Hidden Markov Model

QIP Sponsored Short Term Course on Statistical Methods organized by Computer Engineering Department, SVNIT Surat during 19-23 December 2016

Mukesh A. Zaveri
Computer Engineering Department
Sardar Vallabhbhai National Institute of Technology, Surat
mazaveri@coed.svnit.ac.in

Hidden Markov Models

- the instances that constitute a sample are iid
- the advantage that the likelihood of the sample is the product of likelihoods of the individual instances
- this assumption not valid where successive instances are dependent
- applications: letters in a word, speech recognition - only certain sequences of phonemes are allowed
- a sequence can be characterized as being generated by a parametric random process

Hidden Markov Models

- Discrete Markov Processes
- a system at any time is in one of a set of N distinct states $S_1, \ldots, S_N$
- the state at time t is $q_t = S_i$ means that at time t the system is in state S_i
- the system moves to a state with a given probability, depending on the values of the previous states

$$P(q_{t+1} = S_j | q_t = S_i, q_{t-1} = S_k, \ldots)$$

- First order Markov model,

$$P(q_{t+1} = S_j | q_t = S_i, q_{t-1} = S_k, \ldots) = P(q_{t+1} = S_j | q_t = S_i) \equiv a_{ij}$$

- simplifying - the transition probabilities are independent of time

Hidden Markov Models

- $a_{ij} \geq 0$ and $\sum_{j=1}^{N} a_{ij} = 1$ $\mathbf{A} = [a_{ij}]$ is a $N \times N$ matrix whose rows sum to 1
- it can be seen as stochastic automaton
- initial probabilities π_i is the probability that the first state in the sequence is S_i $\pi_i \equiv P(q_1 = S_i)$ satisfying $\sum_{i=1}^{N} \pi_i = 1$
- $\Pi = [\pi_i]$ is a vector of N elements of that sum to 1
- Observable Markov model - the states are observable
- the system moves from one state to another, results into an observation sequence and that is a sequence of states
- the output of the process is the set of states at each instant of time where each state corresponds to a physical observable event

Hidden Markov Models

- an observance sequence O that is the state sequence $O = Q = \{q_1 q_2 \cdots q_T\}$

$$P(O = Q|\mathbf{A}, \Pi) = P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1}) = \pi_{q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}$$

- π_{q_1} is the probability that the first state q_1, $a_{q_1 q_2}$ is the probability of going from q_1 to q_2
- N urns, each urn contains balls of only one color; there is an urn of red balls, another of blue balls and so forth
- draws balls from urns one by one and their color are shown
- three states: S_1: red, S_2: blue, S_3: green and q_t denote the color of the ball drawn at time t
- initial probabilities $\Pi = [0.5, 0.2, 0.3]^T$

Hidden Markov Models

- a_{ij} is the probability of drawing from urn j (a ball of color j) after drawing a ball of color i from urn i
- the transition matrix

$$\mathbf{A} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

- given Π and $\mathbf{A}$ it is easy to generate K random sequences each of length T
- calculate the probability of a sequence "red, red, green, green", i.e. the observation sequence $O = \{S_1, S_1, S_3, S_3\}$

$$\begin{aligned} P(O|\mathbf{A}, \Pi) &= P(S_1) \cdot P(S_1|S_1) \cdot P(S_3|S_1) \cdot P(S_3|S_3) \\ &= \pi_1 \cdot a_{11} \cdot a_{13} \cdot a_{33} \\ &= 0.5 \cdot 0.4 \cdot 0.3 \cdot 0.8 = 0.048 \end{aligned}$$

Hidden Markov Models

- how to learn the parameters $\Pi, \mathbf{A}$ given K sequences of length T where q_t^k is the state at time t of sequence k
- the initial probability estimate is the number of sequences starting with S_i divided by the number of sequences

$$\hat{\pi}_i = \frac{\{\text{number of sequences starting with} S_i\}}{\{\text{number of sequences}\}} = \frac{\sum_k 1(q_1^k = S_i)}{K}$$

$1(b)$ is 1 if b is true and 0 otherwise

- the transition probabilities, the estimate for a_{ij} is the number of transitions from S_i to S_j divided by the total number of transitions from S_i over all sequences

$$\begin{aligned}\hat{a}_{ij} &= \frac{\{\text{number of transition from} S_i \text{to } S_j\}}{\{\text{number of transitions from} S_i\}} \\ &= \frac{\sum_k \sum_{t=1}^{T} 1(q_t^k = S_i \text{and} q_{t+1}^k = S_j)}{\sum_k \sum_{t=1}^{T} 1(q_t^k = S_i)}\end{aligned}$$

Hidden Markov Models

- a_{12} is the number of times a blue ball follows a red ball divided by the total number of red ball draws over all sequences
- in HMM, the states are not observable, but when a state is visited, an observation is recorded that is a probabilistic function of the state
- a discrete observation in each state from the set $\{v_1, v_2, \ldots, v_M\}$

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

- $b_j(m)$ is the observation, or emission probability that we observe v_m $(m = 1, \ldots, M)$ in state S_j

Hidden Markov Models

- the state sequence Q is not observed, (so called hidden), but it should be inferred from the observation sequence O
- many different state sequences Q that could have generated the same observation sequence but with different probabilities, just as,
- given an iid sample from a normal distribution, there are an infinite number of (μ, σ) value pair possible and interested in the one having the highest likelihood of generating the sample
- there are two sources of randomness: randomly moving from one state to another and the observation in a state
- the hidden case corresponds to the urn-and-ball example where each urn contains balls of different colors
- let $b_j(m)$ denote the probability of drawing a ball of color m from urn j

Hidden Markov Models

- again a sequence of ball colors observed but without knowing the sequence of urns from which the balls were drawn
- as if the urns are placed behind a curtain and somebody picks a ball at random from one of the urns and shows us only the ball, without showing us the urn from which it is picked
- the ball is returned to the urn to keep the probabilities the same
- the number of ball colors may be different from the number of urns
- example: three urns and the observation sequence is $O = \{\text{red, red, green, blue, yellow}\}$
- earlier, knowing the observation (ball color), the state (urn) is known because there are separate urns for separate colors and each urn contained balls of only one color

Hidden Markov Models

- the observable model is a special case of the hidden model where $M = N$ and $b_j(m)$ is 1 if $j = m$ and 0 otherwise
- in case of a hidden model, a ball could have been picked from any urn
- for the same observation sequence O, there may be many possible state sequences Q that could have generated O
- HMM: N states in model $S = \{S_1, \ldots, S_N\}$
- M distinct observation symbols in the alphabet $V = \{v_1, \ldots, v_M\}$
- transition probabilities $\mathbf{A}$
- observation probabilities $\mathbf{B} = [b_j(m)]$
- initial state probabilities $\Pi = [\pi_i]$

Hidden Markov Models

- given λ, the model can be used to generate an arbitrary number of observation sequences of arbitrary length
- interested in estimating the parameters of the model given a training set of sequences
- three basic problems of HMM: given a number of sequences of observations
  1. given a model λ evaluate the probability of any given observation sequence $O = \{O_1 O_2 \cdots O_T\}$, namely, $P(O|\lambda)$
  2. given a λ and O find out the $Q = \{q_1 \cdots q_T\}$ which has the highest probability of generating O, find Q^* that maximizes $P(Q|O, \lambda)$
  3. given a training set of observation sequences, $\mathcal{X} = \{O^k\}_k$, learn the model that maximizes the probability of generating $\mathcal{X}$, namely, find λ^* that maximize $P(\mathcal{X}|\lambda)$

Hidden Markov Models

- evaluation problem

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

- it can not be calculated because the state sequence is not known
- the probability of the state sequence Q

$$P(Q|\lambda) = P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1}) = \pi_{q_1} a_{q_1 q_2}\cdots a_{q_{T-1} q_T}$$

the joint probability is

$$\begin{aligned} P(O,Q|\lambda) &= P(q_1)\prod_{t=2}^{T} P(q_t|q_{t-1})\prod_{t=1}^{T} P(O_t|q_t) \\ &= \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2)\cdots a_{q_{T-1} q_T} b_{q_T}(O_T) \end{aligned}$$

Hidden Markov Models

- $P(O|\lambda)$ can be computed by marginalizing over the joint, namely, by summing up over all possible Q

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

- this is not practical since there are N^T possible Q, assuming that all the probabilities are nonzero
- efficient procedure to calculate $P(O|\lambda)$ - called forward backward procedure
- it is based on the idea of dividing the observation sequence into two parts: the first one starting from time 1 until time t, and the second one from time $t+1$ until T
- define the forward variable $\alpha_t(i)$ as the probability of observing the partial sequence $\{O_1 \cdots O_t\}$ until time t and being in S_i at time t, given the model λ

$$\alpha_t(i) \equiv P(O_1 \cdots O_t, q_t = S_i|\lambda)$$

Hidden Markov Models

- it can be calculated recursively
- initialization

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

- Recursion

$$\begin{aligned}\alpha_{t+1}(j) &\equiv P(O_1 \cdots O_{t+1}, q_{t+1} = S_j|\lambda)\\ &= P(O_1 \cdots O_{t+1}|q_{t+1} = S_j, \lambda)P(q_{t+1} = S_j|\lambda)\end{aligned}$$

$$\begin{aligned}&= P(O_1 \cdots O_t|q_{t+1} = S_j, \lambda)P(O_{t+1}|q_{t+1} = S_j, \lambda)P(q_{t+1} = S_j|\lambda)\\ &= P(O_1 \cdots O_t, q_{t+1} = S_j|\lambda)P(O_{t+1}|q_{t+1} = S_j, \lambda)\\ &= P(O_{t+1}|q_{t+1} = S_j, \lambda)\sum_i P(O_1 \cdots O_t, q_t = S_i, q_{t+1} = S_j|\lambda)\end{aligned}$$

Hidden Markov Models

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

$\alpha_t(i)$ explains the first t observations and ends in state S_i Multiply $\alpha_t(i)$ by the probability a_{ij} to move to state S_j

Hidden Markov Models

- there are N possible previous states, need to sum up over all such possible previous S_i, $b_j(O_{t+1})$ then is the probability for the $(t+1)$st observation while in state S_j at time $t+1$
- it is easy to calculate the probability of the observation sequence

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

- $\alpha_T(i)$ is the probability of generating the full observation sequence and ending up in state S_i and need to sum up over all such possible final states
- computing $\alpha_t(i)$ is $\mathcal{O}(N^2T)$ and solves the evaluation problem
- similarly backward variable $\beta_t(i)$ is the probability of being in S_i at time t and observing the partial sequence $O_{t+1} \cdots O_T$

Hidden Markov Models

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

initialization (arbitrarily to 1) $\beta_T(i) = 1$
Recursion

$$\begin{aligned}\beta_t(i) &\equiv P(O_{t+1} \cdots O_T | q_t = S_i, \lambda) \\ &= \sum_j P(O_{t+1} \cdots O_T, q_{t+1} = S_j | q_t = S_i, \lambda)\end{aligned}$$

$$\begin{aligned}&= \sum_j P(O_{t+1} \cdots O_T | q_{t+1} = S_j, q_t = S_i, \lambda) P(q_{t+1} = S_j | q_t = S_i, \lambda) \\ &= \sum_j P(O_{t+1} | q_{t+1} = S_j, q_t = S_i, \lambda) \\ &\quad P(O_{t+2} \cdots O_T | q_{t+1} = S_j, q_t = S_i, \lambda) P(q_{t+1} = S_j | q_t = S_i, \lambda)\end{aligned}$$

Hidden Markov Models

$$
\begin{aligned}
&= \sum_j P(O_{t+1} | q_{t+1} = S_j, \lambda) \\
&\quad P(O_{t+2} \cdots O_T | q_{t+1} = S_j, \lambda) P(q_{t+1} = S_j | q_t = S_i, \lambda) \\
&= \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)
\end{aligned}
$$

- when in state S_i, we can go to N possible next states S_j, each with probability a_{ij}

Hidden Markov Models

- finding the state sequence $Q = \{q_1 q_2 \cdots q_T\}$ having the highest probability of generating the observation sequence $O = \{O_1 O_2 \cdots O_T\}$ given the model λ
- $\gamma_t(i)$ the probability of being in state S_i at time t given O and λ

$$\gamma_t(i) \equiv P(q_t = S_i | O, \lambda) = \frac{P(O|q_t = S_i, \lambda) P(q_t = S_i|\lambda)}{P(O|\lambda)}$$

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

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

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

Hidden Markov Models

- $\alpha_t(i)$ and $\beta_t(i)$ split the sequence, $\alpha_t(i)$ explains the starting part of the sequence until time t and ends in S_i and the $\beta_t(i)$ takes it from there and explains the ending part until time T
- $\alpha_t(i)\beta_t(i)$ explains the whole sequence given that at time t the system is in state S_i. It is normalized by dividing this over all possible intermediate states that can be traversed at time t, and guarantee that $\sum_i \gamma_t(i) = 1$
- to find the state sequence, for each time step t, choose the state that has the highest probability

$$q_t^* = \arg\max_i \gamma_t(i)$$

- this may choose S_i and S_j as the most probable states at time t and $t+1$ even when $a_{ij} = 0$.
- to find the single best state sequence, the Viterbi algorithm is used