Lecture 7 — 03/18/24
Probability Theory[1]

1 Counting

Example A coin is *fair* if it comes up heads or tails with equal probability. You flip a fair coin three times. What is the probability that exactly one of the flips results in a head?

First, what are the possible ***outcome***s?

...

An ***outcome*** is a possible result of an ***experiment or trial***. Each possible outcome of a particular experiment is ***unique***, and different outcomes are ***mutually exclusive*** (only one and exactly one outcome will occur on each trial of the experiment). All of the possible outcomes of an experiment form the elements of a ***sample space***, which we usually write with the symbol: Ω.

Two useful principles of counting and sets are worth mention here:

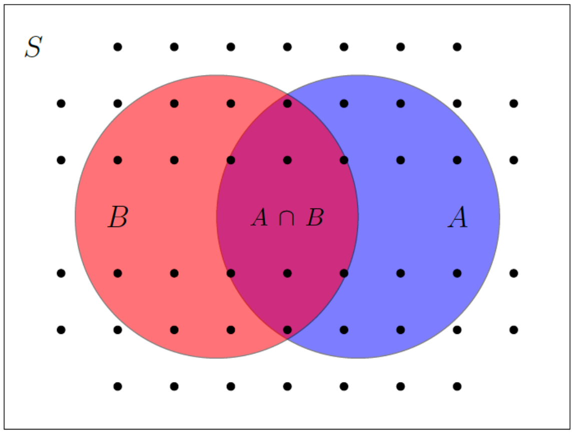
- ***Inclusion-Exclusion Principle***
- ***Rule of Product***

1.1 Inclusion-Exclusion Principle

The inclusion-exclusion principle states that:

$$|A \cup B| = |A| + |B| - |A \cap B| \tag{1}$$

We can illustrate this with a Venn diagram. S is all the dots, A is the dots in the blue circle, and B is the dots in the red circle.



1.2 Rule of Product

The Rule of Product (or *multiplication* rule) states that:

If there are n ways to perform action 1 and then by m ways to perform action 2, then there are $n * m$ ways to perform action 1 followed by action 2.

Important idea here: the rule of product holds even if the ways to perform action 2 depend on action 1, as long as the *number* of ways to perform action 2 is independent of action 1.

[1] **Some of the notation is based on this week's reading, ch. 2 of Manning and Schütze 1999. See also Jelinek (1998).**

Consider the following: There are 5 competitors in the 100m final at the Olympics. In how many ways can the gold, silver, and bronze medals be awarded?

2 Outcomes and Events

Take a page or two from your favorite book[2], cut them into strips each with exactly one word, and dump them into an urn[3]. Now pull out a strip of paper. How likely is it that the word you get is “the”? To make this more precise, what is the probability that the word is “the”?

In order to answer this question, we need to formalize a few things:

- Sample space Ω
  - For a coin toss, Ω = {“head , tail”}
  - For a dice roll, Ω = {1, ..., 6}
  - For a lottery ticket, Ω = {0, ...,$1.5 * 10^8$}
- Each sample $x \in \Omega$ is assigned a probability $P(x)$

Let’s formalize this a bit more with two constraints on what it means to be a probability:

- $P(x) \in [0, 1]$ for all $x \in \Omega$, i.e. probabilities are real numbers between 0 and 1
- $\sum_{x \in \Omega} P(x) = 1$, i.e., the probabilities of all samples sum to one

Of course, many of these strips of paper have the same word. For example, if the pages you cut up are written in English the word “the” is likely to appear on more than 50 of your strips of paper. The 1000 strips of paper each correspond to different word ***tokens*** or occurrence of a word in the urn, so the urn contains 1000 word tokens. But because many of these strips contain the same word, the number of word ***types*** (i.e., distinct words) labeling these strips is much smaller; our urn might contain only 200 word types.

We can formalize the type/token distinction using ***Random Events***. Formally, an event E is a set of samples, i.e., $E \subseteq \Omega$, and the probability of an event is the sum of the probabilities of the samples that constitute it:

$$P(E) = \sum_{x \in E} P(X) \tag{2}$$

We can treat each word type as an event.

Suppose $E_{‘the’}$ is the event of drawing a strip labeled ‘the’, that $|E_{‘the’}|$ = 60 (i.e., there are 60 strips labeled ‘the’) and that P(x) = 0.001 for all samples $x \in \Omega$. Then...

$$P(E_{‘the’}) = 60 * 0.0001 = 0.06$$

3 Probability distirbutions

Probabilities are numbers between 0 and 1 (inclusive) where 0 indicates impossibility and 1 indicates certainty. A ***probability distribution** P* is a function from a set of events Ω to probabilities such that:

$$\sum_{\omega \in \Omega} P(\omega) = 1. \tag{3}$$

Or, in prose, the probabilities of all events in a given probability distribution must sum to one.

[2] Probably better to use one of your least favorite books

[3] If you prefer you can use a trash can, but for some reason all books on probability use urns

3.1 Random Variables

A ***random variable*** is a variable whose value is the outcome of some random or unpredictable phenomenon. These phenomenon may be continuous or discrete. For example, "current windspeed" is a continuous random variable (some positive real number representing speed in miles per hour), and a coin flip is a discrete random variable (where 1 might represent heads and 0 tails). A random variable has a probability distribution specifying the probabilities of its values.

A little more formally, ***Random variables*** are a method for specifying events: a random variable is a function from the sample space Ω to some set of values. For example, to capture the type-token distinction we might introduce a random variable W that maps samples to the words that appear on them, so that $W(x)$ is the word labeling strip $x \in \Omega$.

Given a random variable V and a value v, $P(V = v)$ is the probability of the event that V takes on the value v, i.e.:

$$P(V = v) = P(\{x \in \Omega : V(x) = v\})$$

Note that the standard notation is to capitalize random variables and use lower-cased variables as their values.

Returning to our type-token example, $P(W = `the') = 0.06$

If the random variable intended is clear from the context, sometimes we elide it and just write its value, e.g. P('*the*') — even though the pedantic way to write this is really $P(W = `the')$.

Similarly, the *value* of the random variable may be elided if it is unspecified or clear from the context, e.g., $P(W)$ is sometimes short for $P(W = w)$ where w ranges over words.

Random variables are useful because they let us easily construct a variety of complex events. For example, suppose F is the random variable mapping each sample to the first letter of the word appearing in it and S is the random variable mapping samples to their second letters (or the space character if there is no second letter). Then $P(F = `t')$ is the probability of the event in which the first letter is 't' and $P(S = `h')$ is the probability of the event in which the second letter is 'h.'

4 Maximum likelihood estimation

How do we estimate a probability distribution for a random variable? Intuitively, we want to set the probabilities such that we maximize the ***likelihood***, the probability of the observed data. This can be accomplished with ***maximum likelihood estimate*** (or MLE). In the discrete case, it simply requires us to count observed values of the random variable and dividing them by the number of samples drawn.

Imagine that we have another opaque urn which contains red and blue marbles. We are interested in a discrete random variable X with outcomes $\{r, b\}$ where r represents red and b represents blue. We then sample X by reaching in and drawing out marbles, one at a time. We sample R red marbles and B blue marbles. Then, our maximum likelihood estimate for the probability of drawing a red marble is given by:

$$P(r) = \frac{R}{R + B}$$

and the probability of drawing a blue marble is given by:

$$P(b) = \frac{B}{R + B}.$$

More concretely, imagine that $R = 7$ and $B = 2$. Then:

$$P(r) = \frac{7}{9}$$
$$P(b) = \frac{2}{9}.$$

Similarly, consider a universe that also includes green marbles. Let $R = 7$, $B = 2$, and $G = 4$. Then:

$$P(g) = \frac{3}{7+2+4} = \frac{4}{13}$$

4.1 Joint Probabilities

Given any two events E_1 and E_2, the probability of their conjunction $P(E_1, E_2) = P(E_1 \cap E_2)$ = $P(E_1 \wedge E_2)$ is called the ***joint probability***. This is simply the probability of E_1 and E_2 occurring simultaneously.

Continuing with our letter-urn example, $P(F = `t', S = `h')$ is the joint probability that the first letter is ‘t’ AND that the second letter is ‘h.’

Question: What is the relationship between $P(F = `t', S = `h')$ and $P(`the')$?

$\{\geq, =, \leq\}$?

4.2 Conditional and Marginal Probabilities

Now imagine temporarily moving all the strips whose first letter is ‘q’ into a new urn. Clearly this new urn has a different distribution of words from the old one; for example, $P(F = `q') = 1$ in the sample contained in the new urn. The distributions of the other random variables change as well; if our strips of paper contain only English words then $P(S = `u') \approx 1$ in the new urn.

Conditional probabilities formalize this notion of temporarily setting the sample set to a particular set. The ***conditional probability*** $P(E_2|E_1)$ is the probability of event E_2 given that event E_1 has occurred — you could think of this as the probability of E_2 given that E_1 is the temporary sample set.

$P(E_2|E_1)$ is defined as:

$$P(E_2|E_1) = \frac{P(E_1, E_2)}{P(E_1)} \quad \text{if } P(E_1) > 0 \tag{4}$$

and it’s undefined if $P(E_1) = 0$. This equation relates the ***conditional probability*** $P(E_2|E_1)$ (left-hand side), the ***joint probability*** $P(E_1, E_2)$ (numerator) and the ***marginal probability*** $P(E_1)$ (denominator).

To visualize this, consider the Venn diagram in figure 1. Imagine we have observed E_2 and are interested in whether E_1 has also occurred. We simply need to compute what the percentage of the E_2 circle which overlaps with E_1, and this is given by our above formula definition of conditional probability (eq. 4).

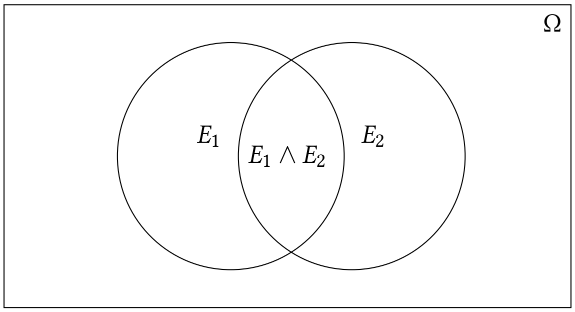


Figure 1: Graphical representation of joint and conditional probability for a discrete random variable, after Manning and Schütze 1999, 42.

Suppose our urn contains just 10 strips of paper (i.e., our sample space Ω has 10 elements) that are labeled with four word types, and the frequency of each word is as follows:

word type	frequency
‘nab’	1
‘no’	2
‘tap’	3
‘top’	4

Let F and S be random variables that map each strip of paper (i.e., sample) to the first and second letters that appear on them, as before. Let’s compute the marginal, joint, and conditional probabilities relating F and S.

Note that $P(A|B) \neq P(B|A)$

4.3 Independence

Two events A and B are said to be (***statistically***) ***independent*** if knowledge of the outcome of one has no influence on the probability of the other. More formally, A and B are independent if their joint probability is the product of their probabilities.

$$P(A \wedge B) = P(B \wedge A) = P(A) \cdot P(B) \tag{5}$$

or equivalently, we can rewrite these using eqs. (4—5):

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A) \tag{6}$$

$$P(B \mid A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B). \tag{7}$$

In prose: the conditional probability of A given B is equal to the prior probability of A, and the conditional probability of B given A is equal to the prior probability of B.

Note The definitions of conditional and joint probability are valid regardless of whether variables are independent or not.

4.4 Expectation

If a random variable X ranges over numerical values (say integers or reals), then its ***expectation*** or ***expected value*** is just a weighted average of these values, where the weight on value X is $P(X)$. More precisely, the ***expected value*** $E[X]$ of a random variable X is:

$$E[X] = \sum_{x \in \chi} x * P(X = x) \tag{8}$$

where χ is the set of values that the random variable X ranges over. Conditional expectations are defined in the same way as conditional probabilities, namely by restricting attention to a particular conditioning event, so:

$$E[X|Y = y] = \sum_{x \in \chi} x * P(X = x|Y = y) \tag{9}$$

Furthermore, expectation is a linear operator, so if $X_1, ..., X_n$ are random variables, then:

$$E[X_1 + ... + X_n] = E[X_1] + ... + E[X_n] \tag{10}$$

Suppose X is a random variable generated by throwing a fair six-sided die. Then the expected value of X is

...

Next, if X_1 and X_2 are random variables generated by throwing two fair six-sided dice. Then the expected value of their sum is:

...

Here's a practice problem:

If $P(A \cup B) = 0.7$ and $P(A \cup B') = 0.9$, what is $P(A)$?

...

5 The chain rule

The ***chain rule*** permits us to compute joint probabilities from conditional probabilities or vis versa. For two random variables A and B, it says that:

$$P(A \wedge B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A) \tag{11}$$

More generally, we can write:

$$P(A_1 \wedge \ldots \wedge A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \wedge A_2) \cdot \ldots P(A_n \mid \wedge_{i=1}^{n-1} A_i). \tag{12}$$

6 The Markov property and the Markov assumption

Now imagine that the events $A_1, \ldots, A_n$ are in fact a sequence of observations such as words in a sentence, or part-of-speech tags. Intuitively, adjacent elements in these sequences have strong statistical dependencies—e.g., in English a determiner is almost always immediately followed by an adjective or a noun—but distant elements are roughly independent. Thus, rather than conditioning A_n on all $n - 1$ previous elements, we may wish to only condition each element of the sequence on just the k preceding elements. That is, we approximate the conditional probability of an element A_j in the sequence $A_1, \ldots, A_n$ as follows:

$$P(A_j \mid \wedge_{i=1}^{j-1} A_i) \approx P(A_j \mid \wedge_{i=j-k}^{j-1} A_i) \tag{13}$$

In prose: we approximate the conditional probability of an element not using the entire prior history but only the previous k elements. This is referred to as a (k-th-order) ***Markov assumption*** (or ***approximation***), and data which obeys a Markov assumption is said to have the ***Markov property***.[4] For instance, if we assume each element in the sequence is conditioned only on the preceding element, we make a first-order Markov assumption. Markov assumptions are crucial for both ***language models*** and ***tagging models*** like part-of-speech taggers.

7 Bayes' rule

Definition ***Bayes' rule*** (or ***theorem***) allows us to swap the "order of dependence", that is, to calculate $P(B \mid A)$ using $P(A \mid B)$. This is particularly useful when one of the these two orders is challenging to estimate directly. Bayes' rule follows directly from the definition of conditional probability and the chain rule. It holds that:

$$P(B \mid A) = \frac{P(A \land B)}{P(A)} = \frac{P(A \mid B) \cdot P(B)}{P(A)}. \tag{14}$$

When dealing with two competing hypotheses it is helpful to spell out the denominator like so:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \neg B) \cdot P(\neg B)}. \tag{15}$$

The fundamental theorem of speech recognition Many speech and language technologies can be described in these terms. For instance, Jelinek (1998, 4f.) formulates the speech recognition problem as follows. Let $\mathbf{A}$ be a sequence of "acoustic symbols" and $\mathbf{W}$ a sequence of words. If $P(\mathbf{W} \mid \mathbf{A})$ denotes the probability that the words $\mathbf{W}$ were spoken given that we observed the acoustic sequence $\mathbf{A}$, then the recognizer should select a transcript $\hat{\mathbf{W}}$ for an acoustic sequence $\mathbf{A}$ according to:[5]

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{A}). \tag{16}$$

Bayes' rule permits us to rewrite the right-hand side of (16) as:

$$P(\mathbf{W} \mid \mathbf{A}) = \frac{P(\mathbf{W}) \cdot P(\mathbf{A} \mid \mathbf{W})}{P(\mathbf{A})}. \tag{17}$$

$P(\mathbf{W})$, the probability of uttering word sequence $\mathbf{W}$, is known as a ***language model***. $P(\mathbf{A} \mid \mathbf{W})$, the probability that the word sequence $\mathbf{W}$ produces the acoustic sequence $\mathbf{A}$, is known as the ***acoustic model***, or more generally, the ***channel model***. Since $\mathbf{A}$ is constant, we can ignore it for the purposes of finding $\hat{\mathbf{W}}$:

$$P(\mathbf{W} \mid \mathbf{A}) \propto P(\mathbf{W}) \cdot P(\mathbf{A} \mid \mathbf{W}), \tag{18}$$

and we can simplify (16) to:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}) \cdot P(\mathbf{A} \mid \mathbf{W}). \tag{19}$$

In prose: we select the transcription which maximizes the product of the language model and acoustic model probabilities.

[4] Naturally, then, a zeroth-order Markov approximation assumes that each element of the sequence is independent of all previous elements and a zeroth-order Markov property obtains when this assumption is correct.

[5] The arg max (***arguments of the maxima***) is the point or points of a function which maximize the value of that function. The arg min function is defined similarly but for minimizing the value of the function.

8 Avoiding underflow

As mentioned in Methods I, floating-point numbers (like Python's `float`) are only approximate representations of actual decimals and fractions. One problem with this is that the products of probabilities may trigger ***arithmetic underflow***. Underflow occurs when a number greater than zero—but smaller in magnitude than the smallest representable positive floating-point number (***machine epsilon***)—is rounded down to zero.

One strategy used to avoid underflow is to store the (negative) logarithm of a probability instead of the raw probability, and then to conduct arithmetic in the (negative) "log space". First, let's define a function computing the negative natural logarithm of a positive real number:

$$\text{nlog}(x) = -\log x$$

or the equivalent Python function:

```
def nlog(x: float) -> float:
    return math.inf if x == 0.0 else -math.log(x)
```

We can also define an inverse function:

$$\text{nlog}^{-1} x = \exp(-x) = e^{-x}$$

or the equivalent Python function:

```
def inv_nlog(x: float) -> float:
    return math.exp(-x)
```

This conversion helps us to avoid underflow because we can perform multiplication using addition! Perhaps you are familiar with the following logarithmic identity:

$$\forall x, y \in \mathbb{R}_+ : \log(x \cdot y) = \log x + \log y.$$

In prose, this says that the product of the log of positive real numbers in equal to their log-sums. Thus we can multiply positive real numbers by adding their logarithms and then exponentiating. More concretely, we can compute the product of .5 and .01 as follows:

$$.5 \cdot .01 \approx \exp(\log .5 + \log .01) = \exp(-5.298) \approx .005.$$

Or, in Python:

```
inv_nlog(nlog(.5) + nlog(.01))
```

References

Jelinek, F. (1998). ***Statistical methods for speech recognition***. MIT Press.

Manning, C., & Schütze, H. (1999). ***Foundations of statistical natural language processing***. MIT Press.