Hidden Markov Models

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

October 29, 2008

Reference: The material in this note is taken from Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

There are three important problems that need to be solved for HMMs to be useful:

1. Given the observation sequence $\mathbf{O} = (o_1 \dots o_T)$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(\mathbf{O}|\lambda)$, the probability of the observation sequence given the model?
2. Given the observation sequence $\mathbf{O} = (o_1 \dots o_T)$, and a model $\lambda = (A, B, \pi)$, how do we choose a corresponding sequence $\mathbf{q} = (q_1 \dots q_T)$ that is optimal in some sense (i.e., best "explains" the observations)?
3. How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(\mathbf{O}|\lambda)$?

Solution to Problem 1: First enumerate every possible state sequence of length T. There are N^T such state sequences (where N is the number of possible states). Let's consider one such state sequence: $\mathbf{q} = (q_1 \dots q_T)$. The probability of observation sequence $\mathbf{O}$ given this state sequence and model λ is

$$P(\mathbf{O}|\mathbf{q}, \lambda) = \prod_{t=1}^{T} P(o_t|q_t, \lambda) \tag{1}$$

where we have assumed conditional independence of the observations. Thus we get

$$P(\mathbf{O}|\mathbf{q}, \lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \tag{2}$$

where $b_{q_i}(o_i) = P(o_i|q_i, \lambda)$. The probability of such a state sequence $\mathbf{q}$ is

$$P(\mathbf{q}|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \tag{3}$$

where $a_{q_i q_j}$ is the probability of a state transition from q_i to q_j. Note that the joint probability of $\mathbf{O}$ and $\mathbf{q}$ is

$$P(\mathbf{O}, \mathbf{q}|\lambda) = P(\mathbf{O}|\mathbf{q}, \lambda)\ P(\mathbf{q}|\lambda) \tag{4}$$

and, thus, the marginal probability of $\mathbf{O}$ is

$$P(\mathbf{O}|\lambda) = \sum_{\text{all } \mathbf{q}} P(\mathbf{O}|\mathbf{q}, \lambda)\ P(\mathbf{q}|\lambda) \tag{5}$$

$$= \sum_{q_1 \dots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T). \tag{6}$$

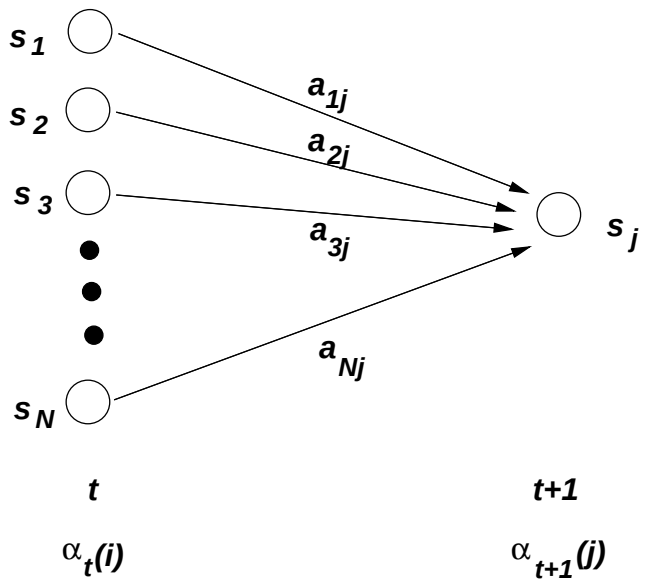

Figure 1: Schematic illustration of the forward procedure.

Importantly, we have expressed $P(\mathbf{O}|\lambda)$ as a mixture model. Unfortunately, this is a very expensive computation because it includes N^T terms in the summation. We, therefore, need a more efficient procedure.

This procedure is known as the forward procedure (see Figure 1). Consider the forward variable $\alpha_t(i)$

$$\alpha_t(i) = P(o_1 o_2 \cdots o_t, q_t = i|\lambda). \tag{7}$$

We can solve for $\alpha_t(i)$ inductively as follows:

- Initialization: $\alpha_1(i) = \pi_i b_i(o_1)$.
- Induction: $\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$.
- Termination: $P(\mathbf{O}|\lambda) = \sum_{i=1}^N \alpha_T(i)$.

That is, the forward procedure uses induction (or recursion) to efficiently solve Problem 1.

In a few moments, we'll also need the backward procedure (see Figure 2). Consider the backward variable $\beta_t(i)$

$$\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_T|q_t = i, \lambda). \tag{8}$$

We can solve for $\beta_t(i)$ inductively:

- Initialization: $\beta_T(i) = 1$.
- Induction: $\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$.

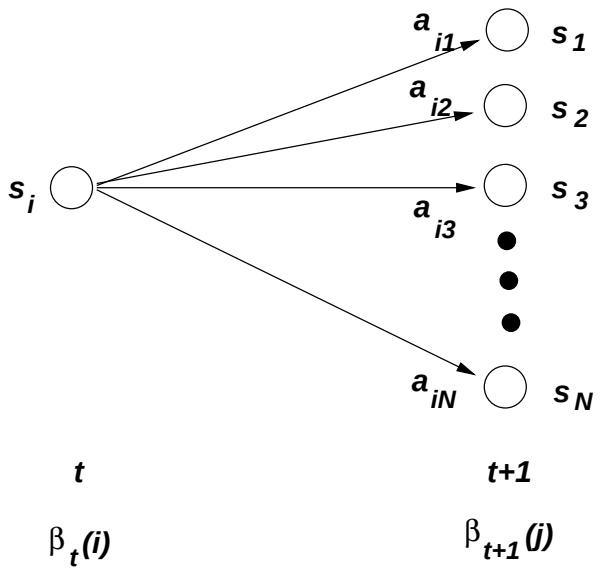

Figure 2: Schematic illustration of the backward procedure.

Solution to Problem 2: The solution to this problem depends on your definition of optimality. Suppose our goal is to choose the states q_t^* that are individually most likely at each time t. Define

$$\begin{aligned} \gamma_t(i) &= P(q_t = i|\mathbf{O}, \lambda) && (9) \\ &= \frac{P(\mathbf{O}, q_t = i|\lambda)}{P(\mathbf{O}|\lambda)} && (10) \\ &= \frac{P(\mathbf{O}, q_t = i|\lambda)}{\sum_{j=1}^N P(\mathbf{O}, q_t = j|\lambda)}. && (11) \end{aligned}$$

Since $P(\mathbf{O}, q_t = i|\lambda) = \alpha_t(i)\beta_t(i)$, we get

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)}. \quad (12)$$

Using $\gamma_t(i)$, we solve for the individually most likely state q_t^* at time t.

A possible problem with this solution is that it is locally optimal in the sense that it finds the q_t^* which is individually most likely at time t, but it is not globally optimal in the sense that it is not guaranteed to find the sequence $\mathbf{q} = (q_1 q_2 \dots q_T)$ that maximizes $P(\mathbf{q}|\mathbf{O}, \lambda)$. There is an algorithm, known as the Viterbi algorithm, which efficiently finds this globally optimal sequence. For the sake of brevity, we will omit it here.

Solution to Problem 3: The solution to this problem is known as the Baum-Welch algorithm. (It is an instance of an EM algorithm.) Define $\epsilon_t(i, j) = P(q_t = i, q_{t+1} = j|\mathbf{O}, \lambda)$. We can re-write it as follows:

$$\epsilon_t(i, j) = \frac{P(q_t = i, q_{t+1} = j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} \quad (13)$$

$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(\mathbf{O}|\lambda)} \tag{14}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_t(k)a_{kl}b_l(o_{t+1})\beta_{t+1}(l)}. \tag{15}$$

There are several points worth noting:

- $\gamma_t(i) = \sum_{j=1}^{N}\epsilon_t(i,j)$
- $\sum_{t=1}^{T-1}\gamma_t(i)$ = expected number of transitions from state i in observation $\mathbf{O}$
- $\sum_{t=1}^{T-1}\epsilon_t(i,j)$ = expected number of transitions from state i to state j in observation $\mathbf{O}$

Using these quantities, we can write the parameter re-estimation equations. The initial state probabilities are

$$\pi_i = \gamma_1(i). \tag{16}$$

The state transition probabilities are

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\epsilon_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}. \tag{17}$$

Here, the numerator is the expected number of transitions from state i to state j, and the denominator is the expected number of transitions from state i. The emission probabilities are

$$b_j(k) = \frac{\sum_{t=1,\ o_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{18}$$

Here the numerator is the expected number of times in state j and observing symbol v_k, and the denominator is the expected number of times in state j. Note that it is necessary to iterate through these equations several times until convergence.