EE 178/278A Probabilistic Systems Analysis
Spring 2014 Tse/Hussami Lecture 4

Conditional Probability Review

In the previous lecture, the conditional probability $\mathbf{P}(A|B)$ was defined as $\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}$. The following table summarizes the differences before and after conditioning on the event B.

Table 1: Outcome probabilities

	Before Conditioning	After Conditioning
Sample Space	Ω	B
Probability Assignment	$\mathbf{P}(\omega)$	$\mathbf{P}(\omega\|B)$
	$\sum_{\omega\in\Omega}\mathbf{P}(\omega)=1$	$\sum_{\omega\in\Omega}\mathbf{P}(\omega\|B)=\sum_{\omega\in B}\mathbf{P}(\omega)/\mathbf{P}(B)=1$

Bayesian Inference [1]

The medical test problem is a canonical example of an *inference* problem: given a noisy observation (the result of the test), we want to figure out the likelihood of something not directly observable (whether a person is healthy). To bring out the common structure of such inference problems, let us redo the calculations in the medical test example but only in terms of events without explicitly mentioning the outcomes of the underlying sample space.

Recall: A is the event the person is affected, B is the event that the test is positive. What are we given?

- $\mathbf{P}(A)=0.05$, (5% of the U.S. population is affected)
- $\mathbf{P}(B|A)=0.9$ (90% of the affected people test positive)
- $\mathbf{P}(B|A^c)=0.2$ (20% of healthy people test positive)

We want to calculate $\mathbf{P}(A|B)$. We can proceed as follows:

$$\mathbf{P}(A|B)=\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}=\frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)} \tag{1}$$

and

$$\mathbf{P}(B)=\mathbf{P}(A\cap B)+\mathbf{P}(A^c\cap B)=\mathbf{P}(B|A)\mathbf{P}(A)+\mathbf{P}(B|A^c)(1-\mathbf{P}(A)) \tag{2}$$

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

Combining equations (1) and (2), we have expressed $\mathbf{P}(A|B)$ in terms of $\mathbf{P}(A), \mathbf{P}(B|A)$ and $\mathbf{P}(B|A^c)$:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)(1 - \mathbf{P}(A)} \tag{3}$$

This equation is useful for many inference problems. We are given $\mathbf{P}(A)$, which is the (unconditional) probability that the event of interest A happens. We are given $\mathbf{P}(B|A)$ and $\mathbf{P}(B|A^c)$, which quantify how noisy the observation is. (If $\mathbf{P}(B|A) = 1$ and $\mathbf{P}(B|A^c) = 0$, for example, the observation is completely noiseless.) Now we want to calculate $\mathbf{P}(A|B)$, the probability that the event of interest happens given we made the observation. Equation (3) allows us to do just that.

Equation (3) is at the heart of a subject called *Bayesian inference*, used extensively in fields such as machine learning, communications and signal processing. The equation can be interpreted as a way to *update knowledge* after making an observation. In this interpretation, $\mathbf{P}(A)$ can be thought of as a *prior* probability: our assessment of the likelihood of an event of interest A *before* making an observation. It reflects our prior knowledge. $\mathbf{P}(A|B)$ can be interpreted as the *posterior* probability of A after the observation. It reflects our new knowledge.

Of course, equations (1), (2) and (3) are derived from the basic axioms of probability and the definition of conditional probability, and are therefore true with or without the above Bayesian inference interpretation. However, this interpretation is very useful when we apply probability theory to study inference problems.

Bayes' Rule and Total Probability Rule

Equations (1) and (2) are very useful in their own right. The first is called **Bayes' Rule** and the second is called the **Total Probability Rule**. Bayes' Rule is useful when one wants to calculate $\mathbf{P}(A|B)$ but one is given $\mathbf{P}(B|A)$ instead, i.e. it allows us to "flip" things around. The Total Probability Rule is an application of the strategy of "dividing into cases" we learned in Note 2 to calculating probabilities. We want to calculate the probability of an event B. There are two possibilities: either an event A happens or A does not happen. If A happens the probability that B happens is $\mathbf{P}(B|A)$. If A does not happen, the probability that B happens is $\mathbf{P}(B|A^c)$. If we know or can easily calculate these two probabilities and also $\mathbf{P}(A)$, then the total probability rule yields the probability of event B.

Independent events

Definition 4.1 (independence): Two events A, B in the same probability space are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A) \times \mathbf{P}(B)$.

The intuition behind this definition is the following. Suppose that $\mathbf{P}(B) > 0$ and A, B are independent. Then we have

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

Thus independence has the natural meaning that "the probability of A is the same whether or not we conditional on B ." (By a symmetrical argument, we also have $\mathbf{P}(B|A) = \mathbf{P}(B)$ provided $\mathbf{P}(A) > 0$.) For events A, B such that $\mathbf{P}(B) > 0$, the condition $\mathbf{P}(A|B) = \mathbf{P}(A)$ is actually *equivalent* to the definition of independence.

The above definition generalizes to any finite set of events:

Definition 4.2 (mutual independence): Events $A_1, \dots, A_n$ are *mutually independent* if for every subset $I \subseteq \{1, \dots, n\}$,

$$\mathbf{P}(\textstyle\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbf{P}(A_i).$$

Note that we need this property to hold for *every* subset I.

For mutually independent events $A_1,\dots,A_n$, it is not hard to check from the definition of conditional probability that, for any $1 \le i \le n$ and any subset $I \subseteq \{1,\dots,n\} \setminus \{i\}$, we have

$$\mathbf{P}(A_i | \textstyle\bigcap_{j\in I} A_j) = \mathbf{P}(A_i).$$

Note that the independence of every pair of events (so-called *pairwise independence*) does *not* necessarily imply mutual independence. For example, it is possible to construct three events A,B,C such that each *pair* is independent but the triple A,B,C is *not* mutually independent.

We now provide several examples to illustrate independence.

Examples

1. **Coin tosses.** Toss a fair coin three times. Let A be the event that all three tosses are heads. Then $A = A_1 \cap A_2 \cap A_3$, where A_i is the event that the ith toss comes up heads. We have

$$\begin{aligned}\mathbf{P}(A) &= \mathbf{P}(A_1) \times \mathbf{P}(A_2|A_1) \times \mathbf{P}(A_3|A_1 \cap A_2)\\ &= \mathbf{P}(A_1) \times \mathbf{P}(A_2) \times \mathbf{P}(A_3)\\ &= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{1}{8}.\end{aligned}$$

The second line here follows from the fact that the tosses are mutually independent. Of course, we already know that $\mathbf{P}(A) = \frac{1}{8}$ from our definition of the probability space in the previous lecture note. The above is really a check that the space behaves as we expect.[2]

It seems reasonable that the tosses should remain mutually independent, even if the coin is biased, since no coin toss is affected by any of the other tosses. If the coin is biased with heads probability p, this independence assumption implies

$$\mathbf{P}(A) = \mathbf{P}(A_1) \times \mathbf{P}(A_2) \times \mathbf{P}(A_3) = p^3.$$

As another example, let B denote the event that the first coin toss comes up tails and the next two coin tosses come up heads. Then $B = A_1^c \cap A_2 \cap A_3$, and these events are independent, so

$$\mathbf{P}(B) = \mathbf{P}(A_1^c) \times \mathbf{P}(A_2) \times \mathbf{P}(A_3) = p^2(1-p),$$

since $\mathbf{P}(A_1^c) = 1 - \mathbf{P}(A_1) = 1-p$. More generally, the probability of any sequence of n tosses containing r heads and $n-r$ tails is $p^r(1-p)^{n-r}$. The notion of independence is the key concept that enables us to assign probabilities to these outcomes.

2. **Balls and bins.** Let A be the event that bin 1 is empty. We saw in the previous lecture note (by counting) that $\mathbf{P}(A) = (1-\frac{1}{n})^m$, where m is the number of balls and n is the number of bins. The product rule gives us a different way to compute the same probability. We can write $A = \bigcap_{i=1}^m A_i$, where A_i is the event that ball i misses bin 1. Clearly $\mathbf{P}(A_i) = 1 - \frac{1}{n}$ for each i. Also, the A_i are mutually independent since ball i chooses its bin regardless of the choices made by any of the other balls. So

$$\mathbf{P}(A) = \mathbf{P}(A_1) \times \cdots \times \mathbf{P}(A_m) = \left(1 - \frac{1}{n}\right)^m.$$

[2] Strictly speaking, we should really also have checked from our original definition of the probability space that $\mathbf{P}(A_1)$, $\mathbf{P}(A_2|A_1)$ and $\mathbf{P}(A_3|A_1 \cap A_2)$ are all equal to $\frac{1}{2}$.

3. **Monty Hall.** We let A bet the event that the car is behind the first door, B be the event that Bob picks the first door, and C be the event that Carol picks the third door,

$$\mathbf{P}(A\cap B\cap C)=P(A\cap B)P(C|A\cap B)=P(A)P(B)P(C|A\cap B)=\tfrac{1}{3}\times\tfrac{1}{3}\times\tfrac{1}{2}=\tfrac{1}{18}.$$

Here we are making the assumption that the contestant's choice is independent of the location of the car. Note that we specifyy they probability of this outcome earlier by counting. Here we use the higher level concept of independent and conditional probabilities to define the probabilities of our outcomes.

4. **Biased coins.** Suppose we have two biased coins, with probabilities p and q of obtaining heads. We consider the experiment where we generate a sequence of length 3 by each time randomly picking one of the two coins and then flipping it. Let A_i be the event that the i*th* flip is a heads. Then the events A_1, A_2 and A_3 are independent. However, if we select a coin initially, and then flip that same coin 3 times, the events A_1, A_2 and A_3 are not independent anymore.