Hidden Markov Models

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2015

Outline

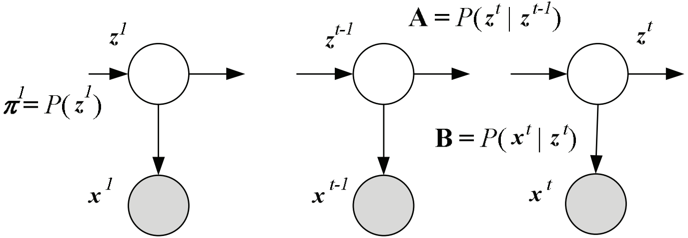
Outline

1. **Hidden Markov Models**
   - Definitions and Usage
2. Learning the Model Parameters
   - Expectation Maximization for HMM
   - The Forward-Backward Procedure
3. Inferring the State Sequences
4. Making Predictions
5. Practical Considerations

Hidden Markov Models

- A ***Hidden Markov Model (HMM)*** is a Markov chain where we don't know which state the process $X^{(t)}$ locates in at any time t
  - Let $\boldsymbol{z}^{(t)} \in \{0,1\}^K$ be a vector where $z_i^{(t)} = 1$ if $X^{(t)} = S_i$; 0 otherwise
  - $P[\boldsymbol{z}^{(t)} = \boldsymbol{e}_i] = P[X^{(t)} = S_i]$ (for brevity, we use the shorthand $P[z_i^{(t)}]$)
  - In HMM, $\boldsymbol{z}^{(t)}$ is hidden (not observable) and is a latent variable
- When a state is visited, however, we can record an observation $\boldsymbol{x}^{(t)}$
  - $P[\boldsymbol{x}^{(t)} | z_i^{(t)}]$ is called the ***emission probability*** of state i at time t
  - Like transition probabilities, the emission probabilities are usually assumed to be ***time homogeneous***
  - If we assume that the emission probability of state i follows some distribution parametrized by θ_i, we can rewrite it as $P[\boldsymbol{x}^{(t)} | z_i^{(t)}, \theta_i]$
- Markov chain is a special case of HMM where
  - $\boldsymbol{x}^{(t)}$ must be one of the $S_1, \cdots, S_K$
  - $P[\boldsymbol{x}^{(t)} = S_j | z_i^{(t)}] = 1$ if $i = j$; 0 otherwise

Graph Representation



- HMM is a candidate for modeling a problem when we are given a sequence $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ of observations of length T, where $\boldsymbol{x}^{(t)}$ are ***not*** i.i.d.

Goals

- HMM is a candidate for modeling a problem when we are given a sequence $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ of observations of length T, where $\boldsymbol{x}^{(t)}$ are ***not*** i.i.d.
- Generally, we want to perform the following tasks:

1. Given $\mathcal{X}$, learn the parameters $\Theta = (\boldsymbol{\pi}^{(1)}, \boldsymbol{A}, \{\theta_i\}_{i=1}^{K})$ maximizing the likelihood $P[\mathcal{X}|\Theta]$
    - $\boldsymbol{\pi}^{(1)}$ is the initial state probability
    - $\boldsymbol{A}$ is the transition matrix
    - θ_i is the parameter of the emission probability of state i
2. Given the learned Θ, infer the hidden state sequence $\mathcal{Z} = \{\boldsymbol{z}^{(t)}\}_{t}^{T}$ that generated $\mathcal{X}$ with the highest probability $P[\mathcal{X}|\mathcal{Z}, \Theta]$
3. Given the learned Θ, evaluate $P[\mathcal{X}^{new}|\Theta]$ for a new sequence $\mathcal{X}^{new}$

Applications

- For classification, we can model each class as an HMM
  - Learn the parameter Θ_i of each class C_i using a training sequence $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ (or a set $\mathcal{X} = \{\boldsymbol{x}^{(n,t)}\}_{n=1,t=1}^{N,T}$ of n training sequences)
  - Predict a new sequence $\mathcal{X}^{new}$ to be in class C_i if the posterior $P[C_i|\mathcal{X}^{new}] \propto P[\mathcal{X}^{new}|\Theta_i]P[C_i]$ is the highest
- Applications:
  - Pattern recognition (speech recognition, gesture recognition, handwritten character recognition, etc.)
  - Sequential data analysis
  - Molecular biology, biochemistry, and genetics, etc.
- One most powerful property of an HMM is that ***it can accommodate the local warping (compression/stretching) in the time axis***
  - E.g., the likelihood $P[\mathcal{X}^{new}|\Theta] = \sum_{\mathcal{Z}} P[\mathcal{X}^{new}, \mathcal{Z}|\Theta]$ of a speech $\mathcal{X}^{new}$ will not change dramatically when it is spoken slowly, as $\mathcal{Z}$ having more transitions to the same state will contribute to the likelihood more

Outline

Problem Formulation

- Problem: given a sequence $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ of observations up to time T, we want to find $\Theta = (\boldsymbol{\pi}^{(1)}, \boldsymbol{A}, \{\theta_i\}_{i=1}^{K})$ that maximizes $P[\mathcal{X}|\Theta]$
- If we know $\mathcal{Z} = \{\boldsymbol{z}^{(t)}\}_{t=1}^{T}$, we have
  - $P[\mathcal{X}|\Theta] = \sum_{\mathcal{Z}} P[\mathcal{X}, \mathcal{Z}|\Theta]$
  - $P[\mathcal{X}, \mathcal{Z}|\Theta] = P[\mathcal{X}|\mathcal{Z}, \Theta] P[\mathcal{Z}|\Theta]$
  - $P[\mathcal{Z}|\Theta] = P[\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(T)}|\Theta] = P[\boldsymbol{z}^{(2)}, \cdots, \boldsymbol{z}^{(T)}|\boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(1)}|\Theta] = P[\boldsymbol{z}^{(3)}, \cdots, \boldsymbol{z}^{(T)}|\boldsymbol{z}^{(2)}, \boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(2)}|\boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] = \cdots = P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] \left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)}, \boldsymbol{A}] \right)$
  - $P[\mathcal{X}|\mathcal{Z}, \Theta] = \prod_{t=1}^{T} P[\boldsymbol{x}^{(t)}|\boldsymbol{z}^{(t)}, \theta_{d(\boldsymbol{z}^{(t)})}]$, where $d(\boldsymbol{z}^{(t)})$ is the index of attribute of $\boldsymbol{z}^{(t)}$ equal to 1
- Unfortunately, we don't know $\mathcal{Z}$ so Θ cannot be solved analytically
- Solution?

Problem Formulation

- Problem: given a sequence $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ of observations up to time T, we want to find $\Theta = (\boldsymbol{\pi}^{(1)}, \boldsymbol{A}, \{\theta_i\}_{i=1}^{K})$ that maximizes $P[\mathcal{X}|\Theta]$
- If we know $\mathcal{Z} = \{\boldsymbol{z}^{(t)}\}_{t=1}^{T}$, we have
  - $P[\mathcal{X}|\Theta] = \sum_{\mathcal{Z}} P[\mathcal{X}, \mathcal{Z}|\Theta]$
  - $P[\mathcal{X}, \mathcal{Z}|\Theta] = P[\mathcal{X}|\mathcal{Z}, \Theta] P[\mathcal{Z}|\Theta]$
  - $P[\mathcal{Z}|\Theta] = P[\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(T)}|\Theta] = P[\boldsymbol{z}^{(2)}, \cdots, \boldsymbol{z}^{(T)}|\boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(1)}|\Theta] = P[\boldsymbol{z}^{(3)}, \cdots, \boldsymbol{z}^{(T)}|\boldsymbol{z}^{(2)}, \boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(2)}|\boldsymbol{z}^{(1)}, \Theta] P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] = \cdots = P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] \left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)}, \boldsymbol{A}] \right)$
  - $P[\mathcal{X}|\mathcal{Z}, \Theta] = \prod_{t=1}^{T} P[\boldsymbol{x}^{(t)}|\boldsymbol{z}^{(t)}, \theta_{d(\boldsymbol{z}^{(t)})}]$, where $d(\boldsymbol{z}^{(t)})$ is the index of attribute of $\boldsymbol{z}^{(t)}$ equal to 1
- Unfortunately, we don't know $\mathcal{Z}$ so Θ cannot be solved analytically
- Solution? Since each $\boldsymbol{z}^{(t)}$ is discrete and corresponds to an instance $\boldsymbol{x}^{(t)}$, we can resort to the EM algorithm

Formulating $\mathcal{Q}(\Theta;\Theta^{old})$ (1/4)

- Recall that $P[\mathcal{X},\mathcal{Z}|\Theta] = P[\mathcal{X}|\mathcal{Z},\Theta]P[\mathcal{Z}|\Theta] = \left(\prod_{s=1}^{T} P[\boldsymbol{x}^{(s)}|\boldsymbol{z}^{(s)},\theta_{d(\boldsymbol{z}^{(s)})}]\right) P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] \left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right)$
- $\mathcal{Q}(\Theta;\Theta^{old}) = E_{\mathcal{Z}}\left[\ln\left(P[\mathcal{X},\mathcal{Z}|\Theta]\right)|\mathcal{X},\Theta^{old}\right]$
$= \sum_{\mathcal{Z}} \ln\left(P[\mathcal{X},\mathcal{Z}|\Theta]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$
$= \sum_{\mathcal{Z}} \ln\left(P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$
$+ \sum_{\mathcal{Z}} \ln\left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$
$+ \sum_{\mathcal{Z}} \ln\left(\prod_{t=1}^{T} P[\boldsymbol{x}^{(t)}|\boldsymbol{z}^{(t)},\theta_{d(\boldsymbol{z}^{(t)})}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$

Formulating $\mathcal{Q}(\Theta;\Theta^{old})$ (2/4)

The first term $\sum_{\mathcal{Z}} \ln\left(P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$

$$
\begin{aligned}
&= \sum_{\mathcal{Z}} \ln\left(\pi^{(1)}_{d(\boldsymbol{z}^{(1)})}\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}] \\
&= \sum_{\boldsymbol{z}^{(1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \cdots \sum_{\boldsymbol{z}^{(T)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln\left(\pi^{(1)}_{d(\boldsymbol{z}^{(1)})}\right) P[\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(T)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{\boldsymbol{z}^{(1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln\left(\pi^{(1)}_{d(\boldsymbol{z}^{(1)})}\right) \sum_{\boldsymbol{z}^{(2)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \cdots \sum_{\boldsymbol{z}^{(T)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} P[\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(T)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{\boldsymbol{z}^{(1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln\left(\pi^{(1)}_{d(\boldsymbol{z}^{(1)})}\right) P[\boldsymbol{z}^{(1)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{i=1}^{K} \ln\left(\pi^{(1)}_i\right) P[z^{(1)}_i|\mathcal{X},\Theta^{old}]
\end{aligned}
$$

Formulating $\mathcal{Q}(\Theta;\Theta^{old})$ (3/4)

The second term $\sum_{\mathcal{Z}} \ln \left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}]$

$$
\begin{aligned}
&= \sum_{\mathcal{Z}} \sum_{t=1}^{T-1} \ln \left(P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}] \\
&= \\
&\sum_{t=1}^{T-1} \sum_{\boldsymbol{z}^{(1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \cdots \sum_{\boldsymbol{z}^{(T)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln \left(P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) P[\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(T)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{t=1}^{T-1} \sum_{\boldsymbol{z}^{(t)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \sum_{\boldsymbol{z}^{(t+1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln \left(P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) \sum_{\boldsymbol{z}^{(1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \cdots \sum_{\boldsymbol{z}^{(t-1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \\
&\qquad\qquad \sum_{\boldsymbol{z}^{(t+2)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \cdots \sum_{\boldsymbol{z}^{(T)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} P[\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(T)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{t=1}^{T-1} \sum_{\boldsymbol{z}^{(t)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \sum_{\boldsymbol{z}^{(t+1)}=\boldsymbol{e}_1}^{\boldsymbol{e}_K} \ln \left(P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right) P[\boldsymbol{z}^{(t)},\boldsymbol{z}^{(t+1)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{t=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{K} \ln \left(P[z_j^{(t+1)}|z_i^{(t)},\boldsymbol{A}]\right) P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}] \\
&= \sum_{t=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{K} \ln (a_{i,j}) P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}]
\end{aligned}
$$

Formulating $\mathcal{Q}(\Theta;\Theta^{old})$ (4/4)

- Similarly, the third term
  $\sum_{\mathcal{Z}} \ln\left(\prod_{t=1}^{T} P[\boldsymbol{x}^{(t)}|\boldsymbol{z}^{(t)},\theta_{d(\boldsymbol{z}^{(t)})}]\right) P[\mathcal{Z}|\mathcal{X},\Theta^{old}] =$
  $\sum_{t=1}^{T}\sum_{i=1}^{K} \ln\left(P[\boldsymbol{x}^{(t)}|z_i^{(t)},\theta_i]\right) P[z_i^{(t)}|\mathcal{X},\Theta^{old}]$ [Proof]
- $\mathcal{Q}(\Theta;\Theta^{old}) = \sum_{i=1}^{K} \ln\left(\pi_i^{(1)}\right) P[z_i^{(1)}|\mathcal{X},\Theta^{old}]$
  $+\sum_{t=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{K} \ln(a_{i,j})\, P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}]$
  $+\sum_{t=1}^{T}\sum_{i=1}^{K} \ln\left(P[\boldsymbol{x}^{(t)}|z_i^{(t)},\theta_i]\right) P[z_i^{(t)}|\mathcal{X},\Theta^{old}]$
- In the E-step, we need to evaluate $P[z_i^{(t)}|\mathcal{X},\Theta^{old}]$ and $P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}]$ for all t (to be discussed later)
- After the evaluation, we denote $\gamma_i^{(t)} = P[z_i^{(t)}|\mathcal{X},\Theta^{old}]$ and $\xi_{i,j}^{(t)} = P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}]$ respectively as constants in the M-step
  - $\gamma_i^{(t)} = \sum_{j=1}^{K} \xi_{i,j}^{(t)}$ and $\sum_{i=1}^{K} \gamma_i^{(t)} = 1$

Solving Θ

- Problem: $\arg_{\Theta=(\boldsymbol{\pi}^{(1)},\boldsymbol{A},\{\theta_i\}_{i=1}^K)} \max \mathcal{Q}(\Theta;\Theta^{old})$, where
  - $\mathcal{Q}(\Theta;\Theta^{old}) = \sum_{i=1}^K \ln\left(\pi_i^{(1)}\right)\gamma_i^{(1)} + \sum_{t=1}^{T-1}\sum_{i=1}^K\sum_{j=1}^K \ln\left(a_{i,j}\right)\xi_{i,j}^{(t)}$ $+\sum_{t=1}^T\sum_{i=1}^K \ln\left(P[\boldsymbol{x}^{(t)}|z_i^{(t)},\theta_i]\right)\gamma_i^{(t)}$
- Subject to
  - $\sum_{i=1}^K \pi_i^{(1)} = 1$
  - $\sum_{j=1}^K a_{i,j} = 1$ for all $1 \leqslant i \leqslant K$
- We can solve $\boldsymbol{\pi}^{(1)}$, $\boldsymbol{A}$, and $\{\theta_i\}_{i=1}^K$ by considering only the first, second, and third terms respectively

Solving $\pi^{(1)}$

- Lagrangian: $L(\boldsymbol{\pi}^{(1)}, \alpha) = \sum_{i=1}^{K} \ln\left(\pi_i^{(1)}\right) \gamma_i^{(1)} - \alpha\left(\sum_{i=1}^{K} \pi_i^{(1)} - 1\right)$
- Taking the partial derivatives of L with respect to $\pi_i^{(1)}$ and α and then setting them to zero we have $\frac{\gamma_i^{(1)}}{\pi_i^{(1)}} - \alpha = 0 \Rightarrow \pi_i^{(1)} = \frac{\gamma_i^{(1)}}{\alpha}$ for all $1 \leqslant i \leqslant K$ and $\sum_{i=1}^{K} \pi_i^{(1)} = 1$
- Summing all equations for $\pi_i^{(1)}$ we have $\alpha = \sum_{i=1}^{K} \gamma_i^{(1)}$, and therefore $\pi_i^{(1)} = \frac{\gamma_i^{(1)}}{\sum_{i=1}^{K} \gamma_i^{(1)}} = \gamma_i^{(1)}$

Solving A

- Lagrangian: $L(\boldsymbol{A}, \{\alpha_i\}_{i=1}^{K}) =$
$\sum_{t=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{K} \ln(a_{i,j})\, \xi_{i,j}^{(t)} - \sum_{i=1}^{K} \alpha_i \left(\sum_{j=1}^{K} a_{i,j} - 1 \right)$
- Taking the partial derivatives of L with respect to $a_{i,j}$ and α_i and then setting them to zero we have $\frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(t)}}{a_{i,j}} - \alpha_i = 0 \Rightarrow a_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(t)}}{\alpha_i}$ and $\sum_{j=1}^{K} a_{i,j} = 1$ for all $1 \leqslant i \leqslant K$
- Summing the equations for $a_{i,j}$ along j we have $\alpha_i = \sum_{j=1}^{K} \sum_{t=1}^{T-1} \xi_{i,j}^{(t)}$, and therefore $a_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(t)}}{\sum_{t=1}^{T-1} \sum_{j=1}^{K} \xi_{i,j}^{(t)}} = \frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(t)}}{\sum_{t=1}^{T-1} \gamma_i^{(t)}}$

Solving $\{\theta_i\}_{i=1}^K$ (1/3)

- Problem: finding $\{\theta_k\}_{k=1}^K$ such that $\sum_{t=1}^T \sum_{k=1}^K \ln\left(P[\boldsymbol{x}^{(t)}|z_k^{(t)},\theta_k]\right)\gamma_k^{(t)}$ is maximized
- Suppose $\boldsymbol{x}^{(t)}$ is discrete such that $x_i^{(t)}=1$ if the predefined value O_i from $\{O_1,\cdots,O_d\}$ is observed; 0 otherwise
- We can assume that the emission probability $P[\boldsymbol{x}^{(t)}|z_k^{(t)},\theta_k]=\prod_{i=1}^d b_{k,i}^{x_i^{(t)}}$ follows the multinomial distribution where $\theta_k=\{b_{k,i}\}_{i=1}^d$ and $b_{k,i}$ is the probability that O_i is observed in state k
    - $\sum_{i=1}^d b_{k,i}=1$

Solving $\{\theta_i\}_{i=1}^K$ (2/3)

- Lagrangian: $L(\{b_{k,i}\}_{k=1,i=1}^{K,d}, \{\alpha_k\}_{k=1}^K) =$
$\sum_{t=1}^T \sum_{k=1}^K \ln \left(\prod_{i=1}^d b_{k,i}^{x_i^{(t)}} \right) \gamma_k^{(t)} - \sum_{k=1}^K \alpha_k \left(\sum_{i=1}^d b_{k,i} - 1 \right) =$
$\sum_{t=1}^T \sum_{k=1}^K \sum_{i=1}^d x_i^{(t)} \ln (b_{k,i}) \gamma_k^{(t)} - \sum_{k=1}^K \alpha_k \left(\sum_{i=1}^d b_{k,i} - 1 \right)$
- Taking the partial derivatives of L with respect to $b_{k,i}$ and α_k and then setting them to zero we have
$\frac{\sum_{t=1}^T x_i^{(t)} \gamma_k^{(t)}}{b_{k,i}} - \alpha_k = 0 \Rightarrow b_{k,i} = \frac{\sum_{t=1}^T x_i^{(t)} \gamma_k^{(t)}}{\alpha_k}$ and $\sum_{i=1}^d b_{k,i} = 1$ for all $1 \leqslant k \leqslant K$
- Summing the equations for $b_{k,i}$ along i we have
$\alpha_k = \sum_{i=1}^d \sum_{t=1}^T x_i^{(t)} \gamma_k^{(t)}$, and therefore
$b_{k,i} = \frac{\sum_{t=1}^T x_i^{(t)} \gamma_k^{(t)}}{\sum_{t=1}^T \gamma_k^{(t)} \sum_{i=1}^d x_i^{(t)}} = \frac{\sum_{t=1}^T x_i^{(t)} \gamma_k^{(t)}}{\sum_{t=1}^T \gamma_k^{(t)}}$

Solving $\{\theta_i\}_{i=1}^K$ (3/3)

- What if $\boldsymbol{x}^{(t)}$ are continuous?

Solving $\{\theta_i\}_{i=1}^K$ (3/3)

- What if $\boldsymbol{x}^{(t)}$ are continuous?
- We can assume that $P[\boldsymbol{x}^{(t)}|z_k^{(t)}, \theta_k]$ follows the multivariate normal distribution where $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- It can be shown that
  - $\boldsymbol{\mu}_k = \frac{\sum_{t=1}^T \gamma_k^{(t)} \boldsymbol{x}^{(t)}}{\sum_{t=1}^T \gamma_k^{(t)}}$
  - $\boldsymbol{\Sigma}_k = \frac{\sum_{t=1}^T \gamma_k^{(t)} (\boldsymbol{x}^{(t)} - \boldsymbol{\mu}_k)(\boldsymbol{x}^{(t)} - \boldsymbol{\mu}_k)^\top}{\sum_{t=1}^T \gamma_k^{(t)}}$ [Homework]

Learning form Multiple Sequences

- Suppose we are given a set $\mathcal{X} = \{\boldsymbol{x}^{(n,t)}\}_{n=1,t=1}^{N,T}$ of observation sequences, where sequences are independent with each other
  - $P[\mathcal{X}|\Theta] = \prod_{n=1}^{N} P[\mathcal{X}^{(n)}|\Theta]$, where $\mathcal{X}^{(n)} = \{\boldsymbol{x}^{(n,t)}\}_{t=1}^{T}$
- Then for discrete $\boldsymbol{x}^{(t)}$ with multinomial emission probability:
  - $\pi_i^{(1)} = \frac{\sum_{n=1}^{N} \gamma_i^{(n,1)}}{N}$
  - $a_{i,j} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \xi_{i,j}^{(n,t)}}{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \gamma_i^{(n,t)}}$
  - $b_{k,i} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T} x_i^{(n,t)} \gamma_k^{(n,t)}}{\sum_{n=1}^{N} \sum_{t=1}^{T} \gamma_k^{(n,t)}}$
- This is analogous to the estimators of a Markov chain we have seen previously, except that $\gamma_i^{(t)} = P[z_i^{(t)}|\mathcal{X}, \Theta^{old}]$ and $\xi_{i,j}^{(t)} = P[z_i^{(t)}, z_j^{(t+1)}|\mathcal{X}, \Theta^{old}]$ are ***soft counts*** of state visits

Outline

We Are Not Done Yet

- Given Θ^{old}, in the E-step we need to evaluate $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$
  - $\gamma_i^{(t)} = P[z_i^{(t)}|\mathcal{X},\Theta^{old}] = \frac{P[\mathcal{X},z_i^{(t)}|\Theta^{old}]}{P[\mathcal{X}|\Theta^{old}]}$
  - $\xi_{i,j}^{(t)} = P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}] = \frac{P[\mathcal{X},z_i^{(t)},z_j^{(t+1)}|\Theta^{old}]}{P[\mathcal{X}|\Theta^{old}]}$
- We can evaluate $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$ by considering all possible state sequences:
  - $P[\mathcal{X}|\Theta^{old}] = \sum_{\mathcal{Z}} P[\mathcal{X},\mathcal{Z}|\Theta^{old}]$
  - $P[\mathcal{X},z_i^{(t)}|\Theta^{old}] = \sum_{\mathcal{Z},\boldsymbol{z}^{(t)}=\boldsymbol{e}_i} P[\mathcal{X},\mathcal{Z}|\Theta^{old}]$
- However, there are exponentially many sequences (specifically, K^T and K^{T-1} for $P[\mathcal{X}|\Theta^{old}]$ and $P[\mathcal{X},z_i^{(t)}|\Theta^{old}]$ respectively)
- The evaluation would be very slow, if not infeasible
- Better idea?

We Are Not Done Yet

- Given Θ^{old}, in the E-step we need to evaluate $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$
  - $\gamma_i^{(t)} = P[z_i^{(t)}|\mathcal{X},\Theta^{old}] = \frac{P[\mathcal{X},z_i^{(t)}|\Theta^{old}]}{P[\mathcal{X}|\Theta^{old}]}$
  - $\xi_{i,j}^{(t)} = P[z_i^{(t)},z_j^{(t+1)}|\mathcal{X},\Theta^{old}] = \frac{P[\mathcal{X},z_i^{(t)},z_j^{(t+1)}|\Theta^{old}]}{P[\mathcal{X}|\Theta^{old}]}$
- We can evaluate $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$ by considering all possible state sequences:
  - $P[\mathcal{X}|\Theta^{old}] = \sum_{\mathcal{Z}} P[\mathcal{X},\mathcal{Z}|\Theta^{old}]$
  - $P[\mathcal{X},z_i^{(t)}|\Theta^{old}] = \sum_{\mathcal{Z},\boldsymbol{z}^{(t)}=\boldsymbol{e}_i} P[\mathcal{X},\mathcal{Z}|\Theta^{old}]$
- However, there are exponentially many sequences (specifically, K^T and K^{T-1} for $P[\mathcal{X}|\Theta^{old}]$ and $P[\mathcal{X},z_i^{(t)}|\Theta^{old}]$ respectively)
- The evaluation would be very slow, if not infeasible
- Better idea? for all $t \Rightarrow$belief propagation

Forward-Backward Procedure

- There is an algorithm, called the ***forward-backward procedure***, that provides an efficient way to evaluate $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$
- Given a $\Theta = (\boldsymbol{\pi}^{(1)}, \boldsymbol{A}, \{\theta_i\}_{i=1}^K)$, define the ***forward variable*** as $\alpha_i^{(t)} = P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}, z_i^{(t)} | \Theta]$
  - $\alpha_i^{(t)}$ denotes the probability that a partial sequence $\{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}\}$ until time t is observed while the state ends in S_i at time t
- Similarly, define the ***backward variable*** as $\beta_i^{(t)} = P[\boldsymbol{x}^{(t+1)}, \cdots, \boldsymbol{x}^{(T)} | z_i^{(t)}, \Theta]$
  - $\beta_i^{(t)}$ denotes the probability that a partial sequence $\{\boldsymbol{x}^{(t+1)}, \cdots, \boldsymbol{x}^{(T)}\}$ after time t will be observed given a starting state S_i at time t
- We can express $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$ using the forward/backward variables:

$$\gamma_i^{(t)} = \frac{\alpha_i^{(t)}\beta_i^{(t)}}{\sum_{j=1}^K \alpha_j^{(t)}\beta_j^{(t)}} \text{ and } \xi_{i,j}^{(t)} = \frac{\alpha_i^{(t)} a_{i,j} P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \theta_j] \beta_j^{(t+1)}}{\sum_{k=1}^K \sum_{l=1}^K \alpha_k^{(t)} a_{k,l} P[\boldsymbol{x}^{(t+1)} | z_l^{(t+1)}, \theta_l] \beta_l^{(t+1)}}$$

Expressing $\gamma_i^{(t)}$ Using $\alpha_i^{(t)}$ And $\beta_i^{(t)}$

- $\gamma_i^{(t)} = P[z_i^{(t)}|\mathcal{X},\Theta] = \frac{P[\mathcal{X},z_i^{(t)}|\Theta]}{P[\mathcal{X}|\Theta]} = \frac{P[\mathcal{X}|z_i^{(t)},\Theta]P[z_i^{(t)}|\Theta]}{P[\mathcal{X}|\Theta]}$
  $= \frac{P[\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(t)}|z_i^{(t)},\Theta]P[\boldsymbol{x}^{(t+1)},\cdots,\boldsymbol{x}^{(T)}|z_i^{(t)},\Theta]P[z_i^{(t)}|\Theta]}{P[\mathcal{X}|\Theta]}$
  $= \frac{P[\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(t)}z_i^{(t)},|\Theta]P[\boldsymbol{x}^{(t+1)},\cdots,\boldsymbol{x}^{(T)}|z_i^{(t)},\Theta]}{P[\mathcal{X}|\Theta]}$
  $= \frac{\alpha_i^{(t)}\beta_i^{(t)}}{P[\mathcal{X}|\Theta]} = \frac{\alpha_i^{(t)}\beta_i^{(t)}}{\sum_{j=1}^K P[\mathcal{X},z_j^{(t)}|\Theta]} = \frac{\alpha_i^{(t)}\beta_i^{(t)}}{\sum_{j=1}^K \alpha_j^{(t)}\beta_j^{(t)}}$
- The numerator $\alpha_i^{(t)}\beta_i^{(t)}$ explains the whole observation sequence $\{\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(t)}\}$ and that at time t, the state is S_i
- $\alpha_i^{(t)}\beta_i^{(t)}$ is normalized by dividing over all possible intermediate states at time t to ensure that $\sum_{i=1}^K \gamma_i^{(t)} = 1$

Expressing $\xi_{i,j}^{(t)}$ Using $\alpha_i^{(t)}$ And $\beta_i^{(t)}$ (1/2)

- $$\xi_{i,j}^{(t)} = P[z_i^{(t)}, z_j^{(t+1)} | \mathcal{X}, \Theta] = \frac{P[\mathcal{X}, z_i^{(t)}, z_j^{(t+1)} | \Theta]}{P[\mathcal{X}|\Theta]}$$
$$= \frac{P[\mathcal{X}|z_i^{(t)}, z_j^{(t+1)}, \Theta] P[z_i^{(t)}, z_j^{(t+1)}|\Theta]}{P[\mathcal{X}|\Theta]}$$
$$= \frac{P[\mathcal{X}|z_i^{(t)}, z_j^{(t+1)}, \Theta] P[z_j^{(t+1)}|z_i^{(t)}, \Theta] P[z_i^{(t)}|\Theta]}{P[\mathcal{X}|\Theta]}$$
$$= \left(\frac{1}{P[\mathcal{X}|\Theta]}\right) P[\boldsymbol{x}^{(1)} \cdots, \boldsymbol{x}^{(t)} | z_i^{(t)}, \Theta] P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta]$$
$$P[\boldsymbol{x}^{(t+2)} \cdots, \boldsymbol{x}^{(T)} | z_j^{(t+1)}, \Theta] a_{i,j} P[z_i^{(t)}|\Theta]$$
$$= \frac{P[\boldsymbol{x}^{(1)} \cdots, \boldsymbol{x}^{(t)}, z_i^{(t)}|\Theta] P[\boldsymbol{x}^{(t+1)}|z_j^{(t+1)}, \Theta] P[\boldsymbol{x}^{(t+2)} \cdots, \boldsymbol{x}^{(T)}|z_j^{(t+1)}, \Theta] a_{i,j}}{P[\mathcal{X}|\Theta]}$$
$$= \frac{\alpha_i^{(t)} a_{i,j} P[\boldsymbol{x}^{(t+1)}|z_j^{(t+1)}, \Theta] \beta_j^{(t+1)}}{P[\mathcal{X}|\Theta]} = \frac{\alpha_i^{(t)} a_{i,j} P[\boldsymbol{x}^{(t+1)}|z_j^{(t+1)}, \Theta] \beta_j^{(t+1)}}{\sum_{k=1}^{K} \sum_{l=1}^{K} P[\mathcal{X}, z_k^{(t)}, z_l^{(t+1)}|\Theta]}$$
$$= \frac{\alpha_i^{(t)} a_{i,j} P[\boldsymbol{x}^{(t+1)}|z_j^{(t+1)}, \theta_j] \beta_j^{(t+1)}}{\sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k^{(t)} a_{k,l} P[\boldsymbol{x}^{(t+1)}|z_l^{(t+1)}, \theta_l] \beta_l^{(t+1)}}$$

Expressing $\xi_{i,j}^{(t)}$ Using $\alpha_i^{(t)}$ And $\beta_i^{(t)}$ (2/2)

- $\xi_{i,j}^{(t)} = \frac{\alpha_i^{(t)} a_{i,j} P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \theta_j] \beta_j^{(t+1)}}{\sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k^{(t)} a_{k,l} P[\boldsymbol{x}^{(t+1)} | z_l^{(t+1)}, \theta_l] \beta_l^{(t+1)}}$
- $\alpha_i^{(t)}$ in the numerator explains the first t observations and ends in state S_i at time t
- At time $t+1$, the process moves on to sate S_j with probability $a_{i,j}$, and generates the $(t+1)$st observation
- Continue from S_j at time $t+1$, $\beta_j^{(t+1)}$ explains the rest observations
- Finally, normalize by dividing all possible pairs of states at time t and $t+1$

Evaluating $\alpha_i^{(t)}$ And $\beta_i^{(t)}$ (1/2)

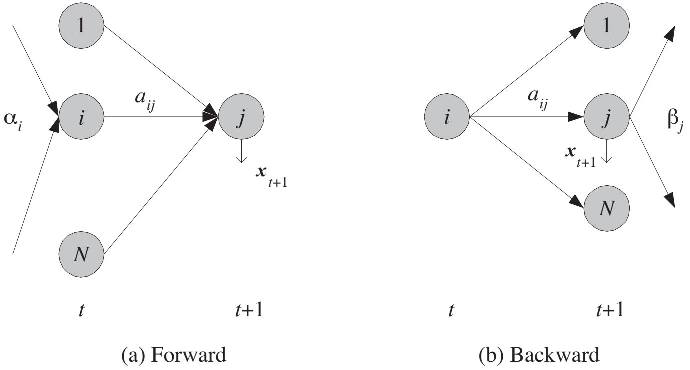
- The merit of the forward-backward procedure is that $\alpha_i^{(t)}$ and $\beta_i^{(t)}$ can be evaluated efficiently
- $\alpha_i^{(1)} = P[\boldsymbol{x}^{(1)}, z_i^{(1)} | \Theta] = P[\boldsymbol{x}^{(1)} | z_i^{(1)}, \Theta] P[z_i^{(1)} | \Theta] = P[\boldsymbol{x}^{(1)} | z_i^{(1)}, \theta_i] \pi_i^{(1)}$
- $\alpha_j^{(t+1)} = P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t+1)}, z_j^{(t+1)} | \Theta] =$

$$
\begin{aligned}
& P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] P[z_j^{(t+1)} | \Theta] \\
&= P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)} | z_j^{(t+1)}, \Theta] P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] P[z_j^{(t+1)} | \Theta] \\
&= P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}, z_j^{(t+1)} | \Theta] P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] \\
&= P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] \sum_{i=1}^{K} P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}, z_i^{(t)}, z_j^{(t+1)} | \Theta] \\
&= P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] \sum_{i=1}^{K} P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}, z_j^{(t+1)} | z_i^{(t)}, \Theta] P[z_i^{(t)} | \Theta] \\
&= P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] \\
&\qquad \sum_{i=1}^{K} P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)} | z_i^{(t)}, \Theta] P[z_j^{(t+1)} | z_i^{(t)}, \Theta] P[z_i^{(t)} | \Theta] \\
&= P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \Theta] \sum_{i=1}^{K} P[\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}, z_i^{(t)} | \Theta] P[z_j^{(t+1)} | z_i^{(t)}, \Theta] \\
&= \left(\sum_{i=1}^{K} \alpha_i^{(t)} a_{i,j} \right) P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \theta_j]
\end{aligned}
$$

Evaluating $\alpha_i^{(t)}$ And $\beta_i^{(t)}$ (2/2)

- Based on the recurrence relation $\begin{cases} \alpha_i^{(1)} = P[\mathbf{x}^{(1)} | z_i^{(1)}, \theta_i] \pi_i^{(1)} \\ \alpha_j^{(t+1)} = \left(\sum_{i=1}^{K} \alpha_i^{(t)} a_{i,j} \right) P[\mathbf{x}^{(t+1)} | z_j^{(t+1)}, \theta_j] \end{cases}$, $\alpha_j^{(t+1)}$ has the ***optimal substructure*** that it can be evaluated efficiently within $O(K)$ time if all $\alpha_i^{(t)}$, $1 \leqslant i \leqslant K$, are known
- We can evaluate all $\alpha_i^{(t)}$, $1 \leqslant i \leqslant K$ and $1 \leqslant t \leqslant T$, within $O(K^2 T)$ time using the dynamic programming from $t = 1$ to T
- Similarly, we can derive the recurrence relation for $\beta_i^{(t)}$ as $\begin{cases} \beta_i^{(T)} = 1 \\ \beta_i^{(t)} = \sum_{j=1}^{K} a_{i,j} P[\mathbf{x}^{(t+1)} | z_j^{(t+1)}, \theta_j] \beta_j^{(t+1)} \end{cases}$ [Proof]
- All $\beta_i^{(t)}$ can also be evaluated within $O(K^2 T)$ time from $t = T$ to 1
- Once obtaining $\alpha_i^{(t)}$ and $\beta_i^{(t)}$, we can derive all $\gamma_i^{(t)}$ and $\xi_{i,j}^{(t)}$ within $O(K)$ and $O(K^2)$ time respectively
- The total time complexity for an E-step is $O(K^2 T)$

The recurrence relations



Outline

Problem Formulation

- Problem: given a sequence $\mathcal{X}$ and parameters Θ, we want to infer the hidden state sequence $\mathcal{Z}^* = \{z^{(t)}\}_t^T$ such that it has the highest posterior $P[\mathcal{Z}|\mathcal{X},\Theta]$
  - This helps us understand the "reason" behind $\mathcal{X}$
  - A common task in time series analysis
- Since $P[\mathcal{Z}|\mathcal{X},\Theta] = \frac{P[\mathcal{X},\mathcal{Z}|\Theta]}{P[\mathcal{X}|\Theta]}$ and $P[\mathcal{X}|\Theta]$ is independent with $\mathcal{Z}$, we only need to find $\mathcal{Z}^*$ maximizing $P[\mathcal{X},\mathcal{Z}|\Theta]$
- Objective: $\arg_{\mathcal{Z}} \max P[\mathcal{X},\mathcal{Z}|\Theta]$
- We can try out all possible $\mathcal{Z}$, at the cost of exponential time complexity
- Efficient solution?

The Optimal Substructure

- $P[\mathcal{X},\mathcal{Z}|\Theta] = P[\mathcal{X}|\mathcal{Z},\Theta]P[\mathcal{Z}|\Theta]$
  $= \left(\prod_{s=1}^{T} P[\boldsymbol{x}^{(s)}|\boldsymbol{z}^{(s)},\theta_{d(\boldsymbol{z}^{(s)})}]\right) P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}] \left(\prod_{t=1}^{T-1} P[\boldsymbol{z}^{(t+1)}|\boldsymbol{z}^{(t)},\boldsymbol{A}]\right)$
  $= \left(P[\boldsymbol{z}^{(1)}|\boldsymbol{\pi}^{(1)}]P[\boldsymbol{x}^{(1)}|\boldsymbol{z}^{(1)},\theta_{d(\boldsymbol{z}^{(1)})}]\right)$
  $\left(\prod_{t=1}^{T-1} a_{d(\boldsymbol{z}^{(t)}),d(\boldsymbol{z}^{(t+1)})} P[\boldsymbol{x}^{(t+1)}|\boldsymbol{z}^{(t+1)},\theta_{d(\boldsymbol{z}^{(t+1)})}]\right)$
- Define
  $\delta_i^{(t)} = \max_{\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(t-1)}} P[\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(t)},\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(t-1)},\boldsymbol{z}^{(t)} = \boldsymbol{e}_i|\Theta]$
  - $\mathcal{Z}^*$ is the sequence having $\delta^{(T)*} = \max_{1\leqslant i\leqslant K} \delta_i^{(T)}$
- Notice that we can calculate $\delta_j^{(T)}$ efficiently if we already know $\delta_i^{(T-1)}$ for all $1 \leqslant i \leqslant K$
  - $\delta_j^{(t)} = \left(\max_{1\leqslant i\leqslant K} \delta_i^{(t-1)} a_{i,j}\right) P[\boldsymbol{x}^{(t)}|z_j^{(t)},\theta_j]$
- $\delta_j^{(t)}$ has the optimal substructure and can be evaluated efficiently using dynamic programming
  - We can obtain $\mathcal{Z}^*$ by backtracking

The Viterbi Algorithm

Input: $\mathcal{X} \leftarrow \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ and $\Theta \leftarrow (\boldsymbol{\pi}^{(1)}, \boldsymbol{A}, \{\theta_i\}_{i=1}^{K})$
Output: $\mathcal{Z} \leftarrow \{\boldsymbol{z}^{(t)}\}_{t=1}^{T}$ resulting the highest $P[\mathcal{X}, \mathcal{Z} | \Theta]$

for $i \leftarrow 1$ **to** K **do**
  $\delta_i^{(1)} \leftarrow \pi_i^{(1)} P[\boldsymbol{x}^{(1)} | z_i^{(1)}, \theta_i]$;
  $\psi_i^{(1)} \leftarrow$ **null**;
end
for $t \leftarrow 2$ **to** T **do**
  for $j \leftarrow 1$ **to** K **do**
    $\delta_j^{(t)} \leftarrow \left(\max_{1 \leqslant i \leqslant K} \delta_i^{(t-1)} a_{i,j} \right) P[\boldsymbol{x}^{(t)} | z_j^{(t)}, \theta_j]$;
    $\psi_j^{(t)} \leftarrow \arg_{1 \leqslant i \leqslant K} \max \delta_i^{(t-1)} a_{i,j}$; `// previous state`
  end
end
$d(\boldsymbol{z}^{(T)}) \leftarrow \arg_{1 \leqslant i \leqslant K} \max \delta_i^{(T)}$;
for $t \leftarrow T-1$ **to** 1 **do**
  $d(\boldsymbol{z}^{(t)}) \leftarrow \psi_{d(\boldsymbol{z}^{(t+1)})}^{(t+1)}$; `// backtracking`
end

Algorithm 1: The Viterbi algorithm of time complexity $O(K^2 T)$.

Outline

Making Predictions

- When HMM is used to model a class, we predict a new sequence $\mathcal{X}^{new}$ to be in class C_i if the posterior $P[C_i|\mathcal{X}^{new}] \propto P[\mathcal{X}^{new}|\Theta_i]P[C_i]$ is the highest
- Problem: given parameters Θ and a sequence $\mathcal{X}$, we want to know $P[\mathcal{X}|\Theta]$
- Again, we can try out all possible $\mathcal{Z}$ using $P[\mathcal{X}|\Theta] = \sum_{\mathcal{Z}} P[\mathcal{X},\mathcal{Z}|\Theta]$, but this is cost prohibitive
- Better way?

Making Predictions

- When HMM is used to model a class, we predict a new sequence $\mathcal{X}^{new}$ to be in class C_i if the posterior $P[C_i|\mathcal{X}^{new}] \propto P[\mathcal{X}^{new}|\Theta_i]P[C_i]$ is the highest
- Problem: given parameters Θ and a sequence $\mathcal{X}$, we want to know $P[\mathcal{X}|\Theta]$
- Again, we can try out all possible $\mathcal{Z}$ using $P[\mathcal{X}|\Theta] = \sum_{\mathcal{Z}} P[\mathcal{X}, \mathcal{Z}|\Theta]$, but this is cost prohibitive
- Better way?
- Notice that $P[\mathcal{X}|\Theta] = \sum_{i=1}^{K} P[\mathcal{X}, z_i^{(t)}|\Theta] = \sum_{i=1}^{K} \alpha_i^{(T)}$
    - Calculate the forward variables $\alpha_i^{(T)}$ for all $1 \leqslant i \leqslant K$ first, which takes $O(K^2 T)$ time
    - Obtain $P[\mathcal{X}|\Theta]$ by summing $\alpha_i^{(T)}$

Outline

Implementation Issues

- When calculating $\alpha_j^{(t)}$, $\beta_i^{(t)}$, and $\delta_j^{(t)}$ in a program, we risk getting the underflow
  - $\alpha_j^{(t)} = \left(\sum_{i=1}^K \alpha_i^{(t-1)} a_{i,j}\right) P[\boldsymbol{x}^{(t)} | z_j^{(t)}, \theta_j]$, $\beta_i^{(t)} = \sum_{j=1}^K a_{i,j} P[\boldsymbol{x}^{(t+1)} | z_j^{(t+1)}, \theta_j] \beta_j^{(t+1)}$, and $\delta_j^{(t)} = \left(\max_{1 \leqslant i \leqslant K} \delta_i^{(t-1)} a_{i,j}\right) P[\boldsymbol{x}^{(t)} | z_j^{(t)}, \theta_j]$ are all multiplication of small numbers
- We can calculate the normalized $\widetilde{\alpha}_i^{(t)}$ and $\widetilde{\beta}_i^{(t)}$ by multiplying $\alpha_i^{(t)}$ and $\beta_i^{(t)}$ by $c_t = \sum_{j=1}^K \frac{1}{\alpha_j^{(t)}}$ (note $\sum_{j=1}^K \beta_j^{(t)} \neq 1$) at each step of the dynamic programming, and then denormalize the related targets
  - E.g., since $\widetilde{\alpha}_i^{(T)} = \alpha_i^{(T)} \prod_{t=1}^T c_t$ and $\sum_{i=1}^K \widetilde{\alpha}_i^{(T)} = 1$, we denormalize $P[\mathcal{X}|\Theta]$ by $P[\mathcal{X}|\Theta] = \sum_{i=1}^K \alpha_i^{(T)} = \frac{1}{\prod_{t=1}^T c_t} \sum_{i=1}^K \widetilde{\alpha}_i^{(T)} = \frac{1}{\prod_{t=1}^T c_t}$
- For $\delta_j^{(t)}$, we can simply calculate $\widetilde{\delta}_j^{(t)} = \log \delta_j^{(t)}$ at each step, and then exponent the related targets

Model Selection

- Reduce the number of states, K
  - The optimal K can be determined using the cross validation
- Or, constrain the model structure
  - Limit the number of states K, $K' < K$, that can be transited to
  - This reduces the complexity of forward-backward procedure and Viterbi algorithm to $O(KK'T)$
  - In particular, the ***left-to-right HMM*** is commonly used (e.g., in speech recognition)

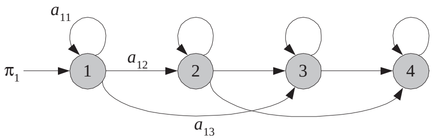


Figure : An example left-to-right HMM. The process never moves to a state with a smaller index (i.e., $a_{i,j} = 0$ if $j < i$), and a big jump in state index is not allowed (i.e., $a_{i,j} = 0$ for $j > i + c$, where $c = 2$ in this case).