Hidden Markov Models

Discrete Markov Processes

Consider a system at any time is in one of a set of N distinct states: S1, S2 ...Sn. The state at time t is denoted by q(t), t=1,2,..., so on.

At regularly spaced discrete time, the system moves to a state with a given probability, depending on the values of all the previous states. For the special case of the first order Markov model, we say that the state at time t+1 depends only on the state at time t, regardless of the states in the previous times.

We also assume in the first order model the transition probabilities are independent of time as

$$a_{ij} = P\left(q_{t+1} = S_j | q_t = S_i \right)$$
$$satisf\ ying$$
$$a_{ij} \geq 0\ and\ \sum_{j=1}^{N} a_{ij} = 1$$

So, going from S1 to S2 has the same probability no matter where it happens in the observation sequence.

This can be seen as a stochastic automaton, where a system moves from state i to state j with probability p, and this probability is the same for any t, except for the first state.

π denotes the initial probabilities, which is the probability that a state is observed first in a sequence. A denotes all the transition probabilities from states i to j.

In an observable Markov model, all the states are observable, and we obtain an observation sequence O that is the state sequence Q, observed under the conditions of A and π.

$$P(O = Q|A, \pi) = P\left(q_1 \right) \prod_{t=2}^{T} P\left(q_t | q_{t-1} \right) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots\ldots a_{qT-1qT}$$

Hidden Markov Models

In a HMM, the states are not observable, but when we visit a state, an observation is recorded that is a probabilistic function of the state.

The emission probability is given by

$$b_j(m) = P\left(O_t = v_m | q_t = S_j \right)$$

Where we observe v_m, m=1...M in state S_j. The state sequence Q is not observed, and that is what makes the model hidden, and should be inferred from the observation sequence O.

The HMM has the following elements

1. N : Number of states in the model

   $S = \{S_1, S_2, \ldots, S_N\}$

2. M : Number of distinct observation symbols in the alphabet

   $V = \{v_1, v_2, \ldots, v_M\}$

3. State transition probabilities

   $A = [a_{ij}]$ where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$

4. Observation probabilities

   $B = [b_j(m)]$ where $b_j(m) = P(O_t = v_m | q_t = S_j)$

5. Initial state probabilities

   $\Pi = [\pi_i]$ where $\pi_i = P(q_1 = S_i)$

From this, we generalise a HMM model to be $\lambda(A, B, \Pi)$.

Three Basic Problems of HMMs

Given a number of sequences of observations

1. Given a model λ, evaluate the probability of any given observation sequence O = $\{O_1,O_2,...O_T\}$, namely $P(O|\lambda)$.
2. Given a model λ and an observation sequence O, find the state sequence Q = $\{q_1,q_2,...q_T\}$, which has the highest probability of generating O, namely, we find Q^* that maximises $P(Q|O,\lambda)$.
3. Given a set of observation sequences $X = \{O^k\}_k$, learn the model that maximises the probability of generating X, namely, to find λ^* that maximises $P(X|\lambda)$.

Evaluation Problem

Given an observation sequence O and a state sequence Q, the probability of observing O given the state sequence Q is simply

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda) = b_{q_1}(O_1).b_{q_2}(O_2)....b_{q_T}(O_T)$$

Which we cannot calculate because we do not know the state sequence. The probability of the state sequence Q is

$$P(Q|\lambda) = P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1}) = \pi_{q1}a_{q1q2}....a_{qT-1qT}$$

Then, the joint probability gives

$$\begin{aligned} P(O,Q|\lambda) &= P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1})\prod_{t=1}^{T} P(O_t|q_t) \\ &= \pi_{q1}b_{q1}(O_1)a_{q1q2}b_{q2}(O_2).....a_{qT-1qT}b_{qT}(O_T) \end{aligned}$$

$P(O|\lambda)$ *can be calculated by marginalising over the joint, by summing over all possible* Q

$$P(O|\lambda) = \sum_{all\ possible\ Q} P(O,Q|\lambda)$$

This is however, not feasible, as it has a large search space of N^T. So we use divide-and-conquer, called the forward-backward procedure. In this, we divide the observation space into two parts: one starting from time 1 until time t, and the second one from time t+1 until T.

1. Forward variable

The forward variable is defined as the probability of observing the partial sequence O until time t and being in state S_i at time t, given by

$$\alpha_t(i) = P\left(O_1 \ldots O_t,\ q_t = S_i | \lambda\right)$$

This value can be calculated recursively as

Initialisation

$$\begin{aligned}\alpha_1(i) &= P\left(O_1, q_1 = S_i | \lambda\right) \\ &= P\left(O_1 | q_1 = S_i, \lambda\right) P\left(q_1 = S_i | \lambda\right) \\ &= \pi_i b_i\left(O_1\right)\end{aligned}$$

Recursion

$$\begin{aligned}&\alpha_{t+1}(j) \\ &= P\left(O_1 \ldots O_{t+1}, q_{t+1} = S_j | \lambda\right) \\ &= P\left(O_1 \ldots O_{t+1} | q_{t+1} = S_j, \lambda\right) P\left(q_{t+1} = S_j | \lambda\right) \\ &= P\left(O_1 \ldots O_t | q_{t+1} = S_j, \lambda\right) P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) P\left(q_{t+1} = S_j | \lambda\right) \\ &= P\left(O_1 \ldots O_t | q_{t+1} = S_j | \lambda\right) P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) \\ &= P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) \sum_i P\left(O_1 \ldots O_t, q_t = S_i,\ q_{t+1} = S_j, \lambda\right) \\ &= P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) \sum_i P\left(O_1 .. O_t, q_{t+1} = S_j | q_t = S_i, \lambda\right) P\left(q_t = S_i | \lambda\right) \\ &= P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) \sum_i P\left(O_1 .. O_t | q_t = S_i, \lambda\right) P\left(q_{t+1} = S_j | q_t = S_i, \lambda\right) P\left(q_t = S_i | \lambda\right) \\ &= P\left(O_{t+1} | q_{t+1} = S_j, \lambda\right) \sum_i P\left(O_1 .. O_t, q_t = S_i | \lambda\right) P\left(q_{t+1} = S_j | q_t = S_i, \lambda\right) \\ &= \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j\left(O_{t+1}\right)\end{aligned}$$

This basically explains how the alpha term explains the first t observations and ends in state S_i. We multiply this with the transition probability a_{ij} to move to state S_j, and sum up over all such combinations. In the end we multiply it with the emission probability.

The probability of the observation sequence is given by

$$P(O|\lambda) = \sum_{i=1}^{N} P\left(O, q_T = S_i|\lambda\right)$$
$$= \sum_{i=1}^{N} \alpha_T(i)$$

Where alpha denotes the probability of generating the full observation sequence and ending up in state S_i. Computing this happens in $O(N^2T)$.

2. Backward variable

We can similarly define a backward variable as the probability of being in S_i at time t by observing the partial sequence $O_{t+1},...O_T$.

$$\beta_t(i) = P\left(O_{t+1}...O_T|q_t = S_i, \lambda\right)$$

Initialisation

$$\beta_T(i) = 1$$

Recursion

$$\beta_t(i)$$
$$= P\left(O_{t+1}...O_T|q_t = S_i, \lambda\right)$$
$$= \sum_j P\left(O_{t+1}...O_T, q_{t+1} = S_j|q_t = S_i, \lambda\right)$$
$$= \sum_j P\left(O_{t+1}...O_T|q_{t+1} = S_j, q_t = S_i, \lambda\right) P\left(q_{t+1} = S_j|q_t = S_i, \lambda\right)$$

$$= \sum_j P\left(O_{t+1}|q_{t+1} = S_j, q_t = S_i, \lambda\right) P\left(O_{t+2}...O_T|q_{t+1} = S_j, q_t = S_i, \lambda\right) P\left(q_{t+1} = S_j|q_t = S_i, \lambda\right)$$
$$= \sum_j P\left(O_{t+1}|q_{t+1} = S_j, \lambda\right) P\left(O_{t+2}...O_T|q_{t+1} = S_j, \lambda\right) P\left(q_{t+1} = S_j|q_t = S_i, \lambda\right)$$
$$= \sum_{j=1}^{N} a_{ij} b_j\left(O_{t+1}\right) \beta_{t+1}(j)$$

When in state S_i, we can go to N possible next states S_j, each with a probability a_{ij}. While there, the next observation beta explains all the observations after time t +1.

Both alpha and beta values are calculated by very small probabilities, risking underflow. To avoid this, we normalise at each step

$$c_t = \frac{1}{\sum_j \alpha_t(j)}$$

Which gives

$$P(O|\lambda) = \frac{1}{\Pi_t c_t} \; or \; \log(P(O|\lambda)) = -\sum_t \log c_t$$

The forward-backward algorithm can be thus summarised as

Forward Step

Initialisation

$$\alpha_1(i) = \pi_i b_i(O_1)$$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1})$$

Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Backward Step

Initialisation

$$\beta_T(i) = 1$$

Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)$$

Finding the State Sequence

This problem is to find the state sequence Q = $\{q_1 q_2 \ldots q_T\}$ having the highest probability of generating the observation sequence O given the model.

The probability of being in state S_i at time t, given O and the parameters is,

$$\varphi_t(i)$$
$$= P(q_t = S_i | O, \lambda)$$
$$= \frac{P(O | q_t = S_i, \lambda) P(q_t = S_i | \lambda)}{P(O|\lambda)}$$
$$= \frac{P(O_1, O_2, \ldots O_t | q_t = S_i, \lambda) P(O_t, \ldots O_T | q_t = S_i, \lambda) P(q_t = S_i | \lambda)}{P(O,\ q_t = S_j | \lambda)}$$

$$= \frac{P(O_1 \ldots O_t,\ q_t = S_i | \lambda) P(O_{t+1} \ldots O_T | q_t = S_i, \lambda)}{\sum_{j=1}^{N} P(O | q_t = S_j, \lambda) P(q_t = S_j | \lambda)}$$
$$= \frac{\alpha_t(i)\, \beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)}$$

The numerator explains the whole sequence given at that time t, the system is in state S.

To find the state sequence, at each time step t, we choose the state with the highest probability

$$q_t^* = \underset{i}{argmax}\ \varphi_t(i)$$

But this may choose S_i and S_j as the most probable states at time t and t+1 even when a_{ij} is 0. To find the best path, we use the Viterbi Algorithm, based on DP, that takes these transition probabilities into account.

Given a state sequence Q and an observation sequence O, we define the highest probability path at time t that accounts for the first t observations and ends in S_i as

$$\delta_t(i) = \max_{q1q2 \ldots qt-1} p(q_1 q_2 \ldots q_{t-1},\ q_t = S_i,\ O_1 \ldots . O_t | \lambda)$$

The algorithm is as follows

Initialisation

$$\delta_1(i) = \pi_i b_i(O_1)$$

$$\psi_1(i) = 0$$

Recursion

$$\delta_t(j) = \max_i \delta_{t-1}(i)\, a_{ij} b_j(O_t)$$

$$\psi_t(j) = \operatorname{argmax}_i \delta_{t-1}(i)\, a_{ij}$$

Termination

$$p^* = \max_i \delta_T(i)$$

$$q_T^* = \operatorname{argmax}_i \delta_T(i)$$

Backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, \ldots 1$$

Learning Model Parameters

In this, we try to calculate the parameters for the model that maximises the likelihood of the sample of the training sequences, $X = \{O^k\}$, namely $P(X|\lambda)$.

We define a probability of being in S_i at time t and S_j at time t+1, given the whole observation O and λ as

$$\xi_t(i,j)$$

$$= P(q_t = S_i,\ q_{t+1} = S_j | O, \lambda)$$

$$= \frac{P(O|q_t = S_i, q_{t+1} = S_j, \lambda) P(q_t = S_i, q_{t+1} = S_j|\lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)\, b_j(O_{t+1})\, \beta_{t+1}(j)\, a_{ij}}{\sum_k \sum_l P(q_t = S_k, q_{t+1} = S_l, O|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)}{\sum_k \sum_l \alpha_t(k)\, a_{kl} b_l(O_{t+1})\, \beta_{t+1}(l)}$$

The probability of being in state S_i at time t can be found by marginalising over all arc probabilities for all possible next states

$$\varphi_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

Baum Welch

Probability of transition

$$a_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T-1} \varphi_t(i)}$$

The probability of observing v_m in S_j is the expected number of times v_m is observed when system is in S_j over the total number of times the system is in S_j.

$$b_j(m) = \frac{\sum_{t=1}^{T} \varphi_t(j)\, 1\left(O_t = v_m\right)}{\sum_{t=1}^{T} \varphi_t(j)}$$

If there are k observation sequences independent of each other

$$a_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \xi_t^k(i,j)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \varphi^k{}_t(j)}$$

$$b_j(m) = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \varphi_t^k(i,j)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \varphi^k{}_t(j)}$$

$$\pi_i = \frac{\sum_{k=1}^{K} \varphi_1^k(i)}{K}$$