Will Monroe
CS 109

Lecture Notes #4
July 3, 2017

Conditional Probability

Based on a chapter by Chris Piech

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:

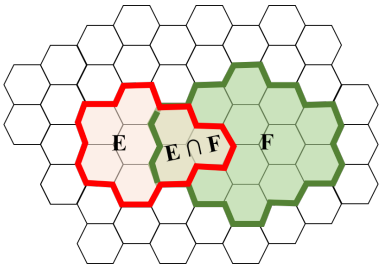

Figure 1: Conditional Probability Intuition

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:

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

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.

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").

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	$P(E \mid F) = \frac{P(F\|E)P(E)}{P(F)}$	$P(E \mid FG) = \frac{P(F\|EG)P(E\|G)}{P(F\|G)}$