Conditional Probability Review

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

<table>
<thead>
<tr>
<th>Sample Space</th>
<th>Probability Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega$</td>
<td>$P(\omega)$</td>
</tr>
<tr>
<td>$B$</td>
<td>$P(\omega</td>
</tr>
</tbody>
</table>

$\sum_{\omega \in \Omega} P(\omega) = 1$  $\sum_{\omega \in \Omega} P(\omega|B) = \sum_{\omega \in B} P(\omega)/P(B) = 1$

Bayesian Inference

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 ($H = 1$), $B$ is the event that the test is positive ($T = 1$). What are we given?

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

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

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

and

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

\(^1\)Part of this note is adapted from the notes of EECS 70 at Berkeley.
Combining equations (1) and (2), we have expressed $P(A|B)$ in terms of $P(A)$, $P(B|A)$ and $P(B|A^c)$:

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

This equation is useful for many inference problems. We are given $P(A)$, which is the (unconditional) probability that the event of interest $A$ happens. We are given $P(B|A)$ and $P(B|A^c)$, which quantify how noisy the observation is. (If $P(B|A) = 1$ and $P(B|A^c) = 0$, for example, the observation is completely noiseless.) Now we want to calculate $P(A|B)$, 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, $P(A)$ 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. $P(A|B)$ 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 $P(A|B)$ but one is given $P(B|A)$ 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 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 $P(B|A)$. If $A$ does not happen, the probability that $B$ happens is $P(B|A^c)$. If we know or can easily calculate these two probabilities and also $P(A)$, then the total probability rule yields the probability of event $B$.

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(B|A)P(A) + P(B|A^c)(1 - P(A)) \quad (4)$$

In general, if $\{A_1, A_2, \ldots, A_n\}$ form a partition of the sample space $\Omega$, i.e., $A_1, \ldots, A_n$ are mutually disjoint ($A_i \cap A_j = \phi$ for $i \neq j$) and exhaustive ($A_1 \cup \cdots \cup A_n = \Omega$), then,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \quad (5)$$

**Independent events**

**Definition 3.1 (independence):** Two events $A, B$ in the same probability space are independent if $P(A \cap B) = P(A) \times P(B)$.

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A).$$
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 $P(B|A) = P(B)$ provided $P(A) > 0$.) For events $A, B$ such that $P(B) > 0$, the condition $P(A|B) = P(A)$ is actually equivalent to the definition of independence.

The above definition generalizes to any finite set of events:

**Definition 3.2 (mutual independence):** Events $A_1, \ldots, A_n$ are **mutually independent** if for every subset $I \subseteq \{1, \ldots, n\}$,

\[
P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i).
\]

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

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

\[
P(A_i | \bigcap_{j \in I} A_j) = P(A_i).
\]

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 $i$th toss comes up heads. We have

\[
P(A) = P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1 \cap A_2)
\]

\[
= P(A_1) \times P(A_2) \times P(A_3)
\]

\[
= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.
\]

The second line here follows from the fact that the tosses are mutually independent. Of course, we already know that $P(A) = \frac{1}{8}$ 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

\[
P(A) = P(A_1) \times P(A_2) \times P(A_3) = p^3.
\]

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

\[
P(B) = P(A_1^c) \times P(A_2) \times P(A_3) = p^2(1 - p),
\]

since $P(A_1^c) = 1 - P(A_1) = 1 - p$. 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.

\(^2\)Strictly speaking, we should really also have checked from our original definition of the probability space that $P(A_1), P(A_2 | A_1)$ and $P(A_3 | A_1 \cap A_2)$ are all equal to $\frac{1}{2}$.
2. **Balls and bins.** Suppose we toss \( m \) (labelled) balls at random into \( n \) bins independently. The event that the first bin is empty is same as the event that each of the \( m \) balls go into a bin other than the first one. The probability that a particular ball goes into a bin other than the first is \( \left(\frac{n-1}{n}\right) \). Thus, by independence, the probability that the first bin is empty is \( \left(1 - \frac{1}{n}\right)^m \).

Now suppose that \( m = 3 \) and \( n = 3 \). We already know that the probability the first bin is empty is \( \left(1 - \frac{1}{3}\right)^3 = \left(\frac{2}{3}\right)^3 = \frac{8}{27} \). What is the probability of this event given that the second bin is empty? Call these events \( A, B \) respectively. To compute \( P(A|B) \) we need to figure out \( P(A \cap B) \). But \( A \cap B \) is the event that the first and second bins are both empty, i.e., all three balls fall in the third bin. So \( P(A \cap B) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9} \) (why? Hint: use independence again). Therefore, \( P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/27}{8/27} = \frac{1}{8} \).

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 also be empty.

3. **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.

### Independent random variables

In the previous section we have defined the notion of independent events. But it seems natural to extend the notion to **random variables** as well. When should we call two random variables independent?

Suppose \( X_1 \) and \( X_2 \) are the results of the first and second roll of a die respectively. Then it is reasonable to model the events \( X_1 = 3 \) and \( X_2 = 5 \) to be independent. But so should the events \( X_1 = 6 \) and \( X_2 = 6 \). In fact, the events \( X_1 = a \) and \( X_2 = b \) should be independent for every \( a, b \in \{1, \ldots, 6\} \). This example leads to a natural definition for two random variables.

**Definition 3.3 (independence):** Two random variables \( X \) and \( Y \) are **independent** if the events \( X = a \) and \( Y = b \) are independent for every value \( a \) that \( X \) can take on and for every value \( b \) that \( Y \) can take on.

The above definition generalizes to any finite set of random variables:

**Definition 3.4 (mutual independence):** Random variables \( X_1, \ldots, X_n \) are **mutually independent** if the events \( X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n \) are mutually independent for every value \( a_i \) \( X_i \) can take on , \( i = 1, \ldots, n \).

Note that independence of random variables drastically reduces the number of parameters needed to specify a probability model. Indeed, in a general probability model with \( n \) random variables each taking on \( m \) possible values, one needs to specific \( m^n - 1 \) numbers to determine the underlying probabilities. But if the random variables are independent, then one needs only \( n(m - 1) \) numbers!