

Hidden Markov Models

Here we assume a finite set of hidden (latent) states. We will represent a hidden state by an integer $i, j \in \{1, \dots, S\}$. We also assume a set of observations. We will represent an observation by an integer $k \in \{1, \dots, O\}$ where O is the number of observations. A run of a hidden Markov model consists of a sequence z_1, \dots, z_N of hidden states and a sequence x_1, \dots, x_N of observations where the first hidden state z_1 is generated from a fixed distribution over state and each successor hidden state z_{t+1} is generated stochastically from the preceding hidden state z_t . Each observation x_t is generated stochastically from the hidden state z_t at that time. More formally an HMM has parameters Θ which defines the probability $P_{\Theta}(z_1, \dots, z_N, x_1, \dots, x_N)$ as follows.

$$\Theta = \langle \pi, A, C \rangle$$

$$\pi_i = P(z_1 = i), \quad \sum_{i=1}^S \pi_i = 1$$

$$A_{i,j} = P(z_{t+1} = i | z_t = j), \quad \sum_{i=1}^S A_{i,j} = 1$$

$$C_{k,j} = P(x_t = k | z_t = j), \quad \sum_{k=1}^O C_{k,j} = 1$$

$$P_{\Theta}(z_1, \dots, z_N, x_1, \dots, x_N) = \pi_{z_1} \left(\prod_{t=1}^{N-1} A_{z_t, z_{t+1}} \right) \left(\prod_{t=1}^N C_{x_t, z_t} \right)$$

Applications of HMMs:

- Speech Recognition. The hidden states are word positions and the observable tokens are acoustic feature vectors.
- Part of speech tagging. The hidden states are the parts of speech (noun, verb, adjective, and so on).
- DNA sequence analysis. The hidden states might be protein secondary structure or a position in a homologous sequence.

1 The Viterbi Algorithm

In typical applications of an HMM we are given x_1, \dots, x_N and must infer z_1, \dots, z_N . The Viterbi algorithm is used to compute the following.

$$\begin{aligned} z_1^*, \dots, z_N^* &= \operatorname{argmax}_{z_1, \dots, z_N} P_{\Theta}(z_1, \dots, z_N \mid x_1, \dots, x_N) \\ &= \operatorname{argmax}_{z_1, \dots, z_N} P_{\Theta}(z_1, \dots, z_N, x_1, \dots, x_N) \end{aligned} \quad (1)$$

In the Viterbi algorithm we define the following $S \times N$ matrix V .

$$V_{i,t} = \max_{z_1, \dots, z_{t-1}} P(z_1, \dots, z_{t-1}, x_1, \dots, x_{t-1}, z_t = i)$$

Given this definition of the matrix V it is possible to prove the following identities.

$$V_{i,1} = \pi_i \quad (2)$$

$$V_{i,t+1} = \max_j V_{j,t} C_{x_t, j} A_{i,j} \quad (3)$$

Equation (2) follows directly from the definition of V . Equation (3) can be derived as follows.

$$\begin{aligned} V_{i,t+1} &= \max_{z_1, \dots, z_t} P_{\Theta}(z_1, \dots, z_t, x_1, \dots, x_t, z_{t+1} = i) \\ &= \max_{z_1, \dots, z_t} \pi_{z_1} \left(\prod_{s=1}^{t-1} A_{z_s, z_{s+1}} \right) \left(\prod_{s=1}^t C_{x_s, z_s} \right) A_{i, z_t} \\ &= \max_{z_1, \dots, z_{t-1}} \max_j \pi_{z_1} \left(\prod_{s=1}^{t-2} A_{z_s, z_{s+1}} \right) \left(\prod_{s=1}^{t-1} C_{x_s, z_s} \right) A_{j, z_{t-1}} C_{x_t, j} A_{i,j} \\ &= \max_j \max_{z_1, \dots, z_{t-1}} \left(\pi_{z_1} \left(\prod_{s=1}^{t-2} A_{z_s, z_{s+1}} \right) \left(\prod_{s=1}^{t-1} C_{x_s, z_s} \right) A_{j, z_{t-1}} \right) C_{x_t, j} A_{i,j} \\ &= \max_j \left(\max_{z_1, \dots, z_{t-1}} \pi_{z_1} \left(\prod_{s=1}^{t-2} A_{z_s, z_{s+1}} \right) \left(\prod_{s=1}^{t-1} C_{x_s, z_s} \right) A_{j, z_{t-1}} \right) C_{x_t, j} A_{i,j} \\ &= \max_j V_{j,t} C_{x_t, j} A_{i,j} \end{aligned} \quad (4)$$

We note that (2) and (3) provide a way of computing the matrix V starting by using (2) to compute $V_{\cdot,1}$ and then using (3) to compute $V_{\cdot,t+1}$ from $V_{\cdot,t}$. We can then compute z_t^* backward from $t = N$ as follows.

$$z_N^* = \operatorname{argmax}_i V_{i,N} C_{x_N,i}$$

$$z_{t-1}^* = \operatorname{argmax}_i V_{i,t-1} C_{x_{t-1},i} A_{z_t^*,i}$$

Rather than compute best predecessors backward in this way, we can record the best predecessor of for each pair $\langle i, t \rangle$ during the forward computation of the matrix $V_{i,t}$.

2 The Forward-Backward Procedure

We now consider the following problem.

$$z_t^* = \operatorname{argmax}_i P_{\Theta}(z_t = i \mid x_1, \dots, x_N) \quad (5)$$

It is important to note that z_t^* as defined by (5) is different from z_t^* as defined by (1). To see this consider a case where $N = 3$ and $S = 100$ and the state transition matrix has the property that 99 states deterministically transition into state 1 and state 1 uniformly transitions into one of the 99 other states. Suppose that the initial distribution π is such that $P(z_1 = 1) = 1/3$ with the remaining probability uniformly distributed among the other states. Suppose that all observations are equally likely for all states so that the observations have no influence on the state probabilities. In this case the most likely sequence has $z_1 = 1$ and $z_2 \neq 1$. So for z_t^* as defined by (1) we have $z_2^* \neq 1$. But $P(x_2 = 1) = 2/3$ so for z_t^* as defined by (5) we have $z_2^* = 1$.

Consider the case where we generate z_1, \dots, z_N and x_1, \dots, x_N from the distribution P_{Θ} and then have to guess z_1, \dots, z_n given only x_1, \dots, x_N . There are two different ways that the guess might be scored. In the first way, which we will call 01-loss, the guess is scored correct only if the entire state sequence is correct. If we are to be scored in this way then the optimal guess is given by (1). The second method of scoring counts the number of times t where our guess for z_t is correct. This score is one minus the hamming distance from our guess to the actual hidden state sequence where the Hamming distance is just the number of times where the guess is different from the actual. This will be called Hamming loss. If we are to be scored by hamming loss then the optimal guess is given by (5). Note that the sequence defined by (5) can be extremely unlikely and can even contain impossible state transition — a time t where $A_{z_t^*, z_{t+1}^*} = 0$.

We can solve (5) using the forward-backward procedure. As in the case of the Viterbi algorithm, we first define matrices and then derive an efficient way of computing them. We define the matrices F (for forward) and B (for backward) as follows.

$$\begin{aligned} F_{i,t} &= P_{\Theta}(x_1, \dots, x_{t-1}, z_t = s) \\ B_{i,t} &= P_{\Theta}(x_t, \dots, x_N \mid z_t = s) \end{aligned}$$

From these definitions we can prove the following equations.

$$F_{i,1} = \pi_i \quad (6)$$

$$F_{i,t+1} = \sum_{j=1}^S F_{j,t} C_{x_t,j} A_{i,j} \quad (7)$$

$$B_{i,N} = C_{x_n,i} \quad (8)$$

$$B_{i,t-1} = \sum_{j=1}^S C_{x_{t-1,i}} A_{j,i} B_{j,t} \quad (9)$$

Equation (6) can be used to compute $F_{\cdot,1}$ and equation (7) can be used to compute $F_{cdot,t+1}$ from $F_{\cdot,t}$. Equation (6) follows directly from the definition of F . Equation (7) can be derived as follows.

$$\begin{aligned} F_{i,t+1} &= P(x_1, \dots, x_t, z_{t+1} = i) \\ &= \sum_{j=1}^S P(x_1, \dots, x_t, z_t = j, z_{t+1} = i) \\ &= \sum_{j=1}^S P(x_1, \dots, x_{t-1}, z_t = j) C_{x_t,j} A_{i,j} \\ &= \sum_{j=1}^S F_{j,t} C_{x_t,j} A_{i,j} \end{aligned}$$

Equation (8) can be used to compute $B_{\cdot,N}$ and equation (9) can be used to compute $B_{\cdot,t-1}$ from $B_{\cdot,t}$. Equation (8) follows directly from the definition of B . Equation (9) can be derived as follows.

$$\begin{aligned}
B_{i,t-1} &= P(x_{t-1}, x_t, \dots, x_N \mid z_{t-1} = i) \\
&= \sum_{j=1}^S P(z_t = j, x_{t-1}, x_t, \dots, x_N \mid z_{t-1} = i) \\
&= \sum_{j=1}^S P(z_t = j, x_{t-1} \mid z_{t-1} = i) P(x_t, \dots, x_N \mid z_{t-1} = i, z_t = j, x_{t-1}) \\
&= \sum_{j=1}^S C_{x_{t-1}, i} A_{j,i} P(x_t, \dots, x_N \mid z_t = j,) \\
&= \sum_{j=1}^S C_{x_{t-1}, i} A_{j,i} B_{j,t} \\
&= C_{x_{t-1}, i} \sum_{j=1}^S A_{j,i} B_{j,t}
\end{aligned}$$

We can now compute z_t^* as defined by (5) using the following.

$$\begin{aligned}
P(z_t = i \mid x_1, \dots, x_N) &= \frac{F_{i,t} B_{i,t}}{P(x_1, \dots, x_N)} \\
P(x_1, \dots, x_N) &= \sum_{j=1}^S \pi_j B_{j,1}
\end{aligned} \tag{10}$$

3 Problems

1. Suppose that we have two hidden states ($S = 2$), two observations ($O = 2$), and Θ is given as follows.

$$\begin{aligned}
\pi_1 &= \pi_2 = \frac{1}{2} \\
A_{1,1} &= A_{2,2} = 1 - \epsilon \\
A_{2,1} &= A_{1,2} = \epsilon \\
C_{1,1} &= C_{2,2} = 1 - \delta \\
C_{2,1} &= C_{1,2} = \delta
\end{aligned}$$

Now suppose that we observe x_1, \dots, x_N with $x_t = 1$ for all $1 \leq t \leq N$.

- a. Give values for $F_{1,1}$ and $F_{2,1}$.
- b. Give expressions for $F_{1,t+1}$ and $F_{2,t+1}$ in terms of $F_{1,t}$, $F_{2,t}$, ϵ and δ .

- c. Give expressions for $B_{1,N}$ and $B_{2,N}$ in terms of δ .
- d. Give expressions for $B_{1,t-1}$ and $B_{2,t-1}$ in terms of $B_{1,t}$, $B_{2,t}$, ϵ and δ .
- Extra Credit:** Give closed form solutions for $F_{1,t}$, $F_{2,t}$, $B_{1,t}$ and $B_{2,t}$ as functions of t , ϵ and δ . Use your answers to give a closed form solution for $P(z_t = 1 \mid x_1, \dots, x_N)$ as a function of t .

2. Give an expression for $P_{\Theta}(z_t = i, z_{t+1} = j \mid x_1, \dots, x_N)$ in terms of the matrices A and C , the forward value $F_{i,t}$, the backward value $B_{j,t+1}$ and the observed data probability $P_{\Theta}(x_1, \dots, x_N)$. Hint: the answer is similar to (10).