Probability Basics for NLP

Suhas Arehalli

COMP394 (NLP)

1 Why Probability

As we've discussed earlier, our goal this semester will be to explore how we can build good models of language. we've also seen that language involves quite a bit of *uncertainty*.

To see this, let's consider a very important task in NLP: **language modeling**. A first approximation of the task description for language modeling would be *next-word prediction* — Given a sequence of words, predict what word comes next. For example, a language model may be given the sequence of words *"An apple a day keeps the "* and the model would have to produce the next word (something like *doctor*, say).

The problem with our first-pass analysis should be somewhat apparent though — there isn't one right answer! "An apple a day keeps *blood...*" forms a valid prefix of a sentence (i.e., "An apple a day keeps the blood flowing"). Of course, there's a number of other continuations (e.g., "An apple a day keeps the proctor away"). But some continuations are better than others: "An apple a day keeps the is" doesn't seem to form any grammatical continuations. And neither "doctor" or "blood" is incorrect, but "doctor" seems *better*, in some way.

To solve this, we need a way to model this uncertainty, and probability is going to be the tool we need!

In the end, we're going to arrive at a definition of language modeling that has a language model provide a *probability distribution* over all possible next words.

2 The Basics

Here's the framework: We are going to want to model the probability of an *outcome* s happening. The set of all possible outcomes is something we'll call the sample space, Ω. A probability distribution $p(\cdot)$ is a function $p : \Omega \to \mathbb{R}$ that assigns each outcome s a *probability*, such that:

$$\forall s \in \Omega, 0 \leq p(s) \leq 1 \tag{1}$$

$$\sum_{s \in \Omega} p(s) = 1 \tag{2}$$

That is, probabilities lie between 0 and 1 (inclusive!) and if we add up the probabilities for each event in the sample space, we get 1.

Consider a 6-sided die. When we roll a die, we're uncertain what the outcome will be, so we can model the die with a probability distribution! We have 6 possible events — one corresponding to each side of the die, so we can define our sample space as $\Omega = \{1, 2, 3, 4, 5, 6\}$ with each outcome representing the number on each side of the die. Now, if we're modeling a fair die (i.e., each face has equal probability of begin rolled), we can derive the appropriate probability distribution p! Let the probability of each face showing up be k. To practice the mathematical notation, we can write this as...

$$\forall s \in \Omega, p(s) = k$$

but by the 2nd part of the definition of a probability distribution, we know that

$$\begin{aligned} \sum_{s \in \Omega} p(s) &= 1 \\ \sum_{s \in \Omega} k &= 1 \\ |\Omega| k &= 1 \\ k &= \frac{1}{|\Omega|} = \frac{1}{6} \end{aligned}$$

And this generalizes quite neatly! If we have a *finite* sample space and we know that each event is equally likely (i.e., the probability distribution is *uniform*, $\forall s \in \Omega, p(s) = \frac{1}{|\Omega|}$).

We also might want to talk about the probability of *events*, which we formally define as subsets $E \subseteq \Omega$. As you might expect, we define the probability of an event E as

$$p(E) = \sum_{s \in E} p(s).$$

For example, we might want to construct the probability that the die rolls an even number, so we construct $E = \{2, 4, 6\} \subset \Omega$, and then we know that $p(E) = \sum_{s \in E} p(s) = 3k = \frac{1}{2}$!

To be fully formal about this, I should introduce the term *random variable*. When we write $p(s)$, this is equivalent to the more technically correct notation $p(X = s)$, where X is a random variable that models our die. The notation is more correct in that it specifies that we are measuring the probability with which the random variable X takes on the value s. These subtle distinctions are important only to avoid getting tricked by notation: I've told you $p(s)$ is a function from $\Omega \to [0, 1]$, but when we consider multiple random variables

at once, p is going to look like it's a bunch of different functions! In actuality, probability distributions correspond to random variables and $p(s)$ is just a convenient shorthand for $p(X = s)$ when we can infer what random variable X is.

Also worth noting that what we're discussing here are *discrete* random variables. Things get trickier with *continuous* random variables, but discrete random variables will be enough to get us through the semester!

3 But what is a probability, actually?

So far, we've mentioned a few rules about what a probability distribution *can be* (i.e., it abides by those two rules stated above). But what *should* the distribution be?

This may seem obvious for a case like a single die, but it's worth digging into the philosophy while things are simple. One way of thinking about this is what's often called a *frequentist* view. Let's return to our 6-sided die example, with the appropriate X, p, and Ω. Let's think of each $s \in \Omega$ as a possible outcome of a *trial* (i.e., a roll of the die). If we think of $p(s)$ as the *likelihood* of s being the outcome, then (informally) if we run infinite number of trials (i.e., roll infinite dice), the correct choice of $p(e)$ is the proportion of trials that have s as the outcome. Of course, we can't run an infinite amount of trials or even compute that proportion, but we can get at that idea by invoking *limits*: As the number of experiments we run, n, approaches infinity, $p(s)$ should be the ratio between the number of times the outcome is s (what we'll call $c(s)$) and n. That is, writ formally,

$$p(e) = \lim_{n \to \infty} \frac{c(e)}{n}. \tag{3}$$

This formulation may seem quite natural (it is!), but your intuitions may break down a bit if you consider other things we might want to assign probabilities to.

Consider a presidential election — how do we appropriately assign a probability to a candidate winning? By definition, we can only observe the election a single time, and thus the only "correct" distribution would be one that assigns the eventual winner probability 1 and the other candidates probability 0. This isn't perfect either, since we have to drop the limit as part of our frequentist definition, since it doesn't make sense to take the limit as $n \to \infty$ if there can only ever be 1 event.

Instead, one can adopt a *Bayesian* view of probability, where probabilities represent the *credence*, or degree of belief assigned to an event happening (assumed to be assigned by a "reasonable" person). If you're economically minded, this is sometimes formulated as *$p(s)$ is the price at which (a rational/reasonable) you would bet on s occurring if the payout is 1 unit of currency.* A natural

conclusion of this view is that probabilities under this view are inherently *subjective* — there is no way of measuring the right probability like a frequentist view would imply!

Of course, that doesn't mean that Bayesian models are entirely untethered from real data. For example, a model that assigns probabilities to *multiple* elections based on the same underlying principles can be evaluated on their *calibration*: Do events that you assign probability p happen with proportion p? That is, do candidates you give, say, a 1 in 3 chance win 1 out of 3 times? There is a subtle trick here worth pointing out: This allows us to evaluate the reasonableness of the model as a whole (say, $p(\text{winning} \mid \text{candidate})$) but not the legitimacy of the probability assigned to a *particular* candidates odds: That kind of evaluation can only live in the realm of repeated trials!

For the most part, these views of probability are interchangeable — in practice, these views will converge on similar numbers for all the cases we'll care about. However, keeping these views on probability in mind will help us think through what the probabilities we assign actually mean, which will in turn help us understand how to assign probabilities to things when we build probabilistic models.

4 More Probability Ideas

Here, we'll quickly jump through some slightly more sophisticated probability ideas that we'll need for the course. Things will be brief, but don't worry — practice will come!

4.1 Joint Probabilities and Independence

Suppose we have 2 events, $A, B \subseteq \Omega$. What is the probability that both events happen? We can formalize this using the notion of a **joint probability**, which we will notate as $p(A \cap B)$. As the notation indicates, we can think of this as being related to intersection of the sets A and B! Since $A \cap B \subseteq \Omega$, $A \cap B$ represent a *joint* event, and it's probability can be treated just like any other!

$$p(A \cap B) = \sum_{s \in A \cap B} p(s)$$

Things get a little more tricky when we think of two different, but potentially related random variables. Suppose we have Random Variables X and Y with sample spaces Ω_X and Ω_Y, and we want to find the probability that $X \in E_X$ and $Y = E_Y$. Since they have different sample spaces, we can't just take the intersection $E_X \cap E_Y$! Instead, we'll have to construct a new sample space that's the *Cartesian product* of Ω_X and Ω_Y, $\Omega_X \times \Omega_Y$. As a reminder from Discrete Math, this is just the set of pairs of $\{(x, y) \mid x \in \Omega_X, y \in \Omega_y\}$. The first component of the pair represents the outcome of X, while the second represents the outcome of Y.

If that's the case, then what should the joint event space be? Well, the space where the X-component is in E_X and the y-component is in E_Y (i.e., events in the joint space where both E_X and E_Y occurred!).

These two formulations are actually equivalent once we note that we can redefine E_X and E_Y in terms of the joint sample space (you can convince yourself by working this out by hand1).

Now joint probabilities are fairly common things to want to model: Suppose we flip 2 coins, and want to model the odds of two heads. Each coin will be modeled with a random variable C_i has sample space $\Omega_i = \{H, T\}$. To get $p(C_1 = H, C_2 = H)$, we need to first construct the joint sample space for the joint random variable C, $\Omega = \Omega_1 \times \Omega_2 = \{HH, HT, TH, TT\}$, and our joint event is $E = \{HH\}$.

One thing we haven't talked about is the relationship between the probability distributions for individual random variables as opposed to their joint distribution. Why? Because it's tricky! In general, we *can't* describe the probability of a joint event based on the marginal distributions of the individual component random variables!

Now, we intuitively know that there *is*, in fact, a relationship in the case of two coins being flipped. The probability of a coin landing on heads in $\frac{1}{2}$, and the probability of two heads being flipped with 2 coins is $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$ — that is, $p(C_1 = H, C_2 = H) = p(C_1 = H) \cdot p(C_2 = H)$!

It turns out this is only true when the two random variables are **independent** — that means that outcome of one event doesn't affect the likelihood of the outcome of the other. Colloquially, one might just say that outcomes of the two events are *unrelated*.

To wrap that concept up formally, we say that two random variables X, Y are independent if and only if

$$p(x, y) = p(x)p(y)$$

4.2 Computing Marginals

Suppose we're given a joint distribution $P(X = x, Y = y)$ and we want to derive the **marginal** distribution $P(X = x)$?

Conceptually, if we assume we have the joint sample space we constructed above, our goal is just to ignore the second component in the pair! If we don't care what the second component is, we can just sum over all possible values of that second component:

$$p(X = x) = \sum_{y \in \Omega_Y} p(X = x, Y = y)$$

This process is called **Marginalization**.

Why all this margin talk? Imagine a table with outcomes of X and Y along either axis and probabilities in each cell. If we sum across each row and each column and notate those sums in the margins, we find the *marginal* distributions!

4.3 Conditional Probabilities

So if two random variables are *not* independent, then we can't write out their joint probability in terms of their marginal probabilities (i.e., you can't write out $p(x, y)$ in terms of $P(X)$ and $P(Y)$. This is because the outcomes of X depends on the outcome of Y and vice versa. Of course, the relationship between the outcomes of X and Y is interesting, so we'll develop some more theory to deal with this!

We define the **conditional probability** $P(X = x \mid Y = y)$ (or, in short, $p(x \mid y)$) is defined as the probability that the random variable X takes on the value x *if we know that* $Y = y$. Now we can define joint probabilities in terms both a marginal and conditional probability!

$$\begin{aligned} p(X = x, Y = y) &= p(X = x \mid Y = y)p(Y = y) \\ &= p(Y = y \mid X = x)p(X = x). \end{aligned}$$

Of course, we can also do this backwards if we want to derive conditional probabilities:

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}.$$

And we can decompose the joint probability the other way to derive an identity called **Bayes Rule**, which has been extraordinarily influential in a number of ways, some of which we'll see later on in this class!

$$p(X = x \mid Y = y) = \frac{p(Y = y \mid X = x)p(X = x)}{p(Y = y)}.$$

And, if we assume X, Y are independent (and assuming $p(Y = y) > 0$) we can see that...

$$\begin{aligned} p(x, y) = p(x)p(y) &= p(x \mid y)p(y) \\ p(x) &= p(x \mid y). \end{aligned}$$

That is, as we might assume intuitively, if X, Y are independent, the conditional probability of X given Y is just the marginal probability of X (i.e., the outcome of Y doesn't affect the probability of X!). Independence, as we might intuit!

5 Back to NLP

Okay, let's bring this back to NLP.

Suppose we're designing the language modeling task. We know we can't expect the model to predict the next word, but what we can have the model do is produce a *probability distribution* over the next word. Formally, we want to construct a probability distribution over a random variable W which represents the next word in the sentence. Our sample space Ω should consist of every possible word — our **vocabulary**. We could say we want to estimate $p(W)$, but that will likely depend on what words came prior, and since those are given to us in the language modeling task, we will want to define language modeling as estimating the *conditional* probability distribution $p(W = w \mid C = c)$, where C is a random variable representing the prior context![1]

[1] Historically, the task of language modeling was actually defined in terms of the joint probability distribution rather than the conditional one. That is, we were tasked with estimating a probability distribution over all word sequences. However, in practice, these formulations of the task are equivalent!