The Inside-Outside Algorithm

Karl Stratos

1 The Setup

Our data consists of Q observations $O^{(1)}, \ldots, O^{(Q)}$ where each observation is a sequence of symbols in some set $X = \{1, \ldots, n\}$. We suspect a Probabilistic Context-Free Grammar (PCFG) $G = (H, X, S, R, q)$ is responsible for what we observe. That is, we assume there are hidden nonterminals $H = \{1, \ldots, m\}$ that has generated the observation symbols, starting with a certain $S \in H$, according to rules $r \in R$ defined for each $i, j, k \in H$ and $x \in X$ as

$$i \to jk$$

and

$$i \to x,$$

with probability dictated by $q : R \to \mathbb{R}$. Note that for any $i \in H$, q must satisfy

$$\sum_{j,k \in H} q(i \to jk) + \sum_{x \in X} q(i \to x) = 1.$$

In other words, $q(i \to jk)$ and $q(i \to x)$ represent the conditional probabilities $P(jk \mid i)$ and $P(m \mid i)$.

2 The Algorithm

The goal of the inside-outside algorithm is to infer the parameters a and b, for all $i, j, k \in H$ and $x \in X$, such that

$$a(i, j, k) \text{ as an estimation of } q(i \to jk)$$
$$b(i, x) \text{ as an estimation of } q(i \to x).$$

Of course, to make a and b valid probabilities, we must also have

$$\sum_{j,k \in H} a(i, j, k) + \sum_{x \in X} b(i, x) = 1. \tag{1}$$

It is a straightforward application of the EM principle. In each iteration, it "counts" the occurrences of rules in the data with the current parameter values:

$$\widehat{C}(i),\ \widehat{C}(i \to jk,\ i),\ \widehat{C}(i \to x,\ i). \tag{2}$$

This will allow us to make an MLE update to the parameters

$$a(i, j, k) = \frac{\widehat{C}(i \to jk,\ i)}{\widehat{C}(i)},\ b(i, x) = \frac{\widehat{C}(i \to x,\ i)}{\widehat{C}(i)}.$$

The algorithm starts by filling a and b with values that obey Eq. (1), and repeats the above updates until a local optimum is reached.

3 How to Count

Hence the whole of the algorithm hinges on the following question:

How to count the rules $i \to jk$ and $i \to x$ when we don't see them?

We will first try to extract the local counts from each observation sequence $O^{(q)}$,

$$\widehat{C}(i \mid O^{(q)}),\ \widehat{C}(i \to jk,\ i \mid O^{(q)}),\ \widehat{C}(i \to x,\ i \mid O^{(q)}) \tag{3}$$

so that the global counts in Eq. (2) can be computed

$$\begin{aligned}
\widehat{C}(i) &= \sum_{q=1}^{Q} \widehat{C}(i \mid O^{(q)}) \\
\widehat{C}(i \to jk,\ i) &= \sum_{q=1}^{Q} \widehat{C}(i \to jk,\ i \mid O^{(q)}) \\
\widehat{C}(i \to x,\ i) &= \sum_{q=1}^{Q} \widehat{C}(i \to x,\ i \mid O^{(q)}).
\end{aligned}$$

Thus we need a way to relate the probabilities to specific observations in the data. For notational simplicity, fix $O = O(1) \dots O(T)$ to be a particular observation sequence, with each $O(1), \dots, O(T) \in X$.

By now, we are familiar with the principle of the expected count, so what we do below shouldn't be a surprise. Given O, we will estimate the following conditional probabilities for all $1 \le s \le t \le T,\ i, j, k \in H,\ x \in X$,

$$P(i \xrightarrow{*} O(s) \dots O(t) \mid O) \tag{4}$$

$$P(i \xrightarrow{*} jk \xrightarrow{*} O(s) \dots O(t) \mid O) \tag{5}$$

$$P(i \to O(s) \mid O). \tag{6}$$

Then the targets in Eq. (3) for O are nothing but

$$\begin{aligned}
\widehat{C}(i \mid O) &= \sum_{s=1}^{T} \sum_{t=s}^{T} P(i \xrightarrow{*} O(s) \dots O(t) \mid O) \\
\widehat{C}(i \to jk,\ i \mid O) &= \sum_{s=1}^{T-1} \sum_{t=s+1}^{T} P(i \to jk \xrightarrow{*} O(s) \dots O(t) \mid O) \\
\widehat{C}(i \to x,\ i \mid O) &= \sum_{s:O(s)=x} P(i \to O(s) \mid O).
\end{aligned}$$

Note we used the fact that the probability of a nonterminal expanding to its children $P(A \to B)$ is implicitly a joint probability $P(A \to B, A)$, because

$$P(A \to B) = P(A \to B) \cdot 1 = P(A \to B) P(A \mid A \to B) = P(A \to B, A).$$

4 Inside and Outside Probabilities

The key quantities that will allow us to estimate the probabilities in Eq. (4–6) spring from an elegant decomposition of the tree space of O with the so-called inside and outside trees.

An inside tree rooted at $i \in H$ spanning $O(s) \dots O(t)$ has an associated (inside) probability

$$e(s,t,i) = P(i \stackrel{*}{\rightarrow} O(s) \dots O(t)).$$

An outside tree rooted at the top spanning all observation symbols except for some $i \in H$ spanning $O(s) \dots O(t)$ has an associated (outside) probability

$$f(s,t,i) = P(S \stackrel{*}{\rightarrow} O(1) \dots O(s-1),\ i\ , O(t+1) \dots O(T)).$$

Let's muse over how precise these knives are. Two crucial facts are that we immediately have the probability of the sequence $P(O) = P(S \stackrel{*}{\rightarrow} O(1) \dots O(T))$, given by $e(1,T,S)$, and that the product of an inside and outside quantity that "click together" (i.e., cover $O(1), \dots, O(T)$) always yields a joint probability of the sequence and a rule. For instance, $e(s,t,i)f(s,t,i) = P(O,\ i \stackrel{*}{\rightarrow} O(s) \dots O(t))$. Therefore, we can devide it by $P(O)$ to have a *conditional probability of a rule given the sequence*, which is the form of Eq. (4–6)!

Eq. (4) is given by

$$P(i \stackrel{*}{\rightarrow} O(s) \dots O(t) \mid O) = \Big(e(s,t,i) f(s,t,i) \Big) / P(O)$$

Eq. (5) is given by

$$P(i \stackrel{*}{\rightarrow} jk \stackrel{*}{\rightarrow} O(s) \dots O(t) \mid O) = \left(\sum_{r=s}^{t-1} a(i,j,k) e(s,r,j) e(r+1,t,k) f(s,t,i) \right) / P(O).$$

Eq. (6) is given by

$$P(i \rightarrow O(s) \mid O) = \Big(e(s,s,i) f(s,s,i) \Big) / P(O).$$

In other words, if we can compute the values of e and f, we are done. We can first compute e bottom-up from scratch, filling

$$e(s,s,i) = b(i, O(s))$$

for all $s = 1, \dots, T$ and $i \in H$, and summing over all possible binarization of $i \rightarrow jk$

$$e(s,t,i) = \sum_{j,k \in H} \sum_{r=s}^{t-1} a(i,j,k) e(s,r,j) e(r+1,t,k).$$

It is worth emphasizing that in practice, the above expression must be implemented in what everybody calls the "CYK style". It means for efficient dynammic programming, we gradually grow the length of the span (s,t) to use the previous spans (s,r) and $(r+1,t)$ for $r = s, \dots, t$. Here is a pseudocode:

Input: A sequence $O(1)\dots O(T)$.
Output: Inside probabilities $e(s,t,i)$ for $1 \le s \le t \le T,\ i \in H$.

1. For $s = 1,\dots,T$, for $i \in H$, set $e(s,s,i) = b(i,O(s))$.
2. For $l = 1,\dots,T-1$, for $s = 1,\dots,T-l$,
   - set $t = s + l$
   - For $i \in H$, compute

$$e(s,t,i) = \sum_{j,k \in H} \sum_{r=s}^{t-1} a(i,j,k)e(s,r,j)e(r+1,t,k).$$

We can then compute f top-down from the values of e. The base case is easy:

$$f(1,T,i) = \begin{cases} 1 & \text{if } i = S \\ 0 & \text{otherwise} \end{cases}.$$

A slight complication in the recursion is that i may have come from its parent j as a left child or a right child, the other child being k. So we must sum over the two possible cases of binarization $j \to ik$ and $j \to ki$.

$$f(s,t,i) = \sum_{j,k \in H} \left(\sum_{r=t+1}^{T} a(j,i,k)e(t+1,r,k)f(s,r,j) + \sum_{r=1}^{s-1} a(j,k,i)e(r,s-1,k)f(r,t,j) \right).$$

Again, similar to the case of e, the above expression must be computed in the right order (of decreasing length (s,t)) to be able to use the previous spans in the recursion. Here is a pseudocode:

Input: A sequence $O(1)\dots O(T)$ and its inside probabilities e.
Output: Outside probabilities $f(s,t,i)$ for $1 \le s \le t \le T,\ i \in H$.

1. For $i \in H$, set

$$f(1,T,i) = \begin{cases} 1 & \text{if } i = S \\ 0 & \text{otherwise} \end{cases}.$$

2. For $l = T-2,\dots,0$, for $s = 1,\dots,T-l$,
   - set $t = s + l$
   - For $i \in H$, compute

$$f(s,t,i) = \sum_{j,k \in H} \Bigg(\sum_{r=t+1}^{T} a(j,i,k)e(t+1,r,k)f(s,r,j) + \sum_{r=1}^{s-1} a(j,k,i)e(r,s-1,k)f(r,t,j) \Bigg).$$

One more tip on the implementation. The following expression can be used as a certificate of correctness of e and f:

$$P(O) = e(1, T, S) = \sum_{i \in H} b(i, O(s)) f(s, s, i)$$

for all $1 \leq s \leq T$. Note that $b(i, O(s)) f(s, s, i) = P(O, i \to O(s))$, and that $\sum_{i \in H} b(i, O(s)) = \sum_{i \in H} P(i \to O(s)) = 1$ because our grammar (implicitly in Chomsky Normal Form) mandates that every observation symbol has a parent nonterminal.

We have everything to carry out the full algorithm now.

1. Initialize a and b without violating Eq. (1).
2. For $q = 1, \ldots, Q$, compute the inside probabilities e and outside probabilities f for $O^{(q)}$, and use them to calculate Eq. (4–6), which would in turn allow us to calculate $\widehat{C}(i \mid O^{(q)})$, $\widehat{C}(i \to jk,\ i \mid O^{(q)})$, $\widehat{C}(i \to x,\ i \mid O^{(q)})$ for all $i, j, k \in H$ and $x \in X$.
3. Compute the expected counts over the whole data

$$\widehat{C}(i) = \sum_{q=1}^{Q} \widehat{C}(i \mid O^{(q)})$$

$$\widehat{C}(i \to jk,\ i) = \sum_{q=1}^{Q} \widehat{C}(i \to jk,\ i \mid O^{(q)})$$

$$\widehat{C}(i \to x,\ i) = \sum_{q=1}^{Q} \widehat{C}(i \to x,\ i \mid O^{(q)}).$$

4. Make an MLE update to a and b

$$a(i, j, k) = \frac{\widehat{C}(i \to jk,\ i)}{\widehat{C}(i)},\ b(i, x) = \frac{\widehat{C}(i \to x,\ i)}{\widehat{C}(i)}.$$

   and return to step 2 unless a local optimum is reached.

References

Lari, K. and Young, S. J. (1990). The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 4:35–56.