^3 = \frac{8}{27}$. What is the probability of this event *given that* the second bin is empty? Call these events A, B respectively. To compute $\mathbf{P}(A|B)$ we need to figure out $\mathbf{P}(A \cap B)$. But $A \cap B$ is the event that both the first two bins are empty, i.e., all three balls fall in the third bin. So $\mathbf{P}(A \cap B) = \frac{1}{27}$ (why?). Therefore,

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{1/27}{8/27} = \frac{1}{8}.$$

Not surprisingly, $\frac{1}{8}$ is quite a bit less than $\frac{8}{27}$: knowing that bin 2 is empty makes it significantly less likely that bin 1 will be empty.