PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 13: SEQUENTIAL DATA

Learning Objectives

1、What are hidden Markov models (HMMs)?

2、What is the EM scheme for HMMs?

3、What are Forward-Backward and Sum-Product Algorithms?

4、What are Viterbi and Max-Product Algorithms?

5、What are linear dynamic systems?

6、What are Kalman and particle filters?

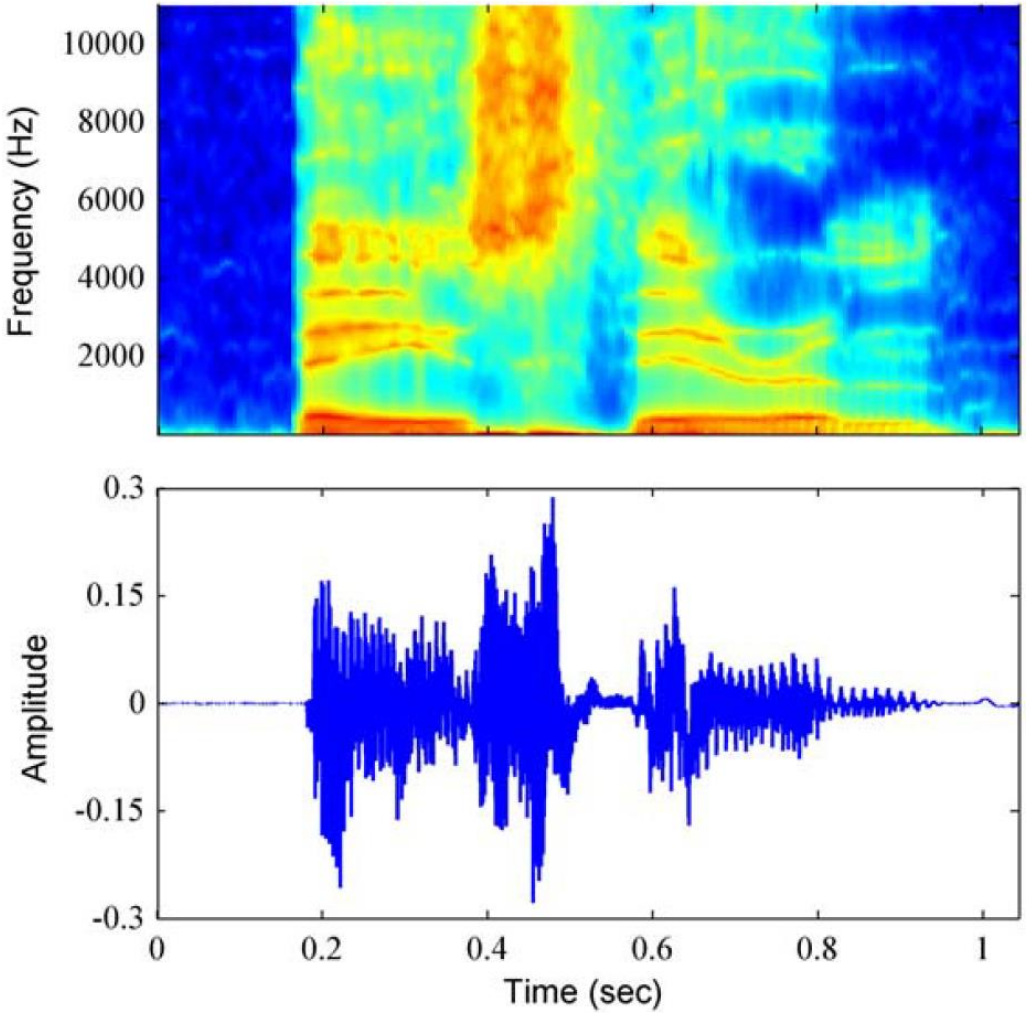
7、How to learn linear dynamic system models?

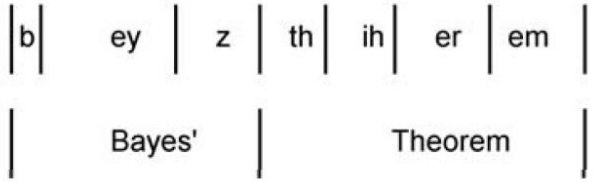
8、What are RNN and LSTM?

Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM

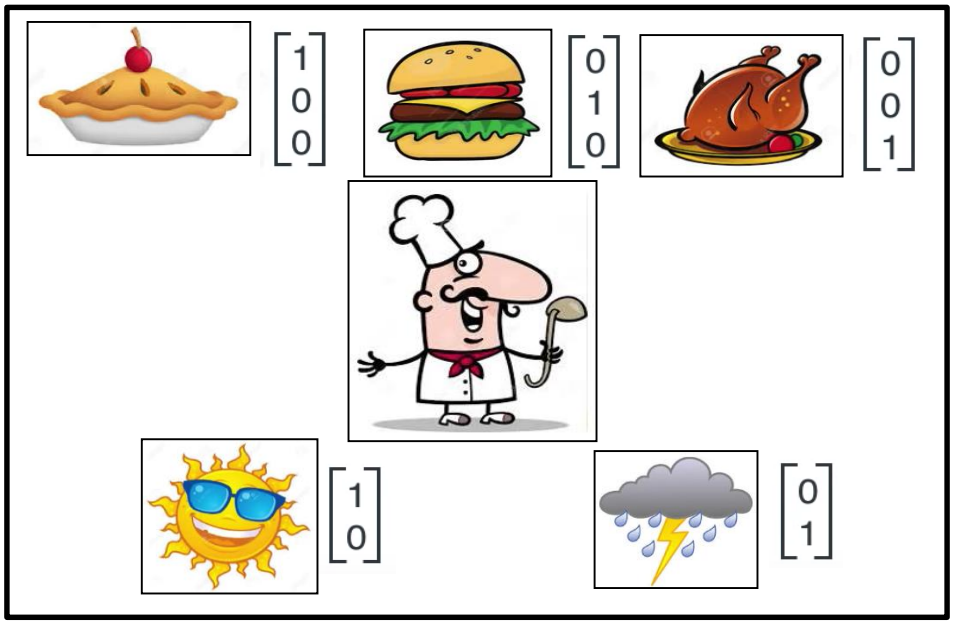
Sequential Data

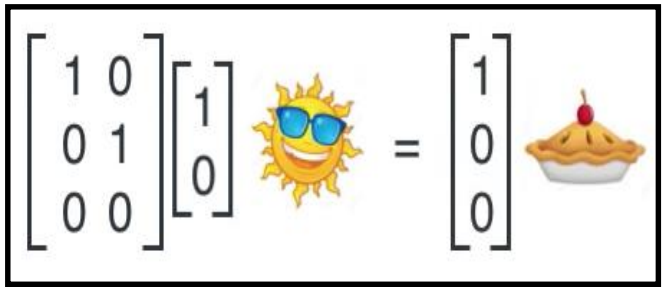


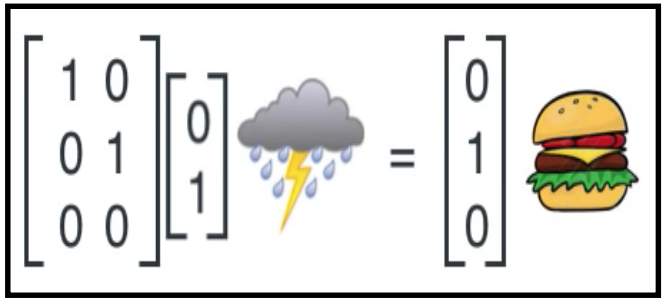


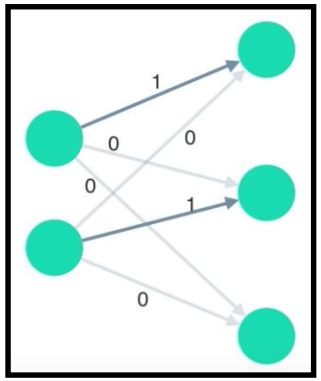
Spectral and temporal signals of Bayes' Theorem

Conditional Data

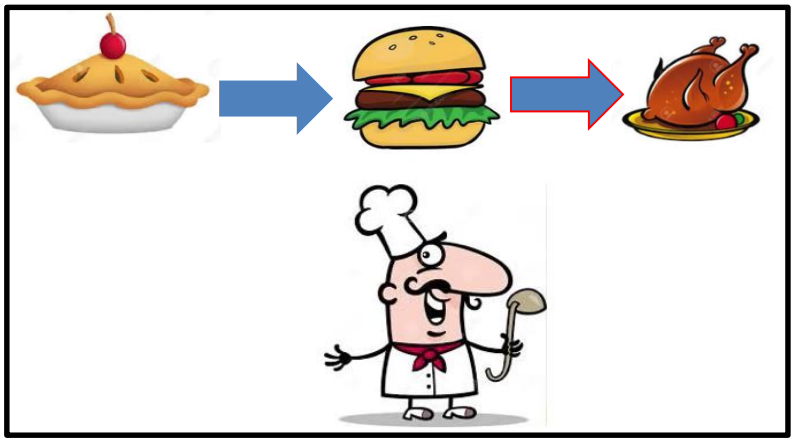


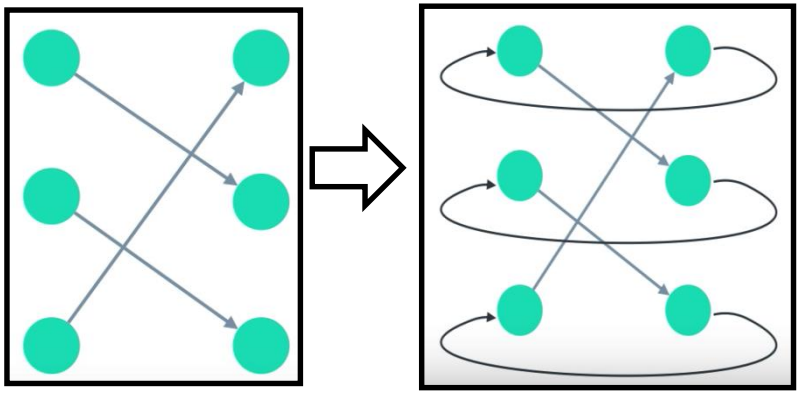






Sequential Data





$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

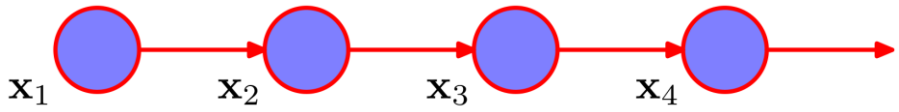
$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

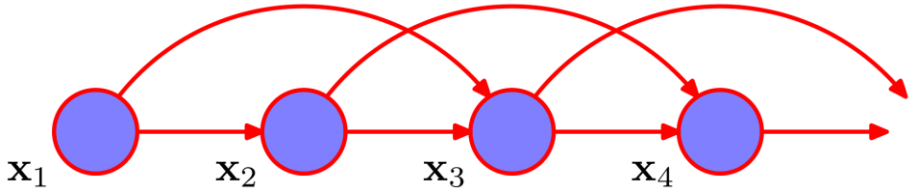
Markov Models

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

$$p(\mathbf{x}_n | \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1})$$



A second-order Markov chain

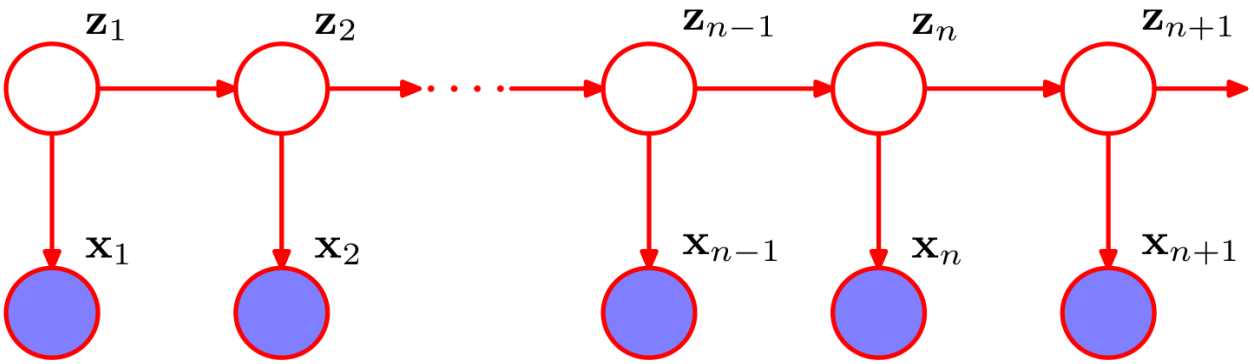


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) p(\mathbf{x}_2 | \mathbf{x}_1) \prod_{n=3}^{N} p(\mathbf{x}_n | \mathbf{x}_{n-1}, \mathbf{x}_{n-2})$$

Hidden Markov Models

Using a Markov chain of latent variables

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n)$$



For continuous variables, we can use linear-Gaussian conditional distributions

Latent Markov Model

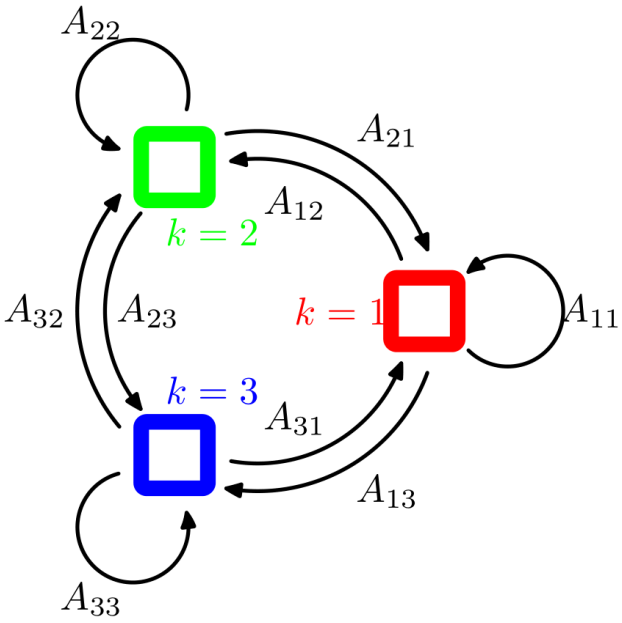
Conditional distribution for latent variable

$$p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}$$

$$p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$$

$$\sum_k \pi_k = 1$$

A means transition probabilities



As in the case of a standard mixture model, the latent variables are the discrete multinomial variables $\mathbf{z}_n$ using the 1-of-K coding scheme

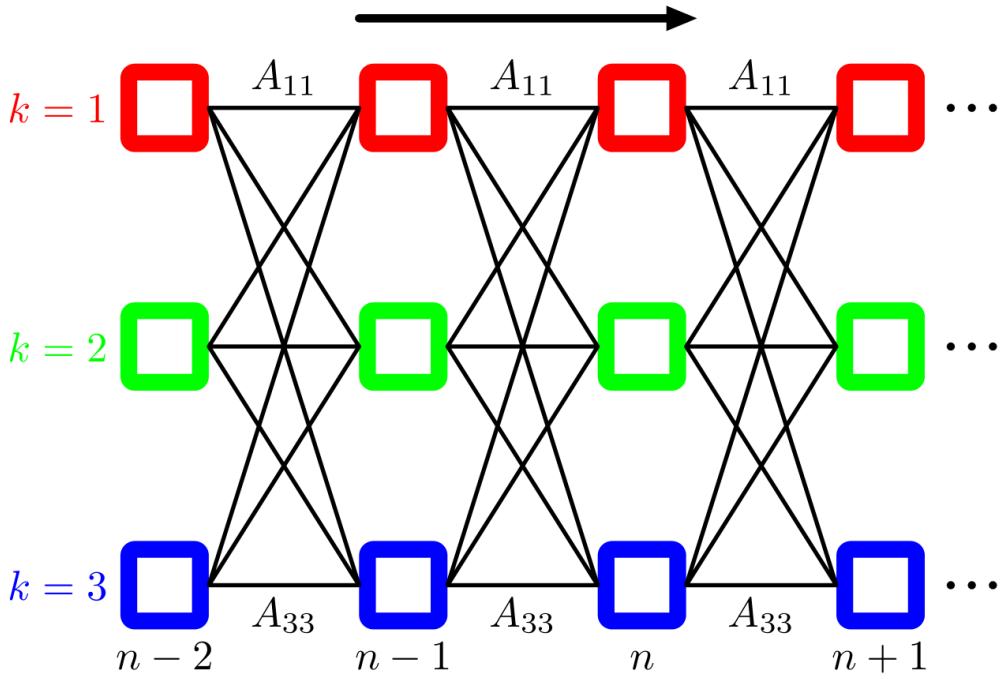
A model whose latent variables have three possible states corresponding to the three boxes. The black lines denote the elements of the transition matrix A_{jk}

Latent State Lattice

Latent states lattice: representing the transitions between latent states

Each column of this diagram corresponds to one of the latent variables $\mathbf{z}_n$

Each row of this diagram corresponds to one state of the latent variables $\mathbf{z}_n$



Hidden Markov Models

Getting emission probabilities from latent variable

$$p(\mathbf{x}_n|\mathbf{z}_n,\boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{z_{nk}} \qquad p(\mathbf{z}_n|\mathbf{z}_{n-1,\mathbf{A}}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}$$

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right]\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\boldsymbol{\phi})$$

$$\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$$

$$\mathbf{Z} = \{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$$

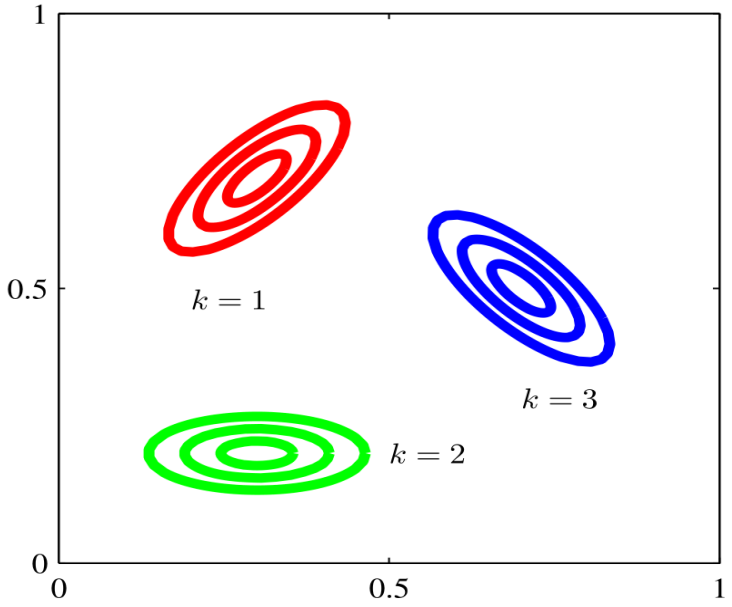
$$\boldsymbol{\theta} = \{\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi}\}$$

Hidden Markov Model Data Sequence

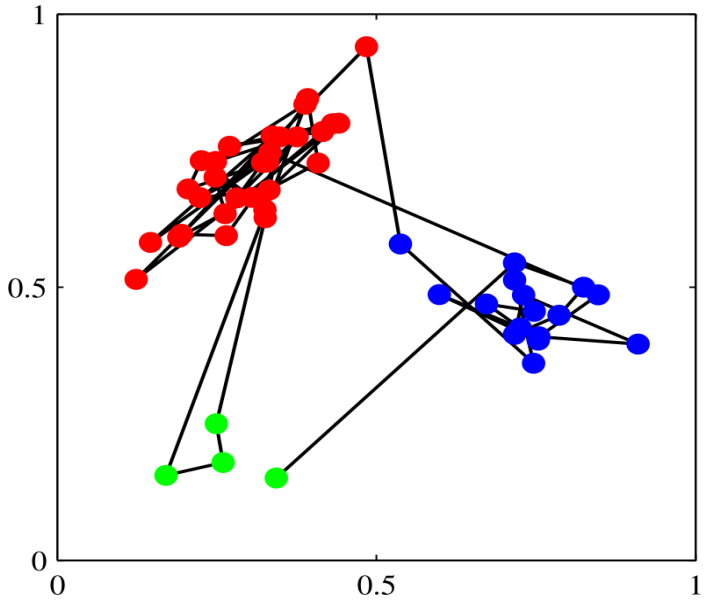
A better understanding of the hidden Markov model

First choose the initial latent variable $\mathbf{z}_1$ $\mathbf{x}_1$ π_k state j

Then, choose the state of the variable $\mathbf{z}_2$ $p(\mathbf{z}_2|\mathbf{z}_1)$ state k A_{jk} $k = 1, \ldots, K$



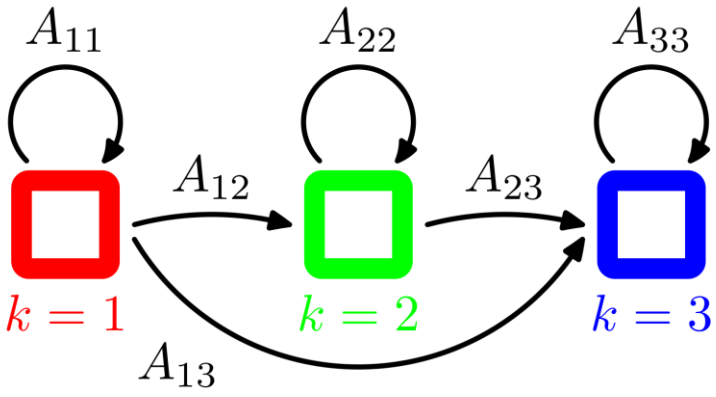
The number of latent variable states is 3;
the dimension of emission variable is 2

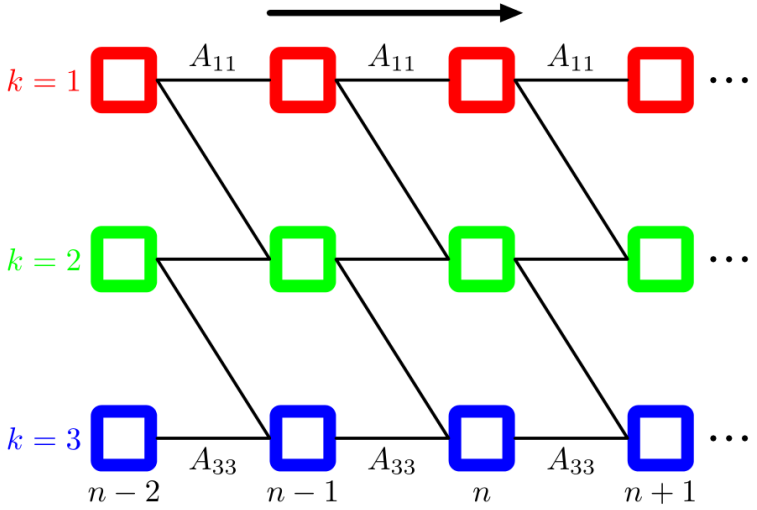


A sequence of 50 samples

Hidden Markov Model Examples

One Example of variants of the standard HMM model





3-state left-to-right hidden Markov model

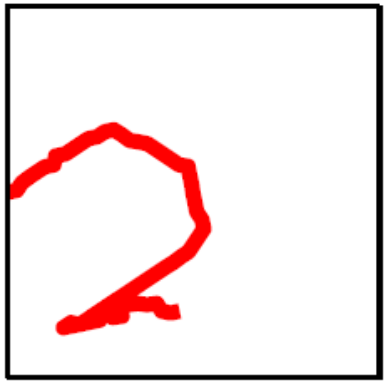
Lattice diagram for a 3-state left- to-right HMM

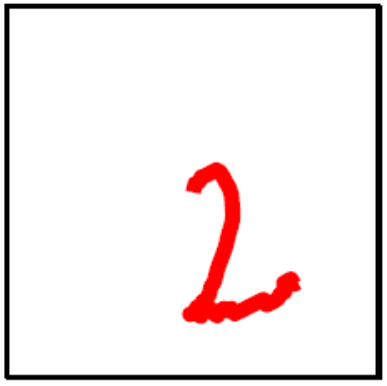
HMM Applications



Handwritten digits

Synthetic digits using HMMs which are trained by using handwritten digits





Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi Algorithm
- Linear Dynamics Systems
- Kalman and Particle Filters
- RNN and LSTM

Three Problems for HMMs

- **Evaluation**:
  Given a HMM model $\theta = \{\pi, \mathbf{A}, \phi\}$, what is likelihood of an observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ generated by that model?

- **Learning**:
  What is the most likely HMM model for an observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$?

- **Decoding**:
  Given a HMM model $\theta = \{\pi, \mathbf{A}, \phi\}$, what is the most likely latent sequence $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ for an observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$?

Expectation of Latent Variables

Expectation of latent variables

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{n-1,j} z_{nk}$$

X: all data samples

Maximum Likelihood of HMMs

Maximum likelihood for the HMM

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

EM algorithm to find an efficient framework

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \quad = \quad & \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} \\ & + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n|\boldsymbol{\phi}_k). \end{aligned}$$

EM Learning of HMMs (I)

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{n-1,j} z_{nk}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

EM Learning of HMMs (II)

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{n-1,j} z_{nk}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{D} \prod_{k=1}^{K} \mu_{ik}^{x_i z_k}$$

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}$$

$$\mu_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

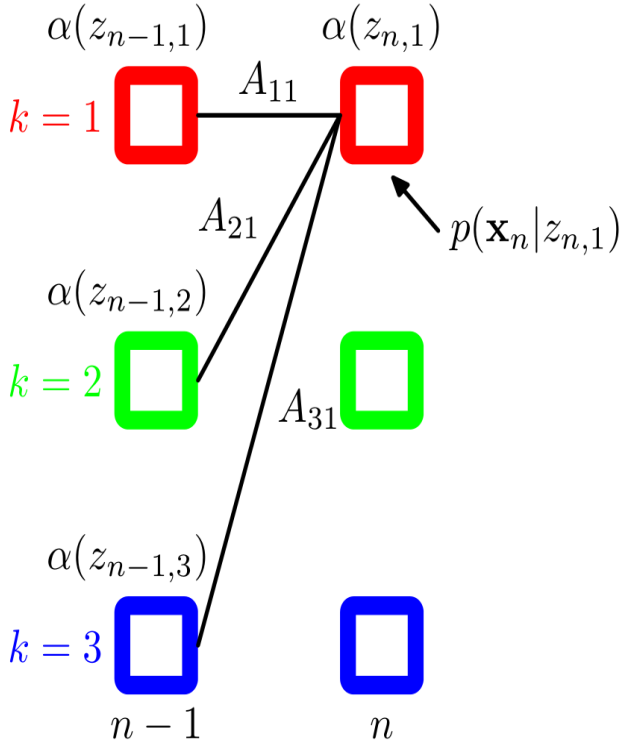
Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM

Forward Recursion (I)

- Forward recursion

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{z}_n) &= p(\mathbf{x}_1,\ldots,\mathbf{x}_n|\mathbf{z}_n) \\
&\quad\; p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n) \\
p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}|\mathbf{x}_n,\mathbf{z}_n) &= p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}|\mathbf{z}_n) \\
p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}|\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}|\mathbf{z}_{n-1}) \\
p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n,\mathbf{z}_{n+1}) &= p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1}) \\
p(\mathbf{x}_{n+2},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1},\mathbf{x}_{n+1}) &= p(\mathbf{x}_{n+2},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1}) \\
p(\mathbf{X}|\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}|\mathbf{z}_{n-1}) \\
&\quad\; p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n) \\
p(\mathbf{x}_{N+1}|\mathbf{X},\mathbf{z}_{N+1}) &= p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \\
p(\mathbf{z}_{N+1}|\mathbf{z}_N,\mathbf{X}) &= p(\mathbf{z}_{N+1}|\mathbf{z}_N)
\end{aligned}
$$



$$
\alpha(\mathbf{z}_1) = p(\mathbf{x}_1,\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) = \prod_{k=1}^{K}\{\pi_k p(\mathbf{x}_1|\boldsymbol{\phi}_k)\}^{z_{1k}}
$$

Forward Recursion (II)

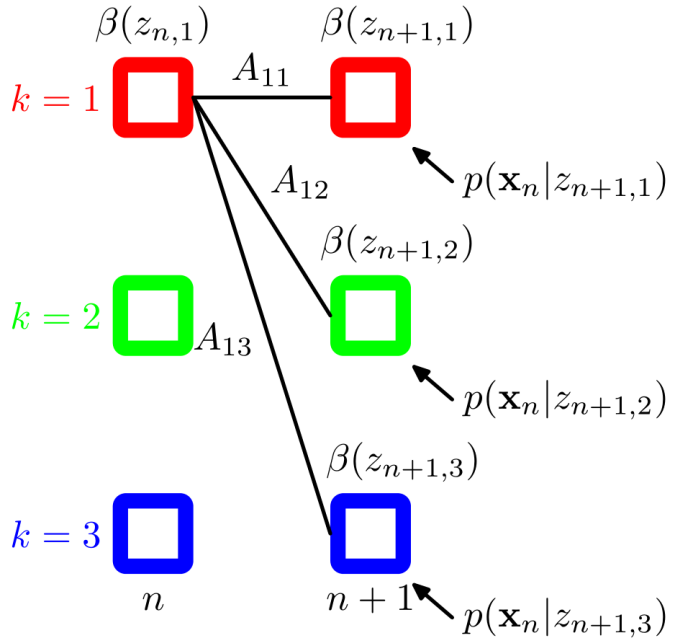
$$
\begin{aligned}
\alpha(\mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_n) p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1})
\end{aligned}
$$

$$
\boxed{\alpha(\mathbf{z}_n) = p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1})}
$$

Backward Recursion

- Backward recursion

$$
\begin{aligned}
\beta(\mathbf{z}_n) &= p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_{n+1} | \mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n, \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1}) p(\mathbf{x}_{n+1} | \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n)
\end{aligned}
$$



$$
\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1} | \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n)
$$

Forward-Backward Estimation

The method of evaluating the quantities of $\gamma(z_{nk})$ $\xi(z_{n-1,j}, z_{nk})$

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X})$$

$$= \frac{\alpha(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$\alpha(\mathbf{z}_n) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$$

$$\beta(\mathbf{z}_n) \equiv p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)$$

$$p(\mathbf{X}) = \sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)$$

$$p(\mathbf{X}) = \sum_{\mathbf{z}_N} \alpha(\mathbf{z}_N)$$

Evaluation

Observation Prediction

$$
\begin{aligned}
p(\mathbf{x}_{N+1}|\mathbf{X}) &= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}, \mathbf{z}_{N+1}|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) p(\mathbf{z}_{N+1}|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}, \mathbf{z}_N|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) p(\mathbf{z}_N|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) \frac{p(\mathbf{z}_N, \mathbf{X})}{p(\mathbf{X})} \\
&= \frac{1}{p(\mathbf{X})} \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) \alpha(\mathbf{z}_N)
\end{aligned}
$$

Sum-Product v.s. Max-Product

- Sum-Product Algorithm (evaluation)
  - ✓ Compute the joint distribution from the Product
  - ✓ Infer marginal distributions from the Sum

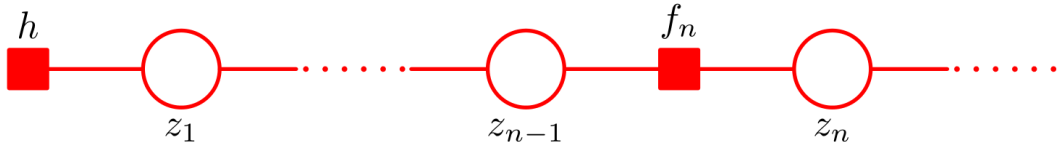
$$p(x_1, x_2) = \sum_{x_3} p(x_1, x_3) p(x_2, x_3)$$

- Max-Product Algorithm (decoding)
  - ✓ Compute the joint distribution from the Product
  - ✓ Perform ML estimation from the Max

$$x_1^* = \max_{x_1} p(x_1, x_3) p(x_2, x_1)$$

Sum-Product for HMMs

- Transforming the directed graph into a factor graph



$$h(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)$$

$$f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)$$

$$\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1}) = \mu_{f_{n-1}\to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1})$$

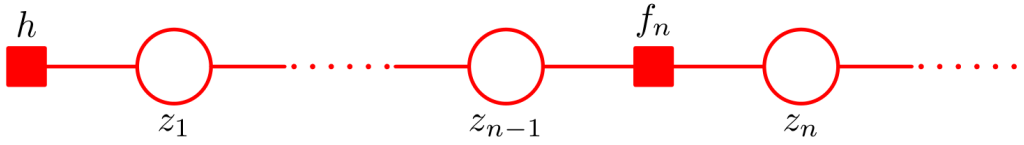
$$\mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} f_n(\mathbf{z}_{n-1}, \mathbf{z}_n)\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1})$$

$$\boxed{\mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} f_n(\mathbf{z}_{n-1}, \mathbf{z}_n)\mu_{f_{n-1}\to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1})}$$

$$\boxed{\alpha(\mathbf{z}_n) = \mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n)}$$

Sum-Product for HMMs

- Transforming the directed graph into a factor graph



$$\mu_{f_{n+1}\to f_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} f_{n+1}(\mathbf{z}_n, \mathbf{z}_{n+1})\mu_{f_{n+2}\to f_{n+1}}(\mathbf{z}_{n+1})$$

$$\boxed{\beta(\mathbf{z}_n) = \mu_{f_{n+1}\to \mathbf{z}_n}(\mathbf{z}_n)}$$

$$\boxed{p(\mathbf{z}_n, \mathbf{X}) = \mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n)\mu_{f_{n+1}\to \mathbf{z}_n}(\mathbf{z}_n) = \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}$$

$$\boxed{\gamma(\mathbf{z}_n) = \frac{p(\mathbf{z}_n, \mathbf{X})}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}}$$

Scaling of HMMs

- Scaling factors for computational practice

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$$

$$\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\alpha(\mathbf{z}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

scaling factors defined by conditional distributions over the observed variables

$$c_n = p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$$

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \widehat{\alpha}(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1})$$

similarly define re-scaled variables

$$\widehat{\beta}(\mathbf{z}_n) = \frac{p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

$$c_{n+1}\widehat{\beta}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \widehat{\beta}(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n)$$

$$\begin{aligned} \gamma(\mathbf{z}_n) &= \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n) \\ \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= c_n^{-1}\widehat{\alpha}(\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) p(\mathbf{z}_n|\mathbf{z}_{-1}) \widehat{\beta}(\mathbf{z}_n) \end{aligned}$$

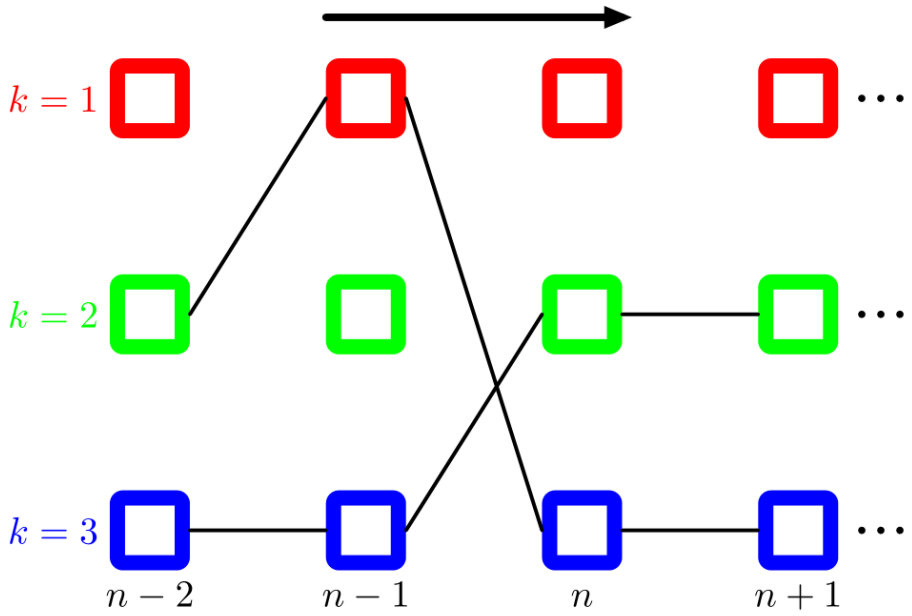
Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM

Latent Sequence Estimation

- **Decoding**:

  Given a HMM model $\theta = \{\pi, \mathbf{A}, \phi\}$, what is the most likely latent sequence $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ for an observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$?



Viterbi Algorithm

$$\omega(\mathbf{z}_n) = \max_{\mathbf{z}_1,\ldots,\mathbf{z}_{n-1}} \ln p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)$$

$$\omega(\mathbf{z}_{n+1}) = \max_{\mathbf{z}_1,\ldots,\mathbf{z}_n} \ln p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_{n+1},\mathbf{z}_1,\ldots,\mathbf{z}_n,\mathbf{z}_{n+1})$$

$$\omega(\mathbf{z}_{n+1}) = \ln p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) + \max_{\mathbf{z}_n} \{\ln p(\mathbf{z}_{n+1}|\mathbf{z}_n) + \omega(\mathbf{z}_n)\}$$

$$\omega(\mathbf{z}_1) = \ln p(\mathbf{z}_1) + \ln p(\mathbf{x}_1|\mathbf{z}_1)$$

Viterbi Algorithm

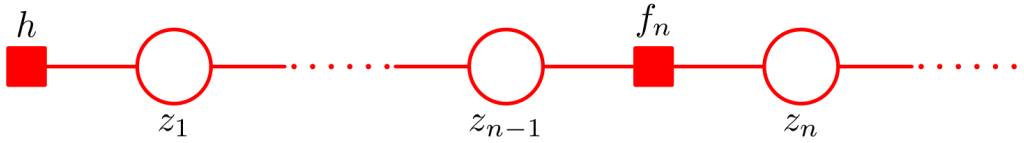
- Note that maximization over $\mathbf{z}_n$ must be performed for each of K possible values of $\mathbf{z}_{n+1}$
- Denote this function by $\psi(k_n)$, where $k \in \{1, \ldots, K\}$
- Once we find the most probable value of $\mathbf{z}_N$, we can trackback along the chain

$$k_n^{\max} = \psi(k_{n+1}^{\max})$$

- Reduce the computational cost from $\mathrm{O}(K^N)$ to $\mathrm{O}(KN)$

Max-Product for HMMs

- Transforming the directed graph into a factor graph



$$
\begin{aligned}
h(\mathbf{z}_1) &= p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) \\
f_n(\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n) \\
\mu_{\mathbf{z}_n\to f_{n+1}}(\mathbf{z}_n) &= \mu_{f_n\to\mathbf{z}_n}(\mathbf{z}_n) \\
\mu_{f_{n+1}\to\mathbf{z}_{n+1}}(\mathbf{z}_{n+1}) &= \max_{\mathbf{z}_n}\left\{\ln f_{n+1}(\mathbf{z}_n,\mathbf{z}_{n+1}) + \mu_{\mathbf{z}_n\to f_{n+1}}(\mathbf{z}_n)\right\}
\end{aligned}
$$

- As such, the $f \to \mathbf{z}$ message will recursively be

$$
\begin{aligned}
\omega(\mathbf{z}_{n+1}) &= \ln p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) + \max_{\mathbf{z}_n}\left\{\ln p(\mathbf{z}_{n+1}|\mathbf{z}_n) + \omega(\mathbf{z}_n)\right\} \\
\omega(\mathbf{z}_n) &\equiv \mu_{f_n\to\mathbf{z}_n}(\mathbf{z}_n)
\end{aligned}
$$

Discriminative HMMs

- Using discriminative rather than Maximum Likelihood techniques

  optimize the cross-entropy of *R* observation sequences, $\mathbf{X}_r$, and labels, m_r

$$\sum_{r=1}^{R} \ln p(m_r|\mathbf{X}_r) \Leftrightarrow \sum_{r=1}^{R} \ln \left\{ \frac{p(\mathbf{X}_r|\boldsymbol{\theta}_r)p(m_r)}{\sum_{l=1}^{M} p(\mathbf{X}_r|\boldsymbol{\theta}_l)p(l_r)} \right\} \begin{array}{l} r = 1, \ldots, R \\ m = 1, \ldots, M \end{array}$$

  for each class there is a HMM, $\boldsymbol{\theta}_m$

- Weakness of the first-order HMM:
  - ✓ distribution of times for which the system remains in a given state
  - ✓ poor at capturing long-range correlations between the observed variables

Example

- Given an HMM and an observation sequence, how to perform evaluation and decoding

Transition A	Emission B	Hidden States Z	Observations X
$\begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}$	$\begin{bmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{bmatrix}$	{bull, bear}	{up, down}

If Z is stationary, then π = [3/7, 4/7]. We also can assume π = [1/2, 1/2]

An observation sequence: {up, up, down}

Evaluation (Sum-Product)

$$\alpha(z_1) = p(z_1, x_1) = p(x_1|z_1)p(z_1)$$

x_1=up, z_1=bull or bear

$$= \begin{bmatrix}0.8\\0.1\end{bmatrix} . \begin{bmatrix}0.5\\0.5\end{bmatrix} = \begin{bmatrix}0.8 \times 0.5\\0.1 \times 0.5\end{bmatrix} = \begin{bmatrix}0.4\\0.05\end{bmatrix}$$

$$\alpha(z_2) = p(z_2, x_1, x_2) = p(x_2|z_2) \sum_{z_1} p(z_2|z_1)\, \alpha(z_1)$$

x_2=up, z_2=bull or bear

$$= \begin{bmatrix}0.8\\0.1\end{bmatrix} . \begin{bmatrix}0.6 \times 0.4 + 0.3 \times 0.05\\0.4 \times 0.4 + 0.7 \times 0.05\end{bmatrix} = \begin{bmatrix}0.204\\0.0195\end{bmatrix}$$

$$\alpha(z_3) = p(z_3, x_1, x_2, x_3) = p(x_3|z_3) \sum_{z_2} p(z_3|z_2)\, \alpha(z_2)$$

x_3=down, z_2=bull or bear

$$= \begin{bmatrix}0.2\\0.9\end{bmatrix} . \begin{bmatrix}0.6 \times 0.204 + 0.3 \times 0.0195\\0.4 \times 0.204 + 0.7 \times 0.0195\end{bmatrix} = \begin{bmatrix}0.02565\\0.085725\end{bmatrix}$$

$$p(x_1, x_2, x_3) = \sum_{z_3} \alpha(z_3) = 0.111375$$

Decoding (Max-Product)

$$\delta(z_1) = p(z_1, x_1) = p(x_1|z_1)p(z_1)$$

x_1=up, z_1=bull or bear

$$= \begin{bmatrix}0.8\\0.1\end{bmatrix} . \begin{bmatrix}0.5\\0.5\end{bmatrix} = \begin{bmatrix}0.8 \times 0.5\\0.1 \times 0.5\end{bmatrix} = \begin{bmatrix}0.4\\0.05\end{bmatrix}$$

$$\delta(z_2) = p(z_2, x_1, x_2) = p(x_2|z_2) \max_{z_1} p(z_2|z_1)\delta(z_1)$$

x_2=up, z_2=bull or bear

$$= \begin{bmatrix}0.8\\0.1\end{bmatrix} . \begin{bmatrix}0.6 \times 0.4\\0.4 \times 0.4\end{bmatrix} = \begin{bmatrix}0.192\\0.016\end{bmatrix} \leftarrow$$

$$\phi(z_2) = \arg \max_{z_1} p(z_2|z_1)\delta(z_1) = \begin{bmatrix}bull \rightarrow bull\\bull \rightarrow bear\end{bmatrix} \leftarrow$$

$$\delta(z_3) = p(z_3, x_1, x_2, x_3) = p(x_3|z_3) \max_{z_2} p(z_3|z_2)\delta(z_2)$$

x_3=down, z_2=bull or bear

$$= \begin{bmatrix}0.2\\0.9\end{bmatrix} . \begin{bmatrix}0.6 \times 0.192\\0.4 \times 0.192\end{bmatrix} = \begin{bmatrix}0.02304\\0.06912\end{bmatrix} \leftarrow$$

$$\phi(z_3) = \arg \max_{z_2} p(z_3|z_2)\delta(z_2) = \begin{bmatrix}bull \rightarrow bull\\bull \rightarrow bear\end{bmatrix} \leftarrow$$

Optimal solution:
bull→bull → bear