2. 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 Event Space and Sample Space

A sample space, S, is set of all possible outcomes of an experiment. For example:

1. Coin flip: S = Head, Tails
2. Flipping two coins: $S = \{$(H, H), (H, T), (T, H), (T, T)$\}$
3. Roll of 6-sided die: $S = \{1, 2, 3, 4, 5, 6\}$
4. # emails in a day: $S = \{x | x \in \mathbb{Z}, x \geq 0\}$ (non-neg. ints)
5. YouTube hours in a day: $S = \{x | x \in \mathbb{R}, 0 \leq x \leq 24\}$

Event Space, E, is some subset of S that we ascribe meaning to. In set notation ($E \subseteq S$).For example:

1. Coin flip is heads: E = Head
2. ≥ 1 head on 2 coin flips = $\{$(H, H), (H, T), (T, H)$\}$
3. Roll of die is 3 or less: $E = \{1, 2, 3\}$
4. # emails in a day ≤ 20: $E = \{x | x \in \mathbb{Z}, 0 \leq x \leq 20\}$ (non-neg. ints)
5. Wasted day (≥ 5 YouTube hours): $E = \{x | x \in \mathbb{R}, 5 \leq x \leq 24\}$

3 Probability

In the 20th century humans figured out a way to precisely define what a probability is:

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

In English this reads: lets say you perform n trials of an experiment. The probability of a desired event E is the ratio of trials that result in E to the number of trials performed (in the limit as your number of trials approaches infinity).

That is mathematically rigorous. You can also apply other semantics to the concept of a probability. One common meaning ascribed is that $P(E)$ is a measure of the chance of E occurring.

I often think of a probability in another way: I don't know everything about the world. So it goes. As a result I have to come up with a way of expressing my belief that E will happen given my limited knowledge. This interpretation acknowledges that there are two sources of probabilities: natural randomness and our own uncertainty.

The different interpretations of probability are reflected in the many origins of probabilities that you will encounter in the wild (and not so wild) world. Some probabilities are calculated analytically using mathematical proofs. Some probabilities are calculated from data, experiments or simulations. Some probabilities are just made up to represent a belief.

Most probabilities are generated from a combination of the above: someone will make up a prior belief, that belief will be mathematically updated using data and evidence.

4 Axioms of Probability

Here are some basic truths about probabilities that we accept as axioms:

Basic Truth 1: $0 \leq P(E) \leq 1$
Basic Truth 2: $P(S) = 1$
Basic Truth 3: $P(E) = 1 - P(E^C)$

You can convince yourself of the first basic truth by thinking about the math definition of probability. As you perform trials of an experiment it is not possible to get more events then trials (thus probabilities are less than 1) and its not possible to get less than 0 occurrences of the event.

The second basic truth makes sense too. If your event space is the sample space, then each trial must produce the event. This is sort of like saying; the probability of you eating cake (event space) if you eat cake (sample space) is 1.

The third basic truth of probability is possibly the most useful. For any event, every element is either in the event or not in the event. A wonderful implication of this is that everything in the world (including you friends, your family, your dinner) is either a potato or not a potato. That holds for all objects.

5 Equally Likely Outcomes

Some sample spaces have equally likely outcomes. We like those sample spaces, because there is a way to calculate probability questions about those sample spaces simply by counting. Here are a few examples where there are discrete outcomes:

1. Coin flip: $S = \{\text{Head, Tails}\}$
2. Flipping two coins: $S = \{\text{(H, H), (H, T), (T, H), (T, T)}\}$
3. Roll of 6-sided die: $S = \{1, 2, 3, 4, 5, 6\}$

Because every outcome is equally likely, and the probability of the sample space must be 1, we can prove that each outcome must have probability:

$$P(\text{an outcome}) = \frac{1}{S}$$

A direct consequence is this formula for the probability of an event that is a subset of a sample space with equally likely outcomes.

Probability of Equally Likely Outcomes If S is a sample space with equally likely outcomes, for an event E that is a subset of the outcomes in S:

$$P(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S} = \frac{|E|}{|S|}$$

Interestingly, this idea also applies to continuous sample spaces. Consider the sample space of all the outcomes of the computer function "random" which produces a real valued number between 0 and 1, where all real valued numbers are equally likely. Now consider the event E that the number generated is in the range [0.3 to 0.7]. Since the sample space is equally likely, $P(E)$ is the ratio of the size of E to the size of S. In this case $P(E) = 0.4$.

6 Probabilities From Computers

The rise in compute power and the abundance of digital data means that many probabilities are calculated by computers, either through simulations or through counting on datasets.

6.1 Probabilities from Data

Machine Learning (sometimes called Data Science) is the love child of: probability, data and computers. Sometimes machine learning involves complex algorithms. But often it's just the core ideas of probability applied to large datasets.

As an example let us consider Netflix, a company that has thrived because of well thought out machine learning. One of the primary probabilities that they calculate is the probability that a user has watched a given movie.

Let E be the event that a user has watched a given movie. Since the number of users is large we can approximate P(E) using the definition of probability, which comes out to some simple counting:

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n} \approx \frac{\text{number of users who watch the movie}}{\text{number of users}}$$

As an important caveat: this only allows you to make a conclusion relative to the data you computed your probability on. If, for example, your dataset is missing users from Uganda, then you are calculating the probability that users of netflix, who are not in Uganda, has watched the movie (as opposed to, say, the probability that a random person in the world has watched the movie). There are elegant ways of dealing with missing data, which we leave for further reading.

6.2 Simulations

Another way to compute probabilities is via simulation. For some complex problems where the probabilities are too hard to compute analytically (for example, because there are complex constraints) you can run simulations using your computers random number generator.

If your simulations generate random instances from the sample space, then the probability of an event E is approximately equal to the fraction of simulations that produced an outcome from E. Again, by the definition of probability, as your number of simulations approaches infinity, the estimate becomes more accurate.

While this is a powerful tool, sometimes a simulation would take too long to generate enough samples that satisfy the event. As a result many probabilistic programs first start with a computer scientist calculating probabilities using probability theory.

7 Probability of OR

How you calculate the probability of either event A or event B happening, written $P(A \cup B)$, is truly analogous to counting the size of outcome spaces. The equation that you can use to calculate the probability of the "or" of two events depends on whether or not they are "mutually exclusive".

7.1 Mutually Exclusive Events

Two, events (E and F) are considered to be mutually exclusive ($E \cap F = \varnothing$) if there are no outcomes that are in both event spaces (recall that all event has a corresponding event space, which are subsets of the sample space).

Mutual exclusion can be visualized. Consider the following visual sample space where each outcome is a hexagon. The set of all the fifty hexagons is the full sample space:

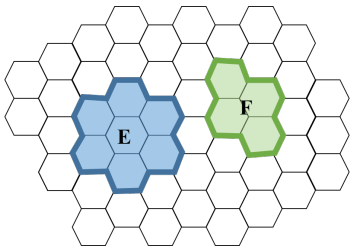

Figure 1: Mutually Exclusive Events E and F

Both events E and F have event spaces that are subsets of the same sample space. Visually, we can note that the two sets do not overlap. They are mutually exclusive: there is no outcome that is in both sets.

Probability of Mutually Exclusive Events
If two events, E and F are mutually exclusive then

$$P(E \cup F) = P(E) + P(F)$$

This property applies regardless of whether or not E and F are from an equally likely sample space. Moreover, the idea extends to more than two events. Lets say you have events $X_1, X_2 \ldots X_n$ where each event is mutually exclusive of one another. Then:

$$P(X_1 \cup X_2 \cup \cdots \cup X_n) = \sum_{i=1}^{n} P(X_i)$$

In other words, if you are told that two events are mutually exclusive (or you can argue that they are), calculating the probability of the intersection ("or") of those events is truly straightforward. In the example in figure 1, if we consider the outcomes in the sample space to be equally likely, then $P(E) = 7/50$ since event E has 7 outcomes and $P(F) = 4/50$ since event F has 4 outcomes. Because they are mutually exclusive, we can calculate that:

$$P(E \cup F) = P(E) + P(F) = \frac{7}{50} + \frac{4}{50} = \frac{11}{50} = 0.22$$

7.2 Inclusion Exclusion

Unfortunately, not all events are mutually exclusive. If you want to calculate $P(E \cup F)$ where the events E and F are **not** mutually exclusive you need to account for your double counting using the inclusion exclusion principle:

The Inclusion Exclusion Principle
For any two events, E and F:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

This property applies regardless of whether or not E and F are from an equally likely sample space and whether or not the events are mutually exclusive. The Inclusion-Exclusion principle extends to more than two events, but becomes considerably more complex. For events $E_1, E_2 \ldots E_n$:

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{r=1}^{n} (-1)^{r+1} Y_r$$

Where Y_r is the sum of the probability of the union, for all subsets of events where the size of the subset is r. Expanded out this becomes: add all the probabilities of each of the events on their own. Then subtract the probability of the union ("and") of all combinations of two events. Then add the probability of the union of all combinations of three events and so on for all powerset sizes until n, changing from add to subtracting each time.

For three events, E, F, and G that formula mathematically expands to:

$$\begin{aligned} P(E \cup F \cup G) = &+ P(E) + P(F) + P(G) \\ &- P(E \cap F) - P(E \cap G) - P(F \cap G) \\ &+ P(E \cap F \cap G) \end{aligned}$$

Phew! Probabilities of the "or" of events gets a ton harder when those events are not mutually exclusive!

8 Probability of AND

How you calculate the probability of event E and event F happening, written $P(E \cap F)$, or sometimes $P(EF)$ depends on whether or not they are "independent".

8.1 Independence

Independence is a big deal for machine learning and probabilistic modelling. Knowing the "joint" probability of many events (the probability of the "and" of the events) requires exponential amounts of data. By making independence and conditional independence claims, computers can essentially decompose how to calculate the joint, making it faster to compute, and meaning the data needs less data to learn probabilities.

Two events E and F are called independent if: $P(E \cap F) = P(E)P(F)$. Otherwise, they are called dependent events.

Independence
If two events, E and F, are independent:

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

This property applies regardless of whether or not E and F are from an equally likely sample space and whether or not the events are mutually exclusive. The independence principle extends to more than two events. For events $E_1, E_2 \ldots E_n$ where all events are independent of one another:

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = \prod_{i=1}^{n} P(E_i)$$

In the same way that the mutual exclusion property makes it easier to calculate the probability of the OR of two events, independence makes it easier to calculate the AND of two events. In the case where two events are not independent, there are still ways to calculate their probability. To do so we are going to need the concept from the next chapter: Conditional Probability.

9 Problem Solving Strategies

Before we move on to gather more probability tools, lets pause and talk about the strategy of solving probability problems that involve events.

Define your events. As a very first step, give a symbol to all the relevant events in the question. Then, as a next step, write the question you are trying to solve in terms of those events.

Be on the lookout for equally likely sample spaces – probabilities are much easier to compute if the sample space has outcomes that are all equally likely.

Sometimes, it may help to think of the underlying objects in your problem as distinct to make outcomes equally likely. For example, if you are considering the probability that a random permutation of the letters in the word "mississippi" have all four "i"s together, thinking of the letters as distinct means that the sample space of their distinct permutations are all equally likely. If you treat letters as indistinct the permutations are not equally likely.

Look for events which are independent or mutually exclusive. Sometimes you are told you can assume that a particular group of events are either independent or mutually exclusive. Other times you can make a logical argument that they are (for example if you are hashing strings into a hashmap,the event that a string hashes to one bucket is mutually exclusive from the event that it hashes to another). Remember that if events are mutually exclusive, it is easy to compute the OR of the events. If events are independent, it is easy to compute the AND of events.

You can use complements and De-Morgan's law to turn ANDs into ORs (and vice versa).

De Morgan's Law for Probability For any two events E and F:

$$P((E \cap F)^C) = P(E^C \cup F^C) \qquad \text{Version 1}$$
$$P((E \cup F)^C) = P(E^C \cap F^C) \qquad \text{Version 2}$$

This often is used in conjunction with the rule that the sum of an event and its complement is 1:

$$
\begin{aligned}
P(E \cap F) &= 1 - P((E \cap F)^C) && \text{Since } P(E) + P(E^C) = 1 \\
&= 1 - P(E^C \cup F^C) && \text{By DeMorgan's law} \\
P(E \cup F) &= 1 - P((E \cup F)^C) && \text{Since } P(E) + P(E^C) = 1 \\
&= 1 - P(E^C \cap B^F) && \text{By DeMorgan's law}
\end{aligned}
$$

Finally, there are often many ways to solve a problem. Try and come up with more than one. In all cases sanity check your answers. All probabilities should be positive and less than or equal to 1.