Review of Basic Probability

Erik G. Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

September 4, 2013

Abstract

This document reviews basic discrete probability theory.

1 Introduction

Artificial intelligence deals with making decisions in the real world, often in the presence of great uncertainty. To make the best decisions, it is important to address this uncertainty head on, and to try to find ways to make decisions which are likely to be beneficial, even when we are not sure.

While there are many ways to model uncertainty, one of the most popular and most successful has grown out of probability theory. As such, an understanding of basic probability theory is critical to the understanding of modern artificial intelligence and related fields, such as computer vision and natural language processing.

This guide is meant to be practical rather than rigorous. While we will attempt to avoid inaccurate statements, the goal is usability rather than completeness or rigor.

2 The Sample Space

In applying probability theory to a particular problem, it is essential to have a solid grasp on the *sample space*, i.e., the set of all possible experimental results, or *outcomes*, in a given experiment. In some experiments, the choice of sample space is clear. In other cases, there may be more than one possible choice, and making the proper choice can significantly effect the ease with which subsequent calculations can be done.

sample space

outcomes

Example 1. In rolling a single six-sided die, the space of outcomes would typically be considered to be the set of six possible faces that might face up on any given roll. We denote the sample space $S = \{1, 2, 3, 4, 5, 6\}$. Letting X be a *random variable* representing the outcome of a roll, we write $X = x$, where $x \in S$.

random variable

An *event* is any subset of a sample space. Examples of events in the space of single die rolls include the single roll 3 ($E = \{3\}$), a roll less than 5 ($E = \{1, 2, 3, 4\}$), a roll in which the result is prime ($E = \{2, 3, 5\}$), the empty event ($E = \{\}$), and the universal event ($E = \{1, 2, 3, 4, 5, 6\}$). If the face on the die is an element of E, then the event is said to have occurred; otherwise it did not occur. It is worth noting that the empty event will never occur, and that any roll will represent the occurrence of the universal event.

event

Example 2. Suppose we roll two dice, one of which is red and one of which is blue. Let R represent the outcome of the red die and B represent the outcome of the blue die. If we are playing a game like backgammon, then we do not care about the distinction between getting a 3 on the red die with a 4 on the blue die and getting a 4 on the red die with a 3 on the blue die. These are considered the same outcome. On the other hand, if we are using the dice to select one of 36 options from a six by six grid of ice cream flavors, then we must treat the result ($R = 3, B = 4$) as being distinct from ($R = 4, B = 3$).

In the latter case, the choice of sample space is clear: the set of events includes all 36 ordered pairs of die rolls $S = \{(1, 1), (1, 2,), (1, 3), ..., (2, 1), (2, 2), ..., (6, 5), (6, 6)\}$.

In the case where we do not care about which die has a particular value, but only on the two values that are obtained, there are two possible choices for the sample space. One choice would be to define the sample space as the set of all 21 possible *unordered* pairs $S = \{(1,1),(1,2),(1,3),...,(2,2),(2,3),...,(5,6),(6,6)\}$, in which we include (x,y) only if $x \leq y$. This may seem like the obvious choice.

However, another choice would be to defined the sample space as the set of all ordered pairs (as in the ice cream selection example), but to define the *events* of interest as being of the form $E = \{(R = x, B = y), (R = y, B = x)\}$. For example, we could define the event "a 3 and a 4" as the subset of events $\{(R = 3, B = 4), (R = 4, B = 3)\}$. One advantage of the latter approach is that the primitive elements of the sample space all have the same probability of occurrence (assuming the dice are fair), and this can make probability calculations easier in many cases. The best choice of sample space is not always obvious in the beginning. However, it is imperative that one makes a clear choice about how the sample space is defined. Many errors in applying probability come from an unclear definition of the sample space.

One more note about defining sample spaces. The events in the sample space must be *mutually exclusive*, i.e., two primitive events in the sample space cannot represent the same event. Hence, defining a sample space as the set of events $\{$(at least one die is a 1), (at least one die is a 2), ..., (at least one die is a 6)$\}$, represents an invalid sample space, since these events are not mutually exclusive.

mutually exclusive

2.1 Calculating probabilities of events

While there are extensive philosophical debates about the meaning of the probability of an event, I will avoid these discussions here. For the purposes of this review, you can think of the probability of an event as the proportion of times you would expect the event to occur in a very large number of trials. For example, you would expect about half of the rolls of a fair die to be even in a very large number of die rolls, and so it would be reasonable (for many, but not all, purposes) to adopt the probability of $\frac{1}{2}$ as the probability of getting an even number in a die roll.

Let us assume that we have chosen a sample space S for an experiment and that we are given the probability of each primitive event, i.e. an event which represents a single element of the sample space. We write $P(E)$ to denote the probability of an event E.

Example 3. Suppose we have an *unfair* die with the probability of each roll given as follows: $P(1) = \frac{1}{16}, P(2) = \frac{1}{16}, P(3) = \frac{1}{16}, P(4) = \frac{1}{16}, P(5) = \frac{1}{4}, P(6) = \frac{1}{2}$.

To calculate the probability of a new event E, we simply add the probabilities of the primitive events that compose it. For example, consider the event $E = \{1, 3\}$. The probability of this event is simply the probability of getting a 1 plus the probability of getting a 3, or $P(E) = P(1) + P(3) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$.

Example 4. Now, using the same unfair die as in the example above, consider a new event defined as $E = \{$"the event that the roll was even or greater

than 3"}. First, let us talk about the *wrong* way to compute this probability.

We must *not* compute this as the probability that the roll was even plus the probability that the roll was greater than 3. This would be tantamount to saying that $P(E) = P(\text{"even"}) + P(\text{"}> 3\text{"}) = P(2) + P(4) + P(6) + P(4) + P(5) + P(6)$. Notice here that we have counted the probability of the primitive event $\{4\}$ twice, and also counted the probability of the primitive event $\{6\}$ twice. We have *overcounted* the probability.

The proper way to calculate this probability is as follows. Consider the set of all primitive events defined by the event of interest. In this case, the set of primitive events that are either even or greater than 3 is $\{2, 4, 5, 6\}$. Once we have this set, which can be obtained by taking the *union* of the primitive elements of the individual events, then we can merely add the probabilities of the primitive events in this union. In other words, $P(E) = P(2) + P(4) + P(5) + P(6) = \frac{1}{16} + \frac{1}{16} + \frac{1}{4} + \frac{1}{2} = \frac{7}{8}$.

This fundamental method, of considering a new event as the union of primitive events in the sample space, makes calculating probabilities very simple and straightforward.

2.2 Joint Probabilities

Suppose that A and B are events defined on a sample space S. When we write $P(A, B)$, we mean the probability that *both* event A occured *and* event B occurred in the same trial of an experiment. This is called the *joint probability* of A and B.

joint probability

Example 5. Using the probabilities from the previous example, Let A be the event that a die roll is even, and B be the event that a die roll is greater than 3. To calculate the probability of both events occurring together, $P(A, B)$, we use the same basic strategy as in the previous example:

1. enumerate the primitive events which satisfy the given criterion,
2. add the probability of these primitive events.

Which primitive events satisfy the criterion that the die roll is even and that it is greater than 3? Clearly, the set of such events is $\{4, 6\}$. The sum of the probability of these primitive events is $P(A, B) = P(4) + P(6) = \frac{1}{16} + \frac{1}{2} = \frac{9}{16}$.

Note that the set of primitive events which satisfies *all* criteria from a set of criteria is the *intersection* of the primitive events of the individual criteria ($\{4, 6\} = \{2, 4, 6\} \cap \{4, 5, 6\}$). On the other hand, the set of primitive events satisfying *any* of a set of criteria is the *union* of the primitive events of the individual criteria ($\{2, 4, 5, 6\} = \{2, 4, 6\} \cup \{4, 5, 6\}$).

intersection

union

We end this subsection with a simple question. Is $P(A, B) = P(B, A)$ for all events A and B? To answer this question, it is sufficient to have a clear understanding of the definition of $P(A, B)$. In particular, we are considering the probability that both events A and B have occurred in a particular trial. Thus, our question reduces to the query: if A and B have occurred, can we also say that B and A have occurred? Since $A \cap B = B \cap A$, we can say that

$P(A, B)$ and $P(B, A)$ are referring to the same subset of primitive events, and hence always have the same probability.[1]

2.3 Marginalization

Let X and Y be two random variables, such as the outcomes of a blue die and a red die which are tossed together. If we are given the probabilities of all events $P(X = x, Y = y)$ in the joint sample space, then we can compute the probability of events involving only a single random variable, such as $P(Y = 3)$, through a process known as *marginalization*. In particular, we can say that *marginalization*

$$\begin{aligned} & P(Y = 3) \\ = \; & P(X = 1, Y = 3) + P(X = 2, Y = 3) + P(X = 3, Y = 3) \\ & \quad + P(X = 4, Y = 3) + P(X = 5, Y = 3) + P(X = 6, Y = 3) \\ = \; & \sum_{x=1}^{6} P(X = x, Y = 3). \end{aligned}$$

While in this example, we have used random variables representing primitive events, this marginalization procedure works for arbitrary random variables.

The preceding analysis discusses marginalization in the context of random variables. It is also of interest to consider how to compute the marginal probability of an event A, i.e., $P(A)$, rather than a random variable. For a pair of events A and B, an experiment can have four possible outcomes: $(A, B), (\tilde{A}, B), (A, \tilde{B}), (\tilde{A}, \tilde{B})$, where $\tilde{A}$ means the event A did not occur. We can then compute $P(A)$ as $P(A, B)+P(A, \tilde{B})$. That is, we have added the probability of the joint events in which A occurred for all possible outcomes of the event B.

2.4 Conditional Probability

When A and B are events on an event space S, we read $P(A|B)$ as *the probability of event A given that the event B has occurred on the same trial*, or more succinctly, *the probability of A given B*. This is also referred to as the *conditional probability* of A given B. *conditional probability*

To understand conditional probability, it is useful to consider the conditioning bar ("|") as defining a new sample space S' that is a subset of the original sample space S. In particular, if we condition on an event B (as in $P(A|B)$), we are defining a new sample space of primitive events from the original space S containing only events that are consistent with the event B. If B is the event that a die roll is greater than 3, then the sample space for $P(A|B)$ is $S' = \{4, 5, 6\}$.

[1] Unfortunately, shorthand notations in applied probability can lead to some confusions here. For example, some authors may write $P(3, 4)$ to mean $P(X = 3, Y = 4)$ where X and Y are random variables. In this case $P(3, 4) \neq P(4, 3)$ since the events $(X = 3, Y = 4)$ and $(X = 4, Y = 3)$ are different events. It is critical to keep in mind the exact meaning of what is written, and whether it should be interpreted as the values of particular random variables, or the occurrence of events. Confusion can be avoided by explicitly naming the random variable of interest. For example, we can still write $P(X = 3, Y = 4) = P(Y = 4, X = 3)$.

2.4.1 Computing conditional probabilities

There are two simple ways of computing the values of conditional probabilities, given an initial sample space S and the probabilities of each of the primitive events. We start with the identity

$$P(A, B) = P(A|B)P(B).$$

Dividing both sides by $P(B)$ yields

$$P(A|B) = \frac{P(A, B)}{P(B)}. \tag{1}$$

Hence, to compute $P(A|B)$ we can simply compute the two quantities on the right hand side and take their ratio. Note that $P(B)$ can be obtained from $P(A, B)$ through the marginalization process described earlier.

A second procedure is to first compute the probabilities of each primitive event in the new sample space S'. Let $P_S(E)$ be the probability of a primitive event E in the original sample space S and $P_{S'}(E)$ be the probability of the same event in the new induced sample space S' which results from conditioning on the event B. Then

$$P_{S'}(E) = \frac{P_S(E)}{P_S(B)},$$

for any event which is consistent with B. For example, assume we roll a fair die and are told that the result is even. Then, conditioned on the event B that the roll was even ($P(B) = \frac{1}{2}$), the probability of each of the rolls in the set $\{2, 4, 6\}$ would be $\frac{1}{3}$ since $\frac{1}{6}/\frac{1}{2} = \frac{1}{3}$.

These two procedures are algebraically equivalent. The first divides only the probability of the single joint event of interest (A, B) by $P(B)$. The second divides the probabilities of the primitive events by $P(B)$ before adding them together to form the joint event.

Example 6. Assume an unfair die with probabilities as in the example above: $P(1) = \frac{1}{16}, P(2) = \frac{1}{16}, P(3) = \frac{1}{16}, P(4) = \frac{1}{16}, P(5) = \frac{1}{4}, P(6) = \frac{1}{2}$. Let A be the event that a die roll is less than 5. Let B be the event that the die roll is prime. Compute $P(A|B)$.

Using Equation 1, we wish to compute $P(A, B)$ and $P(B)$. $P(A, B)$ is the probability of all primitive events (die rolls) which are both less than 5 and prime. This is the set of events $\{2, 3\}$. The probability of these events is $P(A, B) = P(2) + P(3) = \frac{1}{8}$. The probability of B is the set of events for which the die roll is prime, which is the set $\{2, 3, 5\}$. $P(B) = P(2) + P(3) + P(5) = \frac{3}{8}$. Finally, then, $P(A|B) = P(A, B)/P(B) = \frac{1}{8}/\frac{3}{8} = \frac{1}{3}$.

2.5 Bayes' Rule

Bayes' rule

Bayes' rule is used ubiquitously in applied probability to "reverse the direction" of conditioning. That is, it is frequently the case in practice that one has available the quantity $P(A|B)$ and one wishes to find the quantity $P(B|A)$. We will discuss the reasons for this later.

To derive Bayes' rule, we start with the observation that the identity for joint probability can be written in two forms:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A).$$

By simply dividing both sides of the latter equation by $P(B)$, we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

or, alternatively, dividing by $P(A)$, we have

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

These are both statements of Bayes' rule. One notes immediately that in order to derive $P(A|B)$ from $P(B|A)$, one also needs the quantities $P(A)$ and $P(B)$, i.e., the marginal probabilities.

Example 7. Consider two urns, A and B. Urn A contains 9 red balls and 1 black ball. Urn B contains 5 red balls and 5 black balls. Now consider the following experiment. A fair die is rolled. If the die shows a six, then urn A is selected. Otherwise urn B is selected. Now a ball is drawn from the selected urn, and its color is noted.

We can write down the following quantities. $P(\text{red}|A) = 0.9$. $P(\text{red}|B) = 0.5$. $P(A) = \frac{1}{6}$. $P(B) = \frac{5}{6}$. Suppose we are blind-folded and a friend runs the experiment described above. The friend reports that a red ball was selected. Our job is to compute the probability that the ball was drawn from urn A.

To solve this problem, we must first define the quantity of interest, which is $P(A|\text{red})$. To compute this quantity, we use Bayes' rule:

$$P(A|red) = \frac{P(red|A)P(A)}{P(red)}.$$

The two quantities in the numerator are immediately available, but the quantity in the denominator $P(red)$ requires a bit of work to obtain. To obtain $P(red)$, we first write it as a marginalization:

$$P(red) = \sum_{urn \in \{A,B\}} P(red, urn).$$

However we also don't have immediate access to the joint probabilities used in this sum, $P(red, A)$ and $P(red, B)$. To obtain these, we expand the joint probabilities using the conditional probability identity:

$$\sum_{urn \in \{A,B\}} P(red, urn) = \sum_{urn \in \{A,B\}} P(red|urn)P(urn).$$

Now all of the necessary quantities are immediately at hand, and we have merely to perform the necessary arithmetic:

$$P(A|red) = \frac{\frac{9}{10}\frac{1}{6}}{\frac{9}{10}\frac{1}{6} + \frac{1}{2}\frac{5}{6}} = \frac{\frac{9}{60}}{\frac{9}{60} + \frac{25}{60}} = \frac{9}{34}.$$