Chapter 3

Probability theory

In this section, I review the basics of probability theory. This is necessary because the whole point of statistical NLP is to view language as a probabilistic system.

3.1 What is it?

Any event e has some likelihood of occurring: $\mathrm{P}(e)$. If it is certain, then $\mathrm{P}(e) = 1$. If it is impossible, then $\mathrm{P}(e) = 0$. Intermediate likelihoods range between these values.

Let's denote the set of possible outcomes for some experiment as Ω. Then, if all outcomes are equally likely, we can define the probability of some event e as the number of elementary events compatible with e divided by the total number of events possible (Ω). For example, the chances of rolling a six with one role of a fair die is $\frac{1}{6} = .166$. There are six possible events in Ω, and only one of them is compatible with the outcome we are interested in. The chances of rolling a number greater than 4 are $\frac{2}{6} = .333$. Here again $\Omega = 6$, but now there are two elementary events compatible with the outcome we are interested in.

There are actually lots of ways to estimate the probability of some outcome. One possibility is to perform an actual experiment. If we are interested in the probability of rolling a six given a fair die, we roll that die some number of times and then work out the percentage of time that it comes up six. This is called the *maximum likelihood* estimate of the probability of some event. There is a law of probability theory—called the *Law of Large Numbers*—that

says that as the sample gets larger, the maximum likelihood estimate of the probability of some outcome will get closer and closer to the true probability of that outcome.

However we calculate the probability of individual events, it must be the case that the sum of the probabilities of the individual events in some space of outcomes equals one.

$$\sum_{e \in \Omega} e = 1 \tag{3.1}$$

If this condition holds, then these events are said to exhibit a *probability distribution*.

3.2 Combining events

There are a number of ways to think about combining events, all involving notions from set theory. Let's consider four rather simple ones: *not*, *and*, *or*, and *if*.

Consider first negation. As an example, what is the probability of throwing something *other than* six, given a fair die. In this case, there are six outcomes, five of which are consistent with this output: $\frac{5}{6} = .833$. There is a simpler way to do this, however. If the probability of some event e is n, then the probability of e not occurring is $1 - n$.

$$\mathrm{P}(\neg e) = 1 - \mathrm{P}(e) \tag{3.2}$$

This follows from the definition of a probability distribution above.

Let's now consider conjunction (*and*). If two events are independent of each other, then the chances of them co-occurring is the product of their different probabilities.

$$\mathrm{P}(e_1, e_2) = \mathrm{P}(e_1) \times \mathrm{P}(e_2) \text{ if the events are independent} \tag{3.3}$$

Thus the chances of throwing two sixes in two successive throws of a single die are $.166 \times .166 = .027$. When the events are not independent, then it is far more complex to work this out. The probability of two events co-occurring is also referred to as their *joint probability*.

Disjunction—the chance of one of two non-overlapping events occurring—is the *sum* of their independent probabilities.

$$\mathrm{P}(e_1 \cup e_2) = \mathrm{P}(e_1) + \mathrm{P}(e_2) \text{ if } \mathrm{P}(e_1, e_2) = 0 \tag{3.4}$$

Thus the probability of throwing either a one or a two with a single die is $.166 + .166 = .332$. If the events can co-occur, then the probability of them co-occurring is subtracted from the sum of their individual probabilities.

$$\mathrm{P}(e_1 \cup e_2) = \mathrm{P}(e_1) + \mathrm{P}(e_2) - \mathrm{P}(e_1, e_2) \tag{3.5}$$

Consider the probability of throwing an even number (2, 4, 6) or a number over three (4, 5, 6). The probability of an even number is $\frac{3}{6} = .5$. The probability of a number over three is $\frac{3}{6} = .5$. Their joint probability is the probability of throwing a four or six $\frac{2}{6} = .33$. Hence, we get $.5 + .5 - .33 = .66$. Notice how equation 3.5 naturally covers the previous case as well, since the joint probability there is zero, by definition.

3.3 Conditional probability

A notion that shows up a great deal in statistical NLP is *conditional probability*. (We can also think of this as implementing *if*.) This is the probability that some event a might occur given that some event b occurs: $\mathrm{P}(a|b)$. This is defined as the probability of a and b both occurring divided by the probability of b occurring.

$$\mathrm{P}(a|b) = \frac{\mathrm{P}(a, b)}{\mathrm{P}(b)} \tag{3.6}$$

If we want to estimate this from some experimental observation, then we would use observations or counts.[1]

$$\mathrm{P}(a|b) = \frac{|a, b|}{|b|} \tag{3.7}$$

For example, we can use conditional probability to characterize the probability of throwing at least two heads out of three, given that the first coin thrown is heads. First, what is the overall probability of throwing two heads out of three? This can be determined by simple inspection and the definition of probability. There are eight total possibilities: HHH, HHT, HTH,

[1] I use absolute value notation to represent the number of occurrences of some variable, e.g. $|a|$.

HTT, THH, THT, TTH, TTT. Four of these are compatible with the result we are interested in. Hence, P(two-heads+) $= \frac{4}{8} = .5$. There is also a .5 probability that the first coin will come up heads: P(heads-first) = .5. The possibility of both events being true at the same time, that the first coin comes up heads and at least two of the three coins comes up heads, is called the *joint probability*. Here, this is P(both) $= \frac{3}{8} = .375$. The conditional probability of two (or more) heads given the first coin comes up as heads is $\frac{\text{P(both)}}{\text{P(heads-first)}} = \frac{.375}{.5} = .75$.

We will find quite a lot of use for the notion of conditional probability in statistical NLP.

3.4 Bayes' Law

A very important formula in statistical NLP (and much of statistics!) is *Bayes' Law*:

$$\text{P}(a|b) = \frac{\text{P}(a)\text{P}(b|a)}{\text{P}(b)} \tag{3.8}$$

Bayes' Law can actually be derived from the definition of conditional probability.

$$\begin{array}{rcl} \text{P}(a|b) = \frac{\text{P}(a,b)}{\text{P}(b)} & & \frac{\text{P}(a,b)}{\text{P}(a)} = \text{P}(b|a) \\ \text{P}(a|b)\text{P}(b) = \text{P}(a,b) & & \text{P}(a,b) = \text{P}(b|a)\text{P}(a) \\ \text{P}(a|b)\text{P}(b) = \text{P}(b|a)\text{P}(a) & & \\ \text{P}(a|b) = \frac{\text{P}(b|a)\text{P}(a)}{\text{P}(b)} & & \end{array} \tag{3.9}$$

In its bare form above, it may look like gobbledegook, but it can be manipulated algebraically in lots of ways as we will see.

3.5 The Chain Rule

The *Chain Rule* is one simple consequence of the definition of conditional probability which we will also see a lot of in NLP applications. The joint probability of some set of events a_1, a_2, a_3, a_4 can also be expressed as a 'chain' of conditional probabilities, e.g.:

$$\mathrm{P}(a_1, a_2, a_3, a_4) = \mathrm{P}(a_1)\mathrm{P}(a_2|a_1)\mathrm{P}(a_3|a_1, a_2)\mathrm{P}(a_4|a_1, a_2, a_3) \tag{3.10}$$

This follows algebraically from the definition of conditional probability. If we substitute by the definition of conditional probability for each of the conditional probabilities in the preceeding equation and then cancel terms, we get the original joint probability.

$$\begin{aligned} \mathrm{P}(a_1, a_2, a_3, a_4) &= \mathrm{P}(a_1) \times \mathrm{P}(a_2|a_1) \times \mathrm{P}(a_3|a_1, a_2) \times \mathrm{P}(a_4|a_1, a_2, a_3) \\ &= \mathrm{P}(a_1) \times \tfrac{\mathrm{P}(a_1,a_2)}{\mathrm{P}(a_1)} \times \tfrac{\mathrm{P}(a_1,a_2,a_3)}{\mathrm{P}(a_1,a_2)} \times \tfrac{\mathrm{P}(a_1,a_2,a_3,a_4)}{\mathrm{P}(a_1,a_2,a_3)} \\ &= \mathrm{P}(a_1, a_2, a_3, a_4) \end{aligned} \tag{3.11}$$

Notice also that the chain rule can be used to express *any* dependency among the terms of the original joint probability. For example:

$$\mathrm{P}(a_1, a_2, a_3, a_4) = \mathrm{P}(a_4)\mathrm{P}(a_3|a_4)\mathrm{P}(a_2|a_3, a_4)\mathrm{P}(a_1|a_2, a_3, a_4) \tag{3.12}$$