Discrete Probability Refresher

We make quantitative statements about situations with randomness of countable number of events. This presentation of discrete probability is condensed from:
http://www.cis.upenn.edu/~jean/proba.pdf
http://www.cs.yale.edu/homes/aspnes/classes/469/notes.pdf.

Probability Spaces and Events

A *probability space* is a mathematical object that specifies outcomes and their probabilities. A discrete probability space $(\Omega, \mathbb{P})$ is specified by:

- A countable set of *outcomes* Ω, also known as the *sample space*.
- A non-negative *probability* for each outcome in Ω, so that the probabilities of all outcomes sum to 1.

Some examples of probability spaces:

- $\Omega = \{\text{heads}, \text{tails}\}$, $\mathbb{P}[\text{heads}] = \mathbb{P}[\text{tails}] = \frac{1}{2}$. (a fair coin flip)
- $\Omega = \{1,2,3,4,5,6\}$, $\mathbb{P}[\text{roll } 1..6] = \frac{1}{6}$. (a roll of a six-sided die)
- $\Omega = \mathbb{N}$, $\mathbb{P}[k] = \frac{1}{2^k}$. (number of coin flips until you see a heads)
- We address the case of uncountable Ω like $[0,1]$ or $\mathbb{R}$ in the following refresher.

An *event* is a subset of Ω. Compute the probability of an event by adding the probabilities of the outcomes in that event. Examples of such events are respectively:

- $\mathbb{P}[\{\text{heads}\}] = \frac{1}{2}$.
- $\mathbb{P}[\{1,3,5\}] = \frac{1}{2}$.
- $\mathbb{P}[\text{even}] = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}$.

Since events are sets of outcomes, they come with Boolean operations:

- $\neg A$ (“not”). Note that $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$.
- $A \cup B$ (“or”).
- $A \cap B$ (“and”).

Independence

Two events A, B are *independent* if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. For example, consider the probability space where Ω is the set of sequences of 5 coin flips, and each outcome has probability 1/32. Then, it can be verified that the event “first coin comes up heads” and “second coin comes up heads” are independent, but not “first coin comes up heads” and “all coins come up heads”.

We often specify a probability space by combining smaller probability spaces (forming a *joint probability space*), and requiring independence. Above, we have the probability space of “5 independent fair coin flips”. This uniquely determines the probabilities of all outcomes: e.g. $\mathbb{P}[HTHHT] = \mathbb{P}[H] \times \mathbb{P}[T] \times \mathbb{P}[H] \times \mathbb{P}[H] \times \mathbb{P}[T] = \frac{1}{32}$.

Conditioning

The final fundamental operation in a probability space is *conditioning*, which lets us consider smaller cross-sections of probability spaces. For two events A, B, we define the conditional probability of A given B, denoted $\mathbb{P}[A|B]$, as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \ .$$

As a concrete example, consider a fair die roll. Let A be the event "roll an odd number". Let B be the event "roll 3 or less". Then, $\mathbb{P}[A] = \mathbb{P}[B] = 1/2$, and $\mathbb{P}[A \cap B] = 1/3$. So,

$$\mathbb{P}[A|B] = 2/3 \ .$$

That is, given that a fair die roll is a 1, 2, or 3, there's a $2/3$ probability that it's also odd. A and B are certainly not independent.

Random Variables and their Distributions

Even though all objects in probability theory arise from probability spaces, we don't usually need to work with them directly. It's more intuitive to manipulate random variables. The following is an important definition.

A (real-valued) *random variable* on a probability space $(\Omega, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$. (Vector-valued random variables can be obtained by replacing $\mathbb{R}$ with $\mathbb{R}^n$.)

Then, for each $a \in \mathbb{R}$ (a value that X could take), we have a (possibly empty) subset of outcomes ω for which $X(\omega) = a$. Think of this as "the event that $X = a$", whose probability we call $\mathbb{P}[X = a]$, where X depends on some the outcome of $(\Omega, \mathbb{P})$.

For example, consider the probability space $(\Omega, \mathbb{P})$ generated by 5 independent fair coin flips. Then, one example of a random variable on $(\Omega, \mathbb{P})$ is the function X that maps an outcome to the number of heads in that outcome. Then, we have

$$\begin{aligned}\mathbb{P}[X = 3] &= \mathbb{P}[\ \{\text{HHHTT}, \text{HHTHT}, \ldots, \text{TTHHH}\}\ ] \\ &= \mathbb{P}[\text{HHHTT}] + \mathbb{P}[\text{HHTHT}] + \ldots \\ &= \binom{5}{3}\frac{1}{32} = \frac{5}{16}.\end{aligned}$$

We often wish to study $\mathbb{P}[X = a]$ as a function of a, which we call its *distribution*, or *probability mass function*. In the above example, $\mathbb{P}[X = a] = \frac{1}{32}, \frac{5}{32}, \frac{5}{16}, \frac{5}{16}, \frac{5}{32}, \frac{1}{32}$ for $a = 0, \ldots, 5$. We call the set of a with $\mathbb{P}[X = a] > 0$ the *support* of X.

Two random variables X and Y are *independent* if, for any x, y,

$$\mathbb{P}[X = x \wedge Y = y] = \mathbb{P}[X = y] \cdot \mathbb{P}[Y = y] \ .$$

We call the LHS the *joint distribution* of X and Y.

Examples of Distributions

Some distributions of random variables occur so ubiquitously that they have names. The simplest is the *Bernoulli* distribution $\text{Bern}(p)$. X is Bernoulli with parameter p if its support is $\{0, 1\}$, and $\mathbb{P}[X = 1] = p$. They often arise as *indicator* variables of events, like "1 if the first two coins come up heads, 0 otherwise". Examples of other distributions are:

- The *binomial* distribution $B(n,p)$, which has support $\{0,\ldots,n\}$ and
$$\mathbb{P}[X=k]=\binom{n}{k}p^k(1-p)^{n-k}\ .$$
A random variable with such a distribution usually arises by summing n random variables distributed as $\text{Bern}(p)$. $\text{Bern}(p)$ is the same as $B(1,p)$.
- The *geometric* distribution $\text{Geom}(p)$, which has (infinite) support $\mathbb{N}$ and
$$\mathbb{P}[X=k]=(1-p)^{k-1}p\ .$$
These arise as waiting times for a repeatedly flipped coin to come up heads.

We can manipulate random variables algebraically, just as we can manipulate real-valued functions. For example, if $X\sim B(n,p)$ and $Y\sim\text{Bern}(p)$ are independent, then $X+Y$ is a random variable, and its distribution is $B(n+1,p)$. What is done here is constructing a joint probability space on pairs (X,Y), on which we define a new random variable $X+Y$.

Expectations

The *expected value* (or *expectation* or *mean* or *first moment*) of a random variable is the probability-weighted average of its possible values.
$$\mathbb{E}[X]=\sum_{x\in\text{supp}(X)}\mathbb{P}[X=x]\cdot x\ .$$

The *law of large numbers* states that the average of many independent copies of a random variable tends towards the expectation. We will quantify this later on in the core of the machine learning workshop.

For example, if X is the value of die roll, then
$$\mathbb{E}[X]=\frac{1+2+3+4+5+6}{6}=3.5\ .$$
If $X\sim\text{Bern}(p)$, then
$$\mathbb{E}[X]=p\cdot 1+(1-p)\cdot 0=p\ .$$
If $X\sim\text{Geom}(\frac{1}{2})$, then
$$\mathbb{E}[X]=\frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\ldots=2\ .$$

Linearity of Expectation

We make wide use of the fact is that expectation is linear. That is, the expectation of a linear combination of random variables is the linear combination of their expectations, *even if they are not independent*:
$$\mathbb{E}[aX+bY]=a\,\mathbb{E}[X]+b\,\mathbb{E}[Y]\ .$$

Proof:
$$\begin{aligned}\mathbb{E}[aX+bY]&=\sum_{x,y}\mathbb{P}[X=x,Y=y](ax+by)\\&=a\sum_{x,y}\mathbb{P}[X=x,Y=y]x+b\sum_{x,y}\mathbb{P}[X=x,Y=y]y\\&=a\sum_{x,y}\mathbb{P}[X=x]x+b\sum_{x,y}\mathbb{P}[Y=y]y\\&=a\,\mathbb{E}[X]+b\,\mathbb{E}[Y].\end{aligned}$$

This implies that the expected value of a random variable distributed as $B(n,p)$ is np.

Variance

We often want to quantify how *spread-out* a random variable is. For this, it is useful to reason about its *variance*:

$$\text{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$

Think about $|X - \mathbb{E}[X]|$ as a random variable that measures far X is from its expectation. The variance measures the average value of the *square* of this quantity, "penalizing" large deviations quadratically.In some cases, one uses $\mathbb{E}[|X - \mathbb{E}[X]|]$ ("mean absolute deviation") to measure spread, but as it turns out, variance is often easier to compute and manipulate.

For example, consider $X \sim B(5, \frac{1}{2})$, and Y distributed uniformly on the same support $\{0,1,2,3,4,5\}$. Then, $\text{Var}[X] = 1.25$, while $\text{Var}[Y] \approx 1.458$, which agrees with the intuition that X is slightly more concentrated around its mean than Y.

Closely related to the variance is the (raw) second moment $\mathbb{E}[X^2]$. In general, the k-th moment is given by $\mathbb{E}[X^k]$. They tell us different things about a distribution– for example, the third moment measures *skewness* and the fourth measures *kurtosis* (pointiness). More on this later.

Expected Triangle Count

Here is a sample problem that brings together some ideas from this refresher:

A class has 25 students. Each pair of students is friends independently with probability $\frac{1}{10}$. What is the expected number of "triangles" of students who are all mutually friends?

Suppose we try to do this directly. Let Z be the number of such triangles. Then,

$$\mathbb{E}[Z] = \sum_k \mathbb{P}[Z = k] \cdot k .$$

You could do it this way in principle (by writing a simulation), but these probabilities are really hard to compute!

Instead, for each triple $\{a,b,c\}$, let $X_{a,b,c}$ be the indicator variable of the event they are all friends. Then, X_i is Bernoulli with parameter $1/10^3$. But remember that these indicators are not independent, since these triples of students might overlap!

Nonetheless, by linearity of expectation, we have

$$\mathbb{E}[Z] = \binom{25}{3}\mathbb{E}[X_{a,b,c}] = \frac{\binom{25}{3}}{10^3} = 2.3 .$$

Linearity of expectation alone is enough to solve an impressive array of problems.

(Food for thought: how could you compute the variance?)