Hidden Markov Models

A Summary for

“A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, by Lawrence R. Rabiner”

By Seçil Öztürk
Outline

• Signal Models
• Markov Chains
• Hidden Markov Models
• Fundamentals for HMM Design
• Types of HMMs
Signal Models...

- Are used to characterize real world signals.
- Provide a basis for a theoretical description of a signal processing system.
- Tell about the signal source without having the source available.
- Are used to realize practical systems efficiently.
2 Types of Signal Models:

- **Deterministic Models:**
  - Specific properties of the signal are known.
    - eg. The signal is a sine wave
  - Determining values for parameters of the signal, such as frequency, amplitude, etc is required.

- **Statistical Models:**
  - eg. Gaussian processes, Markov processes, Hidden Markov processes
  - Characterizing the statistical properties of the signal is required.
  - Assumption:
    * Signal can be characterized as a parametric random process.
    * Parameters of the random process can be determined in a precise and well defined manner.
Discrete Markov Processes

The system is described by N distinct states: \( S_1, S_2, ..., S_N \).

The system can be in one of these states at any time.

Time instants associated with state changes are: \( t = 1, 2, ... \).

Actual state at time \( t \) is \( q_t \).

Predecessor states must also be known for the probabilistic description.

\( a_{ij} \)'s are state transition probabilities.

Assuming discrete, first order Markov Chain, the probabilistic description of this system is:

\[
P[ q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ... ] = P[ q_t = S_j \mid q_{t-1} = S_i ]
\]
A 3 State Example for Weather

- States are defined as:
  - State 1: rainy/snowy
  - State 2: cloudy
  - State 3: sunny
- The weather at day $t$ should be in one of the states above.
- State transition matrix is $A$.
- $a_{ij}$'s represent the probabilities of going from state $i$ to $j$.
- The observation sequence is denoted with $O$.
  - Say for $t=1$, sun is observed. (initial state)
  - Next observation: sun-sun-rain-rain-cloudy-sun
  - $O = \{S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$
  
  corresponding to
  $t=1, 2, 3, 4, 5, 6, 7, 8$

$$A = \{a_{ij}\} = \begin{bmatrix}
0.4 & 0.3 & 0.3 \\
0.2 & 0.6 & 0.2 \\
0.1 & 0.1 & 0.8 \\
\end{bmatrix}.$$
A 3 State Example for Weather

- The probability of the observation sequence given the model is as follows:

\[
P(O|\text{Model}) = P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|\text{Model})
\]
\[
= P(S_3) \cdot P(S_3|S_3) \cdot P(S_3|S_3) \cdot P(S_1|S_3)
\]
\[
\quad \cdot P(S_1|S_1) \cdot P(S_3|S_1) \cdot P(S_2|S_3) \cdot P(S_3|S_2)
\]
\[
= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23}
\]
\[
= 1 \cdot (0.8) (0.8) (0.1) (0.4) (0.3) (0.1) (0.2)
\]
\[
= 1.536 \times 10^{-4}
\]

where we use the notation

\[
\pi_i = P(q_1 = S_i), \quad 1 \leq i \leq N
\]

Here, \( \pi_i \)'s are the initial state probabilities.
In Markov Models, states corresponded to observable/physical events.

In Hidden Markov Models, observations are probabilistic functions of the state.

So, HMMs are **doubly embedded stochastic processes**.

The underlying stochastic process is not observable/hidden. It can be observed through another set of stochastic processes producing the observation sequences.

*ie., in Markov Models, the problem is finding the probability of the observation to be in a certain state, in HMMs, the problem is still finding the probability of the observation to be in a certain state, but observation is also a probabilistic function of the state.*

eg. Hidden Coin Tossing Experiment, Urn and Ball Model
Elements of an HMM

- **N**: # of states
  Individual states: \( S = \{S_1, \ldots, S_N\} \) 
  State at time \( t \): \( q_t \)

- **M**: (# of distinct observation symbols)/state
  ie. discrete alphabet size in speech processing
  eg. heads & tails in coins experiment
  individual symbols: \( V = \{v_1, v_2, \ldots, v_M\} \)

- **A**: State transition prob. distribution
  \( a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \) where \( 1 \leq i, j \leq N \)

- **B**: the observation symbol probability distribution in state \( j \)
  \( B = b_{j}(k) = P[v_k \text{ at } t | q_t = S_j] \) where \( 1 \leq j \leq N \)
  \( 1 \leq k \leq M \).
  eg. The probability of heads of a certain coin at time \( t \).

- **\( \pi \)** = \{ \( \pi_i \) \} is are the initial state distribution.
  - \( \pi_i = P[q_1 = S_i] \) where \( 1 \leq i \leq N \)

*** Given \( N, M, A, B, \pi \), HMM can be generated for \( O \).

- **O**: Observation sequence \( O = O_1, O_2, \ldots, O_T \).
  \( O_T \)'s are one of \( v_i \)'s. \( T \) is # of total observations.
Complete specification of an HMM requires:

* $A, B, \pi$: probability measures
* $N$ and $M$: model parameters
* $O$: observation symbols

HMM notation: $\lambda \ (A, B, \pi)$
Three Fundamental Questions in Modelling HMMs

1) Evaluation Problem:
   Given Observation sequence: $O = O_1 O_2 \ldots O_T$
   HMM model: $\lambda (A,B,\pi)$

   How to compute $P(O|\lambda)$?

2) Uncover Problem:
   Given Observation sequence: $O = O_1 O_2 \ldots O_T$
   HMM model: $\lambda (A,B,\pi)$

   How to choose corresponding optimal state seq. $Q=q_1q_2\ldots q_T$?

3) Training Problem:
   How to adjust parameters $A,B,\pi$ to maximize $P(O|\lambda)$?
Solution for Problem 1

\[ P(O|\lambda) = ? \]
Enumerate every possible T length state sequence. eg. Assume fixed \( Q = q_1q_2 \ldots q_T \)

\[
P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda)
\]

\[
P(O|Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdots b_{q_T}(O_T).
\]

\[
P(Q|\lambda) = \pi_{q_1} a_{q_1,q_2} a_{q_2,q_3} \cdots a_{q_{T-1},q_T}.
\]

\[
P(O|\lambda) = \sum_{Q} P(O|Q, \lambda) P(Q|\lambda)
\]

\[
= \sum_{q_1,q_2,\ldots,q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1,q_2} b_{q_2}(O_2) \cdots a_{q_{T-1},q_T} b_{q_T}(O_T).
\]

Unfeasible computation time!
On order of \( 2T.N^T \)
Solution for Problem 1

Forward-Backward procedure

Forward variable:

$$\alpha_t(i) = P(O_1O_2\ldots O_t, q_t = S_i | \lambda)$$

(prob. for partial observation sequence $O_1\ldots O_t$ ending at state $S_i$ at time $t$, $\lambda$)

Inductive Solution!

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N.$$

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

$$\quad 1 \leq j \leq N.$$

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

Computation time:

On order of $N^2T$
Trellis
Backward Variable:
\[ \beta_t(i) = P(O_{t+1}O_{t+2}...O_T|q_T=S_i, \lambda) \]
(probability of the partial observation sequence from \(t+1\) to end, given state \(S_i\) at time \(t\), \(\lambda\))

Inductive Solution!

\[ \beta_t(i) = 1, \quad 1 \leq i \leq N. \]

\[ \beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T - 1, T - 2, \ldots, 1, \quad 1 \leq i \leq N. \]

Computation time: 
On order of \(N^2T\)
Solution for Problem 2

* Aim is to find an optimal state sequence for the observation.
* Several solutions exist.
* Optimality criteria must be adjusted.

eg. states individually most likely at time t.
  maximizes expected # of correct individual states.
A posteriori probability variable $\gamma$:

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

(probability of being in state $S_i$ at time $t$ given observation sequence $O$ and $\lambda$)

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

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

$P(O|\lambda)$ is normalization factor to make sure sum of $\gamma_t(i)$'s equals 1

Individually most likely state $q_t$ at time $t$:

$$q_t = \arg \max_{1 \leq i \leq N} [\gamma_t(i)], \quad 1 \leq t \leq T.$$ 

Problem:
This equation finds the most likely state at each $t$ regardless of the probability of occurrence of states, so the resulting sequence may be invalid.
Possible solution to the problem above: Find the state sequence maximizing pairs or triples of states

OR

Find the single best state sequence to maximize $P(Q|O,\lambda)$ equivalent to maximize $P(Q,O|\lambda)$
Viterbi Algorithm

Aim: to find the single best state sequence $Q=\{q_1q_2...q_T\}$ for given observation sequence $O=\{O_1O_2...O_T\}$

Define $\delta$: (the best score, ie. highest probability along a single path, at time $t$) (accounts for the first $t$ observations, ends in state $S_i$)

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_t} \ p(q_1, q_2, \cdots, q_t = i, O_1, O_2, \cdots, O_t | \lambda)$$

$$\delta_{t+1}(j) = \max_i \delta_t(i) a_{ij} \cdot b_j(O_{t+1})$$

For each $t$ and $j$, must keep track of argument maximizing above equation.

Use array $\psi_t(j)$
Viterbi Algorithm

To find the best state sequence:

1. Initialization

\[ \delta_t(i) = \pi_i b_i(O_t), \quad 1 \leq i \leq N \]
\[ \psi_t(i) = 0. \]

2. Recursion

\[ \delta_t(j) = \max_{1 \leq s \leq N} \{ \delta_{t-1}(s) a_{sj} b_j(O_t) \}, \quad 2 \leq t \leq T \]
\[ 1 \leq j \leq N \]
\[ \psi_t(j) = \arg\max_{1 \leq s \leq N} \{ \delta_{t-1}(s) a_{sj} \}, \quad 2 \leq t \leq T \]
\[ 1 \leq j \leq N. \]

3. Termination

\[ p^* = \max_{1 \leq i \leq N} \{ \delta_T(i) \} \]
\[ q_T^* = \arg\max_{1 \leq i \leq N} \{ \delta_T(i) \}. \]

4. Path/State Sequence Backtracking

\[ q_{t+1}^* = \psi_t(q_t^*), \quad t = T - 1, T - 2, \ldots, 1. \]

- Just like forward procedure.
- But finds max instead of summation.
- \( \psi \) Keeps track of maximizing points.
Solution for Problem 3

Aim: Adjusting $A, B, \pi$ to maximize the probability of the training data.

Choose $\lambda (A, B, \pi)$ such that $P(O|\lambda)$ is locally maximized using:

Methods:
* Baum-Welch Method
* Expectation-Modification (EM) Method
* Gradient Techniques
Define Variable $\xi$:

$$
\xi_{t}(i,j) = P(q_{t}=S_{i}, q_{t+1}=S_{j}|O, \lambda)
$$

(the probability of being in state $S_{i}$ at $t$, in $S_{j}$ at $t+1$, given observation and model)

The path satisfying this condition:

\[\xi(i,j) = \frac{P(q_{t}=i, q_{t+1}=j, O | \lambda)}{P(O | \lambda)}\]

\[\xi(i,j) = \frac{\alpha_{t}(i) a_{ij} b_{j}(O_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)}\]

\[= \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{t}(i) a_{ij} b_{j}(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{t}(i) a_{ij} b_{j}(O_{t+1}) \beta_{t+1}(j)}\]

Relate to $\gamma$:

$$
\gamma_{t}(i) = \sum_{j=1}^{N} \xi(i,j).
$$
Baum-Welch Method

Expected number of transitions made from state $S_i$ in $O$:

$$\sum_{t=1}^{T-1} \gamma_t(i)$$

Expected number of transitions made from state $S_i$ to $S_j$ in $O$:

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

Reestimation formulas for $A,B,\pi$:

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time } (t = 1) = \gamma_t(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_j \text{ to state } S_i}{\text{expected number of transitions from state } S_i}$$

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

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j) \text{ s.t. } O_t = v_k}$$
Baum-Welch Method

Current Model: $\lambda (A,B,\pi)$
Reestimation Model: $\bar{\lambda} (\bar{A}, \bar{B}, \bar{\pi})$

Either:
1) $\lambda$ defines critical point of the likelihood function, where $\lambda = \bar{\lambda}$
2) model $\bar{\lambda}$ is more likely than $\lambda$
   in the sense $P(O|\bar{\lambda}) > P(O|\lambda)$

So $\bar{\lambda}$ is the new model matching the observation sequence better.

Using $\bar{\lambda}$ as $\lambda$ iteratively and repeating reestimation calculation, improvement for the probability of $O$ being observed in model is reached.
Final result is called a maximum likelihood estimate of the HMM.
Baum-Welch Method

Reestimation formulas can be derived by maximizing Baum's auxiliary function over \( \lambda \):

\[
Q(\lambda, \bar{\lambda}) = \sum_o P(O|O, \lambda) \log P(O, Q|\bar{\lambda})
\]

Proved that maximizing \( Q(\lambda, \bar{\lambda}) \) leads to increased likelihood.

\[
\max_\lambda [Q(\lambda, \bar{\lambda})] \Rightarrow P(O|\bar{\lambda}) \geq P(O|\lambda).
\]

Eventually likelihood function converges to a critical point.
Baum-Welch Method

Stochastic constraints are satisfied in each reestimation procedure:

\[
\sum_{i=1}^{N} \pi_i = 1
\]
\[
\sum_{j=1}^{N} \tilde{a}_{ij} = 1, \quad 1 \leq i \leq N
\]
\[
\sum_{k=1}^{M} \tilde{b}_j(k) = 1, \quad 1 \leq j \leq N
\]

Also, Lagrange multipliers can be used to find \( \pi, a_{ij}, b_j(k) \) parameters maximizing \( P(O|\lambda) \) (Think of the parameter estimation as a constrained optimization problem for \( P(O|\lambda) \), constrained by above equations)
Using Lagrange Multipliers;

\[
\tau_i = \frac{\pi_i \frac{\partial P}{\partial \pi_i}}{\sum_{k=1}^{N} \pi_k \frac{\partial P}{\partial \pi_k}}
\]

\[
a_{ij} = \frac{\frac{\partial P}{\partial a_{ij}}}{\sum_{k=1}^{N} a_{ik} \frac{\partial P}{\partial a_{ik}}}
\]

\[
b_{j}(k) = \frac{\frac{\partial P}{\partial b_{j}(k)}}{\sum_{\ell=1}^{M} b_{\ell}(\ell) \frac{\partial P}{\partial b_{\ell}(\ell)}}
\]

Manipulating these equations, it can be shown that reestimation formulas are correct at critical points of \(P(O|\lambda)\)
So far, considered only ergodic HMMs:

**Ergodic Model:**
Every state transition is possible. $a_{ij}$'s positive.

**Left-Right (Bakis) Model:**
As time increases, state index increases or stays the same. i.e. states proceed from left to right.

\[ a_{ij} = 0, \quad j < i \]
\[ \pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \]
\[ a_{NN} = 1 \]
\[ a_{Ni} = 0, \quad i < N. \]

\[ A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix} \]
Continuous Observation Densities in HMMs

- Finite alphabet up to now.
- Observations are continuous signals/vectors.
- General representation of the pdf:

\[
b_j(O) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(O, \mu_{jm}, U_{jm}), \quad 1 \leq j \leq N
\]

- \(O\): vector being modeled
- \(c_{jm}\): mixture coeff. for \(m\)th mixture in state \(j\)
- \(\mathcal{N}\): log concave/elliptical symmetric density (e.g., Gaussian) mean: \(\mu_{jm}\), cov: \(U_{jm}\)

- \(c_{jm}\) should satisfy

\[
\sum_{m=1}^{M} c_{jm} = 1, \quad 1 \leq j \leq N
\]
\[
c_{jm} \geq 0, \quad 1 \leq j \leq N, \quad 1 \leq m \leq M
\]

such that pdf is normalized:

\[
\int_{-\infty}^{\infty} b_j(x) \, dx = 1, \quad 1 \leq j \leq N.
\]
Continuous Observation Densities in HMMs

Reestimation formulas:

\[ \tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j, k)} \]

\[ \tilde{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot O_t}{\sum_{t=1}^{T} \gamma_t(j, k)} \]

\[ \tilde{U}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot (O_t - \mu_{jk})(O_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j, k)} \]

\[ \gamma_t(j,k) \text{ prob. Of being in state } j \text{ at time } t \text{ with } k^{th} \text{ mixture component accounting for } O_t \]

\[ \gamma_t(j,k) = \left[ \frac{\alpha_t(j) \beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i) \beta_t(i)} \right] \left[ \frac{c_{jk} S_t(O_t, \mu_{jk}, U_{jk})}{\sum_{m=1}^{M} c_{jm} S_t(O_t, \mu_{jm}, U_{jm})} \right] \]
Autoregressive HMMs

- Particularly applicable to speech processing.
- Observation vectors are drawn from an autoregression process.
- Observation vector \( O: (x_0, x_1, \ldots, x_{k-1}) \)
- \( \theta_k \)'s are related by:

\[
\theta_k = -\sum_{i=1}^{p} a_{ij} \theta_{k-i} + e_k
\]

where \( e_k, k=0, 1, 2, 3, \ldots, p \) are Gaussian, independent, identically distributed rv. with zero mean, variance \( \sigma^2 \)
- \( a_{ij}, i=1,\ldots,p \) are predictor (autoregression) coefficients.
Autoregressive HMMs

- For large $K$, density function $O$ is approximately:

$$f(O) = \frac{(2\pi \sigma^2)^{-K/2}}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \delta(O, a) \right\}$$

where

$$\delta(O, a) = r_a(0) r(0) + 2 \sum_{i=1}^{p} r_a(i) r(i)$$

$$a' = [1, a_1, a_2, \cdots, a_p]$$

- $r(i)$ autocorrelation of observation samples
- $r_a(i)$ autocorr. Of autoreg. coeff.s
Autoregressive HMMs

- Total prediction residual $\alpha$ is
  \[
  \alpha = E \left[ \sum_{i=1}^{K} (e_i)^2 \right] = K \sigma^2
  \]

  $\sigma^2$ is variance per sample of error signal.

- Normalized observation vector:
  \[
  \phi = \frac{O}{\sqrt{\alpha}} = \frac{O}{\sqrt{K \sigma^2}}
  \]

- Samples $x_i$'s are divided by $\sqrt{K \sigma^2}$ (normalized by sample variance)

- Likelihood function
  \[
  f(\Theta) = \left( \frac{2\pi}{K} \right)^{-K/2} \exp \left( -\frac{K}{2} \delta(\Theta, \alpha) \right)
  \]
Autoregressive HMMs

Using Gaussian autoregressive density, assume the mixture density:

\[ b_j(O) = \sum_{m=1}^{M} c_{jm} b_{jm}(O) \]

Each \( b_{jm}(O) \) is density with autoregression vector \( a_{jm} \) (or autocorr. vector \( r_{ajm} \))

\[ b_{jm}(O) = \left( \frac{2\pi}{K} \right)^{-K/2} \exp \left\{ -\frac{1}{2} \delta(O, a_{jm}) \right\} \]

Reestimation formula for sequence autocorrelation \( r(i) \) for the jth state, kth mixture component:

\[ \bar{r}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot r_t}{\sum_{t=1}^{T} \gamma_t(j, k)} \]

Where \( \gamma_t(j,k) \) is the prob. of being in state j at time t, using mixture component k,
NULL Transitions:
Observations are associated with the arcs of the model.
Used for transitions which makes no output. (jumps between states produce no observation)
Eg: a left-right model:

It is possible to omit transitions between states and conclude with 1 observation to account for a path beginning in state 1, ending in state N.
Tied States

• Equivalence relation between HMM parameters in different states.
• # of independent parameters in model is reduced.
• Used in cases where observation density is the same for two or more states. (eg in speech sounds)
• Model becomes simpler for parameter estimation
• Inclusion of Explicit State Duration Density in HMMs
• Optimization Criterion
Bibliography

• A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, by Lawrence R. Rabiner
• Fundamentals of Speech Recognition, by Lawrence R. Rabiner Biign Hwang Juang
Thanks for Listening!