Machine Learning (CS 726) *1 Dec 2011*

Hidden Markov Models II

We now shift back to unsupervised learning. We want to follow the EM framework for inference in *hidden* Markov models. The basic idea is to think of the state lattice, and ask yourself "what is the probability of going through a particular node." This will give us the "fractional count" that we use to re-estimate the parameters of the model.

The solution to this problem has many names: the forward-backward algorithm, the inside-outside algorithm, the Baum-Welch algorithm, etc. (essentially because it's been re-derived a bunch of times). The intuition is that the "forward" probabilities compute the likelihood of *getting* to a node; the "backward" probabilities compute the likelihood of getting from that node to the end. We multiply them together, and voila, we have the probability of *being* at that node.

The "forward" procedure is a recursive process, analogous to the Viterbi algorithm. The difference is that in the Viterbi algorithm we asked "what is the probability of getting to this node *assuming we've followed the optimal path thus far*?" In forward, we ask "what is the probability of getting to this node, period?"

In particular, let:

$$\alpha_i(t) = P(x_1, x_2, \ldots, x_{t-1}, y_t = i \mid \Theta)$$

Here, x is the observed sequence, y is the latent sequence, and Θ are all the model parameters. This is the probability of getting to state i at time t, and observing the sequence up to $t-1$ in doing so.

We compute these recursively as:

$$\alpha_i(1) = \pi_i$$

This is the base case: the probability of observing *nothing* and getting to state i at time 1 is just the probability of starting in state i.

$$\begin{aligned}
\alpha_j(t+1) &= P(x_1, x_2, \ldots, x_t, y_{t+1} = j \mid \Theta) \\
&= P(x_1, x_2, \ldots, x_{t-1}, x_t, y_{t+1} = j \mid \Theta) \\
&= \sum_i P(x_1, x_2, \ldots, x_{t-1}, x_t, y_t = i, y_{t+1} = j \mid \Theta) \\
&= \sum_i P(x_1, x_2, \ldots, x_{t-1}, y_t = i \mid \Theta) P(x_t, y_{t+1} = j \mid y_t = i, \Theta) \\
&= \sum_i \alpha_i(t) P(y_{t+1} = j \mid y_t = i, \Theta) P(x_t \mid y_{t+1} = j, \Theta) \\
&= \sum_i \alpha_i(t) a_{ij} b_{j,x_t}
\end{aligned}$$

The first line is definition, the second is just expanding, the third is marginalizing over y_t, the fourth is chain rule, the fifth is applying the definition of α, the sixth is Markov assumption and the last step is plugging in a and b tables.

The backward procedure is entirely analogous. Define:

$$\beta_i(t) = P(x_t, \ldots, x_T \mid y_t = i, \Theta)$$

This is the probability of the sequence of observations $x_t, \ldots, x_T$ being emitted, given that the machine starts in state i at time t.

As suggested by the name, we compute these right-to-left:

$$\beta_i(T+1) = 1$$

The probability of doing nothing is just 1, no matter what state we're in.

$$\begin{aligned}
\beta_i(t) &= P(x_t, \ldots, x_T \mid y_t = i, \Theta) \\
&= P(x_t, x_{t+1}, \ldots, x_T \mid y_t = i, \Theta) \\
&= \sum_j P(x_t, x_{t+1}, \ldots, x_T, y_{t+1} = j \mid y_t = i, \Theta) \\
&= \sum_j P(x_t, y_{t+1} = j \mid y_t = i, \Theta) P(x_{t+1}, \ldots, x_T \mid y_t = i, x_t, y_{t+1} = j, \Theta) \\
&= \sum_j P(x_t \mid y_t = i, \Theta) P(y_{t+1} = j \mid y_t = i, \Theta) P(x_{t+1}, \ldots, x_T \mid y_{t+1} = j, \Theta) \\
&= \sum_j b_{i,x_t} a_{ij} P(x_{t+1}, \ldots, x_T \mid y_{t+1} = j, \Theta) \\
&= \sum_j b_{i,x_t} a_{ij} \beta_j(t+1)
\end{aligned}$$

Again, first is definition, second is expanding, third is marginalizing over y_{t+1}, fourth is chain rule, fifth is Markov assumption, sixth is definition of model parameters and last is applying the defition of β recursively.

Now, the magic happens. We want to compute the probability that we are in some lattice state at some point in time:

$$\begin{aligned}
P(\boldsymbol{x}, y_t = i \mid \Theta) &= P(x_1, \ldots, x_T, y_t = i \mid \Theta) \\
&= P(x_1, \ldots, x_{t-1}, y_t = i, x_t, \ldots, x_T \mid \Theta) \\
&= P(x_1, \ldots, x_{t-1}, y_t = i \mid \Theta) P(x_t, \ldots, x_T \mid x_1, \ldots, x_{t-1}, y_t = i, \Theta) \\
&= P(x_1, \ldots, x_{t-1}, y_t = i \mid \Theta) P(x_t, \ldots, x_T \mid y_t = i, \Theta) \\
&= \alpha_i(t) \beta_i(t)
\end{aligned}$$

Which gives us an easy way to compute the probability of being in some state at some time.

Note that we can further compute:

$$P(\boldsymbol{x} \mid \Theta) = \sum_{i=1}^{K} \alpha_i(t) \beta_i(t)$$

for any t. (This is a good debugging tip!!!)

Now, if we want to do EM, we need fractional counts for "how many times does state i transition to state j" (to get the transition probabilities) and "how many times does state i emit observation x?".

To get these, define:

$$\begin{aligned}
\gamma_i(t) &= P(y_t = i \mid \boldsymbol{x}, \Theta) \\
&= \frac{P(y_t = i, \boldsymbol{x} \mid \Theta)}{P(\boldsymbol{x} \mid \Theta)} \\
&= \frac{\alpha_i(t) \beta_i(t)}{\sum_j \alpha_j(t) \beta_j(t)}
\end{aligned}$$

This is the probability of hitting a single state.

Similarly, defin:

$$\begin{aligned}
\delta_{i,j}(t) &= P(y_t = i, y_{t+1} = j \mid \boldsymbol{x}, \Theta) \\
&= \frac{P(y_t = i, y_{t+1} = j, \boldsymbol{x} \mid \Theta)}{P(\boldsymbol{x} \mid \Theta)} \\
&= \frac{\alpha_i(t) a_{i,j} b_{i,x_t} \beta_j(t+1)}{\sum_k \alpha_k(t)\beta_k(t)}
\end{aligned}$$

This is the probability of hitting a single transition.

Then, summing, we get:

$$\begin{aligned}
\sum_t \gamma_i(t) &= \text{expected number of transitions from state } i \text{ in } \boldsymbol{x} \\
\sum_t \delta_{i,j}(t) &= \text{expected number of transitions from state } i \text{ to state } j \text{ in } \boldsymbol{x}
\end{aligned}$$

Combining all of this, we can get re-estimated probabilities:

$$\begin{aligned}
\hat{\pi}_i &= \text{expected frequency in state } i \text{ at time 1} \\
&= \gamma_i(1) \\
\hat{a}_{i,j} &= \frac{\text{expected transitions from } i \text{ to } j}{\text{expected transitions from } i} \\
&= \frac{\sum_t \delta_{i,j}(t)}{\sum_t \gamma_i(t)} \\
\hat{b}_{i,x} &= \frac{\text{expected transitions from } i \text{ with } x \text{ observed}}{\text{expected transitions from } i} \\
&= \frac{\sum_t \gamma_i(t) \mathbf{1}[x_t = x]}{\sum_t \gamma_i(t)}
\end{aligned}$$

And that's it for EM!