Hidden Markov Models
(HMM)

A. Enis Cetin

University of Illinois at Chicago

aecyy@uic.edu [1]

April 5, 2018

[1] Based on Prof. Selim Aksoy's notes, Bilkent University, Ankara, Turkey

Overview

1. Applications of HMMs
2. Discrete Markov Processes
3. First-Order Markov Models
4. First-Order Hidden Markov Models
5. Three Fundamental Problems for HMMs
   - HMM Evaluation Problem

Applications of HMMs

- Speech recognition
- Optical character recognition
- Natural language processing (e.g., text summarization)
- Bioinformatics (e.g., protein sequence modeling)
- Image time series (e.g., change detection)
- Video analysis (e.g., story segmentation, motion tracking)
- Robot planning (e.g., navigation)
- Economics and finance (e.g., time series, customer decisions)

Discrete Markov Processes (Markov Chains)

- A sequence of Random Variables (RVs).
- Current RV is influenced by L previous rv's.
- Consider a system that can be described at any time as being in one of a set of N distinct states $w_1, w_2, \ldots, w_N$.
- Let $w(t)$ denote the actual state at time t where $t = 1, 2, \ldots$.
- Markov(L) random process: The probability of the system being in state $w(t)$ is $P(w(t)|w(t-1), \ldots, w(t-L))$.
- Markov(L=1): L=1 in speech recognition and in some image processing applications

First-Order Markov Models (Markov Chain)

- We assume that the state $w(t)$ is conditionally independent of the previous states given the predecessor state $w(t-1)$, i.e.,

$$P(w(t)|w(t-1), \ldots, w(1)) = P(w(t)|w(t-1)).$$

- We also assume that the Markov Process (Markov Chain) defined by $P(w(t)|w(t-1))$ is time homogeneous (independent of the time t).

First-Order Markov Models

- A particular *sequence of states* of length T is denoted by

$$W^T = \{w(1), w(2), \ldots, w(T)\}.$$

- The model for the generation of any sequence of state sequences is described by the *transition probabilities*:

$$a_{ij} = P(w(t) = w_j | w(t-1) = w_i)$$

where $i, j \in \{1, \ldots, N\}$, $a_{ij} \geq 0$, and $\sum_{j=1}^{N} a_{ij} = 1, \forall i$.

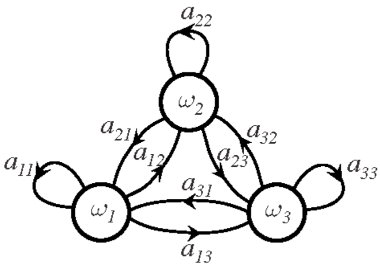
First-Order Markov Models

- There is no requirement that the transition probabilities are symmetric ($a_{ij} \neq a_{ji}$, in general).
- A particular state may be visited in succession ($a_{ii} \neq 0$, in general) and not every state need to be visited.
- If we can observe w(t) the process is called an *observable Markov model* because the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event.
- We may also observe a RV different from the value of the state w(t).

First-Order Markov Model Examples

- Consider the following 3-state first-order Markov model of the weather in Chicago:
  - w_1: rain/snow
  - w_2: cloudy
  - w_3: sunny

$$\Theta = \{a_{ij}\}$$
$$= \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$



- Estimate the transition probabilities from past data!

First-Order Markov Model Examples

- We can use this model to answer the following question: What is the probability that the weather for the next eight days will be "sunny-sunny-sunny-rainy-rainy-sunny-cloudy-sunny" ($W^8 = \{w_3, w_3, w_3, w_1, w_1, w_3, w_2, w_3\}$)?

First-Order Markov Model Examples

- We can use this model to answer the following question: What is the probability that the weather for the next eight days will be "sunny-sunny-sunny-rainy-rainy-sunny-cloudy-sunny" ($W^8 = \{w_3, w_3, w_3, w_1, w_1, w_3, w_2, w_3\}$)?
- Solution:

$$\begin{aligned}
P(W^8|\lambda) &= P(w_3, w_3, w_3, w_1, w_1, w_3, w_2, w_3) \\
&= P(w_3)P(w_3|w_3)P(w_3|w_3)P(w_1|w_3) \\
&\qquad P(w_1|w_1)P(w_3|w_1)P(w_2|w_3)P(w_3|w_2) \\
&= P(w_3)\, a_{33}\, a_{33}\, a_{31}\, a_{11}\, a_{13}\, a_{32}\, a_{23} \\
&= (1/3) \times 0.8 \times 0.8 \times 0.1 \times 0.4 \times 0.3 \times 0.1 \times 0.2 \\
&= 0.52 \times 10^{-4}
\end{aligned}$$

where the initial probabilities are assumed to be equal to each other $P(w_i) = 1/3$, i=1,2,3.

First-Order Hidden Markov Models

- In some cases we cannot observe the state but we observe an output of the system, which is a probabilistic function of the state.
- The resulting model, called a *Hidden Markov Model (HMM)*, has an underlying random "state" process that is not observable (it is hidden), but we can observe another set of random variables.

First-Order Hidden Markov Models

- We denote the observation at time t as $v(t)$ and the probability of producing that observation in state $w(t)$ as $P(v(t)|w(t))$.
- There are many possible state-conditioned observation distributions.
- When the observations are discrete, the distributions

$$b_{jk} = P(v(t) = v_k | w(t) = w_j)$$

are probability mass functions where $j \in \{1, \dots, N\}$, $k \in \{1, \dots, M\}$, $b_{jk} \geq 0$, and $\sum_{k=1}^{M} b_{jk} = 1, \forall j$.

- The random process $v(t)$ can be a Gaussian rv or mixture of Gaussians.

First-Order Hidden Markov Models

- When the observations are continuous, the distributions are typically specified using a parametric model family where the most common family is the Gaussian mixture

$$b_j(x) = \sum_{k=1}^{M_j} \alpha_{jk}\, p(x|\mu_{jk}, \Sigma_{jk})$$

  where $\alpha_{jk} \geq 0$ and $\sum_{k=1}^{M_j} \alpha_{jk} = 1, \forall j$.
- We observe a discrete set of observations of length T denoted by

$$V^T = \{v_1 = O_1, v_2 = O_2, \ldots, v_T = O_T\}.$$

  we do not observe the underlying state sequence in HMM!

First-Order Hidden Markov Models

- An HMM is characterized by:
  - N, the number of hidden states
  - $\{a_{ij}\}$, the state transition probabilities
  - $\{b_{jk}\}$, the observation symbol probability distribution
  - M, the number of distinct observation symbols per state in discrete observation case
  - $\{\pi_i = P(w(1) = w_i)\}$, the initial state distribution
  - HMM:
    $\lambda = (\{a_{ij}\}, \{b_{jk}\}, \{\pi_i\})$, the complete parameter set of the model

Three Fundamental Problems for HMMs

- *Evaluation problem:* Given the model, compute the probability that a particular output sequence was produced by that model (solved by the forward algorithm).
- *Decoding problem:* Given the model, find the most likely sequence of hidden states which could have generated a given output sequence (solved by the Viterbi algorithm).
- *Learning problem:* Given a set of output sequences, find the most likely set of state transition and output probabilities (solved by the Baum-Welch algorithm).

HMM Evaluation Problem

- A particular *sequence of observations* of length T is denoted by

$$V^T = \{v(1) = O_1, v(2) = O_2, \ldots, v(T) = O_T\} = O.$$

- The probability of observing this sequence can be computed by enumerating every possible state sequence of length T as

$$\begin{aligned} P(V^T|\lambda) &= \sum_{\text{all } W^T} P(V^T, W^T|\lambda) \\ &= \sum_{\text{all } W^T} P(V^T|W^T, \lambda)P(W^T|\lambda). \end{aligned}$$

HMM Evaluation Problem

- This summation includes N^T terms in the form

$$P(V^T|W^T)P(W^T) = \left(\prod_{t=1}^{T} P(v(t)|w(t))\right)\left(\prod_{t=1}^{T} P(w(t)|w(t-1))\right)$$
$$= \prod_{t=1}^{T} P(v(t)|w(t))P(w(t)|w(t-1))$$

  where $P(w(t)|w(t-1))$ for $t=1$ is $P(w(1))$.
- It is unfeasible with computational complexity $O(N^T T)$.
- However, a computationally simpler algorithm called the *forward algorithm* computes $P(V^T|\lambda)$ recursively.
- We also use logarithm of probabilities and additions instead of multiplying probabilities.

HMM Evaluation Problem

Given the observation sequence

$$V^T = O = \{v(1) = O_1, v(2) = O_2, \ldots, v(T) = O_T\}.$$

Given a set of states

$$W^T = Q = \{w(1) = q_1, w(2) = q_2, \ldots, w(T) = q_T\}.$$

$$P(O, W^T|\lambda) = \pi_{q1} b_{q1}(O_1) a_{q1q2} b_{q2}(O_2) \ldots a_{q(T-1)qT} b_{qT}(O_T).$$

To cover all possibilities we sum over all possible state sequences

$$P(O|\lambda) = \sum_{all\ \ Q} \pi_{q1} b_{q1}(O_1) a_{q1q2} b_{q2}(O_2) \ldots a_{q(T-1)qT} b_{qT}(O_T).$$

Too many multiplications... However there is a fast algorithm (see the articles by Rabiner)

Computational Cost (from Rabiner's paper)

A little thought should convince the reader that the calculation of $P(O|\lambda)$, according to its direct definition (17) involves on the order of $2T \cdot N^T$ calculations, since at every $t = 1, 2, \cdots, T$, there are N possible states which can be reached (i.e., there are N^T possible state sequences), and for each such state sequence about $2T$ calculations are required for each term in the sum of (17). (To be precise, we need $(2T - 1)N^T$ multiplications, and $N^T - 1$ additions.) This calculation is computationally unfeasible, even for small values of N and T; e.g., for $N = 5$ (states), $T = 100$ (observations), there are on the order of $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$

multiplications.

Forward Algorithm

1) **Initialization:**

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N.$$

2) **Induction:**

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

$$1 \le j \le N.$$

3) **Termination:**

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

Example: $N = 5$, $T = 100 \Rightarrow 3000$ computations instead of 10^{72}.

Recognition Process

Let the observation sequence be

$$V^T = O = \{v(1) = O_1, v(2) = O_2, \ldots, v(T) = O_T\}.$$

and a group of HMMs;

$$\lambda_1, \lambda_2, \ldots, \lambda_K$$

Calculate

$$P(O|\lambda_1), P(O|\lambda_2), \ldots, P(O|\lambda_K)$$

The model producing the highest probability is the "winner":

$$K^* = argmax_k P(O|\lambda_k)$$

Review Article

Reference

- Key
- HMM
- Reference

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

References

Larry Rabiner (1989)
A tutorial on hidden Markov models and selected applications in speech recognition
Proceedings of IEEE 77(2), 257 – 286.

Larry Rabiner, BH Juang (1986)
An introduction to hidden Markov models
IEEE ASSP Magazine 3(1), 4 – 16.

The End