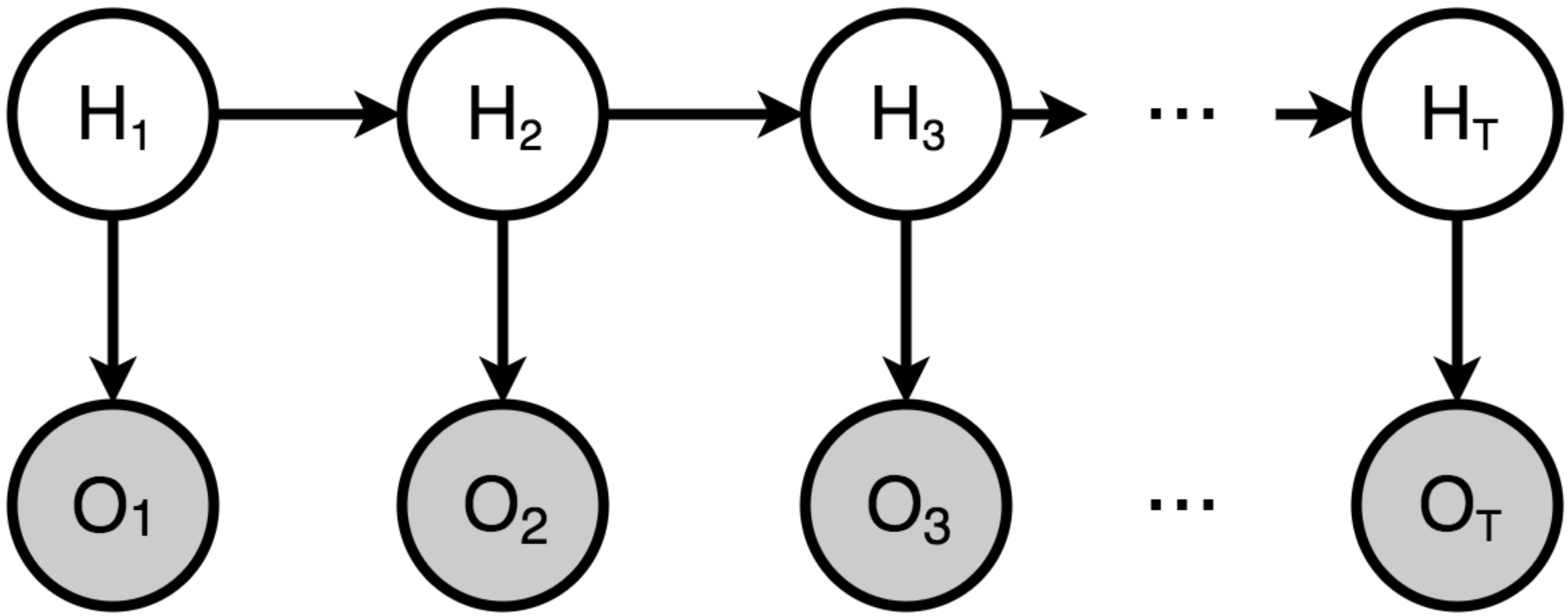
Introduction to Statistical Learning

INF 552, Machine Learning for Data Informatics

University of Southern California

M. R. Rajati, PhD

Lesson 11
Hidden Markov Models



Hidden Markov Models

- What is a hidden Markov model (HMM)?
  - A machine learning technique and…
  - …a discrete hill climb technique
  - Two for the price of one!
- Where are HMMs used?
  - Speech recognition, information security, and too many other things to list
  - Q: Why are HMMs so useful?
  - A: Widely applicable and ***efficient algorithms***

Markov Chain

Markov chain

- “Memoryless random process”
- Transitions depend only on current state (Markov chain of order 1)…
- …and transition probability matrix

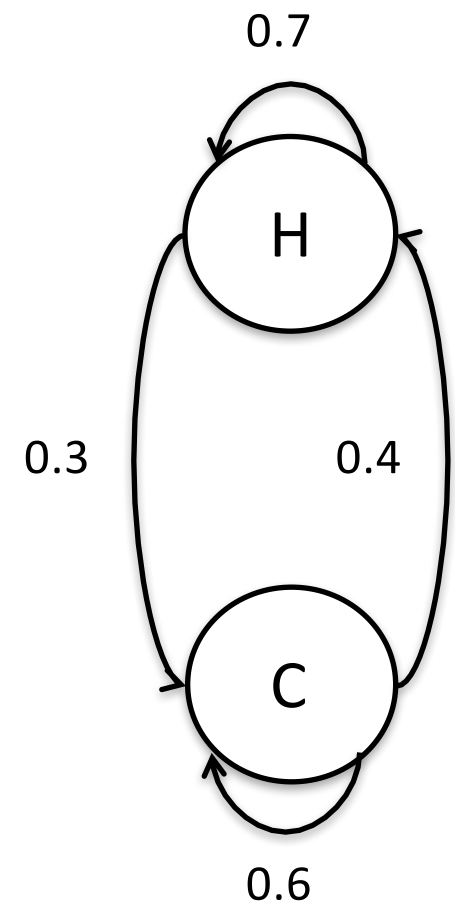
Markov Chain

Suppose we're interested in average annual temperature

- Only consider Hot and Cold

From recorded history, obtain probabilities for…

- Year-to-year transitions
- Based on thermometer readings for "recent" years

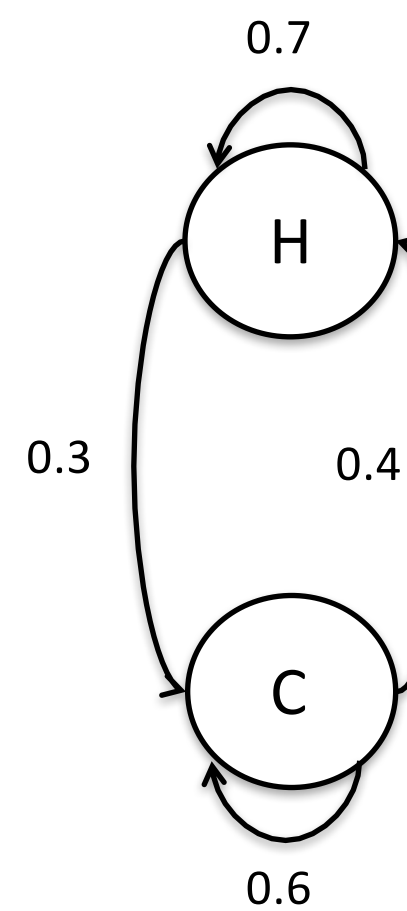


Markov Chain

Transition probability matrix

Matrix is denoted as A

$$\begin{array}{cc} & \begin{array}{cc} H & C \end{array} \\ \begin{array}{c} H \\ C \end{array} & \left[\begin{array}{cc} 0.7 & 0.3 \\ 0.4 & 0.6 \end{array}\right] \end{array}$$



Note, A is "row stochastic"

$$A = \left[\begin{array}{cc} 0.7 & 0.3 \\ 0.4 & 0.6 \end{array}\right]$$

So, each element of A is between 0 and 1 and each row satisfies the definition of a discrete probability distribution, thus the elements of any given row sum to 1.

Markov Chain

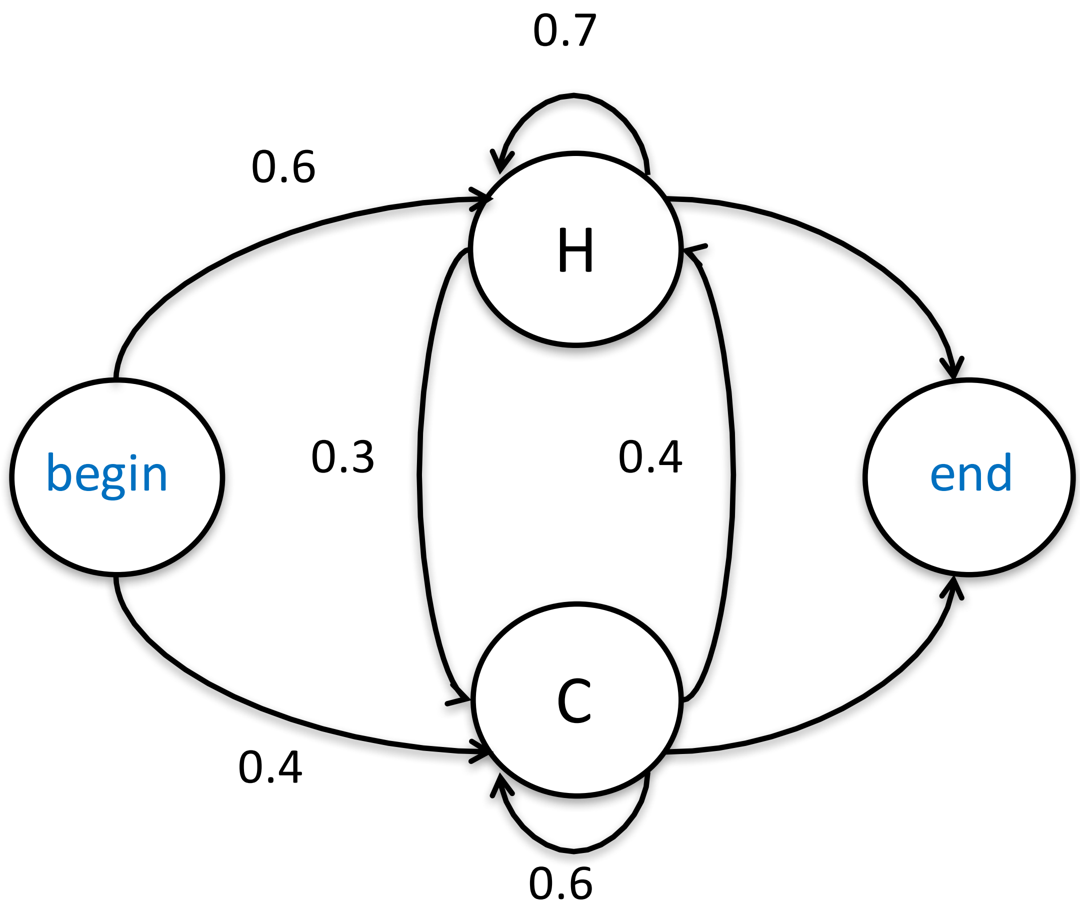
Can also include begin, end states

Matrix for begin state denoted π

In this example,

$$\pi = \left[\begin{array}{cc} 0.6 & 0.4 \end{array} \right]$$

Note that π is also row stochastic



Hidden Markov Model

HMM includes a Markov chain/ process

- But the Markov process is "hidden", i.e., we can't directly observe the Markov process
- Instead, observe things that are probabilistically related to hidden states
- It's as if there is a "curtain" between Markov chain and the observations

HMM Example

Consider H/C annual temp example

Suppose we want to know H or C annual temperature in distant past

- Before thermometers were invented
- Note, we only distinguish between H and C

We assume transition between Hot and Cold years is same as today

Then the A matrix is known

HMM Example

Temps in past follow a Markov process

- But, we cannot observe temperature in past

We find evidence that tree ring size is related to temperature

- Looking at historical data, we find this holds

We only consider 3 tree ring sizes

- Small, Medium, Large (S, M, L, respectively)

Measure tree ring sizes and recorded temperatures to determine relationship

HMM Example

We find that tree ring sizes and temperature related by

$$\begin{array}{cc} & \begin{array}{ccc} S & M & L \end{array} \\ \begin{array}{c} H \\ C \end{array} & \left[\begin{array}{ccc} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{array}\right] \end{array}$$

This is known as the B matrix

The matrix B is also row stochastic

$$B = \left[\begin{array}{ccc} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{array}\right]$$

HMM Example

Can we now find H/C temps in past?
We cannot measure (observe) temps
But we can measure tree ring sizes…
…and tree ring sizes related to temps
  By probabilities in the B matrix
Can we say something intelligent about temps over some interval in the past?

HMM Notation

A lot of notation is required

Notation may be the most difficult part…

$$
\begin{array}{rcl}
T & = & \text{the length of the observation sequence} \\
N & = & \text{the number of states in the model} \\
M & = & \text{the number of observation symbols} \\
Q & = & \{q_0, q_1, \ldots, q_{N-1}\} = \text{the states of the Markov process} \\
V & = & \{0, 1, \ldots, M-1\} = \text{set of possible observations} \\
A & = & \text{the state transition probabilities} \\
B & = & \text{the observation probability matrix} \\
\pi & = & \text{the initial state distribution} \\
\mathcal{O} & = & (\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1}) = \text{observation sequence.}
\end{array}
$$

HMM Notation

Note that for simplicity, observations taken from V = {0,1,…,M-1}

That is, $\mathcal{O}_i \in V \text{ for } i = 0, 1, \ldots, T-1$

The matrix A = $\{a_{ij}\}$ is N x N, where

$$a_{ij} = P(\text{state } q_j \text{ at } t+1 \,|\, \text{state } q_i \text{ at } t)$$

The matrix B = $\{b_j(k)\}$ is N x M, where

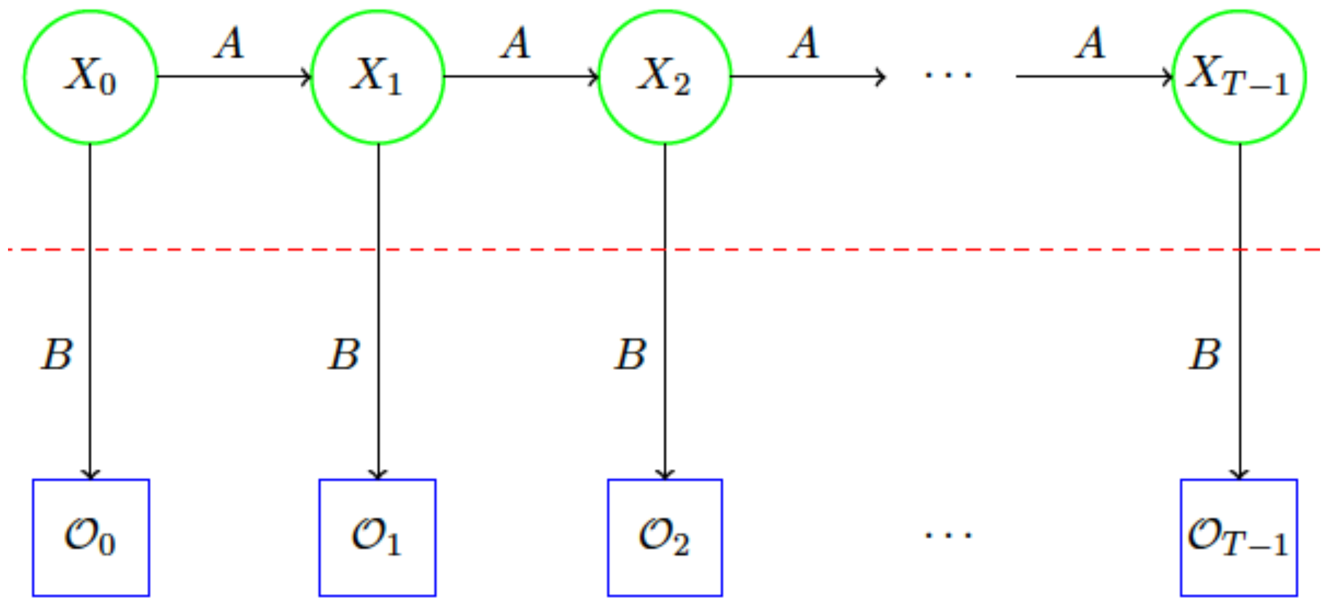
$$b_j(k) = P(\text{observation } k \text{ at } t \,|\, \text{state } q_j \text{ at } t).$$

HMM Example

- Consider our temperature example…
- What are the possible observations?
  - $V = \{0,1,2\}$, corresponding to S, M, L
- What are states of Markov process?
  - $Q = \{H,C\}$
- What are A,B, π, and T?
  - A,B, π on previous slides
  - T is number of tree rings measured
- What are N and M?
  - $N = 2$ and $M = 3$

Generic HMM

Generic view of HMM



HMM defined by A,B, and π

We denote HMM "model" as λ = (A,B,π)

HMM Example

Suppose that we observe tree ring sizes

For a 4 year period of interest: S,M,S,L

Then $\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3)$ = (0, 1, 0, 2)

Most likely (hidden) state sequence?

That is, most likely $X = (X_0, X_1, X_2, X_3)$

Let π_{X0} be prob. of starting in state x_0

Note $b_{x_0}(\mathcal{O}_0)$ prob. of initial observation

And $a_{X0,X1}$ is prob. of transition X_0 to X_1

And so on…

HMM Example

$$\pi = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$$

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array} \quad \begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{cc} H & C \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

Bottom line?
We can compute P(X) for any X
For X = (X_0, X_1, X_2, X_3) we have

$$P(X) = \pi_{x0} b_{x0}(\mathcal{O}_0) a_{x0,x1} b_{x1}(\mathcal{O}_1) a_{x1,x2} b_{x2}(\mathcal{O}_2) a_{x2,x3} b_{x3}(\mathcal{O}_3)$$

Suppose we observe (0,1,0,2), then what is probability of, say, HHCC?
Plug into formula above to find

$$P(HHCC) = 0.6(0.1)(0.7)(0.4)(0.3)(0.7)(0.6)(0.1) = 0.000212.$$

HMM Example

Do same for all 4-state seq's

We find that the winner is…

CCCH

Not so fast!

state	probability	normalized probability
$HHHH$	.000412	.042787
$HHHC$	.000035	.003635
$HHCH$	.000706	.073320
$HHCC$	.000212	.022017
$HCHH$	.000050	.005193
$HCHC$	.000004	.000415
$HCCH$	.000302	.031364
$HCCC$	.000091	.009451
$CHHH$	.001098	.114031
$CHHC$	.000094	.009762
$CHCH$	.001882	.195451
$CHCC$	.000564	.058573
$CCHH$	.000470	.048811
$CCHC$	.000040	.004154
$CCCH$	.002822	.293073
$CCCC$	.000847	.087963

HMM Example

- The ***path*** CCCH scores the highest
- In dynamic programming (DP), we find highest scoring path
- But, in HMM we maximize ***expected number of correct states***
  - Sometimes called “EM algorithm”
  - For “Expectation Maximization”
- How does HMM work in this example?

HMM Example

For first position…
- Sum probabilities for all paths that have H in 1st position, compare to sum of probs for paths with C in 1st position: biggest wins

Repeat for each position and we find

	element			
	0	1	2	3
$P(H)$	0.188182	0.519576	0.228788	0.804029
$P(C)$	0.811818	0.480424	0.771212	0.195971

HMM Example

	element			
	0	1	2	3
$P(H)$	0.188182	0.519576	0.228788	0.804029
$P(C)$	0.811818	0.480424	0.771212	0.195971

So, HMM solution gives us CHCH

While DP solution is CCCH

Which solution is better?

Neither solution is better!

Just using different definitions of "best"

HMM Paradox?

HMM maximizes expected number of correct states

- Whereas DP chooses “best” overall path

Possible for HMM to choose a “path” that is impossible

- Could be a transition probability of 0

Cannot get impossible path with DP

Is this a flaw with HMM?

- No, it’s a feature

Probability of Observations

Table computed for

$\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3)$

= (0,1,0,2)

For this sequence,

P($\mathcal{O}$) = .000412 + .000035 + .000706 + … + .000847

= left to the reader

Similarly for other observations $\mathcal{O}$

state	probability	normalized probability
$HHHH$	.000412	.042787
$HHHC$	.000035	.003635
$HHCH$	.000706	.073320
$HHCC$	.000212	.022017
$HCHH$	.000050	.005193
$HCHC$	.000004	.000415
$HCCH$	.000302	.031364
$HCCC$	.000091	.009451
$CHHH$	.001098	.114031
$CHHC$	.000094	.009762
$CHCH$	.001882	.195451
$CHCC$	.000564	.058573
$CCHH$	.000470	.048811
$CCHC$	.000040	.004154
$CCCH$	.002822	.293073
$CCCC$	.000847	.087963

Probability of Observations

If this calculation is made for all possible 4-observation sequences then the sum of the resulting probabilities (not the normalized probabilities) will be 1.

state	probability	normalized probability
$HHHH$	.000412	.042787
$HHHC$	.000035	.003635
$HHCH$	.000706	.073320
$HHCC$	.000212	.022017
$HCHH$	.000050	.005193
$HCHC$	.000004	.000415
$HCCH$	.000302	.031364
$HCCC$	.000091	.009451
$CHHH$	.001098	.114031
$CHHC$	.000094	.009762
$CHCH$	.001882	.195451
$CHCC$	.000564	.058573
$CCHH$	.000470	.048811
$CCHC$	.000040	.004154
$CCCH$	.002822	.293073
$CCCC$	.000847	.087963

HMM Model

- An HMM is defined by the three matrices, A, B, and π
- Note that M and N are implied, since they are the dimensions of matrices
- So, we denote an HMM "model" as

$$\lambda = (A,B,\pi)$$

The Three Problems

HMMs used to solve 3 problems:

Problem 1: Given a model $\lambda = (A,B,\pi)$ and observation sequence O, find $P(O|\lambda)$

That is, we can ***score*** an observation sequence to see how well it fits a given model

The Three Problems

HMMs used to solve 3 problems

Problem 2: Given $\lambda = (A,B,\pi)$ and O, find an optimal state sequence (in HMM sense)

- Uncover hidden part (like previous example)

In many applications in NLP, the solution to Problem 2 is crucial. Example: Finding the grammatical roles of words in a sentence.

The Three Problems

HMMs used to solve 3 problems

Problem 3: Given O, N, and M, find the model λ that maximizes probability of O

That is, ***train*** a model to fit observations

HMMs in Practice

Often, HMMs used as follows:

Given an observation sequence…

- Assume that (hidden) Markov process exists

Train a model based on observations

- That is, solve Problem 3
- "Best" N can be found by trial and error

Then given a sequence of observations, score it versus the model we trained

- This is Problem 1: high score implies similar to training data, low score says it's not

HMMs in Practice

In this sense, HMM is a “machine learning” technique

- To train a model, we do not need to specify anything except the parameter N
- “Best” N often found by trial and error

So, we don’t need to “think” too much

- Just train HMM and then use it
- Fortunately, there are efficient algorithms for HMMs

The Three Solutions

We give detailed solutions to 3 problems

- Note: We must find ***efficient*** solutions

The three problems:

- **Problem 1**: Score an observation sequence versus a given model
- **Problem 2**: Given a model, “uncover” hidden part
- **Problem 3**: Given an observation sequence, train a model

Recall that we considered example for 2 and 1, but direct solutions are **very inefficient**