

# Python Course: Lecture 11

January 30, 2006

## 1 Stochastic Linguistics

- A major branch of computational linguistics is called “stochastic” or “probabilistic” or “corpus-based” linguistics.
- In this branch we use computers to process large samples of natural language (called *training corpora*<sup>1</sup>) and generate probabilistic mathematical models from them.
- Typical corpora might be a year’s worth of Wall Street Journal text, or transcripts of phone conversations, or news broadcasts.
- Though the details vary, the basic strategy is the same. We take some large training corpora to be representative of natural language as a whole. Then we look at some previously unseen piece of natural language (called a *test corpus*) and use mathematics to determine the degree to which it resembles the training corpus.
- The art lies in coming up with meaningful mathematical definitions of “resemble” in this context.
- This general task is called *language modeling*.
- We use *probability* as the metric for similarity because probability is a well-understood branch of mathematics that has proven useful in other areas.
- The basic stochastic linguistic task then is to assign a probability to a string.

## 2 Motivating Example for Probability

- An automatic speech recognizer takes in a sound file of someone talking and spits out text transcripts that are guesses as to what was said.
- Some of the guesses will be better than others. Say a speech recognizer produced two guesses for some segment of sound.

---

<sup>1</sup>From the Latin “corpus” meaning “body” and pluralized accordingly.

1. a lot of the population of the united states is changing
  2. uh little nation a the united states is plain
- Using your knowledge of the English language, you can conclude that (1) is probably a better transcription than (2) because (1) looks like a reasonable English phrase while (2) looks like gibberish.
  - It would be nice to be able to program a computer to make the same reasonable-English/gibberish distinction automatically.
  - One way to do this is to find a large training corpus of what we know to be a good transcription of English speech, and then automatically determine to degree to which each recognizer output resembles that corpus.

### 3 Basic Probability: Counting Birds

#### 3.1 Probability Distributions

- Before applying the mathematics of probability to language, let's look at some basic concepts in probability.
- A concrete example: birdwatching.
- You sit in your backyard one summer's day and keep track of how many birds you see.

Bird	Count	Probability
sparrow	10	0.40
cardinal	5	0.20
bluejay	2	0.08
pigeon	8	0.32
<b>Total</b>	25	1.00

- This is a simple *probability model*. Given this set of numbers, if you say, "I just saw a bird in your backyard," I can make an educated guess as to what kind of bird it is.
- The probability of seeing a sparrow is the number of sparrows divided by the total number of birds.

$$p(\text{sparrow}) = \frac{10}{25} = 0.4 \tag{1}$$

- The probability of seeing a bluejay is the number of bluejays divided by the total number of birds.

$$p(\text{bluejay}) = \frac{2}{25} = 0.08 \tag{2}$$

And so on.

- A *probability distribution* is a function that assigns a number called a probability to certain events. This number must always fall between 0 and 1 inclusive.

$$0 \leq p(x) \leq 1 \tag{3}$$

And the sum of all the probabilities for all the events that could possibly happen must equal 1.

$$\sum_{x \in X} p(x) = 1 \tag{4}$$

Here  $X$  is the set of all the things that could possibly happen, called the *support* of the probability distribution. In the example above

$$X = \{\text{sparrow, cardinal, bluejay, pigeon}\} \tag{5}$$

- The sum in (4) is called the *normalization*. If this sums to 1 we say that the distribution is *properly normalized*.
- Why normalize? To make separate samples comparable. For example, say you watch for 10 days.

Bird	Count	Probability
sparrow	80	0.34
cardinal	52	0.22
bluejay	26	0.11
pigeon	77	0.32
<b>Total</b>	235	1.0

- The raw counts are different, but the probabilities are almost the same. This indicates that our 1-day sample and our 10-day sample are telling us roughly the same thing about our backyard bird distribution.
- Now say you watch birds for 10 days during the winter and see the following numbers.

Bird	Count	Probability
sparrow	69	0.430
cardinal	5	0.030
bluejay	7	0.043
pigeon	79	0.490
<b>Total</b>	160	1.000

- Note that the proportion of cardinals and bluejays has gone way down.
- There's something different about the distribution as a whole. Maybe most of the cardinals and bluejays had flown south for the winter.
- One interpretation of the "something different" observation is to say that your summer data wasn't a good probability model of your winter data.

- The summer data was representative of something. The winter data was representative of something else.
- More abstractly, you can think of probability distributions as cohesive entities that can be more or less similar to one another.

## 4 Conditional Probability

- Say we combined bird counts from both the summer and the winter but kept track of when each bird was observed.

Bird	Summer	Winter	Total
sparrow	80	69	149
cardinal	52	5	57
bluejay	26	7	33
pigeon	77	79	156
<b>Total</b>	235	160	395

- Let  $C$  be a count function. For example

$$C(\text{sparrow, winter}) = 69 \quad (6)$$

$$C(\text{pigeon, summer}) = 77 \quad (7)$$

- Let  $N$  be the total number of birds observed. In this case  $N = 395$ .
- There are a couple of different ways we could break this data down.
- Could get just the probability of seeing a cardinal regardless of the season.

$$p(\text{cardinal}) = \frac{C(\text{cardinal, summer}) + C(\text{cardinal, winter})}{N} = \frac{57}{395} = 0.14 \quad (8)$$

- Could get the probability of seeing a bird in the summer. (e.g. “I just saw a bird. What is the chance that it is summertime?”)

$$p(\text{summer}) = \frac{\sum_{x \in \text{birds}} C(x, \text{summer})}{N} = \frac{235}{395} = 0.59 \quad (9)$$

Where the set birds is equal to {sparrow, cardinal, bluejay, pigeon}.

- These probabilities are called *marginals*. (Because you typically write the total counts in the margins of the table as I have done above.)
- Could calculate the *joint probability* of seeing a cardinal in the summer.

$$p(\text{cardinal, summer}) = \frac{C(\text{cardinal, summer})}{N} = \frac{52}{395} = 0.13 \quad (10)$$

(e.g. “I just saw a bird. What is the probability that it is a cardinal *and* I’m observing it during the summer?”)

- Could calculate the *conditional probability* of seeing a cardinal given that it is currently the summer.

$$p(\text{cardinal}|\text{summer}) = \frac{C(\text{cardinal, summer})}{\sum_{x \in \text{birds}} C(x, \text{summer})} = \frac{52}{235} = 0.22 \quad (11)$$

(e.g. “I know that it is summer. I just saw a bird. What is the probability that it is a cardinal given my *a priori* knowledge about the current season?”)

- Conditional probabilities are normalized with respect to the conditioning information.

$$\sum_{x \in \text{birds}} p(x|\text{summer}) = 1.0 \quad (12)$$

$$\sum_{x \in \text{birds}} p(x|\text{winter}) = 1.0 \quad (13)$$

## 5 Combining Probabilities

- If you have two independent events  $A$  and  $B$  with probabilities  $p(A)$  and  $p(B)$ , the probability of both  $A$  and  $B$  occurring is the product of their individual probabilities,  $p(A)p(B)$ .
- It’s perhaps easiest to see this with dice. Say you want to know the probability of rolling a 6. If you roll a die, there are six possible outcomes, so the probability of rolling a 6 is  $\frac{1}{6}$ .
- Now what’s the probability of rolling two sixes on two dice? If you roll one die, there are six possible outcomes. If you then roll the second die, there are six possible outcomes for each of those previous six, making  $6 \times 6 = 36$  outcomes. The chance of getting two sixes is then  $\frac{1}{36} = \frac{1}{6} \frac{1}{6}$ .
- Note that this is only true for *independent* probabilities. If the outcome of event  $A$  provides any information about the outcome of event  $B$ , then the probability of both events occurring may not be  $p(A)p(B)$ .
- Often when building probabilistic models we make *independence assumptions*. For instance we know that natural language is a highly context-dependent phenomenon and the appearance of a particular word earlier in a sentence will have an effect on what sentences come later, but for the sake of mathematical ease we may pretend that this is not the case. Often mathematical models have to make this kind of tradeoff between tractability and accuracy.

## 6 Logarithms

- Typically in language applications we will calculate the probabilities of many independent events. For instance we might model the probability of the sentence “The rain in Spain stays mainly in the plane” like so

$$p(\text{The}) \times p(\text{rain}) \times p(\text{in}) \times p(\text{Spain}) \times p(\text{stays}) \times p(\text{mainly}) \times p(\text{in}) \dots \quad (14)$$

where the  $p()$  functions are probabilities that our model assigns to individual words.

- Probabilities are always numbers between 0 and 1, and when you multiply two probabilities together you get an even smaller number between zero and one.
- Because natural language applications deal with large amounts of rare events, the numbers you work with quickly get very small. It’s easy to underflow your computer’s floating point accuracy.
- To get around this problem, we often work with the logarithm of probabilities.
- The log function is the inverse of exponentiation.

$$b^x = y \leftrightarrow \log_b y = x \quad (15)$$

For example

$$2^3 = 8 \leftrightarrow \log_2 8 = 3 \quad (16)$$

- The log of a number between 0 and 1 yields a number between negative infinity (for zero) and 0 (for 1). The magnitude of logs tend to be much smaller than that of their arguments. For example

$$\log_2 10^{-200} = -664.38 \quad (17)$$

$10^{-200}$  is an extremely small number, while  $-664$  is of a much more reasonable magnitude.

- The other reason we use logs is because the log of a product of numbers is the sum of their logs.

$$\log xy = \log x + \log y \quad (18)$$

Or more generally

$$\log \prod_{i=0}^n x_i = \sum_{i=0}^n \log x_i \quad (19)$$

- It probably won’t affect we do in in this class, but this switch from products to sums often makes probability equations easier to work with.
- In natural language applications it is a convention to take all your logarithms to the base 2. Numbers taken to the log base 2 are said to be units of “bits”.

- We do this because there is another branch of mathematics called Information Theory that likes to work in bits, and Information Theory is very useful in modeling natural language. (Though we probably won't touch on it in this class.)