Weekly Notes for EE2012 2014/15 – Week 3

T. J. Lim

February 2, 2015

Book sections covered this week: 2.4 – 2.6.

1 Conditional Probability

1.1 Definition

A central application of probability theory is to answer questions of the form "If event B occurs, how does that affect the probability of A occurring?" For instance,

- $A =$ "Paul committed the crime" and $B =$ "Paul has an alibi". If at the start we already know that Paul is a serial offender of this sort of crime and that he lives in the area, we may say that $P(A)$ is quite large, say 0.5. But after knowing that he has an alibi e.g. he was seen at another place by a number of people, i.e. event B occurred, then $P(A)$ becomes much smaller, probably 0.
- $A =$ "Jim will not feel well tomorrow", and $B =$ "Jim had 5 beers tonight". $P(A)$ would be very different given that B occurred, compared to the case if B did not occur.
- $A =$ "Waiting time at a clinic is more than 20 minutes" and $B =$ "Doctor has not arrived yet". Again, $P(A)$ would be significantly affected by the occurrence of event B.

Conditional probability captures this change in probability of event A after finding out that B occurred in an experiment. Almost always, it is not that B *actually* occurred, but that we *hypothesize* that it did. In other words, we ask "what if B occurs, then what will happen to our belief that A occurred *in the same experiment*"? It is as if we are given a clue as to the outcome obtained that does not pinpoint the exact outcome, but only that it belongs in a certain set B.

Example 1: Suppose we have the sample space $S = \{1, 2, 3, 4, 5, 6\}$, where the elementary events are equi-probable. Let

$$A = \{2, 4, 6\} \quad (1)$$
$$B = \{4, 5, 6\} \quad (2)$$
$$C = \{1, 2\}. \quad (3)$$

If we assume or know that B occurred, then the sample space effectively shrinks to B because the outcome must come from B. Recall that for experiments with

equi-probable outcomes, the probability of an event A is obtained by dividing $|A|$ by $|S|$, its sample space. By this reasoning, and considering the new sample space as B, the so-called conditional probability of A given B is denoted $P(A|B)$ and must be

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{2}{3}.$$

Before assuming/knowing B occurred, then $P(A) = \frac{1}{2}$, so the information that B occurred affected the probability of A.

To compute another conditional probability, say $P(C|A)$, we would also find the number of elements in $A \cap C$, and then divide it by the cardinality of A which is the new sample space. Therefore

$$P(C|A) = \frac{|A \cap C|}{|A|} = \frac{1}{3}$$

which happens to be equal to $P(C)$, the original probability of C. So we see that information that an event occurs does not *always* have an impact on the probability of another event occurring. This is another critical concept called "independence" that we will introduce in the next section. ■

In general, for repeatable experiments, the relative frequency of A conditioned on B is

$$f_{A|B}(n) = \frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n} = \frac{f_{A\cap B}(n)}{f_B(n)}. \tag{4}$$

By extension, for non-repeatable experiments, we can define

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{5}$$

From (5), we have the important expressions

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{6}$$

These are so useful and important in probabilistic applications that we must commit them to memory.

1.2 Theorem of Total Probability

Let the sets $B_1, \ldots, B_n$ be mutually exclusive in such a way that

$$\bigcup_{k=1}^{n} B_k = S. \tag{7}$$

Then B_k, $k = 1, \ldots, n$ are said to *partition* S. For any subset A of S, we can then always write

$$A = \bigcup_{k=1}^{n} A \cap B_k \tag{8}$$

and due to the mutual exclusivity of the B_k's, we can see that $A \cap B_k$, $k = 1, 2, \ldots, n$ are all mutually exclusive too. Therefore by Corollary 4,

$$P(A) = \sum_{k=1}^{n} P(A \cap B_k) \tag{9}$$

But from (6), we arrive at the desired expression

$$P(A) = \sum_{k=1}^{n} P(A|B_k)P(B_k). \tag{10}$$

This result is known as the theorem of total probability.

Example 2: There is an urn with 4 white and 3 black balls. Draw two balls without replacement and note the sequence of colours drawn. We are interested in $A =$ "Second ball is black". How do we find $P(A)$? We can solve this problem using conditional probabilities, by first defining $B_1 =$ "First ball is black" and $W_1 =$ "First ball is white". From (10), we know

$$P(A) = P(A|B_1)P(B_1) + P(A|W_1)P(W_1). \tag{11}$$

But $P(B_1) = 3/7$ while $P(W_1) = 4/7$, and $P(A|B_1) = P$["2nd ball is black given that the first one is black"]. This conditional probability can be found since given B_1, there will be 4 white and 2 black balls left in the bag. Therefore $P(A|B_1) = 1/3$. Similarly, we can rapidly compute $P(A|W_1) = 1/2$. Finally,

$$P(A) = \frac{1}{3} \times \frac{3}{7} + \frac{1}{2} \times \frac{4}{7} = \frac{3}{7}. \tag{12}$$

This example neatly illustrates the key point that *conditional* probabilities are often easier to find than unconditional probabilities. Therefore, a problem can be solved if we cleverly choose the partition $B_1, \ldots, B_n$ so that $P(A|B_k)$ are all known. ■

1.3 Bayes' Rule

As in the previous section, let $B_1, \ldots, B_n$ partition S. Suppose we are interested only in $P(B_j|A)$ but instead know $P(A|B_k)$, $k = 1, \ldots, n$. Bayes' Rule, derived almost trivially from (5), (6) and (10), says that

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}. \tag{13}$$

This little equation is one of the most important results in engineering applications of probability, as illustrated in the next example.

Example 3: This is an example from the medical world. Suppose there is a rare incurable disease that strikes 1 in 10^5 people on average. A test for the disease is 98 percent accurate, meaning that the probability of a false positive is 0.02, and the probability of a false negative is 0.02. If a patient tests positive, what is the probability that he has the disease?

At first glance, given that the test seems pretty accurate, the patient would appear to be doomed. But let's be more careful. Let $A =$ "Patient has the disease" and $B =$ "Test is positive". What we want is $P(A|B)$, but what we have are

$$P(B|A) = 0.98 \quad \text{and} \quad P(B|A^c) = 0.02. \tag{14}$$

We also know that $P(A) = 10^{-5}$ and so $P(A^c) = 1 - 10^{-5}$. Applying Bayes' Rule, we get

$$P(A|B) = \frac{0.98 \times 10^{-5}}{0.98 \times 10^{-5} + 0.02 \times (1 - 10^{-5})} = 4.9 \times 10^{-4}. \tag{15}$$

This is rather surprising! A test that only fails 2 percent on average turns out to be quite useless in predicting the presence of a disease. The reason is that the false alarm rate of the test is much higher than the likelihood of having the disease. Therefore, a good test for a very rare disease must be much more accurate than one for a common disease. ■

2 Independence of Events

2.1 Concept and Definition

Independence between two events is a concept that is easily understood:

1. Whether it's cloudy today has no bearing on how many people are in class tomorrow. "Cloudy today" and "More than 20 people in class tomorrow" are independent events.

2. You flip a coin, and then roll a die. The outcome of the coin flip should have no influence on the outcome of the die roll. Therefore, "Coin turns up Heads" and "Top face of die is 5" are independent events.

On the other hand, some events are clearly not independent:

1. "Paul's dog died yesterday" and "Paul is feeling down today" are not independent because knowing the former event occurred, the latter event is more likely.

2. You roll two dice and note both faces on top. The events "The total is 3" and "One of the faces is 6" are not independent because knowing the former implies that the latter did not occur.

You should be able to see that the concept of independence is closely tied to the notion of conditional probability just introduced in the last section. To be precise, two events A and B are said to be independent, if and only if

$$P(A|B) = P(A) \tag{16}$$

and, equivalently, $P(B|A) = P(B)$. So knowing that A occurred does not change our belief (or certainty) that B occurred, and vice versa.

Equation (16) also means that

$$P(A \cap B) = P(A)P(B) \tag{17}$$

which is a necessary and sufficient condition for independence between two events A and B.

2.2 More Than Two Events

If we say that A, B and C are independent, what do we mean? It must be that given, say $B \cap C$, the probability of A is unchanged, i.e.

$$P(A|B \cap C) = P(A). \tag{18}$$

The events must of course also be pairwise independent, i.e. $P(A|B) = P(A)$, etc. In other words, given the occurrence of some combination of the other two events (union or intersection), or of one of the other events, the probability of A, B or C, individually, remains the same.

So in order to say that three events are independent requires not only that they are *pairwise independent*, but that all unions and intersections of the three events are independent of all other unions and intersections, e.g. $(A \cap B)/C$, $A/(B \cup C)$, $B^c/(A \cap C^c)$, etc. are all independent pairs of events. It can be proven that this condition is met if and only if

$$P(A \cap B \cap C) = P(A)P(B)P(C). \tag{19}$$

In other words, (19) is a necessary and sufficient[1] condition for independence of three events, A, B and C.

This result generalizes as follows: A collection of n events $A_1, \ldots, A_n$ are independent if and only if

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P[A_i]. \tag{20}$$

In many cases, independence is an assumption that can be justified through logic or physics, rather than a property to be discovered, as illustrated below.

Example 4: Three unfair coins are tossed. The probabilities of Heads are 0.2, 0.3 and 0.4 respectively for Coins 1, 2 and 3. Find the probability of all three coins landing Heads.

The sample space of this experiment consists of a set of triples, $S = \{(a, b, c) : a, b, c \in \{h, t\}\}$. We can define the three events

$$H_1 = \text{“First coin is Heads”} \tag{21}$$

$$H_2 = \text{“Second coin is Heads”} \tag{22}$$

$$H_3 = \text{“Third coin is Heads”}. \tag{23}$$

These are equivalent to the subsets $\{(h, \times, \times)\}$, $\{(\times, h, \times)\}$, and $\{(\times, \times, h)\}$, respectively, where each subset has four members. Now because the outcome of the first coin toss would reasonably not have an impact on the outcome of the second or third coin tosses, we can reasonably *assume* that H_1, H_2 and H_3 are independent. Hence,

$$P[H_1 \cap H_2 \cap H_3] = P[H_1]P[H_2]P[H_3] = 0.024. \tag{24}$$

[1]Consider two properties F and G, e.g. F is (19) and G is “A, B and C are independent”. We say that F is a necessary condition for G if F holds whenever G holds, and we write $G \Rightarrow F$. In set notation, $G \subset F$. We say that F is a sufficient condition for G if G holds whenever F holds, and we write $F \Rightarrow G$. Using sets, we have $F \subset G$. Finally, if F is a necessary and sufficient condition for G, then $F \Leftrightarrow G$, and in terms of sets, $F = G$. In the last case, F and G are equivalent, and we can prove that F holds by showing that G holds, or vice versa.

3 Sequential Experiments

These are experiments consisting of a number of sub-experiments performed one after another, as we saw in previous examples of picking k balls from an urn, tossing a coin several times, and so on. We first start with a sequence of *independent sub-experiments*.

3.1 Independent Sub-Experiments

Consider n sub-experiments performed in sequence, $E_1, \ldots, E_n$. The i-th sub-experiment has sample space S_i, so that the complete sample space is $S_1 \times S_2 \times \cdots \times S_n$, and each outcome is an n-tuple $s = (s_1, \ldots, s_n)$, where $s_i \in S_i$, $i = 1, \ldots, n$.

Let A_i be an arbitrary event within the event field of E_i. If any collection of $A_1, \ldots, A_n$ forms a set of independent events, then $E_1, \ldots, E_n$ are said to be *independent sub-experiments*. In this case, we have

$$P\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} P[A_i]. \tag{25}$$

Examples:

- Wait for the bus at the same time in the morning for 10 successive days. Note the waiting time every day. If the waiting time on day i does not affect your uncertainty about the waiting time on day j, $i \neq j$, then we have 10 independent sub-experiments.
- Toss a coin 6 times, and note sequence of heads and tails. Knowledge that the 2nd toss is a heads does not change the probability of the 3rd toss being a tail, etc. Therefore we have a sequence of 6 independent sub-experiments.

3.1.1 Bernoulli Trials and the Binomial Probability Law

A **Bernoulli trial** is an experiment that has a sample space with two elements. We can call these two possible outcomes success/failure, up/down, head/tail, black/white, or anything else that suits the situation. For concreteness we will use the success/failure terminology from now on.

Suppose the probability of success is p. A very important experiment that models many problems consists of performing a fixed number n of independent and identical Bernoulli trials, i.e. every trial has the same probability of success p. For simplicity of notation, let a "success" be denoted by 1, and a "failure" by 0. If $n = 3$, we have

$$P[\{000\}] = (1-p)^3 \tag{26}$$
$$P[\{001\}] = (1-p)^2 p \tag{27}$$
$$P[\{110\}] = (1-p)p^2, \quad \text{etc.} \tag{28}$$

because of the independence of the Bernoulli sub-experiments. Then the probability of 2 successes is

$$P[\text{"Two successes"}] = P[\{011, 101, 110\}] = 3p^2(1-p) \tag{29}$$

because each of the outcomes in the event has the same probability of $p^2(1-p)$.

In general, since the number of ways to obtain k successes in n trials is $\binom{n}{k}$, we have the probability of k successes in n independent identical Bernoulli trials as

$$p_n(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{30}$$

This is a very important probability law, called the **Binomial probability law**.

Example 5: A student did not study for a multiple-choice exam with 10 questions and 4 choices in each question. Find the probability that he answers five or more questions correctly if he randomly picks one of the four choices in each question.

We can reasonably assume that $P[\text{“correct answer in } i\text{-th question”}] = 0.25$. The probability of getting k answers correct is given by the binomial law as

$$p_n(k) = \binom{10}{k} 0.25^k 0.75^{10-k}, \quad k = 0, 1, \ldots, 10. \tag{31}$$

Therefore the probability of passing is

$$\sum_{k=5}^{10} \binom{10}{k} 0.25^k 0.75^{10-k} = 0.0781. \tag{32}$$

This is rather low, and will go even lower if the number of choices is increased of course. ■

3.1.2 Geometric Probability Law

We are often interested in the number of independent Bernoulli trials needed before encountering the first success, for instance:

1. What is the probability that you win the top lottery prize after fewer than 20 tries if the probability of winning is 10^{-6}?
2. What is the probability that we need to re-transmit a packet over the Internet more than once, if the probability of failure of each packet is 10^{-3}?
3. How large does the probability of success p have to be in order for you to meet your first success within 5 attempts?

Suppose success in the k-th trial is denoted by A_k. Then the event “First success at the m-th trial” is equivalent to failing the first $m-1$ times, and then succeeding in the m-th time, i.e. $A_1^c A_2^c \cdots A_{m-1}^c A_m$, where we have dispensed with the intersection notation for convenience.

But since the trials are independent and identical sub-experiments, we can write

$$P[A_1^c A_2^c \cdots A_{m-1}^c A_m] = P[A_m] \prod_{i=1}^{m-1} P[A_i^c] = p(1-p)^{m-1} \tag{33}$$

if p is the probability of success in any one of the Bernoulli trials. Hence the probability of the first success in a sequence of Bernoulli trials occurring at the m-th trial is

$$p_m = p(1-p)^{m-1}, \quad m = 1, 2, \ldots \tag{34}$$

This is known as the geometric probability law.

The probability of the first success occurring later than the N-th trial is given by a simple expression:

$$\begin{aligned} P[\{N+1, N+2, \ldots\}] &= \sum_{m=N+1}^{\infty} p(1-p)^{m-1} && (35) \\ &= p\sum_{k=N}^{\infty}(1-p)^{k} && (36) \\ &= p\frac{(1-p)^N}{1-(1-p)} && (37) \\ &= (1-p)^N. && (38) \end{aligned}$$

We used the formula for the sum of an infinite geometric series to obtain (37) from (36).

Example 6: Let $p = 10^{-6}$, and then we can answer the question at the top of this sub-section. The event "win after fewer than 20 tries" is $\{1, 2, \ldots, 19\}$, which has the probability

$$\begin{aligned} P[\text{“win after fewer than 20 tries”}] &= \sum_{k=1}^{19} p(1-p)^{k-1} \\ &= p\sum_{k=0}^{18}(1-p)^{k} \\ &= p\frac{1-(1-p)^{19}}{1-(1-p)} \\ &= 1-(1-10^{-6})^{19} \\ &= 1.90\times 10^{-5} \end{aligned}$$

which is, unsurprisingly, basically zero. ■

3.2 Non-Independent Sub-Experiments

When we have a sequence of sub-experiments $E_1, E_2, \ldots$ that are not necessarily independent, then events $A_1, A_2, \ldots$ respectively defined within the event fields of these sub-experiments will not necessarily be independent either. We then cannot say e.g. that $P(A_1 \cap A_2) = P(A_1)P(A_2)$, i.e. the probability of a joint event will no longer factorize into a product of individual event probabilities.

However, note that

$$P(A_1A_2A_3) = P(A_3|A_1A_2)P(A_2|A_1)P(A_1) \tag{39}$$

by applying the definition of conditional probability twice. In fact, by recursively applying the rule $P(A \cap B) = P(B|A)P(A)$, we get the general **chain rule** of conditional probability as

$$P\left(A_1A_2\cdots A_n\right) = P(A_1)\prod_{k=2}^{n} P[A_k|A_1A_2\cdots A_{k-1}]. \tag{40}$$

This chain rule is the basis for solving problems involving general sequences of events or experiments.

Closely related to the chain rule is a visualization technique known as the tree diagram. This concept was exemplified in Figure 2.10 of the textbook, and also in class. In a tree diagram, each level of the tree represents one time epoch (e.g. one ball drawn). Every node represents a partial sequence of events, described by the branch labels from the root of the tree to that node. Besides being labelled with an event (e.g. the ball drawn is white), each branch is also labelled with the conditional probability of that event, given the partial sequence of events up to the (upper) node to which it is connected.

For example, in the case of drawing one ball each time without replacement from an urn containing 5 black and 5 white balls initially, we have

$$P(B_3|B_2W_1) = \frac{1}{2}, \tag{41}$$

and therefore the branch sprouting from the node for the partial sequence W_1B_2 that represents the event B_3 is labelled with the value $1/2$, as well as B (or more accurately, B_3).

The probability of any sequence of outcomes is obtained by multiplying together the branch probability labels, as dictated by (40).

3.2.1 Markov Chains

In a surprisingly large number of applications, we find that (40) collapses into a much simpler form:

$$P(A_1A_2\cdots A_n) = P(A_1)\prod_{k=2}^{n} P(A_k|A_{k-1}). \tag{42}$$

In other words, conditioning on the entire set of events A_1 through A_{k-1} is no different from conditioning only on the most recent one A_{k-1}. The example presented in class of picking from two urns filled with different numbers of balls labelled 0 and 1 (with replacement) is a prototypical example. Applications include Kalman filtering, Viterbi decoding, autoregressive modelling of a random signal, queueing analysis, and so on.

A lot of useful results are available for Markov chains, and this motivates us to even introduce additional (auxiliary) variables in order to model a system using a Markov chain. The topic of Markov chains is however quite advanced, and we will not refer to it too much in the remainder of this course.

Diagnostic Questions

1. In Example 1, assume that the probability of $\{1\}$, $\{2\}$, $\{3\}$ and $\{4\}$ are each 0.1, and the probability of $\{5\}$ and $\{6\}$ are each 0.3. Find $P(A|B)$ and $P(C|A)$.

2. In Example 3, let the accuracy of the test be $100p$ percent, rather than 98 percent. How large does p have to be for $P(A|B)$ to exceed 0.9?

3. An automated bus scheduler randomly chooses a time in the range $(0, 10]$ minutes, and sends out the next bus after waiting that amount of time. Find the probability that in eight bus deployments, there are three inter-bus intervals less than 4 minutes long.

4. From a deck of 52 cards, find the probability of drawing four consecutive Hearts.