Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 8: 27 September 2011

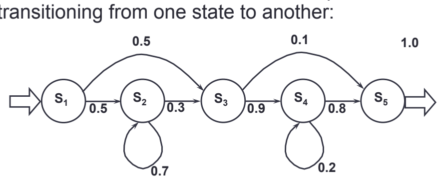
Agenda

- HW2 – due Thursday
- Questions, comments, concerns?
- Markov Chains
- Hidden Markov Models (HMMs)
- Forward Algorithm, Viterbi Algorithm (next time)

Markov Chain, Markov Models

A Markov Model (Markov Chain) is:

- similar to a finite-state automata, with *probabilities* of transitioning from one state to another:



- Transition from state to state at discrete time intervals
- Can only be in 1 state at any given time

Elements of a Markov Model (Chain):

- **clock**
  $t = \{1, 2, 3, \ldots T\}$
- ***N* states**
  $Q = \{1, 2, 3, \ldots N\}$
  the single state *j* at time *t* is referred to as q_t
- ***N* events**
  $E = \{e_1, e_2, e_3, \ldots, e_N\}$
- **initial probabilities**
  $\pi_j = P[q_1 = j]$ $\quad 1 \leq j \leq N$
- **transition probabilities**
  $a_{ij} = P[q_t = j \mid q_{t-1} = i]$ $\quad 1 \leq i, j \leq N$

Elements of a Markov Model (chain):

- **the (potentially) occupied state at time *t* is called q_t**
- **a state can referred to by its index, e.g. $q_t = j$**
- **1 event corresponds to 1 state:**

  At each time *t*, the occupied state outputs ("emits") its corresponding event.
- **Markov model is *generator* of events.**
- **each event is discrete, has single output.**
- **in typical finite-state machine, actions occur at transitions, but in most Markov Models, actions occur at each state.**

Transition Probabilities:

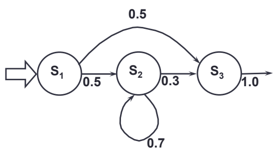
- **no assumptions (full probabilistic description of system):**
  $P[q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots, q_1 = m]$
- **usually use *first-order* Markov Model:**
  $P[q_t = j \mid q_{t-1} = i] = a_{ij}$
- **first-order assumption:**
  transition probabilities depend only on previous state (and time)
- **a_{ij} obeys usual rules:**

$$a_{ij} \geq 0 \quad \forall i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$$

- **sum of probabilities leaving a state = 1 (must leave a state)**

Markov Model Transition Probabilities

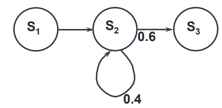


$a_{11} = 0.0$	$a_{12} = 0.5$	$a_{13} = 0.5$	$a_{1Exit}=0.0$	$\Sigma=1.0$
$a_{21} = 0.0$	$a_{22} = 0.7$	$a_{23} = 0.3$	$a_{2Exit}=0.0$	$\Sigma=1.0$
$a_{31} = 0.0$	$a_{32} = 0.0$	$a_{33} = 0.0$	$a_{3Exit}=1.0$	$\Sigma=1.0$

8

Markov Model Transition Probabilities

Probability distribution function:

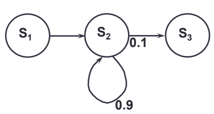


p(being in state S_2 exactly 1 time) = 0.6 = 0.600
p(being in state S_2 exactly 2 times) = 0.4 ·0.6 = 0.240
p(being in state S_2 exactly 3 times) = 0.4 ·0.4 ·0.6 = 0.096
p(being in state S_2 exactly 4 times) = 0.4 ·0.4 ·0.4 ·0.6 = 0.038

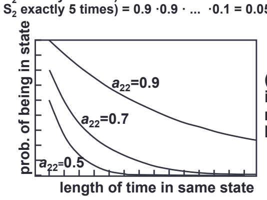
= exponential decay (characteristic of Markov Models)

9

Transition Probabilities



p(being in state S_2 exactly 1 time) = 0.1 = 0.100
p(being in state S_2 exactly 2 times) = 0.9 ·0.1 = 0.090
p(being in state S_2 exactly 3 times) = 0.9 ·0.9 ·0.1 = 0.081
p(being in state S_2 exactly 5 times) = 0.9 ·0.9 · ... ·0.1 = 0.059



(note: in graph, no multiplication by a_{23})

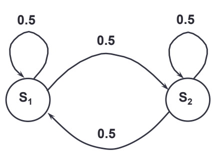
10

Initial Probabilities:

- **probabilities of starting in each state at time 1**
- **denoted by π_j**
- **$\pi_j = P[q_1 = j]$ $\quad 1 \le j \le N$**
- $\sum_{j=1}^{N} \pi_j = 1$

11

Example 1: Single Fair Coin

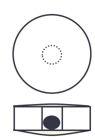


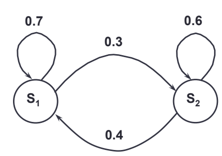
S_1 corresponds to e_1 = Heads	$a_{11} = 0.5$	$a_{12} = 0.5$
S_2 corresponds to e_2 = Tails	$a_{21} = 0.5$	$a_{22} = 0.5$

- **Generate events:**
  H T H H T H T T T H H
  corresponds to state sequence
  S_1 S_2 S_1 S_1 S_2 S_1 S_2 S_2 S_2 S_1 S_1

12

Example 2: Single *Biased* Coin (outcome depends on previous result)



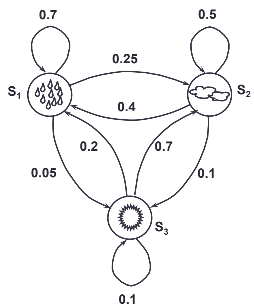


S_1 corresponds to e_1 = Heads	$a_{11} = 0.7$	$a_{12} = 0.3$
S_2 corresponds to e_2 = Tails	$a_{21} = 0.4$	$a_{22} = 0.6$

- **Generate events:**
  H H H T T T H H H T T H
  corresponds to state sequence
  S_1 S_1 S_1 S_2 S_2 S_2 S_1 S_1 S_1 S_2 S_2 S_1

13

Example 3: Portland Winter Weather



Example 3: Portland Winter Weather (con't)

- S_1 = event$_1$ = rain
  S_2 = event$_2$ = clouds
  S_3 = event$_3$ = sun

$$A = \{a_{ij}\} = \begin{pmatrix} .70 & .25 & .05 \\ .40 & .50 & .10 \\ .20 & .70 & .10 \end{pmatrix} \qquad \begin{matrix} \pi_1 = 0.5 \\ \pi_2 = 0.4 \\ \pi_3 = 0.1 \end{matrix}$$

- what is probability of {rain, rain, rain, clouds, sun, clouds, rain}?
  Obs. = {r, r, r, c, s, c, r}
  S = $\{S_1, S_1, S_1, S_2, S_3, S_2, S_1\}$
  time = {1, 2, 3, 4, 5, 6, 7} (days)

$= P[S_1]\ P[S_1|S_1]\ P[S_1|S_1]\ P[S_2|S_1]\ P[S_3|S_2]\ P[S_2|S_3]\ P[S_1|S_2]$

$= 0.5 \cdot 0.7 \cdot 0.7 \cdot 0.25 \cdot 0.1 \cdot 0.7 \cdot 0.4$

$= 0.001715$

Example 3: Portland Winter Weather (con't)

- S_1 = event$_1$ = rain
  S_2 = event$_2$ = clouds
  S_3 = event$_3$ = sunny

$$A = \{a_{ij}\} = \begin{pmatrix} .70 & .25 & .05 \\ .40 & .50 & .10 \\ .20 & .70 & .10 \end{pmatrix} \qquad \begin{matrix} \pi_1 = 0.5 \\ \pi_2 = 0.4 \\ \pi_3 = 0.1 \end{matrix}$$

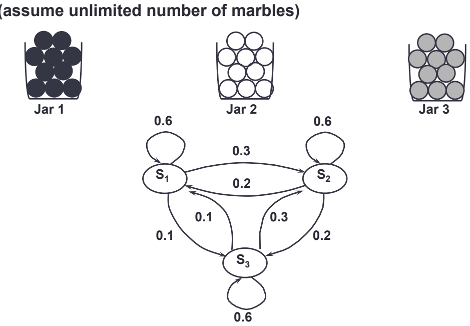
- what is probability of {sun, sun, sun, rain, clouds, sun, sun}?
  Obs. = {s, s, s, r, c, s, s}
  S = $\{S_3, S_3, S_3, S_1, S_2, S_3, S_3\}$
  time = {1, 2, 3, 4, 5, 6, 7} (days)

$= P[S_3]\ P[S_3|S_3]\ P[S_3|S_3]\ P[S_1|S_3]\ P[S_2|S_1]\ P[S_3|S_2]\ P[S_3|S_3]$

$= 0.1 \cdot 0.1 \cdot 0.1 \cdot 0.2 \cdot 0.25 \cdot 0.1 \cdot 0.1$

$= 5.0\times10^{-7}$

Example 4: Marbles in Jars (lazy person)
(assume unlimited number of marbles)



Example 4: Marbles in Jars (con't)

- S_1 = event$_1$ = black
  S_2 = event$_2$ = white
  S_3 = event$_3$ = grey

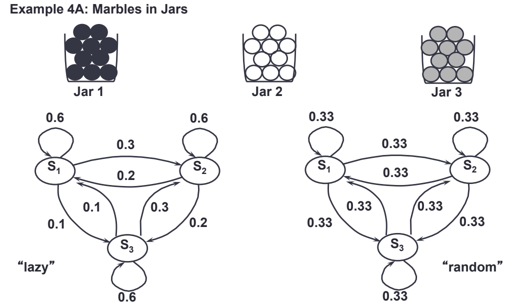
$$A = \{a_{ij}\} = \begin{pmatrix} .60 & .30 & .10 \\ .20 & .60 & .20 \\ .10 & .30 & .60 \end{pmatrix} \qquad \begin{matrix} \pi_1 = 0.33 \\ \pi_2 = 0.33 \\ \pi_3 = 0.33 \end{matrix}$$

- what is probability of {grey, white, white, black, black, grey}?
  Obs. = {g, w, w, b, b, g}
  S = $\{S_3, S_2, S_2, S_1, S_1, S_3\}$
  time = {1, 2, 3, 4, 5, 6}

$= P[S_3]\ P[S_2|S_3]\ P[S_2|S_2]\ P[S_1|S_2]\ P[S_1|S_1]\ P[S_3|S_1]$

$= 0.33 \cdot 0.3 \cdot 0.6 \cdot 0.2 \cdot 0.6 \cdot 0.1$
$= 0.0007128$

Example 4A: Marbles in Jars



- Same data, two different models...

Example 4A: Marbles in Jars

What is probability of:
{w, g, b, b, w}
given each model (“lazy” and “random”)?

S = $\{S_2, S_3, S_1, S_1, S_2\}$
time = {1, 2, 3, 4, 5}

“lazy”	“random”
$= P[S_2]\ P[S_3 \mid S_2]\ P[S_1 \mid S_3]\ P[S_1 \mid S_1]\ P[S_2 \mid S_1]$	$= P[S_2]\ P[S_3 \mid S_2]\ P[S_1 \mid S_3]\ P[S_1 \mid S_1]\ P[S_2 \mid S_1]$
$= 0.33 \cdot 0.2 \cdot 0.1 \cdot 0.6 \cdot 0.3$	$= 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.33$
= 0.001188	= 0.003913

{w, g, b, b, w} has greater probability if generated by “random.”
⇒“random” model more likely to generate {w, g, b, b, w}.

Notes:

- *Independence* is assumed between events that are separated by more than one time frame, when computing probability of sequence of events (for first-order model).
- Given list of observations, we can determine exact state sequence that generated those observations.
  ⇒ state sequence not *hidden.*
- Each state associated with only one event (output).
- Computing probability given a set of observations and a model is straightforward.
- Given multiple Markov Models and an observation sequence, it's easy to determine the M.M. most likely to have generated the data.

Agenda

- HW2 – due Thursday
- Questions, comments, concerns?
- Markov Chains
- Hidden Markov Models (HMMs)

Hidden Markov Models

- ***more than*** **1 event associated with each state.**
- all events have some ***probability*** of emitting at each state.
- given a sequence of observations, we can't determine exactly the state sequence.
- We ***can*** compute the ***probabilities*** of different state sequences given an observation sequence.

Doubly stochastic (probabilities of both emitting events and transitioning between states); exact state sequence is “hidden.”

Elements of a Hidden Markov Model:

• clock	$t = \{1, 2, 3, \ldots T\}$	
• N states	$Q = \{1, 2, 3, \ldots N\}$	
• M events	$E = \{e_1, e_2, e_3, \ldots, e_M\}$	
• initial probabilities	$\pi_j = P[q_1 = j]$	$1 \leq j \leq N$
• transition probabilities	$a_{ij} = P[q_t = j \mid q_{t-1} = i]$	$1 \leq i, j \leq N$
• observation probabilities	$b_j(k)=P[o_t = e_k \mid q_t = j]$ $b_j(o_t)=P[o_t = e_k \mid q_t = j]$	$1 \leq k \leq M$ $1 \leq k \leq M$

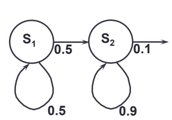
- A = matrix of a_{ij} values, B = set of observation probabilities, π = vector of π_j values.

Entire Model: $\lambda = (A,B,\pi)$

Notes:

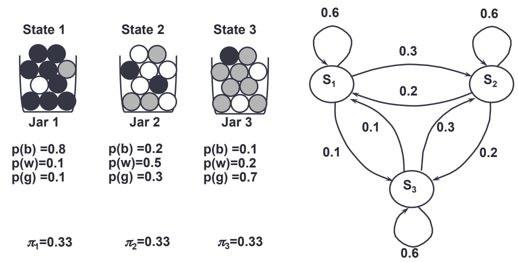
- an HMM still *generates* observations,
  each state is still discrete,
  observations can still come from a finite set (discrete HMMs).
- the number of items in the set of events does not have to be the same as the number of states.
- when in state *S*,
  there's $p(e_1)$ of generating event 1,
  there's $p(e_2)$ of generating event 2, etc.

p_{S1}(black) = 0.3
p_{S1}(white) = 0.7



p_{S2}(black) = 0.6
p_{S2}(white) = 0.4

Example 1: Marbles in Jars (lazy person)
(assume unlimited number of marbles)



27

Example 1: Marbles in Jars (lazy person)
(assume unlimited number of marbles)

- With the following observation:

g w w b b g

- What is probability of this observation, given state sequence $\{S_3\ S_2\ S_2\ S_1\ S_1\ S_3\}$ and the model??

$= b_3(g)\ b_2(w)\ b_2(w)\ b_1(b)\ b_1(b)\ b_3(g)$

$= 0.7 \cdot 0.5 \cdot 0.5 \cdot 0.8 \cdot 0.8 \cdot 0.7$

$= 0.0784$

28

Example 1: Marbles in Jars (lazy person)
(assume unlimited number of marbles)

- With the *same* observation:

g w w b b g

- What is probability of this observation, given state sequence $\{S_1\ S_1\ S_3\ S_2\ S_3\ S_1\}$ and the model??

$= b_1(g)\ b_1(w)\ b_3(w)\ b_2(b)\ b_3(b)\ b_1(g)$

$= 0.1 \cdot 0.1 \cdot 0.2 \cdot 0.2 \cdot 0.1 \cdot 0.1$

$= 4.0\text{x}10^{-6}$

29

Some math…

With an observation sequence $O=(o_1\ o_2\ \ldots\ o_T)$, state sequence $q=(q_1\ q_2\ \ldots\ q_T)$, and model λ:

Probability of O, given state sequence q and model λ, is:

$$P(\mathbf{O}\,|\,\mathbf{q},\lambda) = \prod_{t=1}^{T} P(\mathbf{o}_t\,|\,q_t,\lambda)$$

assuming independence between observations. This expands:

$$P(\mathbf{O}\,|\,\mathbf{q},\lambda) = p(\mathbf{o}_1\,|\,q_1)\cdot p(\mathbf{o}_2\,|\,q_2)\ldots\cdot p(\mathbf{o}_T\,|\,q_T)$$

-- or --

$$P(\mathbf{O}\,|\,\mathbf{q},\lambda) = b_{q_1}(\mathbf{o}_1)\cdot b_{q_2}(\mathbf{o}_2)\ldots b_{q_T}(\mathbf{o}_T)$$

The probability of the state sequence q can be written:

$$P(\mathbf{q}\,|\,\lambda) = \pi_{q_1}\cdot a_{q_1q_2}\cdot a_{q_2q_3}\ldots a_{q_{T-1}q_T}$$

30

The probability of both O and q occurring simultaneously is:

$$P(\mathbf{O},\mathbf{q}\,|\,\lambda) = P(\mathbf{O}\,|\,\mathbf{q},\lambda)\ P(\mathbf{q}\,|\,\lambda)$$

which can be expanded to:

$$P(\mathbf{O},\mathbf{q}\,|\,\lambda) = \pi_{q1}\cdot b_{q_1}(\mathbf{o}_1)\cdot a_{q_1q_2}\cdot b_{q_2}(\mathbf{o}_2)\cdot a_{q_2q_3}\ldots a_{q_{T-1}q_T}\cdot b_{q_T}(\mathbf{o}_T)$$

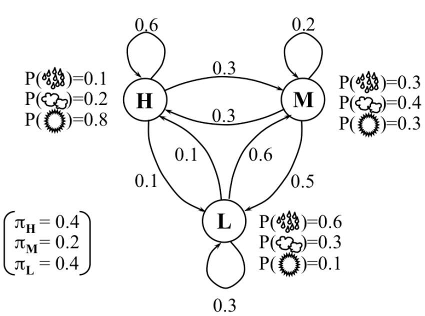
Independence between a_{ij} and $b_j(o_t)$ is NOT assumed:

$$P(\mathbf{O},\mathbf{q}\,|\,\lambda) = P(\mathbf{O}\,|\,\mathbf{q},\lambda)\ P(\mathbf{q}\,|\,\lambda)$$

this is just multiplication rule: P(A∩B) = P(A | B) P(B)

31

HMM Example 2: Weather and Atmospheric Pressure



32

HMM Example 2: Weather and Atmospheric Pressure

If weather observation O={sun, sun, cloud, rain, cloud, sun} what is probability of O, given the model and the sequence {H, M, M, L, L, M}?

$= b_H(\text{sun})\ b_M(\text{sun})\ b_M(\text{cloud})\ b_L(\text{rain})\ b_L(\text{cloud})\ b_M(\text{sun})$

$= 0.8 \cdot 0.3 \cdot 0.4 \cdot 0.6 \cdot 0.3 \cdot 0.3$

$= 5.2\text{x}10^{-3}$

HMM Example 2: Weather and Atmospheric Pressure

What is probability of O={sun, sun, cloud, rain, cloud, sun} and the sequence {H, M, M, L, L, M}, given the model?

$= \pi_H \cdot b_H(s) \cdot a_{HM} \cdot b_M(s) \cdot a_{MM} \cdot b_M(c) \cdot a_{ML} \cdot b_L(r) \cdot a_{LL} \cdot b_L(c) \cdot a_{LM} \cdot b_M(s)$

$= 0.4 \cdot 0.8 \cdot 0.3 \cdot 0.3 \cdot 0.2 \cdot 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.3 \cdot 0.3 \cdot 0.6 \cdot 0.3$

$= 1.12\text{x}10^{-5}$

What is probability of O={sun, sun, cloud, rain, cloud, sun} and the sequence {H, H, M, L, M, H}, given the model?

$= \pi_H \cdot b_H(s) \cdot a_{HH} \cdot b_H(s) \cdot a_{HM} \cdot b_M(c) \cdot a_{ML} \cdot b_L(r) \cdot a_{LM} \cdot b_M(c) \cdot a_{MH} \cdot b_H(s)$

$= 0.4 \cdot 0.8 \cdot 0.6 \cdot 0.8 \cdot 0.3 \cdot 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.3 \cdot 0.6$

$= 2.39\text{x}10^{-4}$

Notes about HMMs:

- **must know all possible states in advance**
- **must know possible state connections in advance**
- **cannot recognize things outside of model**
- **must have some estimate of state emission probabilities and state transition probabilities**
- **make several assumptions (usually so math is easier)**

Agenda

- HW2 – due Thursday
- Markov Chains
- Hidden Markov Models (HMMs)
- Forward Algorithm, Viterbi Algorithm (next time)