Hidden Markov Models (HMM)

ECE/CS 498 DS U/G

Lecture 19: Hidden Markov Models continued, ICA 4

Ravi K. Iyer

Dept. of Electrical and Computer Engineering

University of Illinois at Urbana Champaign

Announcements

- Course Timeline
  - Wed 4/1:
    - Mid-semester feedback
    - Brief review of backwards algorithm
    - ICA 4: HMMs
  - Mon 4/3: Introduction to factor graphs
- MP 3 will be released today
  - Covers data analysis, HMMs, and factor graphs for HPC security
  - Checkpoint 1 (Tasks 0 and 1) will be due **Monday April 13 @ 11:59 PM** on Compass 2G
- Final Project
  - Progress report 2 due **Friday April 17 @ 11:59 PM** on Compass2G
    - There should be *substantial* progress with projects by this point (i.e. meaningful results, ML/AI models)

Mid-Semester Feedback Form

- Please take ~10 minutes to complete the mid-semester feedback form at the following link:

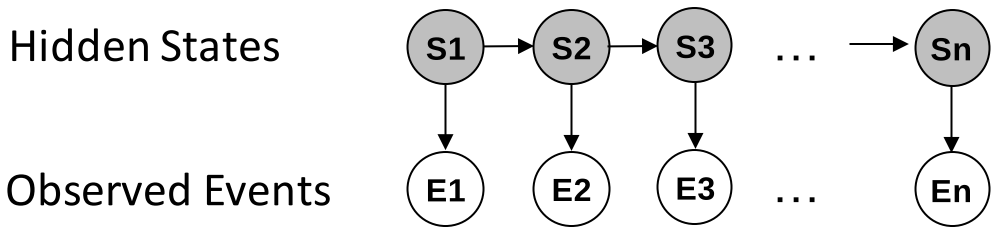
https://forms.gle/CvtaPoMGmTcZZ2BbA

- We are keeping responses anonymous (i.e. not collecting names/emails)
- We will consider your input to shape the remainder of the course

Hidden Markov Models

Model

- Set of hidden states $\boldsymbol{S} = \{\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_N}\}$
- Set of observable events $\boldsymbol{E} = \{\boldsymbol{\epsilon_1}, \dots, \boldsymbol{\epsilon_M}\}$
- Transition probability matrix $\boldsymbol{A}$
- Observation matrix $\boldsymbol{B}$
- Initial distribution of hidden states $\boldsymbol{\pi}$



A Hidden Markov model on observed events and system states

Model assumptions

- An observation depends on its hidden state
- A state variable only depends on the immediate previous state (Markov assumption)
- The future observations and the past observations are conditionally independent given the current hidden state

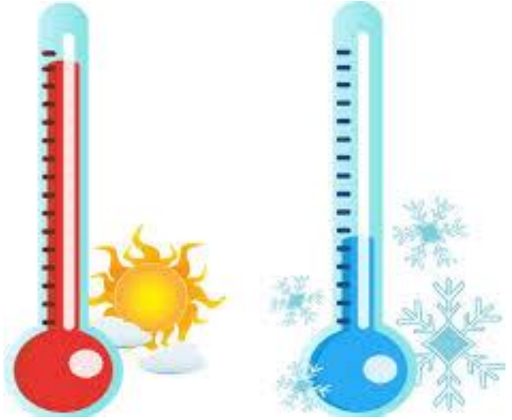
$$P(S_1, \dots, S_n, E_1, \dots, E_n) = P(S_1)P(E_1|S_1)\prod_{i=2}^{n} P(S_i|S_{i-1})P(E_i|S_i)$$

Advantages:

- HMM can model sequential nature of input data (future depends on the past)
- HMM has a linear-chain structure that clearly separates system state and observed events.

HMM Motivating Example: Paleontological Temperature Model

- Want to determine the average temperature at a particular place on earth over a sequence of years in the distant past
- Hidden state Only annual average temperatures -- hot **(H)** and cold (**C**)
  - Probability of a hot year followed by another hot year is 0.7, and the probability of a cold year followed by another cold year is 0.6, independent of the temperature in prior years
- Obervations Correlation between the size of tree growth rings and temperature
  - Three different ring sizes, small (**T**), medium (**D**), and large (**L**)
- Assume that probability values from current period held in paleontological period too
- Determine the most likely temperature state in past years
  - Can't directly observe the temperature in the past
  - We can observe the size of tree rings – can this information be used?



H

C

T

D

L

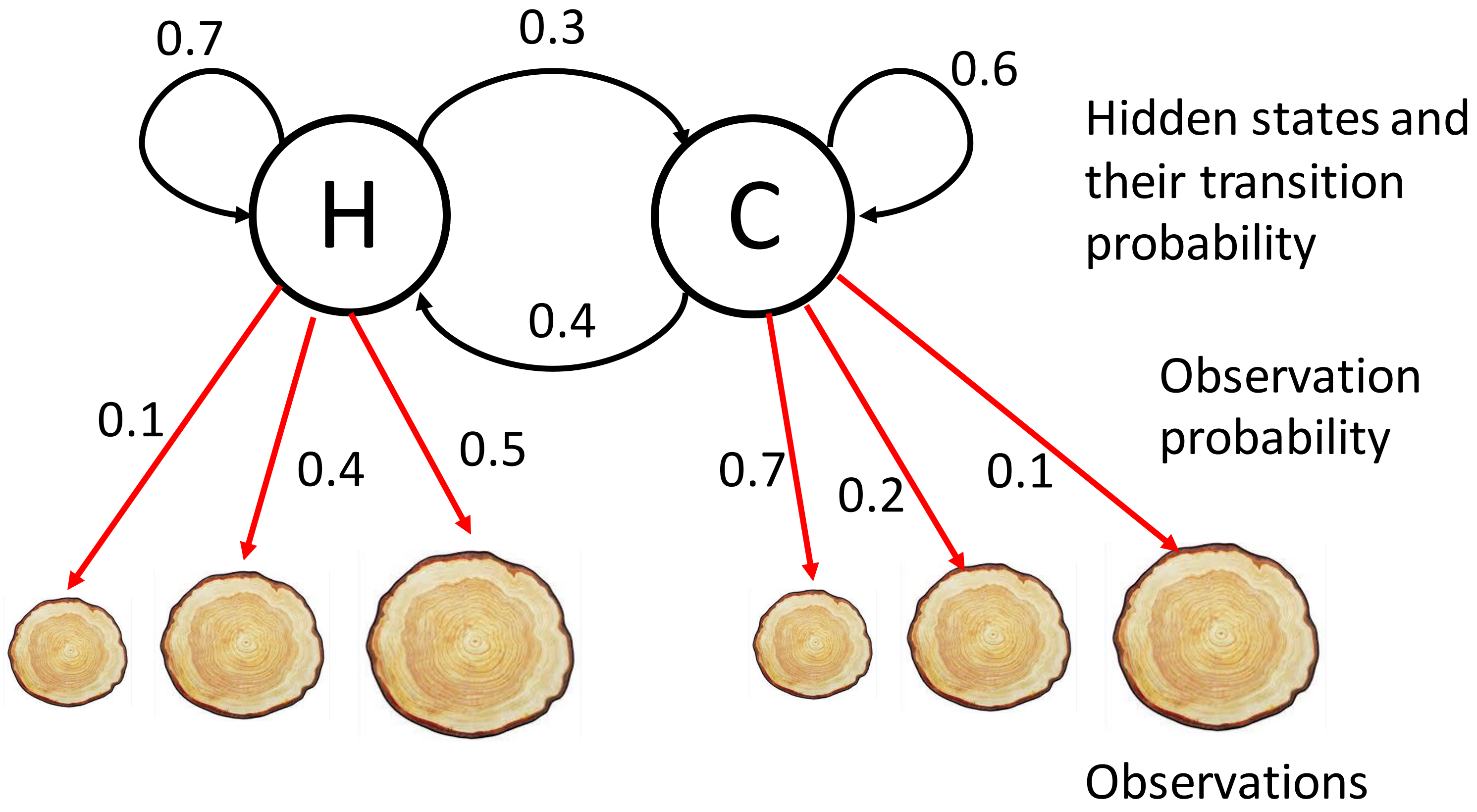
Tree ring size

Source: http://www.math.unl.edu/~sdunbar1/ProbabilityTheory/Lessons/HiddenMarkovModels/Examples/examples.html

Paleontological Temperature Model

- State space of hidden states: $S = \{H, C\}$
- State space of observations: $E = \{T, D, L\}$
- Transition probability matrix: A
- Observation Matrix: B
- Initial distribution for the hidden states: π

Given by an oracle



$$\begin{array}{cc} & \begin{array}{cc} H & C \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

A

$$\begin{array}{cc} & \begin{array}{ccc} T & D & L \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}$$

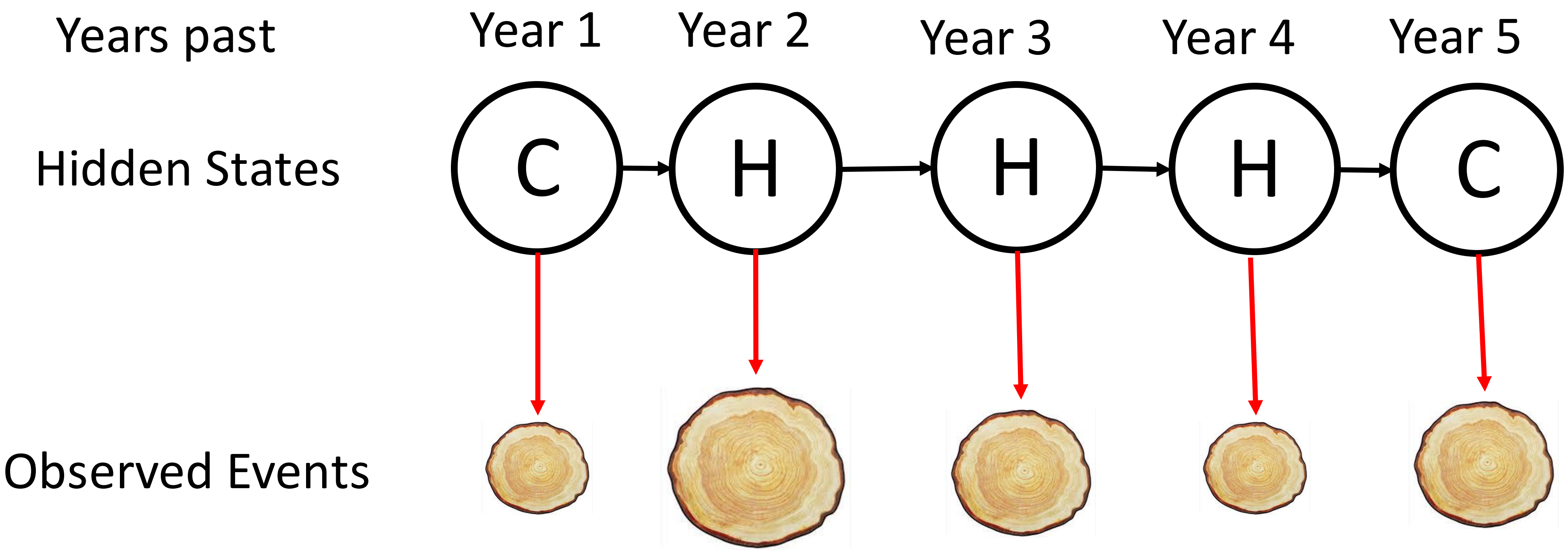
B

$$\begin{array}{cc} H & C \end{array}$$
$$[0.5\ 0.5]$$

π

Paleontological Temperature Model

Example sequence with 5 observations



Determine the sequence of hidden states

Inference question – Paleontological Temperature

Given the sequence of 5 observations T, L, D, T, D and the model (A, B, π), how do we choose a corresponding state sequence $S_1, S_2, \ldots, S_n$ which is optimal in some meaningful sense (i.e., best explains the observations) where $S_t \in \{H, C\}$?

A simpler question: Given the sequence of 5 observations T, L, D, T, D and the model (A, B, π), which of the two is more probable eg., $S_3 = H$ or $S_3 = C$?

General Inference question

Given the sequence of n observations $E_1, E_2, \dots, E_n$, and the model (A, B, π), how do we choose a corresponding state sequence $S_1, S_2, \dots, S_n$ which is optimal in some meaningful sense (i.e., best explains the observations)?

A simpler question: Given the sequence of n observations $E_1, E_2, \dots, E_n$, and the model (A, B, π), what is the most probable state S_t at $t \in \{1, \dots, n\}$?

$$\underset{j \in \{1,\dots,N\}}{\operatorname{argmax}} P(S_t = \sigma_j | E_1, E_2, \dots, E_n)$$

$$S = \{\sigma_1, \dots, \sigma_N\}$$

Breaking down the inference question

$$P(S_t|E_1, E_2, \dots, E_n) = \frac{P(S_t, E_1, \dots, E_n)}{P(E_1, \dots, E_n)} = \frac{P(S_t, E_1, \dots, E_t, E_{t+1}, \dots, E_n)}{P(E_1, \dots, E_n)}$$

Bayes rule

$$= \frac{P(E_{t+1}, \dots, E_n \,|S_t, E_1, \dots, E_t)\, P(S_t, E_1, \dots, E_t)}{P(E_1, \dots, E_n)}$$

Bayes rule

$$= P(E_{t+1}, \dots, E_n \,|S_t, E_1, \dots, E_t)\, P(S_t|E_1, \dots, E_t) \frac{P(E_1, \dots, E_t)}{P(E_1, \dots, E_n)}$$

Markov property

$$= \frac{P(E_{t+1}, \dots, E_n \,|S_t)\, P(S_t|E_1, \dots, E_t)}{P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)}$$

Breaking down the inference question

$$P(S_t|E_1, E_2, \dots, E_n) = \frac{P(E_{t+1}, \dots, E_n\ |S_t)\ P(S_t|E_1, \dots, E_t)}{P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)}$$

$P(S_t|E_1, \dots, E_t)$:
Probability of hidden state at time t given observation up to time t (**Forwards algorithm**)

$P(E_{t+1}, \dots, E_n\ |S_t)$:
Probability of the future observed sequence given the hidden state at time t (**Backwards algorithm**)

$P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)$:
Does not depend on the hidden state (will not affect the maximization because it is just a scaling factor)

Forwards algorithm: Paleontological Temperature

Want to calculate $P(S_t|E_1,\dots,E_t)$

- Let us calculate it for $t = 2$
- In the example, $E_1 = T, E_2 = L$
- Find $P(S_2 = H|E_1 = T, E_2 = L)$?

$$P(S_2 = H|E_1 = T, E_2 = L) = \frac{P(S_2 = H, E_1 = T, E_2 = L)}{P(E_1 = T, E_2 = L)}$$

Adding hidden state S_1

$$= \frac{\sum_{s\in\{H,C\}} P(S_2 = H, E_1 = T, E_2 = L, S_1 = s)}{P(E_1 = T, E_2 = L)}$$

Forwards algorithm: Paleontological Temperature

$$\frac{\sum_{s\in\{H,C\}} P(S_2 = H, E_1 = T, E_2 = L, S_1 = s)}{P(E_1 = T, E_2 = L)}$$

Bayes rule

$$= \frac{\sum_{s\in\{H,C\}} P(E_2 = L | S_2 = H, E_1 = T, S_1 = s) P(S_2 = H, E_1 = T, S_1 = s)}{P(E_1 = T, E_2 = L)}$$

Markov property

Bayes rule

$$= \frac{\sum_{s\in\{H,C\}} P(E_2 = L | S_2 = H) P(S_2 = H | E_1 = T, S_1 = s) P(S_1 = s | E_1 = T) P(E_1 = T)}{P(E_1 = T, E_2 = L)}$$

Markov property

Bayes rule

$$= \frac{\sum_{s\in\{H,C\}} P(E_2 = L | S_2 = H) P(S_2 = H | S_1 = s) P(S_1 = s | E_1 = T)}{P(E_2 = L | E_1 = T)}$$

Forwards algorithm: Paleontological Temperature

Hidden state given all observations up to that point

Observation probability

Transition probability

Hidden state given all observations up to that point

$$P(S_2 = H | E_1 = T, E_2 = L) = \frac{P(E_2 = L | S_2 = H) \sum_{s \in \{H,C\}} P(S_2 = H | S_1 = s) P(S_1 = s | E_1 = T)}{P(E_2 = L | E_1 = T)}$$

Define: $\alpha_t(i) = P(S_t = \sigma_i | E_1, E_2, \ldots, E_t)$ and $Z_t = P(E_t | E_1, \ldots E_{t-1})$

Above equation can be written as,

$$\boxed{\alpha_2(H)} = \frac{1}{Z_2} P(E_2 = L | S_2 = H) \sum_{s \in \{H,C\}} P(S_2 = H | S_1 = s) \boxed{\alpha_1(s)}$$

Where, $Z_2 = P(E_2 | E_1)$

Recursion

Forwards algorithm: General Expression

Define: $\alpha_t(j) = P(S_t = \sigma_j | E_1, E_2, \dots, E_t)$ and $Z_t = P(E_t | E_1, \dots E_{t-1})$

In general,

$$\alpha_t(j) = \frac{1}{Z_t} P\big(E_t \big| S_t = \sigma_j\big) \sum_{i=1}^{N} P(S_t = \sigma_j \mid S_{t-1} = \sigma_i)\, \alpha_{t-1}(i) \qquad Z_t = \sum_{j=1}^{N} b_t \odot (A^T \alpha_{t-1})$$

Transition probability a_{ij}

Above equation can be written as a matrix for all j,

$$\begin{bmatrix} \alpha_t(1) \\ \vdots \\ \alpha_t(j) \\ \vdots \\ \alpha_t(N) \end{bmatrix} \propto \begin{bmatrix} P(E_t|S_t = \sigma_1) \\ \vdots \\ P(E_t|S_t = \sigma_j) \\ \vdots \\ P(E_t|S_t = \sigma_N) \end{bmatrix} \odot \begin{bmatrix} a_{11} & \cdots & \cdots & \cdots & a_{N1} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ a_{1j} & \cdots & a_{ij} & \cdots & a_{Nj} \\ \vdots & \cdots & \cdots & \ddots & \cdots \\ a_{1N} & \cdots & \cdots & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} \alpha_{t-1}(1) \\ \vdots \\ \alpha_{t-1}(i) \\ \vdots \\ \alpha_{t-1}(N) \end{bmatrix}$$

$$\alpha_t \propto b_t \odot (A^T \alpha_{t-1})$$

$\odot$ Represents elementwise product (Hadamard product)

b_t is the column of the observation matrix B corresponding to E_t

Forwards Algorithm: Paleontological Temperature

For observations T, L, D, T, L

$$\begin{array}{c|cc} & H & C \\ \hline H & 0.7 & 0.3 \\ C & 0.4 & 0.6 \end{array}$$

Transition probability matrix

$P(S_2 | E_1 = T, E_2 = L)$ is,

$$\begin{bmatrix} \alpha_2(H) \\ \alpha_2(C) \end{bmatrix} \propto \begin{bmatrix} 0.5 \\ 0.1 \end{bmatrix} \odot \left(\begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} \alpha_1(H) \\ \alpha_1(C) \end{bmatrix} \right)$$

$$\begin{array}{c|ccc} & T & D & L \\ \hline H & 0.1 & \boxed{0.4} & \boxed{0.5} \\ C & 0.7 & 0.2 & 0.1 \end{array}$$

Observation matrix

Similarly, $P(S_3 | E_1 = T, E_2 = L,\ E_3 = D)$ is,

$$\begin{bmatrix} \alpha_3(H) \\ \alpha_3(C) \end{bmatrix} \propto \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix} \odot \left(\begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} \alpha_2(H) \\ \alpha_2(C) \end{bmatrix} \right)$$

Forwards Algorithm

1. Input: (A, B, π) and observed sequence $E_1, \ldots, E_n$
2. $[\alpha_1, Z_1]$ = normalize($b_1 \odot \pi$)
3. **for** $t = 2{:}\,n$ **do**
   $[\alpha_t, Z_t]$ = normalize($b_t \odot (A^T \alpha_{t-1})$)
4. return $\alpha_1, \ldots, \alpha_n$ and $\log\big(P(E_1, \ldots, E_n)\big) = \sum_t \log(Z_t)$

Note:

Subroutine: [v, Z] = normalize(u): $Z = \sum_j u_j; v_j = u_j / Z;$

Breaking down the inference question

$$P(S_t|E_1, E_2, \dots, E_n) = \frac{P(E_{t+1}, \dots, E_n\ |S_t)\ P(S_t|E_1, \dots, E_t)}{P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)}$$

$P(S_t|E_1, \dots, E_t)$:
Probability of hidden state at time t given observation up to time t (**Forwards algorithm**)

$P(E_{t+1}, \dots, E_n\ |S_t)$:
Probability of the future observed sequence given the hidden state at time t (**Backwards algorithm**)

$P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)$:
Does not depend on the hidden state (will not affect the maximization because it is just a scaling factor)

Backwards Algorithm (similar to Forwards Algo.)

Calculate $P(E_{t+1}, \dots, E_n \mid S_t)$

Define: $\beta_t(j) = P(E_{t+1}, \dots, E_n \mid S_t = \sigma_j)$

Include S_t to use information from the one-step future

$$\boxed{\beta_{t-1}(j)} = P(E_t, \dots, E_n \mid S_{t-1} = \sigma_j) = \sum_{i=1}^{N} P(S_t = \sigma_i, E_t, \dots, E_n \mid S_{t-1} = \sigma_j)$$

Chain rule

$$= \sum_{i=1}^{N} P(E_{t+1}, \dots, E_n \mid S_{t-1} = \sigma_j, S_t = \sigma_i, E_t)\, P(E_t \mid S_{t-1} = \sigma_j, S_t = \sigma_i) P(S_t = \sigma_i \mid S_{t-1} = \sigma_j)$$

Markov property

$$= \sum_{i=1}^{N} P(E_{t+1}, \dots, E_n \mid S_t = \sigma_i)\, P(E_t \mid S_t = \sigma_i) P(S_t = \sigma_i \mid S_{t-1} = \sigma_j)$$

By definition of $\beta_t(j)$

$$= \sum_{i=1}^{N} \boxed{\beta_t(i)} P(E_t \mid S_t = \sigma_i) P(S_t = \sigma_i \mid S_{t-1} = \sigma_j)$$

Emission probability

Transition probability

In matrix form, we get,

$$\beta_{t-1} = A(b_t \odot \beta_t)$$

$$\beta_t = \begin{bmatrix} \beta_t(1) \\ \vdots \\ \beta_t(N) \end{bmatrix}$$

Backwards Algorithm

1. Input: (A, B, π) and observed sequence $E_1, \ldots, E_n$
2. $\beta_n = 1$; // initialize $\beta_n(j)$ to 1 for all states σ_j
3. **for** $t = n - 1 : 1$ **do**

   $\beta_{t-1} = A(b_t \odot \beta_t)$
4. return $\beta_1, \ldots, \beta_n$

Breaking down the inference question

$$P(S_t|E_1, E_2, \dots, E_n) = \frac{P(E_{t+1}, \dots, E_n\ |S_t)\ P(S_t|E_1, \dots, E_t)}{P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)}$$

$P(S_t|E_1, \dots, E_t)$:
Probability of hidden state at time t given observation up to time t (**Forwards algorithm**)

$P(E_{t+1}, \dots, E_n\ |S_t)$:
Probability of the future observed sequence given the hidden state at time t (**Backwards algorithm**)

$P(E_{t+1}, \dots, E_n|E_1, \dots, E_t)$:
Does not depend on the hidden state (will not affect the maximization because it is just a scaling factor)

Inference – using Forwards-Backwards expressions

$$P(S_t|E_1, E_2, \ldots, E_n) = \frac{P(E_{t+1}, \ldots, E_n\,|S_t)\,P(S_t|E_1, \ldots, E_t)}{P(E_{t+1}, \ldots, E_n|E_1, \ldots, E_t)}$$

For $S_t = \sigma_j$ and $\gamma_t(j) = P\left(S_t = \sigma_j \middle| E_1, E_2, \ldots, E_n\right)$, the above equation is:

$$P\left(S_t = \sigma_j \middle| E_1, E_2, \ldots, E_n\right) = \frac{P(E_{t+1}, \ldots, E_n\,|S_t = \sigma_j)\,P(S_t = \sigma_j|E_1, \ldots, E_t)}{P(E_{t+1}, \ldots, E_n|E_1, \ldots, E_t)}$$

$$\gamma_t(j) = \frac{\beta_t(j)\alpha_t(j)}{P(E_{t+1}, \ldots, E_n|E_1, \ldots, E_t)} = \frac{\beta_t(j)\alpha_t(j)}{\sum_{i=1}^{N}\beta_t(j)\alpha_t(j)}$$

Theorem of total probability

$$\boxed{\gamma_t(j) \propto \beta_t(j)\alpha_t(j)}$$

Inference: Most likely state

- Forwards-backwards algorithm gives $P(S_t = \sigma_j | E_1, \dots, E_n)$ for all j
- Find the individually most likely state at time t given all observations

$$S_t^* = \operatorname*{argmax}_{j \in \{1,..,N\}} \gamma_t(j)$$

Optimality of inference

- In the inference problem we attempt to uncover the hidden part of HMM, i.e., find the "correct" state sequence
- It is impossible to find the "correct" state sequence (solution)
- Use optimality criterion to find the "best" possible solution
- Several reasonable criteria exist and is a strong function of the intended application
  - Most likely state given observations
    - Application in finding average statistics, expected number of correct states
    - Solved using Forwards-Backwards algorithm
  - Single best sequence that maximises probability of observed events
    - Application in continuous speech recognition
    - Solved using Viterbi algorithm

Resources

Rabiner's (excellent) paper:
https://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf

Begin ICA 4: HMMs For Security