Probabilities and Expectations

Ashique Rupam Mahmood

September 9, 2015

Probabilities tell us about the likelihood of an event in numbers. If an event is certain to occur, such as sunrise, probability of that event is said to be 1. $\mathbf{Pr}(\texttt{sunrise}) = 1$. If an event will certainly not occur, then its probability is 0.

So, probability maps events to a number in $[0, 1]$. How do you specify an event? In the discussions of probabilities, events are technically described as a set. At this point it is important to go through some basic concepts of sets and maybe also functions.

Sets

A *set* is a collection of distinct objects. For example, if we toss a coin once, the set of all possible distinct outcomes will be $S = \{\texttt{head}, \texttt{tail}\}$, where `head` denotes a head and the `tail` denotes a tail. All sets we consider here are finite.

An *element* of a set is denoted as $\texttt{head} \in S$. A *subset* of a set is denoted as $\{\texttt{head}\} \subset S$. What are the possible subsets of S? These are: $\{\texttt{head}\}, \{\texttt{tail}\}, S = \{\texttt{head}, \texttt{tail}\}$, and $\phi = \{\}$. So, note that a set is a subset of itself: $S \subset S$. Also note that, an *empty set* (a collection of nothing) is a subset of any set: $\phi \subset S$. A *union* of two sets A and B is comprised of all the elements of both sets and denoted as $A \cup B$. An *intersection* of two sets A and B is comprised of only the common elements of both sets and denoted as $A \cap B$. A *complement* set of A in B is a set comprising the elements of B that are not in A and denoted as $B \setminus A$ or $B - A$. The *Cartesian product* of two sets A and B is a set denoted as $A \times B$ comprising all ordered pairs (a, b) where $a \in A$ and $b \in B$.

More about Sets (can be skipped)

A *power set* of a set S is a set of all the subsets of S. It is often denoted as 2^S. In our example, it is $2^S = \{\{\texttt{head}\}, \{\texttt{tail}\}, S, \phi\}$. Therefore, $S \in 2^S$. How many distinct elements we have in 2^S? It is 2^n, if n is the number of distinct elements in S.

Functions

A *function* is a map from a one set to another. A function takes an argument and associates it with exactly one quantity, which is often called the *output* of a function. The set of all the arguments a function takes is called the *domain*. The set of all the outputs that a function associates with is called a *codomain*, sometimes also known as *range*.

For example, if we want to associates the outcomes of a coin toss, `head` and `tail`, with numbers 1 and 0, we can do so by using functions. We can say, $f(\texttt{head}) = 1$ and $f(\texttt{tail}) = -1$. Then the domain of the function f is $S = \{\texttt{head}, \texttt{tail}\}$ and the codomain is $V = \{1, -1\}$. This function can also be denoted as $f : S \to V$.

Now, note that if a relation f associates more than one output with an argument, e.g., $f(\texttt{head}) = 1$ or $f(\texttt{head}) = 0$, then f is not a function. However, a function can map two different arguments to a single output. Hence, if $f(\texttt{head}) = 1$ and $f(\texttt{tail}) = 1$, then f is a function.

Sample Space, Events & Probabilities

Let us assume there is an *experiment* that is repeatable, such as rolling a dice. There are two important elements that make up a *probabilistic model*: a sample space and a probability distribution.

A *sample space* is the set of all possible outcomes of an experiment. Therefore, in the dice-rolling experiment, the sample space S is $\{1, 2, 3, 4, 5, 6\}$. An *event* is any subset of the sample space. For example, the event that a number more than two would appear in the dice-rolling experiment is $\{3, 4, 5, 6\}$. It makes sense to find the probability of such an event. An event can comprise only one outcome, for example, the event that 2 will appear: $\{2\}$ or simply 2. A *complementary event* of an event A is $A^c = S - A$.

Now, we are ready to talk about probabilities. A probability is a function that associates an event with a non negative number. Therefore, for any event A

$$\mathbf{Pr}(A) \geq 0. \tag{1}$$

A probability function has certain key properties. For example, the addition of probabilities of all the outcomes is always 1:

$$\sum_{e \in S} \mathbf{Pr}(e) = 1. \tag{2}$$

How the probability is distributed among the outcomes is defined by the *probability distribution*. For example, for the experiment of rolling an unbiased dice, the probability distri-

bution can be uniform, that is, it is the same for all the outcomes: for each outcome $e \in S$, $\mathbf{Pr}(e) = 1/6$.

There are other properties of a probability function. For example, if two events A and B are disjoint (no common elements), then the probability of their union is the addition of their separate probabilities:

$$\mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B); \quad A \cap B = \phi. \tag{3}$$

The outcomes of an experiment are always mutually disjoint.

So, we talked about three properties of a probability function: nonnegativity (equation 1), normalization (equation 2) and additivity (equation 3).

If we are given to find the probability of an event, one way would be to find the sample space and the probability distribution among the outcomes. Then we need to specify the event in question. Then using the additive rule it becomes easy to find the probability of that event.

Now, we are ready to calculate the probability that a number more than 2 will appear in the unbiased dice rolling experiment. Using the properties in the above, we can write:

$$\begin{aligned} \mathbf{Pr}(\text{observe a number more than 2}) &= \mathbf{Pr}(\{3,4,5,6\}) \\ &= \mathbf{Pr}(3) + \mathbf{Pr}(4) + \mathbf{Pr}(5) + \mathbf{Pr}(6);\ \text{additivity} \\ &= 4 \times 1/6 = 2/3. \end{aligned}$$

More about Events & Probabilities (Can be skipped)

How many distinct events are possible in a dice rolling experiment? It is the number of all possible subsets of the sample space. The set of all possible distinct events is the power set of the sample space. The empty set ϕ is a possible event. An empty-set event is the event where nothing appears in a dice rolling experiment. Probability of an empty-set event is zero. This is an event that will certainly not occur. The sample space itself is also an event. It denotes the event that either of the possible outcomes would appear in the dice rolling experiment. This event will certainly occur. Probability of the sample space is always one: $\mathbf{Pr}(S) = 1$.

So, now we understand that probability is in fact a function that takes an event as the argument and gives a number between 0 and 1, inclusively, as an output. We can write $\mathbf{Pr} : 2^S \rightarrow [0, 1]$.

Random Variables

It is often more convenient to express events in relations with random variables. A random variable takes the possible values of the outcomes of an experiment. In the dice-rolling experiment, a random variable X would be such that $X \in \{1, 2, 3, 4, 5, 6\}$.

We can now write the event of having a number more than 2 as $X > 2$. Here, $X > 2$ stands for the event $\{3, 4, 5, 6\}$. For the events with only one outcome, we usually write as $X = 2$ or $X = 3$ or $X = x$ and so on. When we use letters, random variables will be denoted by capital letters and values of a random variable will be denoted by small letters.

Examples of Probabilities

In the dice rolling example, if the random variable X stands for the outcome, what is $\mathbf{Pr}(X > 1)$? Now,

$$\begin{aligned}
\mathbf{Pr}(1 \leq X \leq 6) &= 1 \\
\mathbf{Pr}([X = 1] \cup [2 \leq X \leq 6]) &= 1 \\
\mathbf{Pr}([X = 1] \cup [X > 1]) &= 1 \\
\mathbf{Pr}(X = 1) + \mathbf{Pr}(X > 1) &= 1 \\
\mathbf{Pr}(X > 1) &= 1 - \mathbf{Pr}(X = 1) \\
\mathbf{Pr}(X > 1) &= 5/6.
\end{aligned}$$

Conditional Probabilities

Probability of an event A given another event B is defined as

$$\mathbf{Pr}(A|B) \doteq \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}.$$

Here $\doteq$ stands for equality by definition rather than by derivation. $\mathbf{Pr}(A \cap B)$ is also written as $\mathbf{Pr}(A, B)$ or $\mathbf{Pr}(A \& B)$.

When an event is already given or known to have happened, the uncertainty of another event might change. For example, in the dice rolling experiment, $\mathbf{Pr}(3)$ is $1/6$. However, if we already know that the outcome is an odd number, does the event of observing 3 remains uncertain in the same amount? Actually not. It is now more likely to see 3 than before. Conditional probability reflects such a case. It can be calculated in the following way:

$$\mathbf{Pr}(X = 3|X \text{ is odd}) = \frac{\mathbf{Pr}(X = 3 \cap X \text{ is odd})}{\mathbf{Pr}(\text{X is odd})} = \frac{\mathbf{Pr}(3 \cap \{1, 3, 5\})}{\mathbf{Pr}(\{1, 3, 5\})} = \frac{\mathbf{Pr}(3)}{\mathbf{Pr}(\{1, 3, 5\})} = \frac{1/6}{3/6} = \frac{1}{3}.$$

The uncertainty of observing 3 has indeed changed. It has become more likely to be observed.

Another example. What is the probability to observe an odd number, when we know that it is a multiple of 3?

$$\mathbf{Pr}(X \text{ is odd}|X \text{ is a multiple of 3}) = \frac{\mathbf{Pr}(X \text{ is odd} \cap X \text{ is a multiple of 3})}{\mathbf{Pr}(\text{ is a multiple of 3})} = \frac{\mathbf{Pr}(\{1,3,5\} \cap \{3,6\})}{\mathbf{Pr}(\{3,6\})} = \frac{\mathbf{Pr}(3)}{\mathbf{Pr}(\{3,6\})} = \frac{1/6}{2/6} = \frac{1}{2}.$$

When the likelihood of an event does not change after knowing another event, then those two events are said to be independent:

$$[A \text{ is independent of } B] \Leftrightarrow [\mathbf{Pr}(A|B) = \mathbf{Pr}(A)]\,.$$

Following are also equivalent to the above:

$$\begin{aligned} \mathbf{Pr}(A,B) &= \mathbf{Pr}(A)\mathbf{Pr}(B) \\ \mathbf{Pr}(B|A) &= \mathbf{Pr}(B). \end{aligned}$$

When more than two events are involved in a conditional probability, then the following holds:

$$\mathbf{Pr}(A,B|C) = \mathbf{Pr}(A|B,C)\mathbf{Pr}(B|C).$$

Conditional probabilities are particularly important for compound experiments.

Examples

Let us say we have two urns. The first one contains 2 red balls and 1 black balls, and the second one contains 1 red ball and 1 black ball. If we pick one of the urns randomly and then choose one of the balls randomly, what is the probability that it will be red? What is the sample space here? It is $S = \{1r, 1b, 2r, 2b\}$. Here, $1r$, for example, stands for observing a red ball from urn 1.

What is the probability distribution among the outcomes? It is not straightforward to find as before anymore. In fact, this is a compound experiment comprising two experiments: 1) choosing the urn with sample space $S_1 = \{1, 2\}$ and 2) choosing a ball within an urn with sample space $S_2 = \{r, b\}$. Here, S is a compound sample space constructed from the Cartesian product of S_1 and S_2 as $S = S_1 \times S_2 = \{1, 2\} \times \{r, b\} = \{1r, 1b, 2r, 2b\}$.

Let us consider X_1 be the random variable for the outcome of choosing an urn (over S_1)

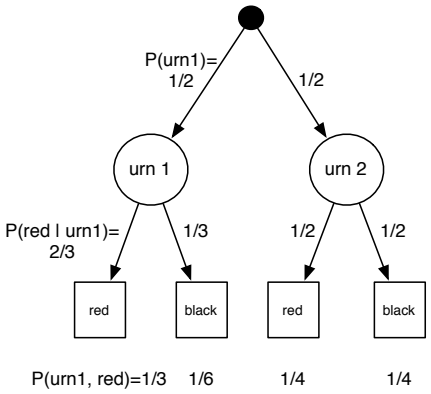
and X_2 be the random variable for the outcome of choosing a ball (over S_2). Then

$$\begin{aligned}\mathbf{Pr}(1r) = \mathbf{Pr}(X_1 = 1, X_2 = r) &= \mathbf{Pr}(X_1 = 1)\mathbf{Pr}(X_2 = r|X_1 = 1) \\ &= \mathbf{Pr}(1)\mathbf{Pr}(r|1) \\ &= 1/2 \times 2/3 \\ &= 1/3.\end{aligned}$$

In the same way, $\mathbf{Pr}(1b) = 1/2 \times 1/3 = 1/6, \mathbf{Pr}(2r) = 1/2 \times 1/2 = 1/4$ and $\mathbf{Pr}(2b) = 1/2 \times 1/2 = 1/4$.

Now, how do we specify the event in question (what is the probability that it will be red)? It is $A = \{1r, 2r\}$. So, $\mathbf{Pr}(A) = \mathbf{Pr}(1r) + \mathbf{Pr}(2r) = 7/12$.

It is often more convenient to use trees for conditional probabilities:



Bayes Theorem

Let us say we know the probability of an event B given another event A. Now we want to know the probability of A given B. Bayes theorem helps to find that. It is simply written as this:

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(B|A)\mathbf{Pr}(A)}{\mathbf{Pr}(B)}. \tag{4}$$

It can be easily derived from the definition of conditional probability.

Now, let us say, there are n mutually disjoint events A_j, where $j = 1, 2, \ldots, n$ and their union is the sample space: $\cup_{j=1}^{n} A_j = S$. Then probability of any event $B \subset S$ can be written in the following way:

$$\begin{aligned}\mathbf{Pr}(B) &= \mathbf{Pr}(B \cap S); \text{ as } B \text{ is completely contained in } S \\ &= \mathbf{Pr}(B \cap (\cup_j A_j)); \text{ as given above}\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{Pr}(\cup_j (B \cap A_j)) \\
&= \sum_j \mathbf{Pr}(B \cap A_j);\ \text{as all } B \cap A_j \text{ are disjoint} \\
&= \sum_j \mathbf{Pr}(B|A_j)\mathbf{Pr}(A_j).\ \text{definition of conditional probability.}
\end{aligned}
$$

The above result is sometimes called *the law of total probability*. Replacing $P(B)$ in 4 with the above, we get a more general Bayes theorem:

$$
\mathbf{Pr}(A_i|B) = \frac{\mathbf{Pr}(B|A_i)\mathbf{Pr}(A_i)}{\sum_j \mathbf{Pr}(B|A_j)\mathbf{Pr}(A_j)}.
$$

Let us look at a classic example of Bayes theorem. Let us say that for a drug test, it returns positive result for a drug user 99% of the time and produces a negative result for a non-user 95% of the time. Suppose that 1% of the population uses drug. Then what is the probability that an individual is a drug user given that she tests positive?

Here, the sample space is constituted of $\{\texttt{user}+, \texttt{user}-, \texttt{nonuser}+, \texttt{nonuser}-\}$. Here, $\texttt{user}-$, for example, stands for the event that an individual is a drug user and he tests negative. We actually do not know the probability distribution here. Instead, we only know some of the conditional probabilities: $\mathbf{Pr}(+|\texttt{user}) = 0.99, \mathbf{Pr}(-|\texttt{nonuser}) = 0.95, \mathbf{Pr}(\texttt{user}) = 0.01$. We want to find $\mathbf{Pr}(\texttt{user}|+)$. Using Bayes theorem, we get

$$
\begin{aligned}
\mathbf{Pr}(\texttt{user}|+) &= \frac{\mathbf{Pr}(+|\texttt{user})\mathbf{Pr}(\texttt{user})}{\mathbf{Pr}(+|\texttt{user})\mathbf{Pr}(\texttt{user}) + \mathbf{Pr}(+|\texttt{nonuser})\mathbf{Pr}(\texttt{nonuser})} \\
&= \frac{0.99 \times 0.01}{0.99 \times 0.01 + (1 - \mathbf{Pr}(-|\texttt{nonuser})) \times (1 - \mathbf{Pr}(\texttt{user}))} \\
&= \frac{0.0099}{0.0099 + (1 - 0.95) \times (1 - 0.01)} \\
&= \frac{0.0099}{0.0099 + 0.05 \times 0.99} \\
&= \frac{0.0099}{0.0099 + 0.0495} \\
&\approx 0.167.
\end{aligned}
$$

this is somewhat surprising and interesting. Even though the stats on the test sounded good, it is still rather unlikely that the person is a drug user even when she tests positive. Bayes theorem helped us realize that.

Expectations

Random variables can have values that are numbers, like an urn number, or they may have non-numeric values, like the color of a ball. If a random variable has exclusively numeric outcomes, then we can talk about its expected value, or expectation. The expected value of a numeric random variable is a weighted average of its possible numeric outcomes, where the weights are the probabilities of the outcome occurring.

When all the outcomes are equally likely, the expectation is the same as the average of possible outcomes. For a discrete random variable (sample space is a finite set), its expectation is defined as

$$\mathbf{E}[X] \doteq \sum_{x \in S} x \mathbf{Pr}(X = x).$$

As any function of a random variable is also a random variable, it is possible to calculate the expectation of a function of a random variable, too. It can be calculated in the following way:

$$\mathbf{E}[g(X)] \doteq \sum_{x \in S} g(x) \mathbf{Pr}(X = x).$$

There are certain properties of expectations:

$$\begin{aligned} \mathbf{E}[X + c] &= \mathbf{E}[X] + c; \text{ where } c \text{ is not a r.v. of the model} \\ \mathbf{E}[X + Y] &= \mathbf{E}[X] + \mathbf{E}[Y] \\ \mathbf{E}[aX] &= a\mathbf{E}[X]; \text{ where } a \text{ is not a r.v. of the model.} \end{aligned}$$

In a certain lottery, it is 0.01% likely to win and the prize is 1000 dollars. The ticket price is 10 dollars. The expected monetary gain from the lottery is

$$\begin{aligned} \mathbf{E}[X] &= (1000 - 10) \times 0.0001 + (-10) \times 0.9999 \\ &= 0.099 - 9.999 \\ &= -9.9. \end{aligned}$$

It is almost the same as giving ten dollars away.

Conditional Expectations

A conditional expectation of a random variable is the expected value of the variable given that an event is already known to have happened. For discrete variables:

$$\mathbf{E}[X|Y=y] \doteq \sum_{x\in S} x\mathbf{Pr}(X=x|Y=y).$$

Examples

Let us say, in the two urns example, we get 10 dollars for observing a red ball and we get 5 dollars for observing a black ball. What is the expected gain of this experiment? Here, the sample space is $S=\{1r,1b,2r,2b\}$.

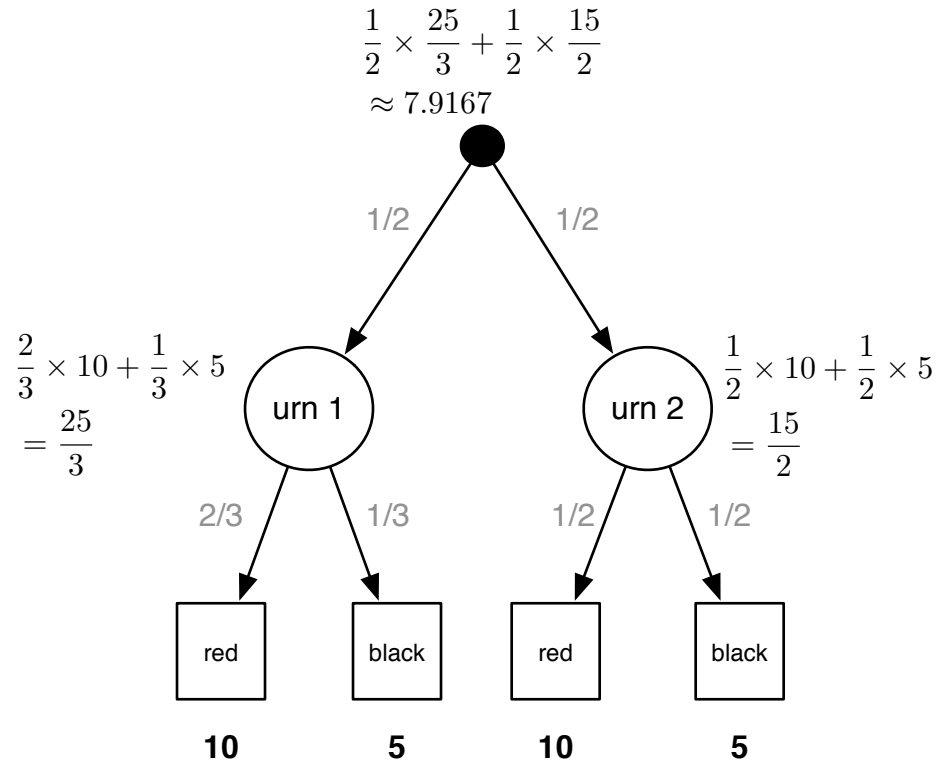
$$\begin{aligned}
\mathbf{E}[gain(X)] &= \sum_{x\in S} gain(x)\mathbf{Pr}(X=x) \\
&= gain(1r)\mathbf{Pr}(X=1r) + gain(1b)\mathbf{Pr}(X=1b) \\
&\quad + gain(2r)\mathbf{Pr}(X=2r) + gain(2b)\mathbf{Pr}(X=2b) \\
&= 10\times\frac{1}{3} + 5\times\frac{1}{6} + 10\times\frac{1}{4} + 5\times\frac{1}{4} \\
&\approx 7.9167.
\end{aligned}$$

Now, let us consider that we can decide which urn to choose. What is the expected gain if we choose urn 1? Let us define another random variable Y for the urn chosen.

$$\begin{aligned}
\mathbf{E}[gain(X)|Y=1] &= \sum_{x\in S} gain(x)\mathbf{Pr}(X=x|Y=1) \\
&= gain(1r)\mathbf{Pr}(X=1r|Y=1) + gain(1b)\mathbf{Pr}(X=1b|Y=1) \\
&\quad + gain(2r)\mathbf{Pr}(X=2r|Y=1) + gain(2b)\mathbf{Pr}(X=2b|Y=1) \\
&= 10\times 2/3 + 5\times 1/3 + 0 + 0 \\
&= 25/3.
\end{aligned}$$

Similarly, we can calculate that the expected gain of choosing urn 2 is $10\times\frac{1}{2}+5\times\frac{1}{2}=15/2$. So, it tells that choosing the first urn is more profitable.

Once again, it is often easier to use trees to find such values. The following tree shows the expected gain at each decision point:



Law of Total Expectation

Following is a very useful proposition often known as the law of total expectation:

$$\mathbf{E}[X] = \sum_y \mathbf{E}\left[X|Y=y\right] \mathbf{Pr}(Y=y).$$

It is analogous to the law of total probability that we derived for the Bayes theorem. The derivation of it as follows:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_x x\mathbf{Pr}(X=x) \\
&= \sum_x x \sum_y \mathbf{Pr}\{[X=x] \cap [Y=y]\}; \text{see the derivation of the law of total probability} \\
&= \sum_x x \sum_y \mathbf{Pr}(X=x|Y=y)\mathbf{Pr}(Y=y); \text{definition of conditional probability} \\
&= \sum_y \left[\sum_x x\mathbf{Pr}(X=x|Y=y)\right] \mathbf{Pr}(Y=y) \\
&= \sum_y \mathbf{E}\left[X|Y=y\right] \mathbf{Pr}(Y=y).
\end{aligned}$$