

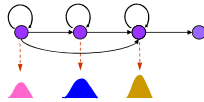
Hidden Markov Models

Class 15. 12 Oct 2010

Administrivia

- HW2 – due Tuesday
- Is everyone on the “projects” page?
 - Where are your project proposals?

Recap: What is an HMM



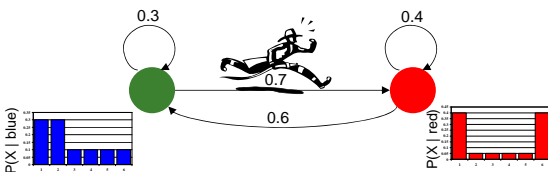
- “Probabilistic function of a markov chain”
- Models a dynamical system
- System goes through a number of states
 - Following a Markov chain model
- On arriving at any state it generates observations according to a state-specific probability distribution

A Thought Experiment



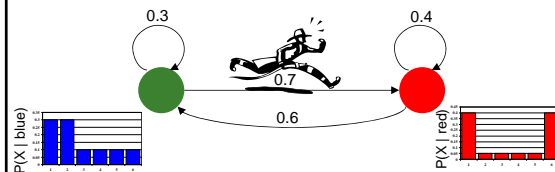
- Two “shooters” roll dice
- A caller calls out the number rolled. We only get to hear what he calls out
- The caller behaves randomly
 - If he has just called a number rolled by the blue shooter, his next call is that of the red shooter 70% of the time
 - But if he has just called the red shooter, he has only a 40% probability of calling the red shooter again in the next call
- How do we characterize this?

A Thought Experiment



- The dots and arrows represent the “states” of the caller
 - When he’s on the blue circle he calls out the blue dice
 - When he’s on the red circle he calls out the red dice
 - The histograms represent the probability distribution of the numbers for the blue and red dice

A Thought Experiment



- When the caller is in any state, he calls a number based on the probability distribution of that state
 - We call these state output distributions
- At each step, he moves from his current state to another state following a probability distribution
 - We call these transition probabilities
- The caller is an HMM!!!

What is an HMM

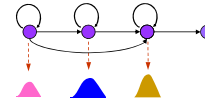
- HMMs are statistical models for (causal) processes
- The model assumes that the process can be in one of a number of states at any time instant
- The state of the process at any time instant depends only on the state at the previous instant (causality, Markovian)
- At each instant the process generates an observation from a probability distribution that is specific to the current state
- The generated observations are all that we get to see
 - the actual state of the process is not directly observable
 - Hence the qualifier hidden

12 Oct 2010

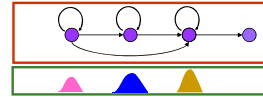
11755/18797

7

Hidden Markov Models



- A Hidden Markov Model consists of two components
 - A state/transition backbone that specifies how many states there are, and how they can follow one another
 - A set of probability distributions, one for each state, which specifies the distribution of all vectors in that state



Markov chain

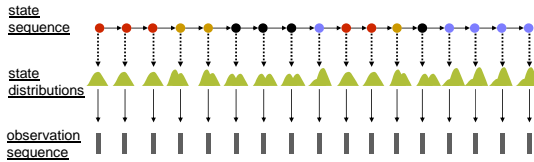
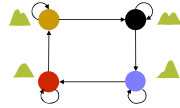
Data distributions

- This can be factored into two separate probabilistic entities
 - A probabilistic Markov chain with states and transitions
 - A set of data probability distributions, associated with the states

11755/18797

How an HMM models a process

HMM assumed to be generating data



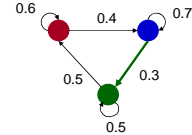
12 Oct 2010

11755/18797

9

HMM Parameters

- The *topology* of the HMM
 - Number of states and allowed transitions
 - E.g. here we have 3 states and cannot go from the blue state to the red
- The transition probabilities
 - Often represented as a matrix as here
 - T_{ij} is the probability that when in state i , the process will move to j
- The probability π_i of beginning at any state s_i
 - The complete set is represented as π
- The *state output distributions*



$$T = \begin{pmatrix} .6 & .4 & 0 \\ 0 & .7 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$



12 Oct 2010

10

HMM state output distributions

- The state output distribution is the distribution of data produced from any state
- Typically modelled as Gaussian

$$P(x | s_i) = \text{Gaussian}(x; \mu_i, \Theta_i) = \frac{1}{\sqrt{(2\pi)^d |\Theta_i|}} e^{-0.5(x-\mu_i)^T \Theta_i^{-1} (x-\mu_i)}$$

- The parameters are μ_i and Θ_i
- More typically, modelled as Gaussian mixtures

$$P(x | s_i) = \sum_{j=0}^{K-1} w_{i,j} \text{Gaussian}(x; \mu_{i,j}, \Theta_{i,j})$$

- Other distributions may also be used
- E.g. histograms in the dice case

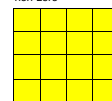
12 Oct 2010

11755/18797

11

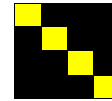
The Diagonal Covariance Matrix

Full covariance: all elements are non-zero



$$-0.5(x-\mu)^T \Theta^{-1} (x-\mu)$$

Diagonal covariance: off-diagonal elements are zero



$$-\Sigma_i (x-\mu_i)^2 / 2\sigma_i^2$$

- For GMMs it is frequently assumed that the feature vector dimensions are all *independent* of each other
- *Result.* The covariance matrix is reduced to a diagonal form
 - The determinant of the diagonal Θ matrix is easy to compute

12 Oct 2010

11755/18797

12

Three Basic HMM Problems

- What is the probability that it will generate a specific observation sequence
- Given a observation sequence, how do we determine which observation was generated from which state
 - The state segmentation problem
- How do we *learn* the parameters of the HMM from observation sequences

12 Oct 2010

11755/18797

13

Computing the Probability of an Observation Sequence

- Two aspects to producing the observation:
 - Progressing through a sequence of states
 - Producing observations from these states

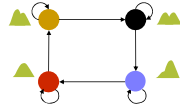
12 Oct 2010

11755/18797

14

Progressing through states

HMM assumed to be generating data



state sequence



- The process begins at some state (red) here
- From that state, it makes an allowed transition
 - To arrive at the same or any other state
- From that state it makes another allowed transition
 - And so on

12 Oct 2010

11755/18797

15

Probability that the HMM will follow a particular state sequence

$$P(s_1, s_2, s_3, \dots) = P(s_1)P(s_2|s_1)P(s_3|s_2)\dots$$

- $P(s_1)$ is the probability that the process will initially be in state s_1
- $P(s_i|s_j)$ is the transition probability of moving to state s_i at the next time instant when the system is currently in s_j
 - Also denoted by T_{ij} earlier

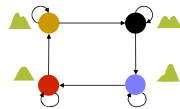
12 Oct 2010

11755/18797

16

Generating Observations from States

HMM assumed to be generating data



state sequence



state distributions



observation sequence



- At each time it generates an observation from the state it is in at that time

12 Oct 2010

11755/18797

17

Probability that the HMM will generate a particular observation sequence given a state sequence (state sequence known)

$$P(o_1, o_2, o_3, \dots | s_1, s_2, s_3, \dots) = P(o_1|s_1)P(o_2|s_2)P(o_3|s_3)\dots$$

Computed from the Gaussian or Gaussian mixture for state s_i

- $P(o_i|s_i)$ is the probability of generating observation o_i when the system is in state s_i

12 Oct 2010

11755/18797

18

Proceeding through States and Producing Observations

HMM assumed to be generating data

state sequence

state distributions

observation sequence

- At each time it produces an observation and makes a transition

12 Oct 2010 11755/18797 19

Probability that the HMM will generate a particular state sequence and from it, a particular observation sequence

$$P(o_1, o_2, o_3, \dots, s_1, s_2, s_3, \dots) =$$

$$P(o_1, o_2, o_3, \dots | s_1, s_2, s_3, \dots) P(s_1, s_2, s_3, \dots) =$$

$$P(o_1 | s_1) P(o_2 | s_2) P(o_3 | s_3) \dots P(s_1) P(s_2 | s_1) P(s_3 | s_2) \dots$$

12 Oct 2010 11755/18797 20

Probability of Generating an Observation Sequence

- The precise state sequence is not known
- All possible state sequences must be considered

$$P(o_1, o_2, o_3, \dots) = \sum_{\text{all possible state sequences}} P(o_1, o_2, o_3, \dots, s_1, s_2, s_3, \dots) =$$

$$\sum_{\text{all possible state sequences}} P(o_1 | s_1) P(o_2 | s_2) P(o_3 | s_3) \dots P(s_1) P(s_2 | s_1) P(s_3 | s_2) \dots$$

12 Oct 2010 11755/18797 21

Computing it Efficiently

- Explicit summing over all state sequences is not tractable
 - A very large number of possible state sequences
- Instead we use the forward algorithm
- A dynamic programming technique.

12 Oct 2010 11755/18797 22

Illustrative Example

- Example: a generic HMM with 5 states and a "terminating state".
 - Left to right topology
 - $P(s) = 1$ for state 1 and 0 for others
 - The arrows represent transition for which the probability is not 0
- Notation:
 - $P(s_j | s_i) = T_{ij}$
 - We represent $P(o_i | s_i) = b_i(t)$ for brevity

12 Oct 2010 11755/18797 23

Diversions: The Trellis

- The trellis is a graphical representation of all possible paths through the HMM to produce a given observation
- The Y-axis represents HMM states, X axis represents observations
- Every edge in the graph represents a valid transition in the HMM over a single time step
- Every node represents the event of a particular observation being generated from a particular state

12 Oct 2010 11755/18797 24

The Forward Algorithm

$$\alpha(s, t) = P(x_1, x_2, \dots, x_t, \text{state}(t) = s)$$

- $\alpha(s, t)$ is the total probability of ALL state sequences that end at state s at time t , and all observations until x_t

12 Oct 2010 11755/18797 25

The Forward Algorithm

$$\alpha(s, t) = P(x_1, x_2, \dots, x_t, \text{state}(t) = s)$$

$$\alpha(s, t) = \sum_{s'} \alpha(s', t-1) P(s | s') P(x_t | s)$$

- $\alpha(s, t)$ can be recursively computed in terms of $\alpha(s', t)$, the forward probabilities at time $t-1$

12 Oct 2010 11755/18797 26

The Forward Algorithm

$$\text{Totalprob} = \sum_s \alpha(s, T)$$

- In the final observation the alpha at each state gives the probability of all state sequences ending at that state
- General model: The total probability of the observation is the sum of the alpha values at all states

12 Oct 2010 11755/18797 27

The absorbing state

- Observation sequences are assumed to end only when the process arrives at an absorbing state
 - No observations are produced from the absorbing state

12 Oct 2010 11755/18797 28

The Forward Algorithm

$$\text{Totalprob} = \alpha(s_{\text{absorbing}}, T+1)$$

$$\alpha(s_{\text{absorbing}}, T+1) = \sum_{s'} \alpha(s', T) P(s_{\text{absorbing}} | s')$$

- Absorbing state model: The total probability is the alpha computed at the absorbing state after the final observation

12 Oct 2010 11755/18797 29

Problem 2: State segmentation

- Given only a sequence of observations, how do we determine which sequence of states was followed in producing it?

12 Oct 2010 11755/18797 30

The HMM as a generator

HMM assumed to be generating data

state sequence

state distributions

observation sequence

- The process goes through a series of states and produces observations from them

12 Oct 2010 11755/18797 31

States are hidden

HMM assumed to be generating data

state sequence

state distributions

observation sequence

- The observations do not reveal the underlying state

12 Oct 2010 11755/18797 32

The state segmentation problem

HMM assumed to be generating data

state sequence

state distributions

observation sequence

- State segmentation: Estimate state sequence given observations

12 Oct 2010 11755/18797 33

Estimating the State Sequence

- Many different state sequences are capable of producing the observation
- Solution: Identify the most *probable* state sequence
 - The state sequence for which the probability of progressing through that sequence and generating the observation sequence is maximum
 - i.e. $P(o_1, o_2, o_3, \dots, s_1, s_2, s_3, \dots)$ is maximum

12 Oct 2010 11755/18797 34

Estimating the state sequence

- Once again, exhaustive evaluation is impossibly expensive
- But once again a simple dynamic-programming solution is available

$$P(o_1, o_2, o_3, \dots, s_1, s_2, s_3, \dots) = P(o_1 | s_1) P(o_2 | s_2) P(o_3 | s_3) \dots P(s_1) P(s_2 | s_1) P(s_3 | s_2) \dots$$

- Needed:

$$\arg \max_{s_1, s_2, s_3, \dots} P(o_1 | s_1) P(s_1) P(o_2 | s_2) P(s_2 | s_1) P(o_3 | s_3) P(s_3 | s_2)$$

12 Oct 2010 11755/18797 35

Estimating the state sequence

- Once again, exhaustive evaluation is impossibly expensive
- But once again a simple dynamic-programming solution is available

$$P(o_1, o_2, o_3, \dots, s_1, s_2, s_3, \dots) = P(o_1 | s_1) P(o_2 | s_2) P(o_3 | s_3) \dots P(s_1) P(s_2 | s_1) P(s_3 | s_2) \dots$$

- Needed:

$$\arg \max_{s_1, s_2, s_3, \dots} P(o_1 | s_1) P(s_1) P(o_2 | s_2) P(s_2 | s_1) P(o_3 | s_3) P(s_3 | s_2)$$

12 Oct 2010 11755/18797 36

The HMM as a generator

HMM assumed to be generating data

state sequence

state distributions

observation sequence

- Each enclosed term represents one forward transition and a subsequent emission

12 Oct 2010 11755/18797 37

The state sequence

- The probability of a state sequence $?, ?, ?, ?, s_x, s_y$ ending at time t , and producing all observations until o_t
 - $P(o_{1:t-1}, ?, ?, ?, ?, s_x, o_t, s_y) = P(o_{1:t-1}, ?, ?, ?, ?, s_x) P(o_t | s_y) P(s_y | s_x)$
- The *best* state sequence that ends with s_x, s_y at t will have a probability equal to the probability of the best state sequence ending at $t-1$ at s_x times $P(o_t | s_y) P(s_y | s_x)$

12 Oct 2010 11755/18797 38

Extending the state sequence

state sequence

state distributions

observation sequence

- The probability of a state sequence $?, ?, ?, ?, s_x, s_y$ ending at time t and producing observations until o_t
 - $P(o_{1:t-1}, o_t, ?, ?, ?, s_x, s_y) = P(o_{1:t-1}, ?, ?, ?, ?, s_x) P(o_t | s_y) P(s_y | s_x)$

12 Oct 2010 11755/18797 39

Trellis

- The graph below shows the set of all possible state sequences through this HMM in five time instants

12 Oct 2010 11755/18797 40

The cost of extending a state sequence

- The cost of *extending* a state sequence ending at s_x is only dependent on the transition from s_x to s_y , and the observation probability at s_y

$P(o_t | s_y) P(s_y | s_x)$

s_y

s_x

time

12 Oct 2010 11755/18797 41

The cost of extending a state sequence

- The best path to s_y through s_x is simply an extension of the best path to s_x

Best $P(o_{1:t-1}, ?, ?, ?, ?, s_x) P(o_t | s_y) P(s_y | s_x)$

s_y

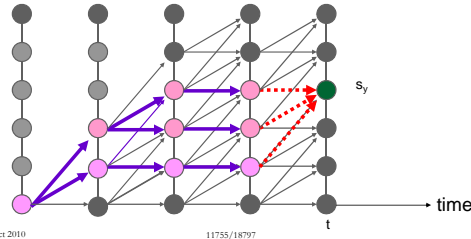
s_x

time

12 Oct 2010 11755/18797 42

The Recursion

- The overall best path to s_y is an extension of the best path to one of the states at the previous time



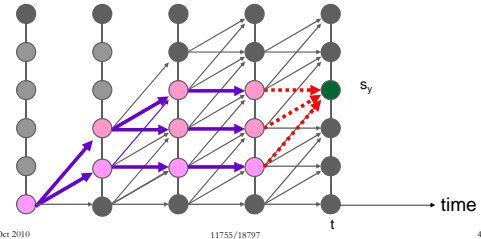
12 Oct 2010

11755/18797

43

The Recursion

- Prob. of best path to $s_y = \text{Max}_{s_x} \text{BestP}(o_{1..t-1}, ?, ?, ?, s_x) P(o_t | s_y) P(s_y | s_x)$



12 Oct 2010

11755/18797

44

Finding the best state sequence

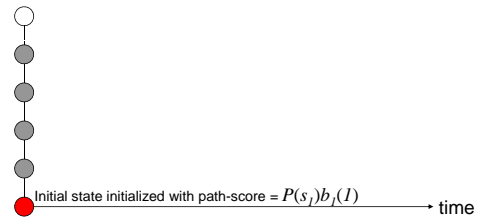
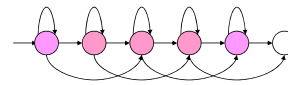
- The simple algorithm just presented is called the VITERBI algorithm in the literature
 - After A.J.Viterbi, who invented this dynamic programming algorithm for a completely different purpose: decoding error correction codes!

12 Oct 2010

11755/18797

45

Viterbi Search (contd.)

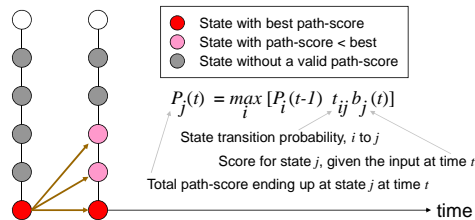
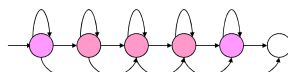


12 Oct 2010

11755/18797

46

Viterbi Search (contd.)

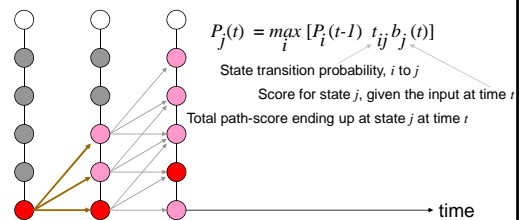
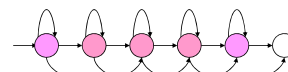


12 Oct 2010

11755/18797

47

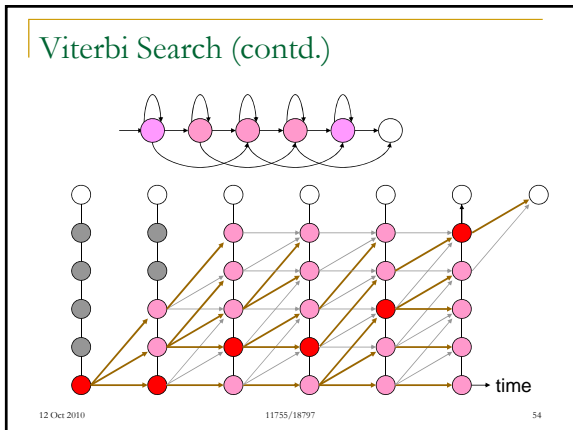
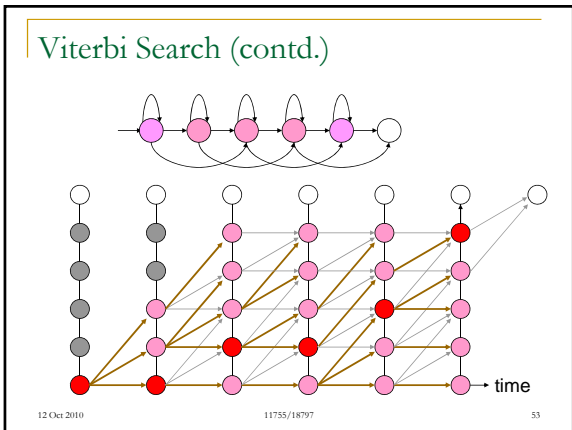
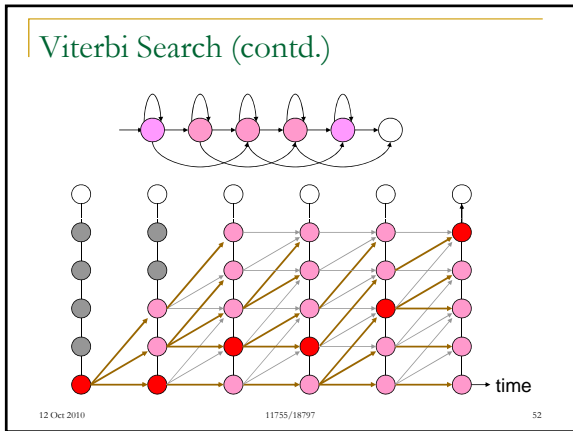
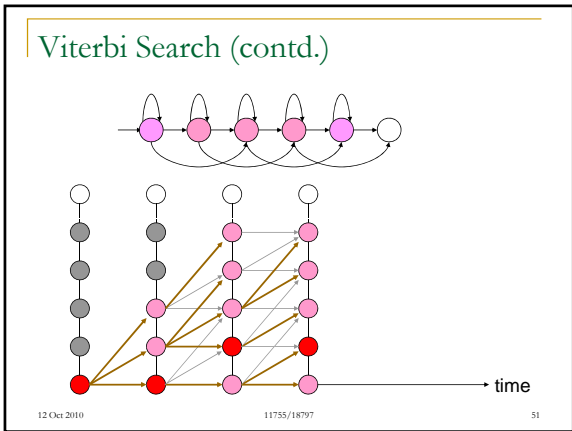
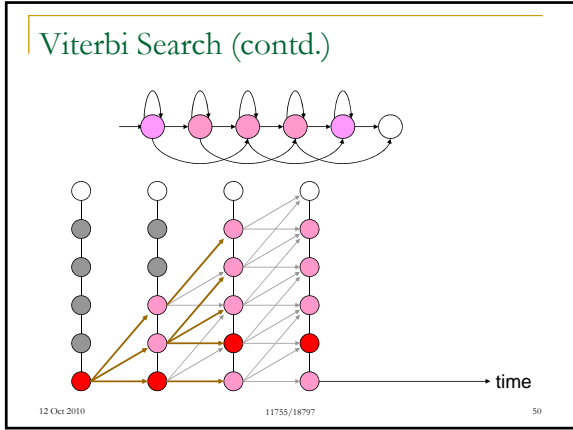
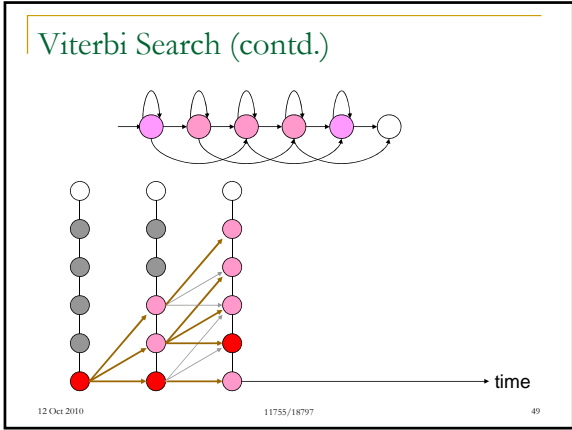
Viterbi Search (contd.)

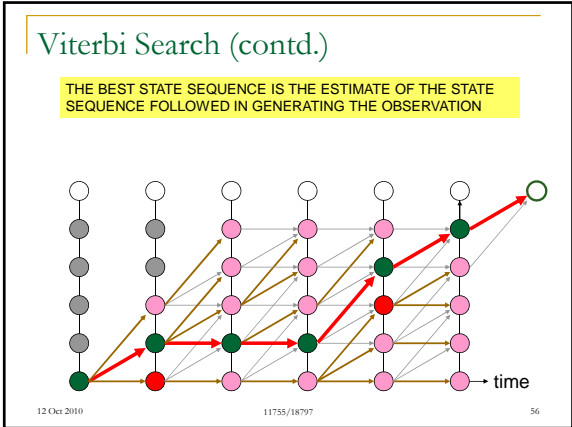
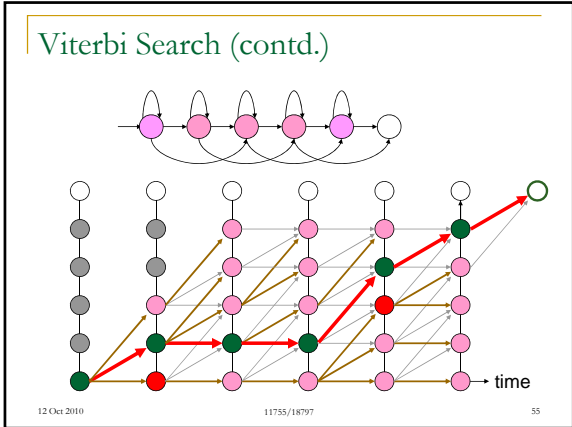


12 Oct 2010

11755/18797

48





- ### Problem3: Training HMM parameters
- We can compute the probability of an observation, and the best state sequence given an observation, using the HMM's parameters
 - But where do the HMM parameters come from?
 - They must be learned from a collection of observation sequences
- 12 Oct 2010 11755/18797 57

- ### Learning HMM parameters: Simple procedure – counting
- Given a set of training instances
 - Iteratively:
 1. Initialize HMM parameters
 2. Segment all training instances
 3. Estimate transition probabilities and state output probability parameters by counting
- 12 Oct 2010 11755/18797 58

- ### Learning by counting example
- Explanation by example in next few slides
 - 2-state HMM, Gaussian PDF at states, 3 observation sequences
 - Example shows ONE iteration
 - How to count after state sequences are obtained
-
- 12 Oct 2010 11755/18797 59

Example: Learning HMM Parameters

- We have an HMM with two states s_1 and s_2 .
- Observations are vectors x_{ij}
 - i -th sequence, j -th vector
- We are given the following three observation sequences
 - And have already estimated state sequences

Time	1	2	3	4	5	6	7	8	9	10
state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
Obs	x_{c1}	x_{c2}	x_{c3}	x_{c4}	x_{c5}	x_{c6}	x_{c7}	x_{c8}	x_{c9}	x_{c10}

Observation 1

Time	1	2	3	4	5	6	7	8	9
state	S2	S2	S1	S1	S2	S2	S2	S2	S1
Obs	x_{c1}	x_{c2}	x_{c3}	x_{c4}	x_{c5}	x_{c6}	x_{c7}	x_{c8}	x_{c9}

Observation 2

Time	1	2	3	4	5	6	7	8
state	S1	S2	S1	S1	S1	S2	S2	S2
Obs	x_{c1}	x_{c2}	x_{c3}	x_{c4}	x_{c5}	x_{c6}	x_{c7}	x_{c8}

Observation 3

12 Oct 2010 11755/18797 60

Example: Learning HMM Parameters

- Initial state probabilities (usually denoted as π):
 - We have 3 observations
 - 2 of these begin with S1, and one with S2
 - $\pi(S1) = 2/3$, $\pi(S2) = 1/3$

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S2	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		

12 Oct 2010

11755/18797

61

Example: Learning HMM Parameters

- Transition probabilities:
 - State S1 occurs 11 times in non-terminal locations

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		

12 Oct 2010

11755/18797

62

Example: Learning HMM Parameters

- Transition probabilities:
 - State S1 occurs 11 times in non-terminal locations
 - Of these, it is followed immediately by S1 6 times

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		

12 Oct 2010

11755/18797

63

Example: Learning HMM Parameters

- Transition probabilities:
 - State S1 occurs 11 times in non-terminal locations
 - Of these, it is followed immediately by S1 6 times
 - It is followed immediately by S2 5 times

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		

12 Oct 2010

11755/18797

64

Example: Learning HMM Parameters

- Transition probabilities:
 - State S1 occurs 11 times in non-terminal locations
 - Of these, it is followed immediately by S1 6 times
 - It is followed immediately by S2 5 times
 - $P(S1 | S1) = 6/11$; $P(S2 | S1) = 5/11$

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		

12 Oct 2010

11755/18797

65

Example: Learning HMM Parameters

- Transition probabilities:
 - State S2 occurs 13 times in non-terminal locations

Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Observation 2	Time	1	2	3	4	5	6	7	8	9	
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1	
	Obs	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	
Observation 3	Time	1	2	3	4	5	6	7	8		
	state	S1	S2	S1	S1	S1	S2	S2	S2		
	Obs	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}		


12 Oct 2010

11755/18797

66

Example: Learning HMM Parameters

- Transition probabilities:
 - State S2 occurs 13 times in non-terminal locations
 - Of these, it is followed immediately by S1 5 times



Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s1}	X _{s1}


Observation 2	Time	1	2	3	4	5	6	7	8	9
	state	S2	S1	S1	S1	S2	S2	S2	S2	S1
	Obs	X _{s2}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s2}	X _{s1}

Observation 3	Time	1	2	3	4	5	6	7	8
	state	S1	S1	S1	S1	S1	S2	S2	S2
	Obs	X _{s1}	X _{s1}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}

12 Oct 2010 11755/18797 67

Example: Learning HMM Parameters

- Transition probabilities:
 - State S2 occurs 13 times in non-terminal locations
 - Of these, it is followed immediately by S1 5 times
 - It is followed immediately by S2 8 times



Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S1	S2	S2	S1	S1	S2	S1	S1
	Obs	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s1}	X _{s1}


Observation 2	Time	1	2	3	4	5	6	7	8	9
	state	S2	S1	S1	S1	S2	S2	S2	S2	S1
	Obs	X _{s2}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s2}	X _{s1}

Observation 3	Time	1	2	3	4	5	6	7	8
	state	S1	S2	S1	S1	S1	S2	S2	S2
	Obs	X _{s1}	X _{s2}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}

12 Oct 2010 11755/18797 68

Example: Learning HMM Parameters

- Transition probabilities:
 - State S2 occurs 13 times in non-terminal locations
 - Of these, it is followed immediately by S1 5 times
 - It is followed immediately by S2 8 times
 - $P(S1 | S2) = 5 / 13$; $P(S2 | S2) = 8 / 13$



Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s1}	X _{s1}

Observation 2	Time	1	2	3	4	5	6	7	8	9
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1
	Obs	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s2}	X _{s1}

Observation 3	Time	1	2	3	4	5	6	7	8
	state	S1	S2	S1	S1	S1	S2	S2	S2
	Obs	X _{s1}	X _{s2}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}

12 Oct 2010 11755/18797 69

Parameters learnt so far

- State initial probabilities, often denoted as π
 - $\pi(S1) = 2/3 = 0.66$
 - $\pi(S2) = 1/3 = 0.33$
- State transition probabilities
 - $P(S1 | S1) = 6/11 = 0.545$; $P(S2 | S1) = 5/11 = 0.455$
 - $P(S1 | S2) = 5/13 = 0.385$; $P(S2 | S2) = 8/13 = 0.615$
 - Represented as a transition matrix


$$A = \begin{pmatrix} P(S1|S1) & P(S2|S1) \\ P(S1|S2) & P(S2|S2) \end{pmatrix} = \begin{pmatrix} 0.545 & 0.455 \\ 0.385 & 0.615 \end{pmatrix}$$

Each row of this matrix must sum to 1.0

12 Oct 2010 11755/18797 70

Example: Learning HMM Parameters

- State output probability for S1
 - There are 13 observations in S1



Observation 1	Time	1	2	3	4	5	6	7	8	9	10
	state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
	Obs	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s1}	X _{s1}


Observation 2	Time	1	2	3	4	5	6	7	8	9
	state	S2	S2	S1	S1	S2	S2	S2	S2	S1
	Obs	X _{s2}	X _{s2}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}	X _{s2}	X _{s1}

Observation 3	Time	1	2	3	4	5	6	7	8
	state	S1	S2	S1	S1	S1	S2	S2	S2
	Obs	X _{s1}	X _{s2}	X _{s1}	X _{s1}	X _{s1}	X _{s2}	X _{s2}	X _{s2}

12 Oct 2010 11755/18797 71

Example: Learning HMM Parameters

- State output probability for S1
 - There are 13 observations in S1
 - Segregate them out and count
 - Compute parameters (mean and variance) of Gaussian output density for state S1



Time	1	2	6	7	9	10
state	S1	S1	S1	S1	S1	S1
Obs	X _{s1}	X _{s1}	X _{s1}	X _{s1}	X _{s1}	X _{s1}

$$P(X | S_1) = \frac{1}{\sqrt{(2\pi)^d} |\Theta_1|} \exp(-0.5(X - \mu)^T \Theta_1^{-1} (X - \mu))$$

Time	3	4	9
state	S1	S1	S1
Obs	X _{s1}	X _{s1}	X _{s1}

$$\mu_1 = \frac{1}{13} (X_{s1} + X_{s2} + X_{s6} + X_{s7} + X_{s9} + X_{s10} + X_{s13} + X_{s4} + X_{s9} + X_{s1} + X_{s2} + X_{s4} + X_{s5})$$

Time	1	3	4	5
state	S1	S1	S1	S1
Obs	X _{s1}	X _{s1}	X _{s1}	X _{s1}

$$\Theta_1 = \frac{1}{13} \begin{pmatrix} (X_{s1} - \mu_1)(X_{s1} - \mu_1)^T + (X_{s2} - \mu_1)(X_{s2} - \mu_1)^T + \dots \\ (X_{s3} - \mu_1)(X_{s3} - \mu_1)^T + (X_{s4} - \mu_1)(X_{s4} - \mu_1)^T + \dots \\ (X_{s11} - \mu_1)(X_{s11} - \mu_1)^T + (X_{s12} - \mu_1)(X_{s12} - \mu_1)^T + \dots \end{pmatrix}$$

12 Oct 2010 11755/18797 72

Example: Learning HMM Parameters

- State output probability for S2
 - There are 14 observations in S2



Observation 1

Time	1	2	3	4	5	6	7	8	9	10
state	S1	S1	S2	S2	S2	S1	S1	S2	S1	S1
Obs	X_{s1}	X_{s1}	X_{s2}	X_{s2}	X_{s2}	X_{s1}	X_{s1}	X_{s2}	X_{s1}	X_{s1}

Observation 2

Time	1	2	3	4	5	6	7	8	9
state	S2	S2	S1	S1	S2	S2	S2	S2	S1
Obs	X_{s2}	X_{s2}	X_{s1}	X_{s1}	X_{s2}	X_{s2}	X_{s2}	X_{s2}	X_{s1}

Observation 3

Time	1	2	3	4	5	6	7	8
state	S1	S2	S1	S1	S2	S2	S2	S1
Obs	X_{s1}	X_{s2}	X_{s1}	X_{s1}	X_{s2}	X_{s2}	X_{s2}	X_{s1}

12 Oct 2010

11755/18797

73

Example: Learning HMM Parameters

- State output probability for S2
 - There are 14 observations in S2
 - Segregate them out and count
 - Compute parameters (mean and variance) of Gaussian output density for state S2

Time	3	4	5	8
state	S2	S2	S2	S2
Obs	X_{s2}	X_{s2}	X_{s2}	X_{s2}

$$P(X | S_2) = \frac{1}{\sqrt{(2\pi)^d |\Theta_2|}} \exp(-0.5(X - \mu_2)^T \Theta_2^{-1} (X - \mu_2))$$

Time	1	2	5	6	7	8
state	S2	S2	S2	S2	S2	S2
Obs	X_{s2}	X_{s2}	X_{s2}	X_{s2}	X_{s2}	X_{s2}

$$\mu_2 = \frac{1}{14} (X_{s2,1} + X_{s2,2} + X_{s2,5} + X_{s2,6} + X_{s2,7} + X_{s2,8} + X_{s2,3} + X_{s2,4} + X_{s2,9} + X_{s2,10} + X_{s2,11} + X_{s2,12} + X_{s2,13} + X_{s2,14})$$

Time	2	6	7	8
state	S2	S2	S2	S2
Obs	X_{s2}	X_{s2}	X_{s2}	X_{s2}

$$\Theta_2 = \frac{1}{14} ((X_{s2,1} - \mu_2)(X_{s2,1} - \mu_2)^T + \dots)$$

12 Oct 2010

11755/18797

74

We have learnt all the HMM parameters

- State initial probabilities, often denoted as π
 - $\pi(S1) = 0.66$ $\pi(S2) = 1/3 = 0.33$
- State transition probabilities

$$A = \begin{pmatrix} 0.545 & 0.455 \\ 0.385 & 0.615 \end{pmatrix}$$

- State output probabilities

State output probability for S1

State output probability for S2

$$P(X | S_1) = \frac{1}{\sqrt{(2\pi)^d |\Theta_1|}} \exp(-0.5(X - \mu_1)^T \Theta_1^{-1} (X - \mu_1))$$

$$P(X | S_2) = \frac{1}{\sqrt{(2\pi)^d |\Theta_2|}} \exp(-0.5(X - \mu_2)^T \Theta_2^{-1} (X - \mu_2))$$

12 Oct 2010

11755/18797

75

Update rules at each iteration

$$\pi(s_i) = \frac{\text{No. of observation sequences that start at state } s_i}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{\text{obs } t: \text{state}(t)=s_i, \text{state}(t+1)=s_j} 1}{\sum_{\text{obs } t: \text{state}(t)=s_i} 1}$$

$$\mu_i = \frac{\sum_{\text{obs } t: \text{state}(t)=s_i} X_{\text{obs},t}}{\sum_{\text{obs } t: \text{state}(t)=s_i} 1}$$

$$\Theta_i = \frac{\sum_{\text{obs } t: \text{state}(t)=s_i} (X_{\text{obs},t} - \mu_i)(X_{\text{obs},t} - \mu_i)^T}{\sum_{\text{obs } t: \text{state}(t)=s_i} 1}$$

- Assumes state output PDF = Gaussian
 - For GMMs, estimate GMM parameters from collection of observations at any state

12 Oct 2010

11755/18797

76

Training by segmentation: Viterbi training



- Initialize all HMM parameters
- Segment all training observation sequences into states using the Viterbi algorithm with the current models
- Using estimated state sequences and training observation sequences, reestimate the HMM parameters
- This method is also called a "segmental k-means" learning procedure

12 Oct 2010

11755/18797

77

Alternative to counting: SOFT counting

- Expectation maximization
- Every observation contributes to every state

12 Oct 2010

11755/18797

78

Update rules at each iteration

$$\pi(s_i) = \frac{\sum_{Obs} P(\text{state}(t=1) = s_i | Obs)}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i, \text{state}(t+1) = s_j | Obs)}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

$$\mu_i = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs) X_{Obs,t}}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

$$\Theta_i = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs) (X_{Obs,t} - \mu_i)(X_{Obs,t} - \mu_i)^T}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

- Every observation contributes to every state

12 Oct 2010 11755/18797 79

Update rules at each iteration

$$\pi(s_i) = \frac{\sum_{Obs} P(\text{state}(t=1) = s_i | Obs)}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i, \text{state}(t+1) = s_j | Obs)}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

$$\mu_i = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs) X_{Obs,t}}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

$$\Theta_i = \frac{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs) (X_{Obs,t} - \mu_i)(X_{Obs,t} - \mu_i)^T}{\sum_{Obs} \sum_t P(\text{state}(t) = s_i | Obs)}$$

- Where did these terms come from?

12 Oct 2010 11755/18797 80

$P(\text{state}(t) = s | Obs)$

- The probability that the process was at s when it generated X_t given the entire observation
 - Dropping the "Obs" subscript for brevity

$$P(\text{state}(t) = s | X_1, X_2, \dots, X_T) \propto P(\text{state}(t) = s, X_1, X_2, \dots, X_T)$$

- We will compute $P(\text{state}(t) = s, x_1, x_2, \dots, x_T)$ first
 - This is the probability that the process visited s at time t while producing the entire observation

12 Oct 2010 11755/18797 81

$P(\text{state}(t) = s, x_1, x_2, \dots, x_T)$

- The probability that the HMM was in a particular state s when generating the observation sequence is the probability that it followed a state sequence that passed through s at time t

12 Oct 2010 11755/18797 82

$P(\text{state}(t) = s, x_1, x_2, \dots, x_T)$

- This can be decomposed into two multiplicative sections
 - The section of the lattice leading into state s at time t and the section leading out of it

12 Oct 2010 11755/18797 83

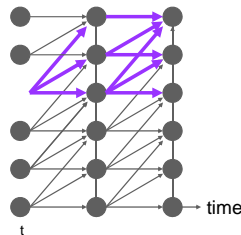
The Forward Paths

- The probability of the red section is the total probability of all state sequences ending at state s at time t
 - This is simply $\alpha(s, t)$
 - Can be computed using the forward algorithm

12 Oct 2010 11755/18797 84

The Backward Paths

- The blue portion represents the probability of all state sequences that began at state s at time t
 - Like the red portion it can be computed using a *backward recursion*



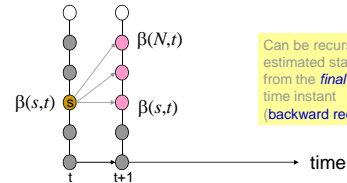
12 Oct 2010

11755/18797

85

The Backward Recursion

$$\beta(s, t) = P(x_{t+1}, x_{t+2}, \dots, x_T | \text{state}(t) = s)$$



$$\beta(s, t) = \sum_{s'} \beta(s', t+1) P(s' | s) P(x_{t+1} | s')$$

- $\beta(s, t)$ is the total probability of ALL state sequences that depart from s at time t , and all observations after x_t
 - $\beta(s, T) = 1$ at the final time instant for all valid final states

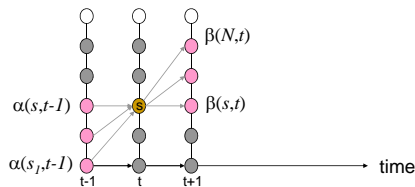
12 Oct 2010

11755/18797

86

The complete probability

$$\alpha(s, t) \beta(s, t) = P(x_{t+1}, x_{t+2}, \dots, x_T, \text{state}(t) = s)$$



12 Oct 2010

11755/18797

87

Posterior probability of a state

- The probability that the process was in state s at time t , given that we have observed the data is obtained by simple normalization

$$P(\text{state}(t) = s | \text{Obs}) = \frac{P(\text{state}(t) = s, x_1, x_2, \dots, x_T)}{\sum_{s'} P(\text{state}(t) = s', x_1, x_2, \dots, x_T)} = \frac{\alpha(s, t) \beta(s, t)}{\sum_{s'} \alpha(s', t) \beta(s', t)}$$

- This term is often referred to as the gamma term and denoted by $\gamma_{s,t}$

12 Oct 2010

11755/18797

88

Update rules at each iteration

$$\pi(s_i) = \frac{\sum_{\text{Obs}} P(\text{state}(t=1) = s_i | \text{Obs})}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_j, \text{state}(t+1) = s_i | \text{Obs})}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

$$\mu_i = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs}) X_{\text{Obs},t}}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

$$\Theta_i = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs}) (X_{\text{Obs},t} - \mu_i)(X_{\text{Obs},t} - \mu_i)^T}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

- These have been found

12 Oct 2010

11755/18797

89

Update rules at each iteration

$$\pi(s_i) = \frac{\sum_{\text{Obs}} P(\text{state}(t=1) = s_i | \text{Obs})}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_j, \text{state}(t+1) = s_i | \text{Obs})}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

$$\mu_i = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs}) X_{\text{Obs},t}}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

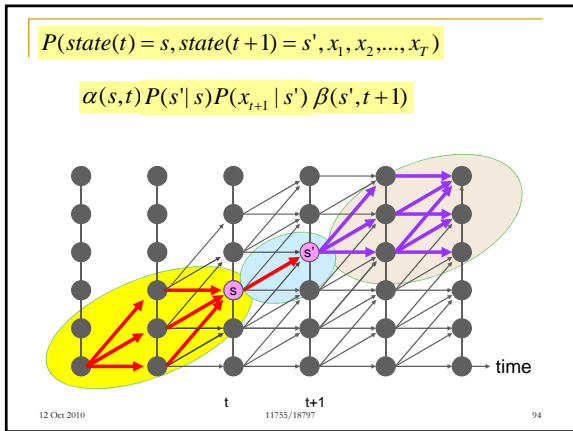
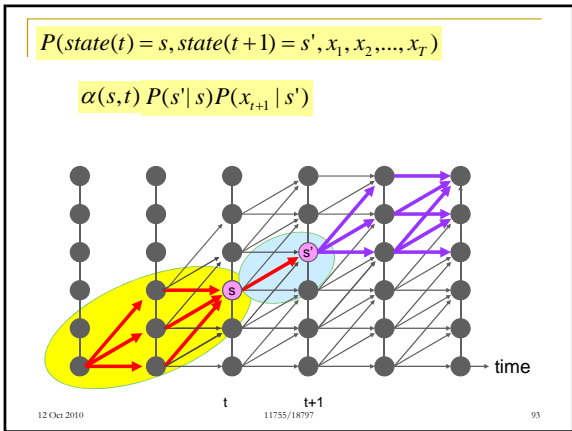
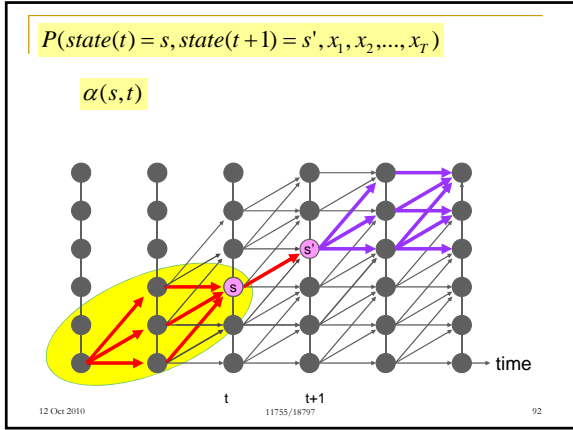
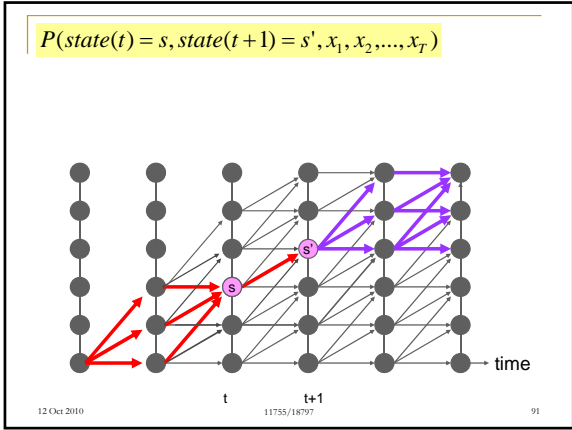
$$\Theta_i = \frac{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs}) (X_{\text{Obs},t} - \mu_i)(X_{\text{Obs},t} - \mu_i)^T}{\sum_{\text{Obs}} \sum_{t-1} P(\text{state}(t) = s_i | \text{Obs})}$$

- Where did these terms come from?

12 Oct 2010

11755/18797

90



The a posteriori probability of transition

$$P(\text{state}(t) = s, \text{state}(t+1) = s' | \text{Obs}) = \frac{\alpha(s, t) P(s' | s) P(x_{t+1} | s') \beta(s', t+1)}{\sum_{s_1} \sum_{s_2} \alpha(s_1, t) P(s_2 | s_1) P(x_{t+1} | s_2) \beta(s_2, t+1)}$$

- The a posteriori probability of a transition given an observation

12 Oct 2010 11755/18797 95

Update rules at each iteration

$$\pi(s_j) = \frac{\sum_{\text{Obs}} P(\text{state}(t=1) = s_j | \text{Obs})}{\text{Total no. of observation sequences}}$$

$$P(s_j | s_i) = \frac{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_i, \text{state}(t+1) = s_j | \text{Obs})}{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_i | \text{Obs})}$$

$$\mu_i = \frac{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_i | \text{Obs}) X_{\text{Obs}, t}}{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_i | \text{Obs})}$$

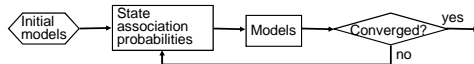
$$\Theta_j = \frac{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_j | \text{Obs}) (X_{\text{Obs}, t} - \mu_i)(X_{\text{Obs}, t} - \mu_i)^T}{\sum_{\text{Obs}} \sum_{t'} P(\text{state}(t) = s_j | \text{Obs})}$$

- These have been found

12 Oct 2010 11755/18797 96

Training without explicit segmentation: Baum-Welch training

- ◆ Every feature vector associated with every state of every HMM with a probability



- ◆ Probabilities computed using the forward-backward algorithm
- ◆ Soft decisions taken at the level of HMM state
- ◆ In practice, the segmentation based Viterbi training is much easier to implement and is much faster
- ◆ The difference in performance between the two is small, especially if we have lots of training data

12 Oct 2010

11755/18797

HMM Issues

- How to find the best state sequence: Covered
- How to learn HMM parameters: Covered
- How to compute the probability of an observation sequence: Covered

12 Oct 2010

11755/18797

98

Magic numbers

- How many states:
 - No nice automatic technique to learn this
 - You choose
 - For speech, HMM topology is usually left to right (no backward transitions)
 - For other cyclic processes, topology must reflect nature of process
 - No. of states – 3 per phoneme in speech
 - For other processes, depends on estimated no. of distinct states in process

12 Oct 2010

11755/18797

99

Applications of HMMs

- Classification:
 - Learn HMMs for the various classes of time series from training data
 - Compute probability of test time series using the HMMs for each class
 - Use in a Bayesian classifier
- Speech recognition, vision, gene sequencing, character recognition, text mining, topic detection...

12 Oct 2010

11755/18797

100

Applications of HMMs

- Segmentation:
 - Given HMMs for various events, find event boundaries
 - Simply find the best state sequence and the locations where state identities change
- Automatic speech segmentation, text segmentation by topic, genome segmentation, ...

12 Oct 2010

11755/18797

101

Implementation Issues

- For long data sequences arithmetic underflow is a problem
 - Scores are products of numbers that are all less than 1
- The Viterbi algorithm provides a workaround – work only with *log* probabilities
 - Multiplication changes to addition – computationally faster too
 - Underflow almost completely eliminated
- For the forward algorithm complex normalization schemes must be implemented to prevent underflow
 - At some computational expense
 - Often not worth it – go with Viterbi

12 Oct 2010

11755/18797

102

Classification with HMMs

HMM for Yes

$P(\text{Yes}) P(X|\text{Yes})$

HMM for No

$P(\text{No}) P(X|\text{No})$

- Speech recognition of isolated words:
- Training:
 - Collect training instances for each word
 - Learn an HMM for each word
- Recognition of an observation X
 - For each word compute $P(X|\text{word})$
 - Using forward algorithm
 - Alternately, compute $P(X, \text{best.state.sequence} | \text{word})$
 - Computed using the Viterbi segmentation algorithm
 - Compute $P(\text{word}) P(X|\text{word})$
 - $P(\text{word})$ = a priori probability of word
 - Select the word for which $P(\text{word}) P(X|\text{word})$ is highest

12 Oct 2010 11755/18797 103

Creating composite models

HMM for Open

HMM for Close

HMM for File

HMM for Open File

HMM for File Close

- HMMs with absorbing states can be combined into composites
 - E.g. train models for open, close and file
 - Concatenate them to create models for "open file" and "file close"
 - Can recognize "open file" and "file close"

12 Oct 2010 11755/18797 104

Model graphs

- Models can also be composed into graphs
 - Not just linearly
- Viterbi state alignment will tell us which portions of the graphs were visited for an observation X

12 Oct 2010 11755/18797 105

Recognizing from graph

- ◆ Trellis for "Open File" vs. "Close File"
- ◆ The VITERBI best path tells you what was spoken

12 Oct 2010 11755/18797 106

Recognizing from graph

- ◆ Trellis for "Open File" vs. "Close File"
- ◆ The VITERBI best path tells you what was spoken

12 Oct 2010 11755/18797 107

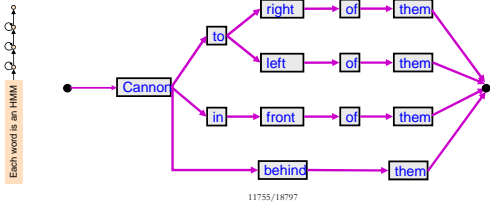
"Language" probabilities can be incorporated

- Transitions between HMMs can be assigned a probability
 - Drawn from properties of the language
 - Here we have shown "Bigram" probabilities

12 Oct 2010 11755/18797 108

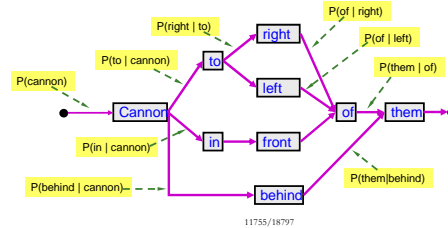
This is used in speech recognition

- Recognizing one of four lines from "charge of the light brigade"
 - Cannon to right of them
 - Cannon to left of them
 - Cannon in front of them
 - Cannon behind them
- Each "word" is an HMM



Graphs can be reduced sometimes

- Recognizing one of four lines from "charge of the light brigade"
 - Graph reduction does not impede recognition of what was spoken



Speech recognition: An aside

- In speech recognition systems models are trained for *phonemes*
 - Actually "triphones" – phonemes in context
- Word HMMs are composed from phoneme HMMs
- Language HMMs are composed from word HMMs
- The graph is "reduced" using automated techniques
 - John McDonough talks about WFSTs on Thursday

12 Oct 2010

11755/18797

111