Lisa Yan and Jerry Cain
CS109

Lecture Notes #4
September 21, 2020

Conditional Probability

Based on a chapter by Chris Piech and Lisa Yan

Pre-recorded lecture: All sections except Section 4. Section 5 up to and including spam email with Bayes' Theorem.
In-lecture: Section 5: Rest of exercises.
Covered later: Section 4 (as part of pre-lecture Lecture 6)

1 Conditional Probability

In English, a conditional probability answers the question: "What is the chance of an event E happening, given that I have already observed some other event F?" Conditional probability quantifies the notion of updating one's beliefs in the face of new evidence.

When you condition on an event happening you are entering the universe where that event has taken place. Mathematically, if you condition on F, then F becomes your new sample space. In the universe where F has taken place, all rules of probability still hold!

The definition for calculating conditional probability is:

Definition of Conditional Probability
The probability of E given that (aka conditioned on) event F already happened:

$$P(E \mid F) = \frac{P(EF)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

(As a reminder, EF means the same thing as $E \cap F$—that is, E "and" F.)

A visualization might help you understand this definition. Consider events E and F which have outcomes that are subsets of a sample space with 50 equally likely outcomes, each one drawn as a hexagon:

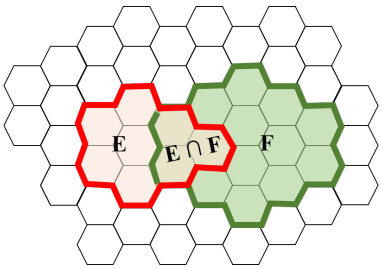

Figure 1: Conditional Probability Intuition

Conditioning on F means that we have entered the world where F has happened (and F, which has 14 equally likely outcomes, has become our new sample space). Given that event F has occured, the conditional probability that event E occurs is the subset of the outcomes of E that are consistent with F. In this case we can visually see that those are the three outcomes in $E \cap F$. Thus we have the:

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{3/50}{14/50} = \frac{3}{14} \approx 0.21$$

Even though the visual example (with equally likely outcome spaces) is useful for gaining intuition, the above definition of conditional probability applies regardless of whether the sample space has equally likely outcomes.

The Chain Rule
The definition of conditional probability can be rewritten as:

$$P(EF) = P(E \mid F)P(F)$$

which we call the Chain Rule. Intuitively it states that the probability of observing events E and F is the probability of observing F, multiplied by the probability of observing E, given that you have observed F. Here is the general form of the Chain Rule:

$$P(E_1E_2 \ldots E_n) = P(E_1)P(E_2 \mid E_1) \ldots P(E_n \mid E_1E_2 \ldots E_{n-1})$$

2 Law of Total Probability

An astute person once observed that in a picture like the one in Figure 1, event F can be thought of as having two parts, the part that is in E (that is, $E \cap F = EF$), and the part that isn't ($E^C \cap F = E^CF$). This is true because E and E^C are mutually exclusive sets of outcomes which together cover the entire sample space. After further investigation this was proved to be a general mathematical truth, and there was much rejoicing:

$$P(F) = P(EF) + P(E^CF)$$

This observation is called the **law of total probability**; however, it is most commonly seen in combination with the chain rule:

The Law of Total Probability
For events E and F,

$$P(F) = P(F \mid E)P(E) + P(F \mid E^C)P(E^C)$$

There is a more general version of the rule. If you can divide your sample space into any number of events $E_1, E_2, \ldots E_n$ that are *mutually exclusive* and *exhaustive*—that is, *every* outcome in sample space falls into *exactly one* of those events—then:

$$P(F) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i)$$

The word "total" refers to the fact that the events in E_i must combine to form the totality of the sample space.

3 Bayes' Theorem

Bayes' Theorem (or **Bayes' Rule**) is one of the most ubiquitous results in probability for computer scientists. Very often we know a conditional probability in one direction, say $P(E \mid F)$, but we would like to know the conditional probability in the other direction. Bayes' theorem provides a way to convert from one to the other. We can derive Bayes' theorem by starting with the definition of conditional probability:

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

Now we can expand $P(F \cap E)$ using the chain rule, which results in Bayes' Theorem.

Bayes' Theorem
The most common form of Bayes' Theorem is:

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)}$$

Each term in the Bayes' Rule formula has its own name. The $P(E \mid F)$ term is often called the **posterior**; the $P(E)$ term is often called the **prior**; the $P(F \mid E)$ term is called the **likelihood** (or the "update"); and $P(F)$ is often called the **normalization constant**.

If the normalization constant (the probability of the event you were initially conditioning on) is not known, you can expand it using the law of Total Probability:

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F \mid E)P(E) + P(F \mid E^C)P(E^C)} = \frac{P(F \mid E)P(E)}{\sum_i P(F \mid E_i)P(E_i)}$$

Again, for the last version, all the events E_i must be *mutually exclusive* and *exhaustive*.

A common scenario for applying the Bayes Rule formula is when you want to know the probability of something "unobservable" given an "observed" event. For example, you want to know the probability that a student understands a concept, given that you observed them solving a particular

problem. It turns out it is much easier to first estimate the probability that a student can solve a problem given that they understand the concept and then to apply Bayes' Theorem.

The "expanded" version of Bayes' Rule (at the bottom of the Bayes' Theorem box) allows you to work around not immediately knowing the denominator $P(F)$. It is worth exploring this in more depth, because this "trick" comes up often, and in slightly different forms. Another way to get to the exact same result is to reason that because the posterior of Bayes Theorem, $P(E \mid F)$, is a probability, we know that $P(E \mid F) + P(E^C \mid F) = 1$. If you expand out $P(E^C \mid F)$ using Bayes, you get:

$$P(E^C \mid F) = \frac{P(F \mid E^C)P(E^C)}{P(F)}$$

Now we have:

$$\begin{aligned}
1 &= P(E \mid F) + P(E^C \mid F) && \text{since } P(E|F) \text{ is a probability} \\
1 &= \frac{P(F \mid E)P(E)}{P(F)} + \frac{P(F \mid E^C)P(E^C)}{P(F)} && \text{by Bayes' rule (twice)} \\
1 &= \frac{1}{P(F)}\left[P(F \mid E)P(E) + P(F \mid E^C)P(E^C)\right] \\
P(F) &= P(F \mid E)P(E) + P(F \mid E^C)P(E^C)
\end{aligned}$$

We call $P(F)$ the normalization constant because it is the term whose value can be calculated by making sure that the probabilities of all outcomes sum to 1 (they are "normalized").

4 Conditional Paradigm

As we mentioned above, when you condition on an event you enter the universe where that event has taken place, all the laws of probability still hold. Thus, as long as you condition consistently on the same event, every one of the tools we have learned still apply. Let's look at a few of our old friends when we condition consistently on an event (in this case G):

Name of Rule	**Original Rule**	**Conditional Rule**
First axiom of probability	$0 \leq P(E) \leq 1$	$0 \leq P(E \mid G) \leq 1$
Corollary 1 (complement)	$P(E) = 1 - P(E^C)$	$P(E \mid G) = 1 - P(E^C \mid G)$
Chain Rule	$P(EF) = P(E \mid F)P(F)$	$P(EF \mid G) = P(E \mid FG)P(F \mid G)$
Bayes Theorem	$P(E \mid F) = \frac{P(F\|E)P(E)}{P(F)}$	$P(E \mid FG) = \frac{P(F\|EG)P(E\|G)}{P(F\|G)}$

5 Exercises

Example: Spam emails

Suppose 24 distinct emails are sent, 6 each to 4 distinct users. 10 of the 24 emails are spam. All possible outcomes of email allocations are equally likely.

Problem 1: Let event E be where user 1 receives 3 spam emails. What is $P(E)$?

Solution 1: We first note that the variations in the ways users 2, 3, and 4 can receive the rest of emails will be counted in both the event and sample space, and when calculating probability, these factors will cancel out—so we only need to consider how to allocate emails to user 1.

There are $\binom{24}{6}$ ways to allocate emails to user 1, where each outcome is equally likely. To count the outcomes in event E, we first choose the 3 of the 10 spam emails to send to user 1, then we choose the remaining 3 emails to be from the 14 non-spam emails. $P(E) = \binom{10}{3}\binom{14}{3}/\binom{24}{6}$.

Problem 2: Let event F be where user 2 receives 6 spam emails. What is $P(E|F)$?

Solution 2: We now define an outcome to be an allocation of emails to users 1 and 2. We would like to compute $P(E|F) = \frac{|EF|}{|F|}$. Starting with the size of F, we choose 6 spam emails to send to user 2, then choose 6 of the remaining 18 emails to send to user 1: $|F| = \binom{10}{6}\binom{18}{6}$. To compute the size of EF, we choose 6 spam emails to send to user 2, then choose 3 of the 4 remaining spam emails and 3 of the 14 non-spam emails to sent to user 1: $|EF| = \binom{10}{6}\binom{4}{3}\binom{14}{3}$. Therefore $P(E|F) = \binom{4}{3}\binom{14}{3}/\binom{18}{6}$.

Problem 3: Let event G be where user 3 receives 5 spam emails. What is $P(G|F)$?

Solution 3: We now define an outcome to be an allocation of emails to users 2 and 3. We note that it is impossible to simultaneously allocate 6 spam emails to user 2 and 5 spam emails (and 1 non-spam email) to user 3, as there are only 10 spam emails total. Therefore $|GF| = 0$ and $P(G|F) = 0$.

Example: Netflix and Learn

One of the key tenants of Netflix and other video streaming platforms is the recommendations feature: Given that you watched this movie, what is the probability that you will watch another movie? We will come back to this application many times throughout the quarter, but here is a taste of how conditional probability comes into play.

Suppose we are looking for the probability that a user watches the movie *Life is Beautiful* (call this event E). One possible approach to calculating $P(E)$ is from the frequentist definition of probability: Let n be the number of users on Netflix, and let $n(E)$ be the number of users who have watched this particular movie. Since n is large (there are a lot of people on Netflix!) we can compute $P(E) \approx \frac{n(E)}{n}$.

Now suppose we are looking for $P(E|F)$, the probability that a user watches *Life is Beautiful* given they watched another movie, *Amelie*. By the definition of conditional probability, $P(E|F) = \frac{P(EF)}{P(F)}$, and by the frequentist definition of probability $P(F) \approx \frac{n(F)}{n}$ and $P(EF) \approx \frac{n(EF)}{n}$, where $n(F)$ and $n(EF)$ are the numbers of users who have watched *Life is Beautiful* and both movies, respectively. Therefore $P(E|F) \approx \frac{n(EF)}{n(F)}$.

The purpose of this example is to show that this large denominator of people on Netflix will go away in conditional probability, and furthermore that the statistics of $P(E)$ and $P(E|F)$ can be different.

Example: Detecting spam email with Bayes'

Suppose 60% of all email in 2016 is spam. 20% of spam has the word "Dear," and 1% of non-spam (aka ham) has the word "Dear." You get an email with the word "Dear" in it. What is the probability that the email is spam?

Solution: Define event E as where you receive an email with the word "Dear" in it, and define event F as where you receive a spam email. We would like to compute $P(F|E)$. We note that $P(F) = 0.60$, $P(E|F) = 0.20$, and $P(E|F^C) = 0.01$. Using the expanded form of Bayes' rule,

$$\begin{aligned} P(F|E) &= \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F^C)} \\ &= \frac{(0.20)(0.6)}{(0.20)(0.6) + (0.01)(0.4)} \approx 0.967. \end{aligned}$$

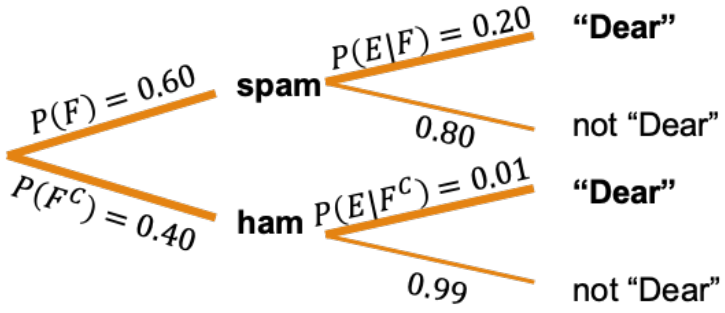

Figure 2: A "probability tree" of the spam email experiment. Possible paths (i.e., paths where we encounter the evidence E) are bolded.

To help us understand the experiment, we draw a "probability tree" of the experiment in Figure 2. Intuitively, we draw out all possibilities, ordered by the conditional probabilities as given in the problem. Since we are given probabilities $P(E|F)$ and $P(E|F^C)$, the first level of our tree decides between F and F^C. Then, the next level of our tree decides between E and E^C, given the events F or F^C. The probability of us reaching the leaves of this tree can be computed using the product rule. Finally, to calculate our target probability, we note that there are only two paths possible: where we observe E, and we would like to compute the probability of reaching the topmost leaf (where we have a spam email with the word "Dear") out of the two possible leaves.

5.1 Taking tests

When performing experiments like disease testing, we often distinguish between the events concerning the evidence E and fact F. We only have the ability to observe the evidence (e.g., testing positive or negative), and the evidence may or may not reflect the fact (e.g., having the disease or not).

Suppose we have the disease, and then we test for the disease. There is some probability $P(E|F)$ that our test returns a *true positive*—that our test accurately reflects the given presence of the disease. However, there is some probability $P(E^C|F)$ that our test returns a *false negative*—that our test will inaccurately say we don't have the disease, even though we actually do. We can enumerate the possibilities in a confusion matrix:

	F, disease +	F^C, disease −
E, Test +	True positive	False positive
E^C, Test −	False negative	True negative

These probabilities—probability of true positive, false positive, etc.—are inherent to the testing protocol and the disease, and are often given as statistics in our problem. However, the primary challenge of disease testing is that in the real world, we only have the ability to observe the evidence, E. Given this evidence, we must draw conclusions about F. To compute $P(F|E)$—the probability that we have the disease given that we test positive—we must use Bayes' Rule in combination with the probabilities from the confusion matrix, as well as $P(F)$, the likelihood of encountering the disease in our population.

Example: Zika Testing

A test is 98% effective at detecting Zika ("true positive"). However, the test has a "false positive" rate of 1%. Suppose that 0.5% of the population has Zika.

Problem 1: What is the likelihood that you have Zika if you test positive?

Solution 1: Let E be the event where you test positive, and F be the event where you actually have the disease. We are given the probabilities of a true positive $P(E|F) = 0.98$ and a false positive, $P(E|F^C) = 0.01$ respectively. Since $P(F) = 0.005$, we can use Bayes' Rule to compute $P(F|E)$:

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F)} = \frac{(0.98)(0.005)}{(0.98)(0.005) + (0.01)(0.995)} \approx 0.330$$

Problem 2: What is the probability that you have Zika if you test negative?

Solution 2: Using the same definitions of E and F as above, we would like to compute $P(F|E^C)$. We can compute the false negative rate as $P(E^C|F) = 1 - 0.98 = 0.02$ and the true negative rate $P(E^C|F^C) = 1 - 0.01 = 0.99$ and use Bayes' Rule:

$$P(F|E^C) = \frac{P(E^C|F)P(F)}{P(E^C|F)P(F) + P(E^C|F^C)P(F^C)} = \frac{(0.02)(0.005)}{(0.02)(0.005) + (0.99)(0.995)} \approx 0.0001$$

In conclusion, a test result—whether positive or negative—updates your belief of having the disease. Prior to the experiment (i.e, the test), your **prior belief** of having Zika was $P(F) = 0.005$. Your **posterior belief** is drastically different: With a positive test result, you update your belief of having Zika to $P(F|E) \approx 0.330$, which may motivate you to get tested a second time. With a negative test result, $P(F|E^C) \approx 0.0001$, meaning that you are very unlikely to have the disease!

Example: Monty Hall

Let's switch gears a bit and talk about a classic application of conditional probability (and, by extension, the Law of Total Probability). Monty Hall was a host of the game show *Let's Make a Deal*, where a participant could play a game of guessing—which we shall soon see is actually a game of probability.

Problem: You are shown three doors: behind one door is a prize (equally likely to be any door), and behind the other two doors is nothing (in the original show, there were goats). You choose a door, and then the host (Monty Hall at the time, Wayne Brady nowadays) opens one of the other two doors, revealing nothing (or a goat). You are then given the option to change to the other door. Should you switch?

Solution: Yes, you should switch. Without loss of generality, suppose you pick Door A (out of Doors A, B, and C). The probability that you win without switching is $1/3$, which is the probability of door A holding the prize.

However, consider the case where you switch. There are three cases to consider: (1) if A is the prize and you switch away from A, you lose, meaning you win with 0% probability. (2) if B is the prize, the host must open door C, in which case you switch to B and win with 100% probability; and (3) if C is the prize, the host must open door B, in which case you switch to C and win with 100% probability. There is an equal likelihood that you are in either of the three cases; we can use the Law of Total probability to calculate the probability of you winning in the case where you switch as $(1/3) \cdot 0 + (1/3) \cdot 1 + (1/3) \cdot 1 = 2/3$. In other words, there is a higher probability that you win if you follow the switching strategy.

Extension: To better understand this problem, consider the case of having 1000 doors, of which exactly 1 is the prize (with equal likelihood). You again have the ability to choose 1 doors. The host opens 998 of the remaining doors to reveal nothing (or a goat), and then you are given the option to switch between the door you chose and the other remaining door. Should you switch?

Yes, you should! While the problem is framed as a choice of "switch" or "don't switch," it may be easier to consider it a choice of "choose one door" (probability of winning = $1/1000$) or "choose 999 doors" (probability of winning = $999/1000$). With this framing, your choice of choosing 999 doors actually enables you to avoid 998 empty doors, since you could open them yourself and verify that they are invalid choices. No matter how you frame it, it is in your best interest to switch!