기계학습

11강 순차데이터(2)

장필훈 교수

한국방송통신대학교 프라임칼리지

학습목차

1. 순차데이터(2) - hmm
2. 중심극한정리

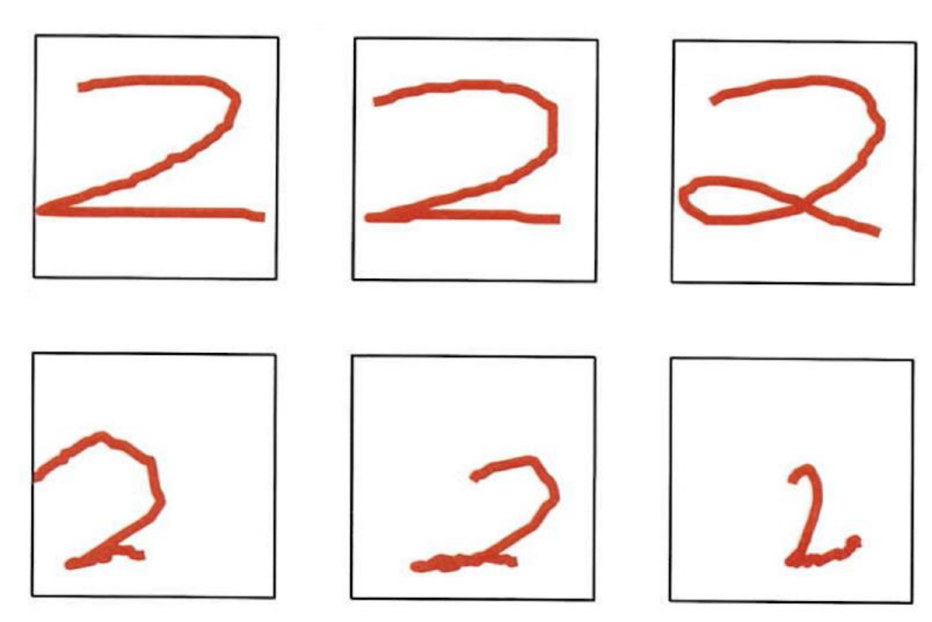
한국방송통신대학교 프라임칼리지

01
순차데이터(2)

한국방송통신대학교 프라임칼리지

1-1 은닉 마르코프 모델(HMM)

- 시간축의 뒤틀림(압축/늘림)에 강하다.



Bishop. fig13.11

- 음성인식에 많이 쓰였음.

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- 관측값 $\mathbf{X}$를 이용해서 매개변수 θ를 정함

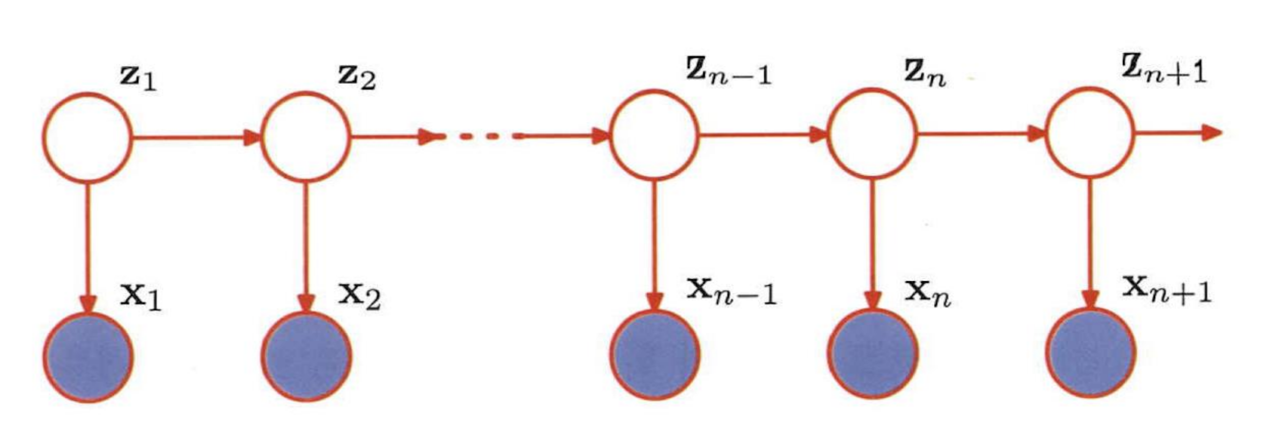
$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

- 우변이 인수분해되지 않아서, 닫힌해를 구하는 것이 어렵다
  - 혼합모델의 경우와 동일(8강 34p)
- EM을 사용한다.

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- X_i가 관찰됨. Z_i가 hidden.(10강 35p)



- Z, X모두 discrete. 가능한 경우의 수를 m_z, m_X라 하자.
  - $w_{m_z \times 1}$: Z_1의 분포
  - A : transition probability. $m_z \times m_z, Z_t \rightarrow Z_{t+1}$

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

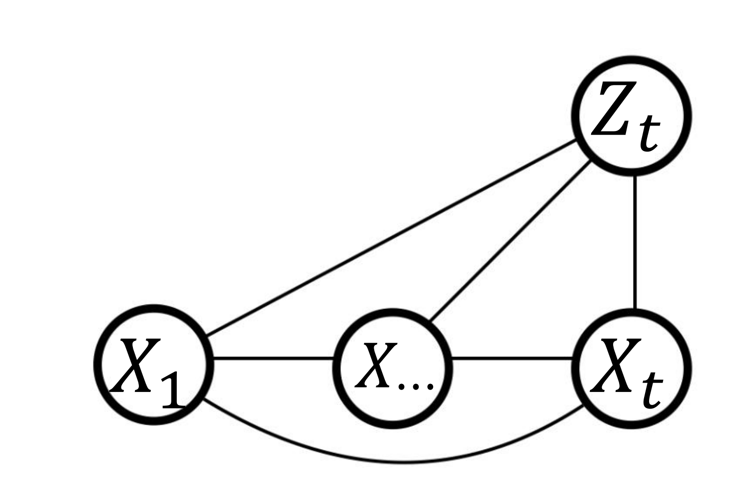
- B : emission probability. $m_z \times m_x, Z_t \to X_t$
- Forward probabilities
  - $\alpha_t(i) = p_\theta(x_1, \dots, x_t, Z_t = i)$
- Backward probabilities
  - $\beta_t(i) = p_\theta(x_{t+1}, \dots, x_n | Z_t = i)$
    - t에서 latent state가 주어졌을 때 미래의 data값?

(ref) Feng Liang, Stat542_W7_HMM

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- w, A, B를 이용해서 θ의 MLE를 어떻게구할 것인가.
- 모든 관측값이 있을 때, latent states의 가능도는?
  - $p_\theta(Z_1, \dots, Z_n|\mathbf{x})$
- forward prob.



한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- forward prob. 계산

$$\alpha_t(i) = p_\theta(x_1, \dots, x_t, Z_t = i)$$

$$\alpha_1(i) = p_\theta(x_1, Z_1 = i)$$

$$= p_\theta(Z_1 = i) \times p_\theta(x_1 | Z_1 = i)$$

$$= w(i)B(i, x_1)$$

(ref) Feng Liang, Stat542_W7_HMM

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- forward prob. 계산(cont.)

$$\alpha_2(i) = p_\theta(x_1, x_2, Z_2 = i)$$

$$= \sum_j p_\theta(x_1, x_2, Z_1 = j, Z_2 = i)$$

$$= \sum_j p_\theta(x_1, Z_1 = j) \times p_\theta(Z_2 | \cancel{x_1}, Z_1 = j) \times p_\theta(x_2 | \cancel{x_1}, \cancel{Z_1 = j}, Z_2 = i)$$

$$= \sum_j \alpha_1(j) A(j, i) B(i, x_2)$$

(ref) Feng Liang, Stat542_W7_HMM

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

- forward prob. 계산(cont.)

$$\alpha_{t+1}(i) = p_\theta(x_1, \cdots, x_{t+1}, Z_{t+1} = i)$$

$$= \sum_j p_\theta(x_1, \cdots, x_{t+1}, Z_t = j, Z_{t+1} = i)$$

$$= \sum_j p_\theta(x_1, \cdots, x_t, Z_t = j) \times p_\theta(Z_{t+1} | \cancel{x_1, \cdots, x_{t+1}}, Z_t = j) \times p_\theta(x_{t+1} | \cancel{x_1, \cdots, x_t, Z_t = j}, Z_{t+1} = i)$$

$$= \sum_j \alpha_1(j) A(j, i) B(i, x_2)$$

(ref) Feng Liang, Stat542_W7_HMM

1-2 HMM에서 최대가능도법

- backward prob. 계산

$$\beta_{n-1}(i) = p_\theta(x_n|Z_{n-1}=i)$$

$$= \sum_j p_\theta(x_n, Z_n = j|Z_{n-1}=i)$$

$$= \sum_j p_\theta(Z_n = j|Z_{n-1}=i) \times p_\theta(x_n|Z_n = j, \cancel{Z_{n-1}=i})$$

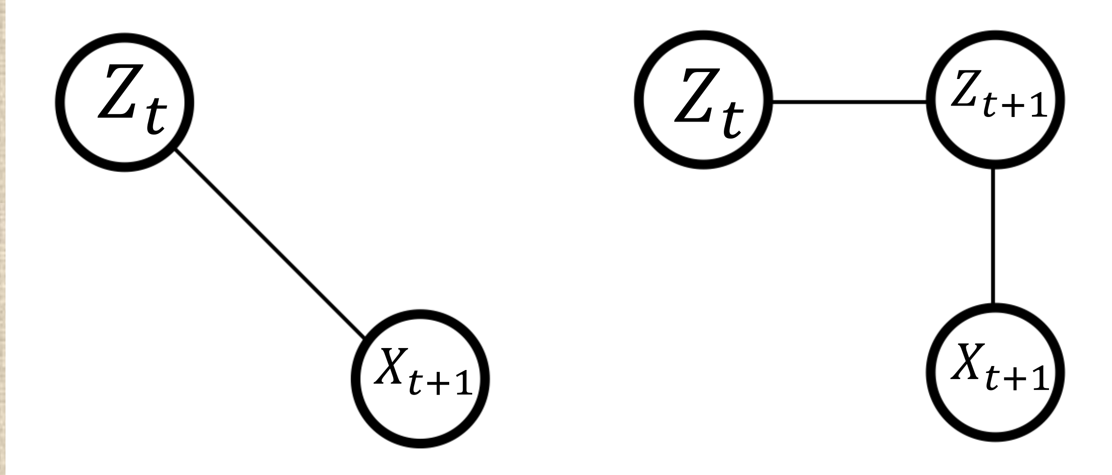
$$= A(i,j)B(j,x_n)$$

$$= A(i,j)B(j,x_n)\beta_n(j) \qquad (\beta_n(*) \equiv 1)$$

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법

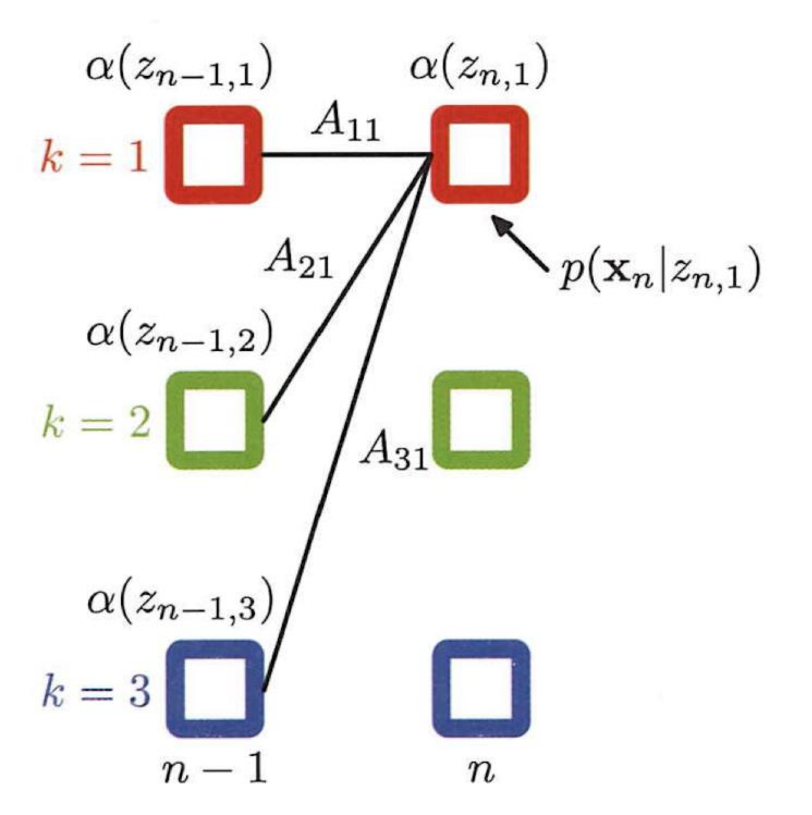
- backward prob. 계산(계속)



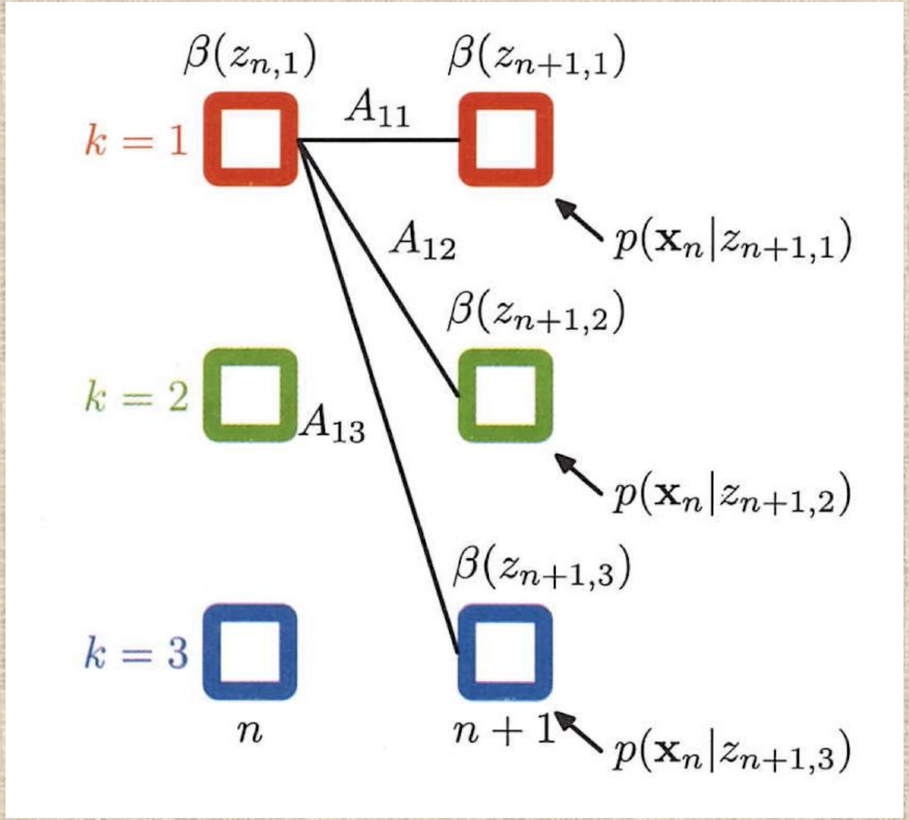
X가 여러개일 때도 동일하게 확장가능

한국방송통신대학교 프라임칼리지

1-2 HMM에서 최대가능도법



Bishop. Fig13.12



Bishop. Fig13.13

1-3 Baum-Welch

- 관찰된 X에 대한 가능도함수

$$\log[p(\mathbf{x}|\theta)] = \log\left[\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z})|\theta)\right]$$

  - 로그 안에 합산이 있다 – 계산이 어려움

- 전체에 대한 가능도 함수(10강 35p)

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

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- 전체에 대한 가능도 함수(계속)

$$\log\left[w(Z_1)\prod_{t=1}^{n-1}A(Z_t,Z_{t+1})\prod_{t=1}^{n}B(Z_t,x_t)\right]$$

$$=\log w(Z_1)+\sum_{t=1}^{n-1}\log A(Z_t,Z_{t+1})+\sum_{t=1}^{n}\log B(Z_t,x_t)$$

- 이것을 최대화하는 알고리즘.

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- γ를 다음과같이 정의

$$\gamma_t(i,j) = p_\theta(Z_t = i, Z_{t+1} = j|\mathbf{x}),$$

$$\gamma_t(i) = p_\theta(Z_t = i|\mathbf{x}) = \sum_j \gamma_t(i,j)$$

- E step

$$\mathbb{E}_{(\mathbf{Z}|\mathbf{x},\theta)} \log p(\mathbf{Z},\mathbf{x}|\theta)$$

$$= \mathbb{E}_{(\mathbf{Z}|\mathbf{x},\theta)} \left[\log w(Z_1) + \sum_{t=1}^{n-1} \log A(Z_t, Z_{t+1}) + \sum_{t=1}^{n} \log B(Z_t, x_t) \right]$$

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- E step(계속)

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{t=1}^{n-1} \sum_{i,j=1}^{m_z} \gamma_t(i,j) \log A(i,j) + \sum_{t=1}^{n} \sum_{i=1}^{m_z} \gamma_t(i) \log B(i,x_t)$$

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{i,j=1}^{m_z} \sum_{t=1}^{n-1} \gamma_t(i,j) \log A(i,j) + \sum_{i=1}^{m_z} \sum_{t=1}^{n} \gamma_t(i) \log B(i,x_t)$$

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- E step(계속)

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{i,j=1}^{m_z} \sum_{t=1}^{n-1} \gamma_t(i,j) \log A(i,j) + \sum_{i=1}^{m_z} \sum_{t=1}^{n} \gamma_t(i) \log B(i, x_t)$$

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{i}^{m_z} \left[\sum_{j=1}^{m_z} \left[\sum_{t=1}^{n-1} \gamma_t(i,j) \right] \log A(i,j) \right]$$

$\because x_t \in \{m_x\}$

$$+ \sum_{i=1}^{m_z} \sum_{l=1}^{m_x} \left(\sum_{\text{at } x_t = l} \gamma_t(i) \right) \log B(i, l)$$

(ref) Feng Liang, Stat542_W7_HMM

1-3 Baum-Welch

- 어떻게 최대화할 것인가?(M step)
- 다음 함수를 최대화 하려면?

$$a_1 \log b_1 + a_2 \log b_2 + \cdots + a_m \log b_m$$

where $a_i > 0,\ b_i > 0,\ \sum_i a_i = 1,\ \sum_i b_i = 1$

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- 먼저, $\sum a_i \log b_i \leq \sum a_i \log a_i$를 보일 수 있다.

$$\sum a_i \log b_i - \sum a_i \log a_i = \sum a_i \log \frac{b_i}{a_i}$$

$\log x \leq x - 1$ for any $x > 0$ 이므로,

$$\sum a_i \log b_i - \sum a_i \log a_i \leq \sum a_i \left(\frac{b_i}{a_i} - 1\right) = \sum b_i - \sum a_i = 0$$

$a_i = b_i$일때만 등호 성립. 그때가 최대.

$\sum a_i > 1$이면 $b_i = a_i / \sum a_i$

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- 따라서 첫번째 항의 경우,

  $\sum_{i}^{m_z} \gamma_1(i) \log w(i)$ 를 최대화 하려면, $\gamma_1(i) = w(i)$

- 두번째 항 $\sum_{i}^{m_z} \left[\sum_{j=1}^{m_z} \left[\sum_{t=1}^{n-1} \gamma_t(i,j) \right] \log A(i,j) \right]$ 은,

$$A(i,j) = \frac{\sum_{t=1}^{n-1} \gamma_t(i,j)}{\sum_{j'} \sum_{t=1}^{n-1} \gamma_t(i,j')}, \qquad i,j = 1, \dots, m_z$$

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- 세번째 항

$$\sum_{i=1}^{m_z} \sum_{l=1}^{m_x} \left(\sum_{\text{at } x_t = l} \gamma_t(i) \right) \log B(i, l)$$ 은,

$$B(i,l) = \frac{\sum_{\text{at } x_t = l} \gamma_t(i)}{\sum_t \gamma_t(i)}$$ 일 때 최대.

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- $\gamma_t(i,j)$

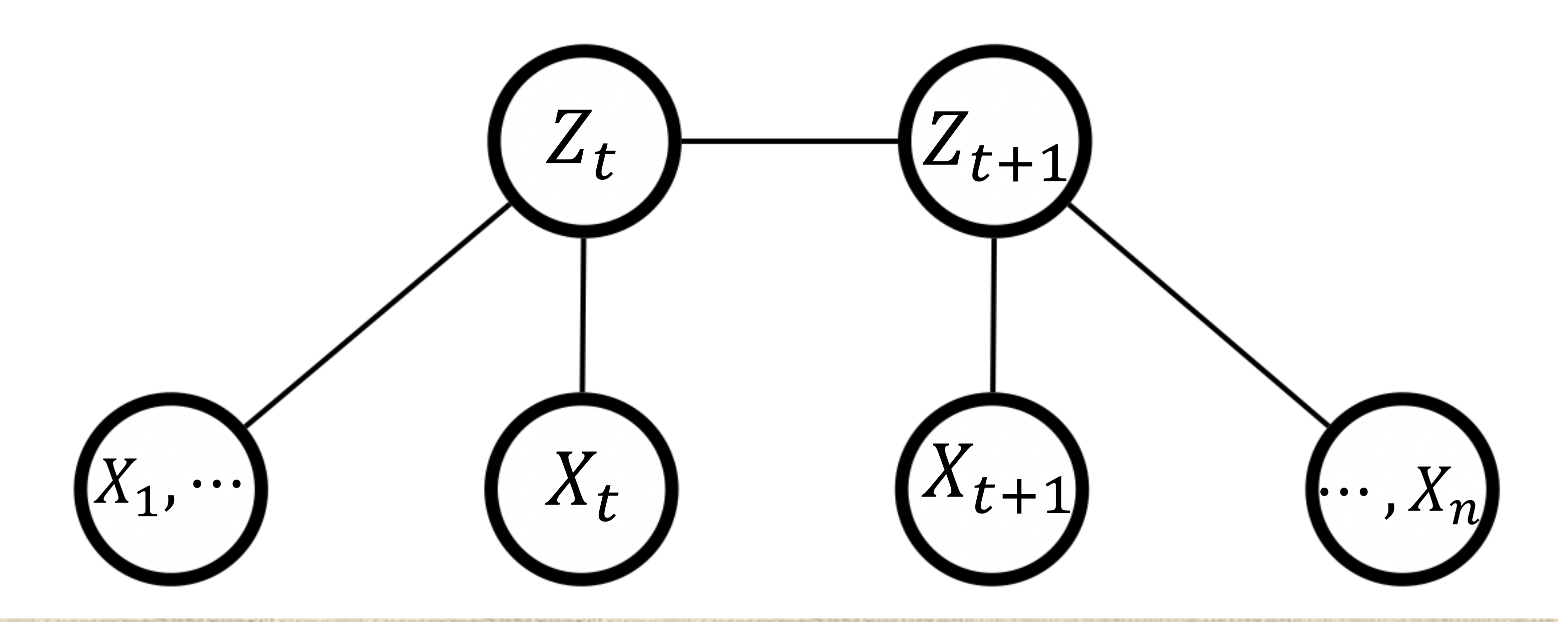
$$p_\theta(Z_t = i, Z_{t+1} = j|\mathbf{x})$$
$$\propto p_\theta(x_1, \dots, x_t, Z_t = i, Z_{t+1} = j, x_{t+1}, \dots, x_n)$$
$$= p_\theta(x_1, \dots, x_t, Z_t = i) \times p_\theta(Z_{t+1} = j|Z_t = i)$$
$$\times p_\theta(x_{t+1}|Z_{t+1} = j) \times p_\theta(x_{t+2}, \dots, x_n|Z_{t+1} = j)$$
$$= \alpha_t(i)A(i,j)B(j, x_{t+1})\beta_{t+1}(j)$$

(ref) Feng Liang, Stat542_W7_HMM

한국방송통신대학교 프라임칼리지

1-3 Baum-Welch

- $\gamma_t(i,j)$



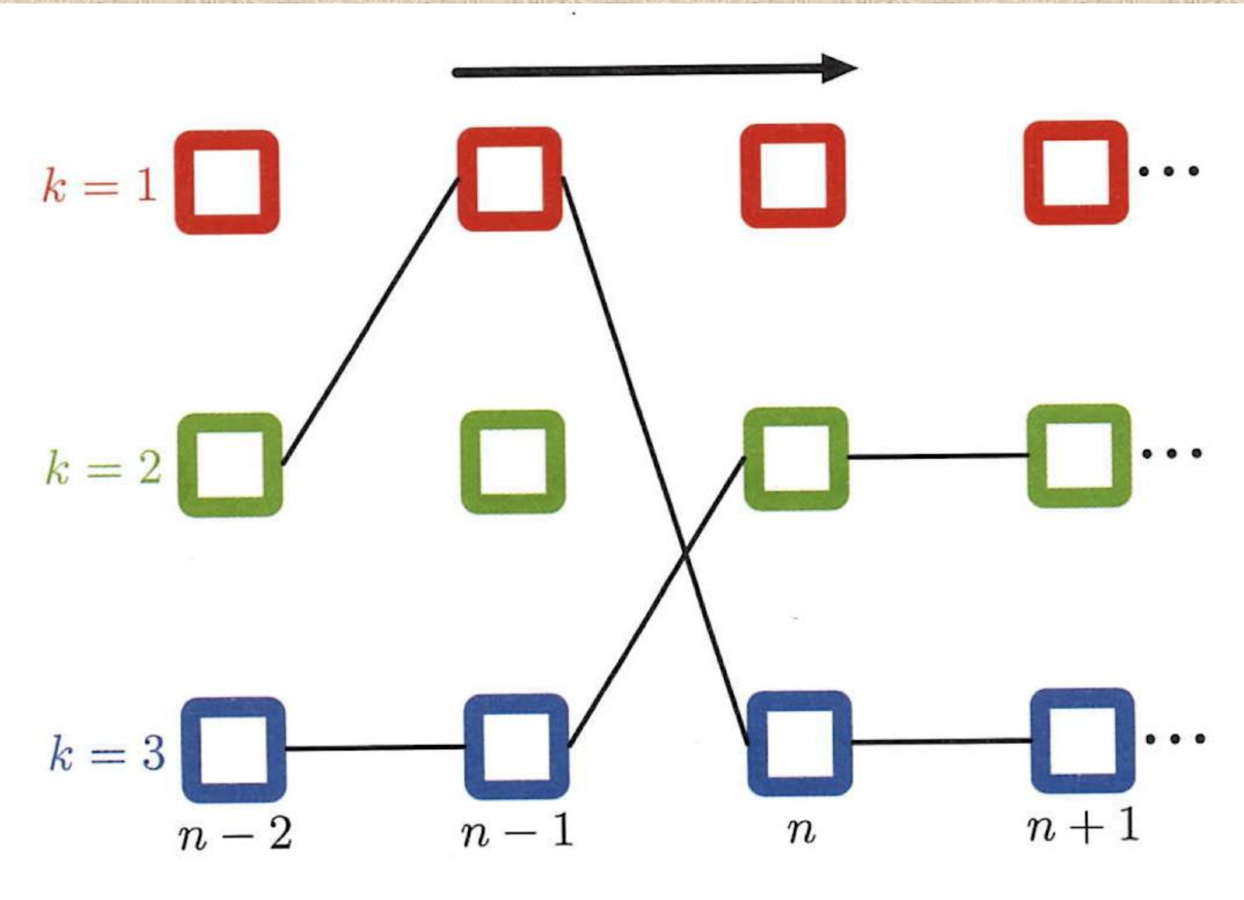
한국방송통신대학교 프라임칼리지

1-4 Viterbi

- Z_t각각에 대해 가장 optimal한 값을 선택하는 방식은 문제가 있다.
    - $Z_t^* = \arg\max_i p_\theta(Z_t = i|\mathbf{x}) = \arg\max_i \gamma_t(i)$
    - 각각의 t에 대해 최적의 state를 선택하므로, sequence가 유효하기 않을 수 있다. 예를들어, $Z_t^* = 1$이고 $Z_{t+1}^* = 2$지만, $A_{12} = 0$인 경우.

한국방송통신대학교 프라임칼리지

1-4 Viterbi



Bishop. Fig13.16

한국방송통신대학교 프라임칼리지

1-4 Viterbi

- 따라서 single sequence를 찾아야 함.

$$Z^* = \arg\max_{i_1,\cdots,i_n} p_\theta(Z_1 = i_1, \cdots, Z_n = i_n | \mathbf{x})$$

- dynamic programming = Viterbi algorithm

$$\delta_t(i) = \max_{j_1,\cdots,j_{t-1}} p_\theta(Z_1 = j_1, \dots, Z_{t-1} = j_{t-1}, Z_t = i, x_1, \cdots, x_t)$$

$$\delta_1(i) = p_\theta(Z_1 = i, x_1) = w(i)B(i, x_1)$$

(ref) Feng Liang, Stat542_W7_HMM

한국방송통신대학교 프라임칼리지

1-4 Viterbi

$$\delta_{t+1}(i) = \max_{j_1,\cdots,j_{t-1},j} \begin{array}{c} p_\theta(Z_1 = j_1, \cdots, Z_{t-1} = j_{t-1}, Z_t = j, \\ Z_{t+1} = i, x_1, \cdots, x_{t+1}) \end{array}$$

$$= \max_{j_1,\cdots,j_{t-1},j} \begin{array}{c} [p_\theta(Z_1 = j_1, \cdots, Z_{t-1} = j_{t-1}, Z_t = j, x_1, \cdots, x_t) \times \\ p_\theta(Z_{t+1} = i | Z_t = j) \times p_\theta(x_{t+1} | Z_{t+1} = i)] \end{array}$$

$$= \left[\max_j \delta_t(j) A(j,i)\right] B(i, x_{t+1})$$

(ref) Feng Liang, Stat542_W7_HMM

1-4 Viterbi

- δ를 이용해서 Z^*를 찾는다.

$$Z^* = \arg\max_{i_1,\cdots,i_n} p_\theta(Z_1 = i_1, \cdots, Z_n = i_n | \mathbf{x})$$

- $Z_n^* = \arg\max_i \delta_n(i)$
  - $\delta_n(i)$는 가장 가능성 높은 sequence중,

    i로 끝나는 것의 확률값이기 때문.

한국방송통신대학교 프라임칼리지

1-4 Viterbi

- Z_{n-1}을 고려해보면,

$$Z_{n-1}^{*} = \arg\max_{i}[\delta_{n-1}(i)A(i, j_{n}^{*})]$$

(Z_{n-1}^{*}값은 $i \in m_z$가 됨. 확률값이 아님에 주의)

- 따라서, 앞서 했던 방식(recursion)과 동일하게,

$$Z_{t-1}^{*} = \arg\max_{i}[\delta_{t}(i)A(i, j_{t}^{*})]$$

한국방송통신대학교 프라임칼리지

1-5 hmm의 확장

- 단점을 보완하기 위한 확장들이 많다.
- 단점

1. 생성모델로 사용했을 때 성능이 좋지 않다.(p4)
2. 주어진 상태를 얼마나 유지하는지 나타내는 시간분포에 부적합

한국방송통신대학교 프라임칼리지

1-5 hmm의 확장

2. (계속)

   배열이 k상태에 T만큼 머물러 있다가 전이할 확률:

$$p(T) = (A_{kk})^T (1 - A_{kk})$$

   T에 대해 기하급수적으로 감쇠하는 함수

   (계속 유지되는 것을 표현하기 힘들다)

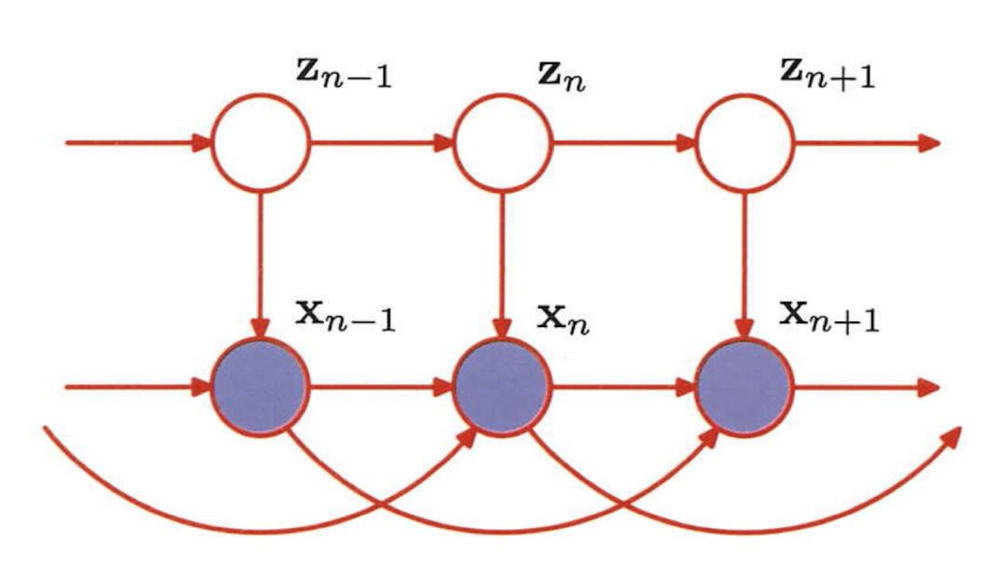
3. 거리가 먼 관측변수들간 상관관계를 잘 잡지 못한다.

한국방송통신대학교 프라임칼리지

1-5 hmm의 확장

3. (계속) 추가적인 링크를 추가하여 해결

예 - 자기회귀적(autoregressive) 은닉 마르코프 모델



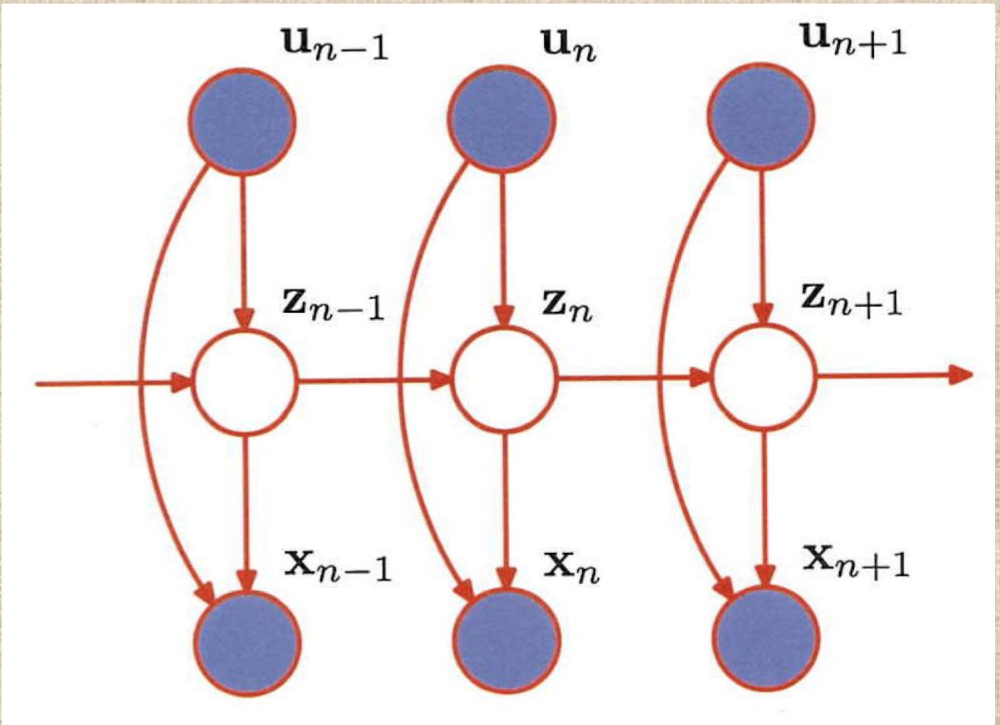
Bishop. Fig13.17

x_n이 앞의 둘(x_{n-1}, x_{n-2})에 의존한다.

한국방송통신대학교 프라임칼리지

1-5 hmm의 확장

- 입출력(input-output) hmm
  - 관측변수($\mathbf{u}$)가 하나 더 있다.
  - 마르코프성질(z_n이 관측되면, z_{n-1}, z_{n+1}이 서로 독립이다)을 만족.



Bishop. Fig13.18

한국방송통신대학교 프라임칼리지

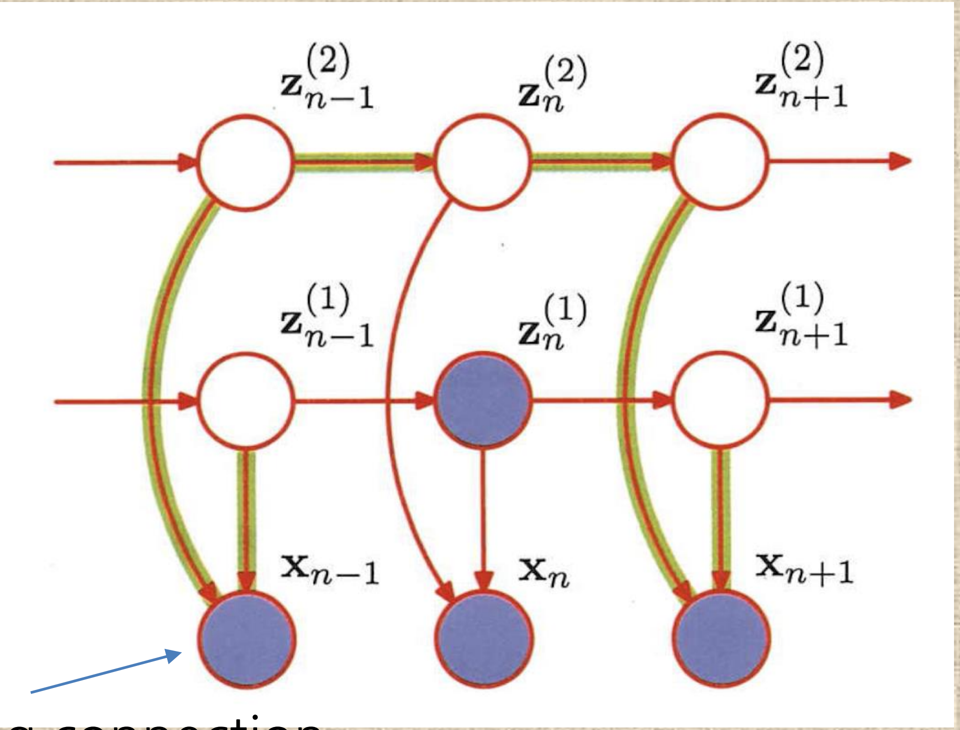
1-5 hmm의 확장

- factorial hmm
  - t 에서 10비트의 정보를 표현해야 한다면, $2^{10} = 1024$개의 잠재변수가 필요
  - factorial hmm은 10개의 연쇄를 사용
  - 마르코프조건을 바로 만족시키지 않아서, 표준형태의 hmm으로 치환한 뒤에 계산 (계산량이 많다)

한국방송통신대학교 프라임칼리지

1-5 hmm의 확장

- factorial hmm



converging connection
(head to head)

Bishop. Fig13.20

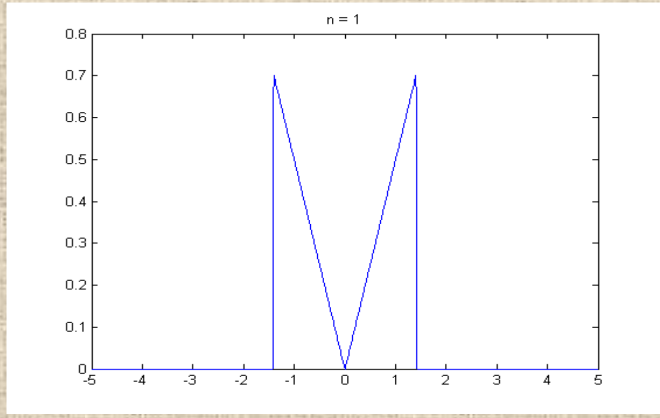
02
중심극한정리

한국방송통신대학교 프라임칼리지

1-2 CLT(Central Limit Theorem)

- 확률변수 $X_1, X_2, \ldots$ 들이 서로 독립이고 같은 확률분포를 가지며 그 확률분포의 기댓값μ과 표준편차σ가 유한하면,

$S_n = \frac{\sum_n X_n}{n}$의 분포는 $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$으로 수렴한다.



$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{N} X_i\right) - \mu\right) \rightarrow \mathcal{N}(0, \sigma^2)$$

https://ko.wikipedia.org/wiki/중심_극한_정리

한국방송통신대학교 프라임칼리지

다음시간

12강

- 칼만필터
- 부스팅