Probabilistic context-free grammars (PCFGs)

1 Looking back at PFSAs/HMMs

1.1 Backward probabilities

A backward probability is the probability, conditioned upon beginning at a particular given state x, that the derivation/run of the machine will be completed in a way that produces a particular given sequence of output symbols $w_0 w_1 \dots w_n$. So the general form, in notation familiar from last week, looks like this:

$$\Pr(W_k = w_0, W_{k+1} = w_1, \dots, W_{k+n} = w_n, \text{end at } S_{k+n+1} \mid S_k = x) \tag{1}$$

As a bit of a shortcut we can imagine writing this instead as:

$$\Pr(W_k W_{k+1} \dots W_{k+n} = w_0 w_1 \dots w_n, \text{end at } S_{k+n+1} \mid S_k = x)$$

Some new notation will be useful now: whereas W_k is the random variable representing the single output symbol that is emitted on the transition out of the state at index k, I'll write $W_{\widehat{k}}$ for the random variable representing the *sequence of output symbols* that appears to the right of state-position k in a complete derivation. So now backwards probabilities can simply be written as:

$$\Pr(W_{\widehat{k}} = w_0 w_1 \dots w_n \mid S_k = x) \tag{2}$$

We calculated these probabilities via recursion on the sequence-of-output-symbols argument:

$$\Pr(W_{\widehat{k}} = \epsilon \mid S_k = x) = \Pr(\text{end at } x) \tag{3}$$

$$\begin{aligned}\Pr(W_{\widehat{k}} = w_0 w_1 \dots w_n \mid S_k = x) = \sum_{next} \big[& \Pr(S_{k+1} = next \mid S_k = x) \\ & \times \Pr(W_k = w_0 \mid S_k = x, S_{k+1} = next) \\ & \times \Pr(W_{\widehat{k+1}} = w_1 \dots w_n \mid S_{k+1} = next) \big]\end{aligned} \tag{4}$$

1.2 Forward probabilities

A forward probability is the probability that a derivation/run of the machine will produce a particular given sequence of output symbols $w_0 w_1 \dots w_n$ and thereafter end up in a particular given state x. We might write this in either of these ways:

$$\Pr(W_0 = w_0, W_1 = w_1, \dots, W_n = w_n, S_{n+1} = x) \tag{5}$$

$$\Pr(W_0 W_1 \dots W_n = w_0 w_1 \dots w_n, S_{n+1} = x)$$

Analogous to the new notation $W_{\widehat{k}}$, I'll write $W_{\overleftarrow{k}}$ for the random variable representing the sequence of output symbols that appears to the left of state-position k in a complete derivation. So now forward probabilities can be written as:

$$\Pr(W_{\overleftarrow{k}} = w_0 w_1 \dots w_n, S_k = x) \tag{6}$$

We calculated these probabilities recursively like this:

$$\Pr(W_{\overleftarrow{k}} = \epsilon, S_k = x) = \Pr(S_0 = x) \tag{7}$$

$$\begin{aligned}\Pr(W_{\overleftarrow{k}} = w_0 w_1 \dots w_n, S_k = x) = \sum_{prev} \big[& \Pr(W_{\overleftarrow{k-1}} = w_0 \dots w_{n-1}, S_{k-1} = prev) \\ & \times \Pr(S_k = x \mid S_{k-1} = prev) \\ & \times \Pr(W_{k-1} = w_n \mid S_{k-1} = prev, S_k = x) \big]\end{aligned} \tag{8}$$

1.3 A difference between backward and forward probabilities

Notice that the way the index k is being used differs slightly in backward probabilities from forward probabilities. If we're told a particular backwards probability, say for example we're given

$$\Pr(W_{\widehat{k}} = \text{ran quickly} \mid S_k = 20) = 0.2 \tag{9}$$

then this is a fact that's completely independent of any particular index k: it's a fact about a partial-derivation that might contribute the last two words of a hundred-word word-sequence, or the last two words of a three-word word-sequence (or even of a two-word word-sequence). Put differently, it's fundamentally a statement about the relationship that 'ran quickly' and state 20 stand in to each other. So more specifically, what this actually says is:

$$\forall k \in \mathbb{N}, \quad \Pr(W_{\widehat{k}} = \text{ran quickly} \mid S_k = 20) = 0.2 \tag{10}$$

On the other hand, if we're given a value for a particular forward probability such as

$$\Pr(W_{\overleftarrow{k}} = \text{the cat}, S_k = 20) = 0.3 \tag{11}$$

this is a bit different: it says there is a k (namely 2) such that 'the cat' and state 20 and k stand in a certain relationship to each other, but there are many other indices k' (namely anything other than 2) for which this three-way relationship does not hold. This suggests that what the equation above really says is:

$$\exists k \in \mathbb{N}, \quad \Pr(W_{\overleftarrow{k}} = \text{the cat}, S_k = 20) = 0.3 \tag{12}$$

or actually, what will be more useful is to view it as

$$\sum_{k \in \mathbb{N}} \big[\Pr(W_{\overleftarrow{k}} = \text{the cat}, S_k = 20) \big] = 0.3 \tag{13}$$

which is a sum of natural-number-many terms, all but one of which (the one where $k = 2$) is zero.

So the recursive specification in (7) and (8) above might instead be rewritten more carefully as follows, where we make use of the fact that the only k for which it's possible that $W_{\overleftarrow{k}} = \epsilon$ is 0, and the only k for which it's possible that $W_{\overleftarrow{k}} = w_0 w_1 \dots w_n$ is $n+1$.

$$\begin{aligned}\sum_{k \in \mathbb{N}} \big[\Pr(W_{\overleftarrow{k}} = \epsilon, S_k = x) \big] &= \Pr(W_{\overleftarrow{0}} = \epsilon, S_0 = x) \\ &= \Pr(S_0 = x)\end{aligned} \tag{14}$$

$$\begin{aligned}\sum_{k \in \mathbb{N}} \big[\Pr(W_{\overleftarrow{k}} = w_0 w_1 \dots w_n, S_k = x) \big] &= \Pr(W_{\overleftarrow{n+1}} = w_0 w_1 \dots w_n, S_{n+1} = x) \\ &= \sum_{prev} \big[\Pr(W_{\overleftarrow{n}} = w_0 \dots w_{n-1}, S_n = prev) \\ &\qquad \times \Pr(S_{n+1} = x \mid S_n = prev) \\ &\qquad \times \Pr(W_n = w_n \mid S_n = prev, S_{n+1} = x) \big]\end{aligned} \tag{15}$$

Notice that this makes it look a bit more like recursion on the natural number subscripts, rather than recursion on the symbol-sequences themselves.

When we come to *outside probabilities* in PCFGs, we'll see an analogous case where there is more than one non-zero contribution to a sum like this.

2 Probabilistic context-free grammars

An interesting way to think about PCFGs is in terms of the probabilities of certain categories (e.g. perhaps N, V, VP, etc.) appearing at certain *addresses* in a tree. This corresponds to the probabilities of certain states appearing at certain positions (indexed by natural numbers) in a state-sequence.

Recall that ϵ is the address of the root node, $\alpha 0$ is the left daughter of the node at address α, and $\alpha 1$ is the right daughter of the node at address α. It will be important to think of the positions $\alpha 0$ and $\alpha 1$ as both "following" the position α, in the same way that position $k+1$ in a state-sequence follows position k. (In all of the trees we consider here, each node will have either exactly two daughters or no daughters.)

A PCFG is usually presented like this:

1.0	S	$\rightarrow$	NP VP	0.3	VP	$\rightarrow$	VP ADV
0.6	NP	$\rightarrow$	John	0.6	VP	$\rightarrow$	left
0.4	NP	$\rightarrow$	D N	0.1	VP	$\rightarrow$	arrived
0.7	D	$\rightarrow$	the	0.2	ADV	$\rightarrow$	quickly
0.3	D	$\rightarrow$	a	0.8	ADV	$\rightarrow$	slowly
0.5	N	$\rightarrow$	dog				
0.5	N	$\rightarrow$	cat				

Each of the probabilities here says something about what will appear *under* a particular node in a tree, conditioned upon the label of that node.

To understand what these are saying in terms of tree positions, we'll introduce a bit of notation:

- C_α will be the random variable representing the category at position α in the tree. (Think of this as analogous to S_k.)
- $W_{\widehat{\alpha}}$ will be the random variable representing the sequence of output symbols dominated by the node at position α. (Think of this as analogous to $W_{\widehat{k}}$.)

Then what the grammar above is saying is this:

$$\begin{aligned}
\Pr(C_{\alpha 0} = \text{NP}, C_{\alpha 1} = \text{VP} \mid C_\alpha = \text{S}) &= 1.0 \\
\Pr(W_{\widehat{\alpha}} = \text{John} \mid C_\alpha = \text{NP}) &= 0.6 \\
\Pr(C_{\alpha 0} = \text{D}, C_{\alpha 1} = \text{N} \mid C_\alpha = \text{NP}) &= 0.4 \\
\Pr(W_{\widehat{\alpha}} = \text{the} \mid C_\alpha = \text{D}) &= 0.7 \\
\Pr(W_{\widehat{\alpha}} = \text{a} \mid C_\alpha = \text{D}) &= 0.3 \\
\Pr(W_{\widehat{\alpha}} = \text{dog} \mid C_\alpha = \text{N}) &= 0.5 \\
\Pr(W_{\widehat{\alpha}} = \text{cat} \mid C_\alpha = \text{N}) &= 0.5
\end{aligned}
\qquad
\begin{aligned}
\Pr(C_{\alpha 0} = \text{VP}, C_{\alpha 1} = \text{ADV} \mid C_\alpha = \text{VP}) &= 0.3 \\
\Pr(W_{\widehat{\alpha}} = \text{left} \mid C_\alpha = \text{VP}) &= 0.6 \\
\Pr(W_{\widehat{\alpha}} = \text{arrived} \mid C_\alpha = \text{VP}) &= 0.1 \\
\Pr(W_{\widehat{\alpha}} = \text{quickly} \mid C_\alpha = \text{ADV}) &= 0.2 \\
\Pr(W_{\widehat{\alpha}} = \text{slowly} \mid C_\alpha = \text{ADV}) &= 0.8
\end{aligned}$$

And we'll make the usual implicit assumption that the (only) start symbol is S, which in probabilistic terms amounts to the assumption that $\Pr(C_\epsilon = \text{S}) = 1.0$.

There are three kinds of questions we might want to ask about these grammars:

(16)
a. Given a PCFG, what is the probability of a particular symbol-sequence $w_1 w_2 \dots w_n$ being generated?
b. Given a PCFG and a particular symbol-sequence $w_1 w_2 \dots w_n$, what tree structure best explains an observation of $w_1 w_2 \dots w_n$?
c. Given some symbol-sequence $w_1 w_2 \dots w_n$ a plain CFG, what values for the rules' probabilities best explains an observation of $w_1 w_2 \dots w_n$?

It's helpful to think about these questions as generalizations of the analogous questions for PFSAs/HMMs. For PFSAs/HMMs, a major part of answering these questions is being able to compute forward and backward probabilities. The relevant generalization of a backward probability is known as an *inside probability*, and the relevant generalization of a forward probability is known as an *outside probability*.

3 Inside and outside probabilities

These two kinds of probabilities can't be treated completely separately the way forward and backward probabilities can for HMMs: in order to calculate outside probabilities, we must already know (or be able to calculate) inside probabilities.

3.1 Inside probabilities

An inside probability has the following general form:

$$\Pr(W_{\widehat{\alpha}} = w_0 w_1 \dots w_n \mid C_\alpha = x) \tag{17}$$

For example, $\Pr(W_{\widehat{\alpha}} = \text{the cat} \mid C_\alpha = \text{NP})$ is the probability of generating 'the cat' as the words dominated by a particular node, given that that node has the category NP.

These can be calculated in a relatively straightforward manner via recursion on the sequence-of-symbols argument, similarly to the equations for backward probabilities in (3) and (4). A difference is that the base case is the length-one sequences (rather than the length-zero sequence), because a node must dominate at least one terminal symbol.

$$\Pr(W_{\widehat{\alpha}} = w \mid C_\alpha = x) \text{ is a probability directed specified by the grammar} \tag{18}$$

$$\begin{aligned} \Pr(W_{\widehat{\alpha}} = w_0 w_1 \dots w_n \mid C_\alpha = x) = \sum_i \sum_\ell \sum_r \big[& \Pr(C_{\alpha 0} = \ell, C_{\alpha 1} = r \mid C_\alpha = x) \\ & \times \Pr(W_{\alpha 0} = w_0 \dots w_i \mid C_{\alpha 0} = \ell) \\ & \times \Pr(W_{\alpha 1} = w_{i+1} \dots w_n \mid C_{\alpha 1} = r) \big] \end{aligned} \tag{19}$$

Notice that the value of an inside probability is "for all addresses α", in the same way that a backwards probability is "for all indices k". So strictly speaking, if we're told that the inside probability of generating 'the cat' from the category NP is 0.4, what we're really being told is:

$$\forall \alpha \in \{0,1\}^*, \quad \Pr(W_{\widehat{\alpha}} = \text{the cat} \mid C_\alpha = \text{NP}) = 0.4$$

3.2 Outside probabilities

We need to introduce a bit more notation at this point: $W_{\overleftarrow{\alpha}}$ and $W_{\overrightarrow{\alpha}}$ are the random variables representing the sequence of output symbols to the left of the subtree rooted at position α, and the sequence of output symbols to the right of the subtree rooted as position α, respectively.

An outside probability has the following general form:

$$\sum_{\alpha \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\alpha}} = w_0 \dots w_n, W_{\overrightarrow{\alpha}} = v_0 \dots v_m, C_\alpha = x) \big] \tag{20}$$

We're summing across possible addresses here in the same way that we summed across indices with forward probabilities in (13). But notice that even if we fix the output sequence $w_0 \dots w_n$ and the output sequence $v_0 \dots v_m$ and the category x, there might still be multiple different addresses α for which this probability is non-zero; in contrast to the way, if you fix an output sequence $w_0 \dots w_n$ and a state x, there is a unique index k for which $\Pr(W_{\overleftarrow{k}} = w_0 \dots w_n, S_k = x)$ can be non-zero.

The recursion here roughly mirrors what we saw with forward probabilities in (14) and (15). What we make use of here is that (i) the only address α for which it's possible that $W_{\overleftarrow{\alpha}}$ and $W_{\overrightarrow{\alpha}}$ are both empty is the empty address, and (ii) the address α in all other cases will either be of the form $\beta 0$ or $\beta 1$.

Here's the base case:

$$\sum_{\alpha \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\alpha}} = \epsilon, W_{\overrightarrow{\alpha}} = \epsilon, C_\alpha = x) \big] = \Pr(C_\epsilon = x)$$

Here's the recursive step:

$$\begin{aligned}
&\sum_{\alpha \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\alpha}} = w_0 \dots w_n, W_{\overrightarrow{\alpha}} = v_0 \dots v_m, C_\alpha = x) \big] \\
&\qquad = \sum_{\beta \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\beta 0}} = w_0 \dots w_n, W_{\overrightarrow{\beta 0}} = v_0 \dots v_m, C_{\beta 0} = x) \big] \\
&\qquad\quad + \sum_{\beta \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\beta 1}} = w_0 \dots w_n, W_{\overrightarrow{\beta 1}} = v_0 \dots v_m, C_{\beta 1} = x) \big]
\end{aligned}$$

$$\begin{aligned}
&\sum_{\beta \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\beta 0}} = w_0 \dots w_n, W_{\overrightarrow{\beta 0}} = v_0 \dots v_m, C_{\beta 0} = x) \big] \\
&\qquad = \sum_{\beta \in \{0,1\}^*} \sum_i \sum_p \sum_r \big[\Pr(C_{\beta 0} = x, C_{\beta 1} = r \mid C_\beta = p) \\
&\qquad\qquad \times \Pr(W_{\widehat{\beta 1}} = v_0 \dots v_i \mid C_{\beta 1} = r) \\
&\qquad\qquad \times \Pr(W_{\overleftarrow{\beta}} = w_0 \dots w_n, W_{\overrightarrow{\beta}} = v_{i+1} \dots v_m, C_\beta = p) \big] \\
&\qquad = \sum_i \sum_p \sum_r \big[\Pr(C_{\beta 0} = x, C_{\beta 1} = r \mid C_\beta = p) \\
&\qquad\qquad \times \Pr(W_{\widehat{\beta 1}} = v_0 \dots v_i \mid C_{\beta 1} = r) \\
&\qquad\qquad \times \sum_{\beta \in \{0,1\}^*} \Pr(W_{\overleftarrow{\beta}} = w_0 \dots w_n, W_{\overrightarrow{\beta}} = v_{i+1} \dots v_m, C_\beta = p) \big]
\end{aligned}$$

$$\begin{aligned}
&\sum_{\beta \in \{0,1\}^*} \big[\Pr(W_{\overleftarrow{\beta 1}} = w_0 \dots w_n, W_{\overrightarrow{\beta 1}} = v_0 \dots v_m, C_{\beta 1} = x) \big] \\
&\qquad = \sum_{\beta \in \{0,1\}^*} \sum_i \sum_p \sum_\ell \big[\Pr(C_{\beta 0} = \ell, C_{\beta 1} = x \mid C_\beta = p) \\
&\qquad\qquad \times \Pr(W_{\widehat{\beta 0}} = w_{i+1} \dots w_n \mid C_{\beta 0} = \ell) \\
&\qquad\qquad \times \Pr(W_{\overleftarrow{\beta}} = w_0 \dots w_i, W_{\overrightarrow{\beta}} = v_0 \dots v_m, C_\beta = p) \big] \\
&\qquad = \sum_i \sum_p \sum_\ell \big[\Pr(C_{\beta 0} = \ell, C_{\beta 1} = x \mid C_\beta = p) \\
&\qquad\qquad \times \Pr(W_{\widehat{\beta 0}} = w_{i+1} \dots w_n \mid C_{\beta 0} = \ell) \\
&\qquad\qquad \times \sum_{\beta \in \{0,1\}^*} \Pr(W_{\overleftarrow{\beta}} = w_0 \dots w_i, W_{\overrightarrow{\beta}} = v_0 \dots v_m, C_\beta = p) \big]
\end{aligned}$$

Once you understand this, you can be confident that you really, *really* understand context-free grammars.