3. Conditional Probability

Chris Piech and Mehran Sahami

May 2017

1 Introduction

It is that time in the quarter (it is still week one) when we get to talk about probability. Again we are going to build up from first principles. We will heavily use the counting that we learned earlier this week.

2 Conditional Probability

In English, a conditional probability states "what is the chance of an event E happening given that I have already observed some other event F". It is a critical idea in machine learning and probability because it allows us to update our 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:

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

Lets turn to our visualization friend to get an intuition for why this is true. Again 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:

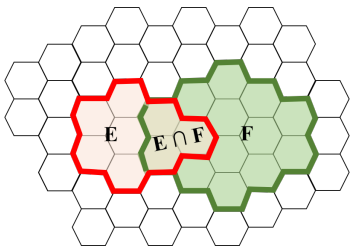

Figure 1: Conditional Probability Intuition

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:

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

Even though the visual example (with equally likely outcome spaces) is useful for gaining intuition, 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:

$$P(E \cap F) = P(E|F)P(F)$$

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:

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1)\ldots P(E_n|E_1 \ldots E_{n-1})$$

3 Law of Total Probability

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

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

This observation proved to be particularly useful when it was combined with the chain rule and gave rise to a tool so useful, it was given the big name, law of total probability.

The Law of Total Probability
If we combine our above observation with the chain rule, we get a very useful formula:

$$P(E) = P(E|F)P(F) + P(E|F^C)P(F^C)$$

There is a more general version of the rule. If you can divide your sample space into any number of mutually exclusive events: $B_1, B_2, \ldots B_n$ such that every outcome in sample space fall into one of those events, then:

$$P(E) = \sum_{i=1}^{n} P(E|B_i)P(B_i)$$

The name total (I assume) comes from the fact that the events in B_i have to make up the total sample space.

4 Bayes Theorem

Bayes Theorem is one of the most ubiquitous results in probability for computer scientists. Very often we know a conditional probability in one direction, say $P(E|F)$, 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:

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

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:

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

The different terms in the Bayes Rule formula have different terms. The $P(E|F)$ term is often called the "posterior" the $P(E)$ term is often called the "prior". The $P(F|E)$ term is called the update and $P(F)$ is often called the normalization constant.

In the case where the denominator is not known (the probability of the event you were initially conditioning on), you can expand it using the law of Total Probability:

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

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 that "trick" comes up often, and in slightly different forms. Another way to get to the exact same result is to think: well because the posterior of Bayes Theorem, $P(E|F)$, is a probability, I know that $P(E|F) + P(E^C|F) = 1$. If you expand out $P(E^C|F)$ using Bayes you get:

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

Now we have:

$$1 = P(E|F) + P(E^C|F) \qquad \text{Since } P(E|F) \text{ is a probability}$$

$$1 = \frac{P(F|E)P(E)}{P(F)} + \frac{P(F|E)P(E)}{P(F)} \qquad \text{By Bayes Rule (twice)}$$

$$1 = \frac{1}{C}P(F|E)P(E) + P(F|E)P(E) \qquad \text{Where I rewrote } P(F) \text{ as } C$$

We call $P(F)$ the normalization term because it is the term whose value can be calculated by making sure that the probability of all outcome sum to 1.

5 Conditional Paradigm

When you condition on an event you enter the universe where that event has taken place. In that new universe all the laws of probability 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**
Axiom of probability	$0 \leq P(E) \leq 1$	$0 \leq P(E\|G) \leq 1$
Axiom of probability	$P(E) = 1 - P(E^C)$	$P(E\|G) = 1 - P(E^C\|G)$
Chain Rule	$P(E \cap F) = P(E\|F)P(F)$	$P(E \cap F\|G) = P(E\|FG)P(FG)$
Bayes Theorem	$P(E\|F) = \frac{P(F\|E)P(E)}{P(F)}$	$P(E\|FG) = \frac{P(F\|EG)P(E\|G)}{P(F\|G)}$

Other properties like mutual exclusion and independence also exist in the world where we have conditioned on an event (and the rules for calculating probabilities when these conditions hold still apply). However, as a word of caution, conditioning on a new event could change the independence or mutual/exclusion relationship between two events.

6 Independence Revisited

Our new understanding of conditional probability can give us a fresh insight into the independence. It also introduces a new concept: conditional independence.

First, lets consider the definition of conditional probability for independent events:

$$
\begin{aligned}
P(E|F) &= \frac{P(E \cap F)}{P(F)} && \text{Definition of conditional probability} \\
&= \frac{P(E)P(F)}{P(F)} && \text{Since } E \text{ and } F \text{ are independent} \\
&= P(E) && \text{Since the } P(F) \text{ terms cancel out}
\end{aligned}
$$

Similarly $P(F|E) = P(E)$. Not only does this give us a new formula when working with independent events, it gives another angle for understanding what independence means. If two events are independent, knowing that one event has occurred gives you no additional information that the other event will occurred. The do not affect one another.

6.1 Conditional Independence

Conditioning on any event can have a dramatic effect on the independence relationships of any pair of events. Events that were previously independent can become dependent. Events that were previously independent can become dependent.

Two events E and F are called conditionally independent given G, if

$$P(E \cap F|G) = P(E|G)P(F|G)$$

Or, equivalently:

$$P(E|F,G) = P(E|G)$$

The two events can be conditionally independent, even if they are dependent without conditioning.

6.2 Breaking Independence

This point does have a large impact on the rest of CS109. But I wanted to pass on this surprising insight so that you have a more full understanding of probability. If two events E and F are independent, it is possible that there exists another event G such that $E|G$ is no longer independent of $F|G$.

As an example, Let's say a person has a fever (G) if they either have malaria (E) or have an infection (F). We are going to assume that getting malaria (E) and having an infection (F) are independent: knowing if a person has malaria does not tell us if they have an infection. Now, a patient walks into a hospital with a fever (G). Your belief that the patient has malaria, given that they have a fever, $P(E|G)$, is high and your belief that the patient has an infection given that they have a fever, $P(F|G)$, is high. Both explain the fever.

Interestingly, at this point in time (conditioned on G), given our knowledge that the patient has a fever, gaining the knowledge that the patient has malaria will change your belief the patient has an infection! The malaria **explains** why the patient has a fever, and so the alternate explanation becomes less likely. The two events (which were previously independent) are dependent now that we have conditioned on the patient having a fever.