

07. 순차 데이터의 인식

- 특징들의 시간성 (temporal property)

ex) 지진파, 음성, 주식 거래량, 온라인 필기 문자 등

= sequential data, context-dependent data
순차 데이터 문맥 의존 데이터

- 순차 데이터를 갖는 패턴 인식

- 시간성의 표현과 정보 주는 방법 필요

- 대표 모델: 은닉 마코프 모델(hidden Markov model, HMM)

- HMM

- 19세기 마코프 모델에 토대를 둘

- 1960 은닉 추가하여 HMM으로 확장

- 많은 분야의 문제 해결 도구

ex) 패턴인식, 컴퓨터 비전, 데이터 마이닝, 정보검색, 생물정보학,
신호처리, 데이터베이스 . . .

응용사례: 음성 인식, 온라인 팔기 인식, DNA 열 찾기,

제스처 인식, 영어 발음 교정, 음악 인식 - -

7.1 순차 데이터

- 시간성이 없는 데이터
 - 특징들의 선후관계는 무의미

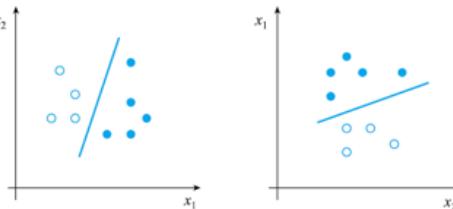


그림 7.1 시간성이 없는 데이터에서 특징의 위치를 바꿈

- 시간성이 있는 데이터
 - 특징들의 선후관계는 매우 중요

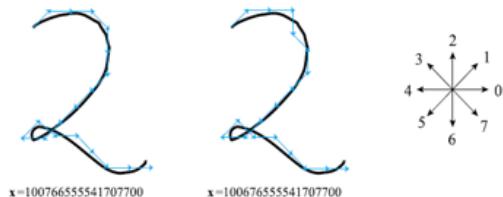


그림 7.2 시간성을 갖는 패턴에서 특징 순서를 바꾸었을 때 나타나는 물리적 왜곡 현상

• 순차 데이터

- 대부분 가변길이를 갖음
- 관측벡터로 표현

$$O = (o_1, o_2, \dots, o_{t-1}, o_t, o_{t+1}, \dots, o_T)^T = o_1, o_2, \dots, o_{t-1}, o_t, o_{t+1}, \dots, o_T$$

- 관측 o_i 가 가질 수 있는 값의 집합을 알파벳이라 함

$$\text{알파벳 } V = \{v_1, v_2, \dots, v_m\} \quad (v_i \text{ 를 기호라 함})$$

- 기호들이 시간에 따라 의존성 가짐

ex) 그림 7.2의 온라인 숫자

$$P(o_t=2 | o_{t-1}=6) \approx 0$$

$$P(o_t=5 | o_{t-1}=6) > P(o_t=4 | o_{t-1}=6)$$

7.2 마코프 모델

- 러시아 수학자 Andrey Markov가 제안
- 시간 t 에서의 관측은 가장 최근 r 개 관측에만 의존한다는 가정 하의 확률 추론

$$\left(\begin{array}{l} r=0 \text{이면 } 0\text{차 마코프 체인}, P(O_t | O_{t-1}, O_{t-2}, \dots, O_1) = P(O_t) \\ r=1 \text{이면 } 1\text{차 마코프 체인}, P(O_t | O_{t-1}, O_{t-2}, \dots, O_1) = P(O_t | O_{t-1}) \\ r=2 \text{이면 } 2\text{차 마코프 체인}, P(O_t | O_{t-1}, O_{t-2}, \dots, O_1) = P(O_t | O_{t-1}, O_{t-2}) \end{array} \right)$$

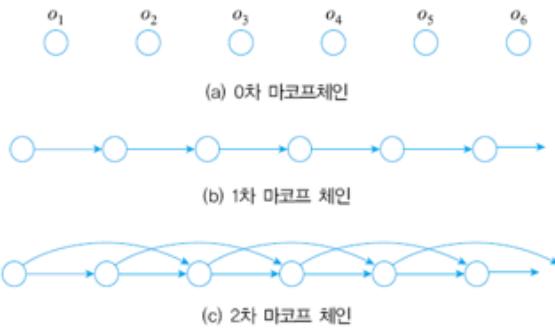


그림 7.3 마코프 체인

- 마코프 모델은 1차 마코프 체인 사용

- 날씨 예

- $V = \{\text{비}, \text{구름}, \text{해}\}$

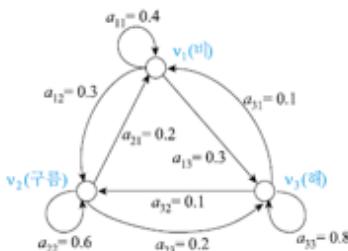
- 기후 관측에 의해 얻은 날씨 변화 확률

표 7.1 날씨 변화 확률

오늘 \ 내일	비	구름	해
비	0.4	0.3	0.3
구름	0.2	0.6	0.2
해	0.1	0.1	0.8

· 상태 전이

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$



(a) 상태 전이 확률 행렬

(b) 상태 전이도

그림 7.4 마크프 모델을 위한 상태 전이 확률과 상태 전이도

$$\left(\begin{array}{l} a_{ij} = P(O_t = o_j | O_{t-1} = o_i) \\ \text{여기서 } a_{ij} \geq 0 \text{ 와 } \sum_{j=1}^m a_{ij} = 1 \text{ 을 만족} \end{array} \right)$$

· 관측 벡터 o의 확률 구하기

$$P(O| \text{마코프모델}) = P(O|A)$$

$$= P(o)$$

$$= P(O_1, O_2, \dots, O_T)$$

$$= P(O_1) P(O_2|O_1) P(O_3|O_2, O_1) \dots P(O_T|O_{T-1}, O_{T-2}, \dots, O_1) \dots P(O_T|O_{T-1}, \dots, O_1)$$

$$= P(O_1) P(O_2|O_1) P(O_3|O_2) \dots P(O_T|O_{T-1}) \dots P(O_T|O_{T-1})$$

$$= P(O_1) \prod_{i=1}^{T-1} (O_{i+1}|O_i)$$

예제 7.1 “오늘 해가 떴는데 내일부터 7일 간의 날씨가
해-해-비-비-해-구름-해 일 확률은 얼마인가?”

$$\begin{aligned}
 P(O|A) &= P(O_1 = \text{해}, O_2 = \text{해}, O_3 = \text{해}, O_4 = \text{비}, O_5 = \text{비}, O_6 = \text{해}, O_7 = \text{구름}, \\
 &\quad O_8 = \text{해} | A) \\
 &= P(\text{해}) P(\text{해}|\text{해}) P(\text{해}|\text{해}) P(\text{비}|\text{비}) P(\text{비}|\text{비}) P(\text{해}|\text{비}) P(\text{구름}|\text{해}) \\
 &\quad P(\text{해}|\text{구름}) \\
 &= \pi_3 \alpha_{33} \alpha_{33} \alpha_{31} \alpha_{11} \alpha_{43} \alpha_{32} \alpha_{23} \\
 &= 1 \times 0.8 \times 0.8 \times 0.1 \times 0.4 \times 0.3 \times 0.1 \times 0.2 \\
 &= 1.536 \times 10^{-4}
 \end{aligned}$$

예제 7.2

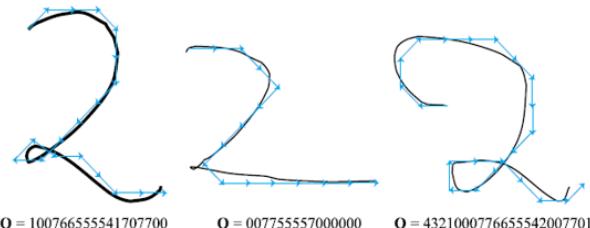


그림 7.5 온라인 필기 숫자 인식에서 부류 2의 출현 샘플들

· 상태전이 확률 구하면,

표 7.2 상태 전이의 발생 회수

$i-1$	0	1	2	3	4	5	6	7
0	11	1	0	0	0	0	0	5
1	2	0	0	0	0	0	0	1
2	1	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0
4	0	1	1	1	0	0	0	0
5	0	0	0	0	2	8	0	1
6	0	0	0	0	0	2	2	0
7	4	0	0	0	0	1	2	4

$A =$	11/17	1/17	0	0	0	0	0	5/17
	2/3	0	0	0	0	0	0	1/3
	1/2	1/2	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	0	1/3	1/3	1/3	0	0	0	0
	0	0	0	0	2/11	8/11	0	1/11
	0	0	0	0	0	1/2	1/2	0
	4/11	0	0	0	0	1/11	2/11	4/11

· 초기 확률 행렬 구하면,

$$\Pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8) = \left(\frac{1}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}, 0, 0, 0\right)$$

“체인코드 $1 \rightarrow 0 \rightarrow 7 \rightarrow 6 \rightarrow 5 \rightarrow 5 \rightarrow 7 \rightarrow 0 \rightarrow 0$ 으로

표현되는 샘플이 발생할 확률은 얼마인가?”

$$\begin{aligned} P(O|A) &= P(o_1=1, o_2=0, o_3=7, o_4=6, o_5=5, o_6=5, o_7=7, o_8=0, o_9=0|A) \\ &= P(1) \cdot P(0|1) \cdot P(7|0) \cdot P(6|7) \cdot P(5|6) \cdot P(5|5) \cdot P(7|5) \cdot P(0|7) \cdot P(0|0) \\ &= 0.9243 \times 10^{-4} \end{aligned}$$

· 마코프 모델과 0차, 2차 마코프 체인과의 비교

표 7.3 0차 마코프 체인을 위한 확률 표

	0	1	2	3	4	5	6	7
t	19 (19/55)	4 (4/55)	2 (2/55)	1 (1/55)	3 (3/55)	11 (11/55)	4 (4/55)	11 (11/55)

$$\begin{aligned} P(\mathbf{O}|\mathbf{A}_0) &= P(o_1=1, o_2=0, o_3=7, o_4=6, o_5=5, o_6=5, o_7=7, o_8=0, o_9=0|\mathbf{A}_1) \\ &= P(1)P(0)P(7)P(6)P(5)P(5)P(7)P(0)P(0) \\ &= (4/55)^*(19/55)^*(11/55)^*(4/55)^*(11/55)^*(11/55)^*(11/55)^*(19/55)^* \\ &\quad *(19/55) \\ &= 0.3489 \times 10^{-6} \end{aligned}$$

표 7.4 2차 마코프 체인을 위한 전이 확률 표

$(t-2, t-1) \setminus t$	0	1	2	3	4	5	6	7
00	5 (5/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	4 (4/9)
01	...							
...								
ij								
...								
77								...

· 주정할 매개변수가
64x8개로 증가

* MM과 HMM의 근본적인 차이점

- 마코프 모델에서는 상태를 나타내는 노드에 관측(비, 해, 구름) 그 자체가 들어 있다. 즉 상태를 볼 수 있다.
- HMM에서는 상태를 볼 수 없다. 즉, 상태가 은닉된다.

7.3 은닉 마코프 모델로의 발전

- 마코프 모델은 한계를 가진다.
 - 보다 복잡한 현상이나 과정에 대한 모델링 능력의 한계
 - 모델의 용량을 키우기 위해
 - 상태를 감춘다.

7.3.1 동기

- 마코프 체인의 차수와 추정할 매개 변수의 수
 - 0차: $m-1$ 개의 확률값 추정해야 함 (합이 1이 아니 나머지 한 개 자동)
 - 1차: $m(m-1)$
 - r차: $m^r(m-1)$ → 차수에 따라 기하급수적으로 모델 크기 증가
- HMM
 - 차수를 미리 고정하지 않고 모델 자체가 확률 프로세스에 따라 적응적으로 정함
 - 따라서 아무리 먼 과거의 관측도 현재에 영향을 미침
 - 즉 $P(\text{해}|\text{비}, \text{비})$, $P(\text{해}|\text{비}, \text{비}, \text{비})$, $P(\text{해}|\text{비}, \text{비}, \dots, \text{해})$ 가 모두 다름
 - 이런 뛰어난 능력에도 불구하고 모델 크기는 감당할 정도임

- 상태를 감추면, 마코프 모델이 은닉 마코프 모델이 된다.

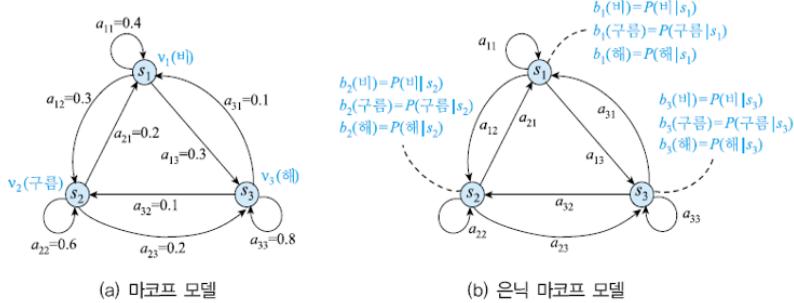


그림 7.7 마코프 모델을 은닉 마코프 모델로 확장

- 그림 7.7의 해석 (예를들어, 해 \rightarrow 해 \rightarrow 비 \rightarrow 구름이 관측되었다.)
- 마코프 모델
 - 그것은 상태 $S_3 \rightarrow S_3 \rightarrow S_1 \rightarrow S_2$ 에서 관측되었다.
- 은닉 마코프 모델
 - 모든 상태에서 {비, 해, 구름}이 관측 가능하므로 $S_3 \rightarrow S_3 \rightarrow S_1 \rightarrow S_2$ 일 수도 있고, $S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_1$ 일 수도 있다.
 - 해 \rightarrow 해 \rightarrow 비 \rightarrow 구름이 $S_3 \rightarrow S_3 \rightarrow S_1 \rightarrow S_2$ 일 확률은?

$$[\pi_3 \times b_3(\text{해})] \times [a_{33} \times b_3(\bar{\text{해}})] \times [a_{31} \times b_1(\text{비})] \times [a_{12} \times b_2(\text{구름})]$$

예제 7.3) 공을 담은 향아리

- n 개의 향아리, 공의 색깔 m 개
 - 향아리가 상태이고 공의 색깔을 고침
- 그림 7.8은 $n=3, m=4$

$$V = \{ \text{하양}(W), \text{검정}(B), \text{연파랑}(L), \text{전파랑}(D) \}$$

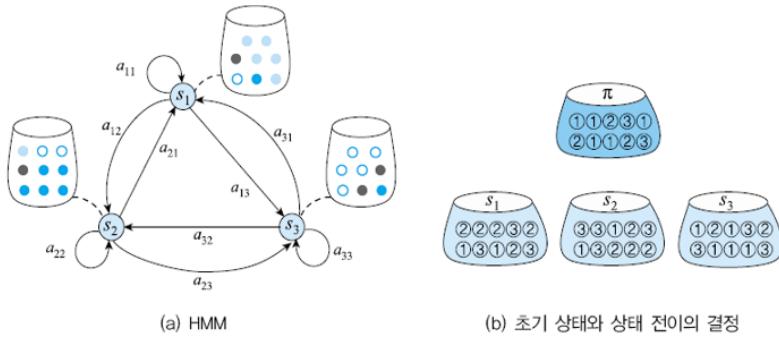


그림 7.8 공을 담은 향아리를 HMM으로 모델링

• 실험 시나리오

i) 예에서의 실험을 정리해 보자. 실험이 시작되면 π -향아리에서 카드를 하나 뽑아 번호를 보고 다시 집어 넣는다. 카드 번호의 상태로 들어간다. 그 상태에 해당하는 공 향아리에서 공을 하나 뽑고 색깔을 확인하고 다시 집어 넣는다. 색깔을 o_1 에 기록한다. 그 상태에 해당하는 카드 향아리에서 카드를 하나 뽑고 번호를 보고 다시 집어 넣는다. 카드 번호의 상태로 이동한다. 그 상태에 해당하는 공 향아리에서 공을 하나 뽑고 색깔을 o_2 에 기록한다. 이런 과정을 반복한다.

ii) 실험이 우리 눈에 다 보이는 것이 아니라는 사실을 기억해야 한다. 보이는 것은 공의 색깔, 즉 $\mathbf{O} = o_1 o_2 \cdots o_T$ 뿐이다. 실험이 커튼 뒤에서 이루어지고 실험하는 사람이 단지 공의 색깔만 불러 준다고 생각하면 된다. 우리가 가진 것은 \mathbf{O} 뿐이다. ■■■

예제 7.4) 여자 친구의 삶

- 여자 친구의 일상을 관찰, $V = \{\text{산책}, \text{쇼핑}, \text{청소}\}$
- 날씨는 해와 비(상태)
- 내가 가진 정보

날씨

$$P(\text{비} | \text{온 다음 날 비}) = P(q_t = \text{비} | q_{t-1} = \text{비}) = 0.7$$

$$P(\text{비} | \text{온 다음 날 해}) = P(q_t = \text{해} | q_{t-1} = \text{비}) = 0.3$$

$$P(\text{해} | \text{뜬 다음 날 비}) = P(q_t = \text{비} | q_{t-1} = \text{해}) = 0.4$$

$$P(\text{해} | \text{뜬 다음 날 해}) = P(q_t = \text{해} | q_{t-1} = \text{해}) = 0.6$$

$$P(\text{비}) = 0.6$$

$$P(\text{해}) = 0.4$$

날씨에 따른 그녀의 행동

$$P(\text{비} | \text{오는 날 산책}) = P(o_t = \text{산책} | q_t = \text{비}) = 0.1$$

$$P(\text{비} | \text{오는 날 쇼핑}) = P(o_t = \text{쇼핑} | q_t = \text{비}) = 0.4$$

$$P(\text{비} | \text{오는 날 청소}) = P(o_t = \text{청소} | q_t = \text{비}) = 0.5$$

$$P(\text{해} | \text{뜬 날 산책}) = P(o_t = \text{산책} | q_t = \text{해}) = 0.6$$

$$P(\text{해} | \text{뜬 날 쇼핑}) = P(o_t = \text{쇼핑} | q_t = \text{해}) = 0.3$$

$$P(\text{해} | \text{뜬 날 청소}) = P(o_t = \text{청소} | q_t = \text{해}) = 0.1$$

HMM을 사용하면 확률 추론할 수 있다.

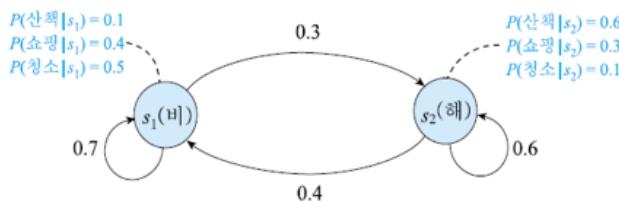


그림 7.9 그녀의 삶을 HMM으로 모델링

7.3.3 구성 요소와 세가지 문제

- 어느것이 상태인가? 상태는 몇 가지인가? (아키텍처 설계)
- 확률 분포는 어떻게 구하나? (학습)
- 예) 온라인 필기 인식: 무엇이 상태이고, 상태를 몇가지로 할까?
가진 것은 오로지 훈련집합

- 아키텍처
 - HMM은 가중치 방향 그래프로 표현
 - 노드가 상태
 - 상태로 사용할 것이 명확한 경우도 있지만 그렇지 않은 경우도 있다.
 - 대표적인 아키텍처
 - 어고딕 모델과 좌우 모델
- 음성 인식에서 주로 사용

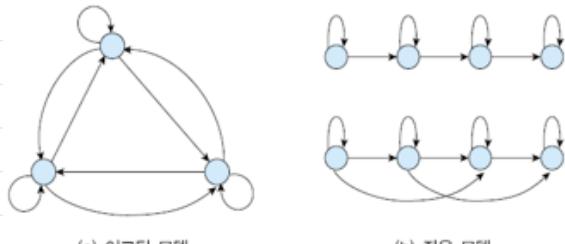


그림 7.10 HMM의 대표적인 아키텍처

• 매개변수

$$\Theta = (A, B, \pi)$$

1. 상태 전이 확률 행렬 $A = |a_{ij}|$

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq n$$

이때 $\sum_{j=1}^n a_{ij} = 1$ 상태의 개수

2. 관측 행렬 $B = |b_j(v_k)|$

$$b_j(v_k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq n, 1 \leq k \leq M$$

이때 $\sum_{k=1}^M b_j(v_k) = 1$ 기호의 개수

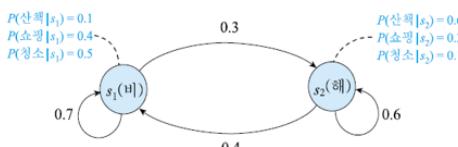


그림 7.9 그녀의 삶을 HMM으로 모델링

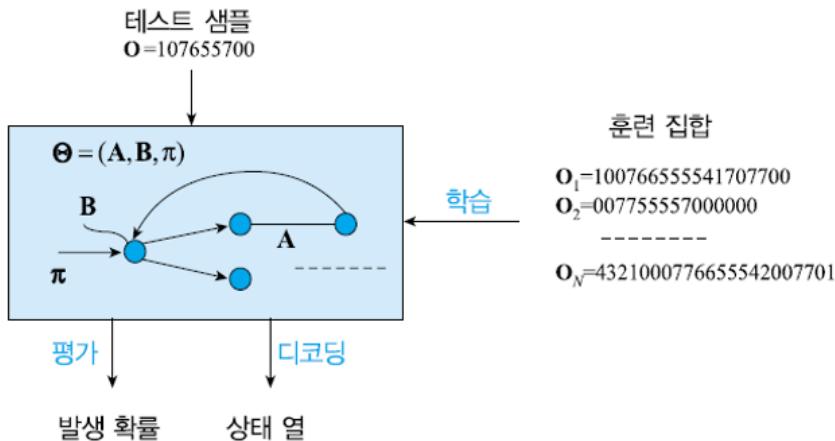
3. 초기 확률 벡터 $\pi = |\pi_i|$

$$\pi_i = p(q_i = s_i), 1 \leq i \leq n$$

$$\text{이 때 } \sum_{i=1}^n \pi_i = 1$$

• 세 가지 문제

1. 평가: 모델 θ 가 주어진 상황에서, 관측벡터 O 를 얻었을 때 $P(O|\theta)$ 는?
2. 디코딩: 모델 θ 가 주어진 상황에서, 관측벡터 O 를 얻었을 때 최적의 상태열은?
3. 학습: 훈련집합 $X = \{o_1, \dots, o_N\}$ 이 주어져 있을 때 HMM의 매개변수 θ 는?



7.4 알고리즘

평가, 디코딩 → 동적 프로그래밍 이용

7.4.1 평가

- 평가 문제란?
 - 모델 θ 가 주어진 상황에서, 관측벡터 O 를 얻었을 때 $P(O|\theta)$ 는?
 - 예) 그녀가 일주일 연속으로 쇼핑만 할 확률은?
- 평가 문제를 풀어 보자.
 - HMM에서는 O 를 관측한 상태 열을 모른다.
 - 우선 O 를 관측한 상태 열을 있다고 가정하고 그것을 $Q = (q_1, \dots, q_T)$ 라 하자.
 - 그럼 아래식을 유도할 수 있다.

$$P(O, Q | \theta) = \prod_{t=1}^T b_{q_t}(O_t) \alpha_{q_1, q_2} \dots \alpha_{q_{T-1}, q_T} b_{q_T}(O_T) \quad (7.10)$$

- 원래 문제) $P(O|\theta)$ 는 모든 상태 열에 대해 (7.10)의 식을 더하여 구할 수 있다.

$$P(O|\theta) = \sum_{\text{모든 } Q} P(O, Q | \theta)$$

예제 7.5 여자친구의 삶

“그녀가 오늘 산책, 내일 산책, 모레 청소, 그리고 글피 쇼핑할 확률은?”

즉, $O = (O_1 = \text{산책}, O_2 = \text{산책}, O_3 = \text{청소}, O_4 = \text{쇼핑})$ 일 때, $P(O|I\theta)$ 는?

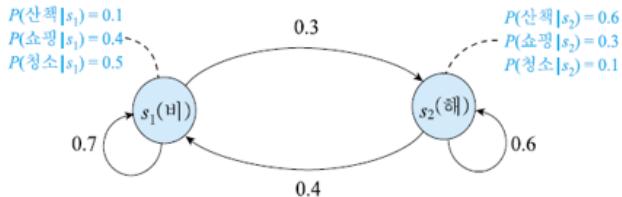


그림 7.9 그녀의 삶을 HMM으로 모델링

초기 확률

$$P(H) = 0.6$$

$$P(\bar{H}) = 0.4$$

모든 상태열을 나열하면,

$$Q_1 = \text{비비비비}, \quad Q_2 = \text{비비비해}, \quad Q_3 = \text{비비해비}, \quad Q_4 = \text{비비해해},$$

$$Q_5 = \text{비해비비}, \quad Q_6 = \text{비해비해}, \quad Q_7 = \text{비해해비}, \quad Q_8 = \text{비해해해},$$

$$Q_9 = \text{해비비비}, \quad Q_{10} = \text{해비비해}, \quad Q_{11} = \text{해비해비}, \quad Q_{12} = \text{해비해해},$$

$$Q_{13} = \text{해해비비}, \quad Q_{14} = \text{해해비해}, \quad Q_{15} = \text{해해해비}, \quad Q_{16} = \text{해해해해}$$

- 상태열 Q_1 에 대해 구해 보면,

$$\begin{aligned} P(O, Q_1 | I\theta) &= \pi_1 b_1(\text{산책}) a_{11} b_1(\text{산책}) a_{11} b_1(\text{청소}) a_{11} b_1(\text{쇼핑}) \\ &= 0.6 \times 0.1 \times 0.7 \times 0.1 \times 0.7 \times 0.5 \times 0.7 \times 0.4 \\ &= 4.116 \times 10^{-4} \end{aligned}$$

- 모든 상태열에 대한 값을 구하여 답을 구해 보면,

$$P(O|I\theta) = \sum_{i=1}^{16} P(O, Q_i | I\theta)$$

$$\begin{aligned} &= 0.4116 \times 10^{-3} + 0.1323 \times 10^{-3} + 0.02216 \times 10^{-3} + 0.02268 \times 10^{-3} \\ &+ 0.6048 \times 10^{-3} + 0.1944 \times 10^{-3} + 0.10368 \times 10^{-3} + 0.11664 \times 10^{-3} \\ &+ 0.9408 \times 10^{-3} + 0.3024 \times 10^{-3} + 0.04608 \times 10^{-3} + 0.05184 \times 10^{-3} \\ &+ 4.8384 \times 10^{-3} + 1.5552 \times 10^{-3} + 0.82944 \times 10^{-3} + 0.93312 \times 10^{-3} \\ &= 0.0111 \end{aligned}$$

- 답은 구할 수 있다. but, 상태 열의 개수는 $2^4 = 16$ 가지, 일반적으로 상태의 수가 n 이고 관측 열 0의 길이가 T 라면 N^T 가지의 상태열 각각의 상태 열은 $2T - 1$ 번의 곱셈. 따라서 시간 복잡도는 $O(N^T T)$
ex) $n=5, T=30$ 이라면 5.4948×10^{22} 번의 곱셈 필요. 계산 풀릴!

→ 효율적인 알고리즘 필요

- 기본적인 아이디어는 동적 프로그래밍에 의한 중복 계산 제거

ex) 예제 7.5의 $Q_1 =$ '비비비비', $Q_2 =$ '비비비해'의 계산

- $P(O, Q_1 | \theta) = \prod_i b_{1i}(\text{산책}) a_{11} b_{1i}(\text{산책}) a_{11} b_{1i}(\text{청소}) a_{11} b_{1i}(\text{쇼핑})$
- $P(O, Q_2 | \theta) = \prod_i b_{2i}(\text{산책}) a_{11} b_{2i}(\text{산책}) a_{11} b_{2i}(\text{청소}) a_{12} b_{2i}(\text{쇼핑})$
- 빨간 부분의 계산은 같다.

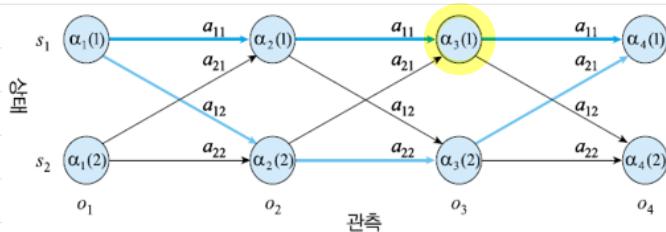


그림 7.13 전방 계산을 위해 상태도를 격자 모양으로 펼침

$\alpha_t(i)$ 는 관측 벡터의 일부 o_1, o_2, \dots, o_t 을 관측하고 시간 t 에 s_i 에 있을 확률

$$\alpha_t(i) = [\alpha_1(i) \times a_{11} + \alpha_2(i) \times a_{21}] \times b_1(o_3)$$

$$\alpha_t(i) = P(O_1, O_2, \dots, O_t, q_t = s_i | \theta)$$

$$= \left[\sum_{j=1}^n \alpha_{t-1}(j) a_{ji} \right] \times b_i(o_t)$$

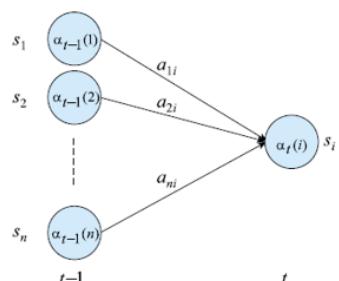


그림 7.14 $\alpha_t(i)$ 의 순환 계산

* 전방 알고리즘

$$\alpha_i(i) = \pi_i b_i(o_1), 1 \leq i \leq n \quad (7.13)$$

$$\alpha_t(i) = \left[\sum_{j=1}^n \alpha_{t-1}(j) a_{ji} \right] b_i(o_t), 2 \leq t \leq T, 1 \leq i \leq n \quad (7.14)$$

$$P(O|\Theta) = \prod_{i=1}^n \alpha_T(i) \quad (7.15)$$

알고리즘 [7.1]

전방 계산에 의한 관측 벡터의 발생 확률 계산

입력: HMM $\Theta = (\mathbf{A}, \mathbf{B}, \pi)$, 관측 벡터 $O = o_1 o_2 \cdots o_T$

출력: O 의 발생 확률 $P(O | \Theta)$

알고리즘:

1. 배열 $\alpha[1 \dots T][1 \dots n]$ 을 생성하라.
2. **for** ($i = 1$ to n) $\alpha[1][i] = \pi_i * b_i(o_1); // (7.13)$
3. **for** ($t = 2$ to T) $T-1$ 번
 4. **for** ($i = 1$ to n) { n 번
 5. $sum = 0;$
 6. **for** ($j = 1$ to n) $sum = sum + \alpha[t-1][j] * a_{ji};$ n 번
 7. $\alpha[t][i] = sum * b_i(o_t); //$ 라인 5-7이 (7.14)
 8. }
 9. $sum = 0;$
10. **for** ($j = 1$ to n) $sum = sum + \alpha[T][j]; // (7.15)$
11. $P(O | \Theta) = sum;$

상태에서의 합

상태

$\Rightarrow \Theta(n^2 T)$

$\rightarrow n=5, T=30$ 이면

약 750번의 곱셈이면

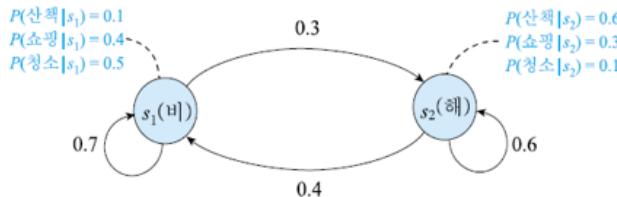
답을 구할 수 있음

각 상태에서의 확률들의 합

예제 7.6 앞의 예를 전방 알고리즘으로 계산하면?

"그녀가 오늘 산책, 내일 산책, 모래 청소, 그리고 글피 쇼핑할 확률은?"

즉, $\Theta = \{O_1=\text{산책}, O_2=\text{산책}, O_3=\text{청소}, O_4=\text{쇼핑}\}$ 일 때, $P(O | \Theta)$ 는?



초기 확률

$$P(H) = 0.6$$

$$P(\bar{H}) = 0.4$$

그림 7.9 그녀의 삶을 HMM으로 모델링

$$t=1 \text{ 일 때 } \alpha[1][1] = \pi_1 * b_1(\text{산책}) = 0.6 * 0.1 = 0.06$$

$$\alpha[1][2] = \pi_2 * b_2(\text{산책}) = 0.4 * 0.6 = 0.24$$

$$t=2 \text{ 일 때 } \alpha[2][1] = (\alpha[1][1] * a_{11} + \alpha[1][2] * a_{21}) * b_1(\text{산책}) = (0.06 * 0.7 + 0.24 * 0.4) * 0.1 = 0.0138$$

$$\alpha[2][2] = (\alpha[1][1] * a_{12} + \alpha[1][2] * a_{22}) * b_2(\text{산책}) = (0.06 * 0.3 + 0.24 * 0.6) * 0.6 = 0.0972$$

$$t=3 \text{ 과 } 4 \text{ 일 때 } \alpha[3][1] = (\alpha[2][1] * a_{11} + \alpha[2][2] * a_{21}) * b_1(\text{청소}) = (0.0138 * 0.7 + 0.0972 * 0.4) * 0.5 = 0.02427$$

$$\alpha[3][2] = (\alpha[2][1] * a_{12} + \alpha[2][2] * a_{22}) * b_2(\text{청소}) = (0.0138 * 0.3 + 0.0972 * 0.6) * 0.1 = 0.00625$$

$$\alpha[4][1] = (\alpha[3][1] * a_{11} + \alpha[3][2] * a_{21}) * b_1(\text{쇼핑}) = (0.02427 * 0.7 + 0.00625 * 0.4) * 0.4 = 0.00780$$

$$\alpha[4][2] = (\alpha[3][1] * a_{12} + \alpha[3][2] * a_{22}) * b_2(\text{쇼핑}) = (0.02427 * 0.3 + 0.00625 * 0.6) * 0.3 = 0.00331$$

$$\text{답은 } P(\mathbf{O} | \Theta) = \alpha[4][1] + \alpha[4][2] = 0.0111$$

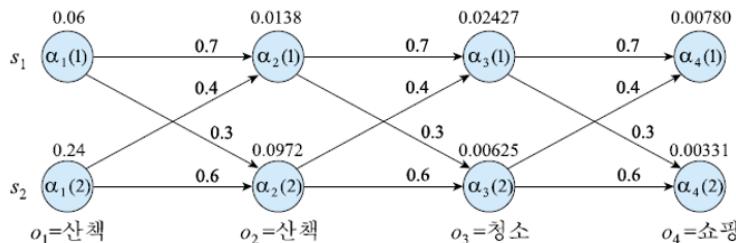


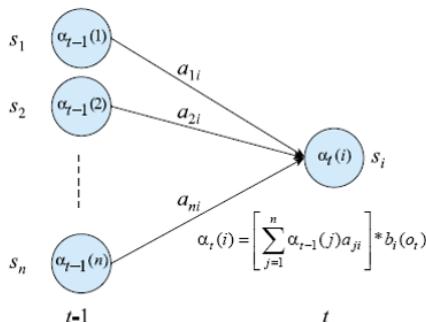
그림 7.15 전방 계산 예

7.4.2 디코딩

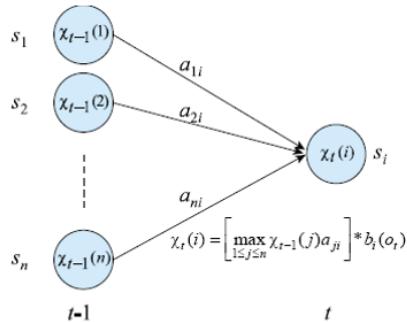
모델 θ 하에서 관측 벡터 O 를 얻었을 때, O 에 해당하는 최적의 상태열 $Q = (q_1, q_2, \dots, q_T)$ 을 찾는 문제가 디코딩이다.

- 험리적 생각은 $P(O, Q | \theta)$ 를 기준 향수로 채택하고, 이것을 최대화하는 \hat{Q} 를 찾아야 한다. 즉,

$$\hat{Q} = \arg \max_{\text{ 모든 } Q} P(O, Q | \theta)$$



(a) 평가 문제 (합 연산)



(b) 디코딩 문제 (최대 선택 연산)

그림 7.16 평가 문제와 디코딩 문제의 연산 차이

• Viterbi 알고리즘

- 전방 계산 (가는 매 단계에서 최적 경로 기록)

$$\chi_1(i) = \pi_i b_i(o_1) \quad 1 \leq i \leq n$$

$$\chi_t(i) = \left[\max_{1 \leq j \leq n} \chi_{t-1}(j) a_{ji} \right] b_i(o_t), \quad 2 \leq t \leq T, \quad 1 \leq i \leq n$$

$$T_t(i) = \arg \max_{1 \leq j \leq n} [\chi_{t-1}(j) a_{ji}], \quad 2 \leq t \leq T, \quad 1 \leq i \leq n$$

의 상태값 저장

• 최적 경로 역추적

$$\hat{q}_T = \arg \max_{1 \leq j \leq n} \chi_T(j)$$

$$\hat{q}_t = \tau_{t+1}(\hat{q}_{t+1}), \quad t=T-1, T-2, \dots, 1$$

• Viterbi 알고리즘

알고리즘 [7.2] 디코딩을 위한 Viterbi 알고리즘

입력: HMM $\Theta = (A, B, \pi)$, 관측 벡터 $O = o_1 o_2 \cdots o_T$

출력: 최적 경로 $\hat{Q} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_T)$

알고리즘:

1. 배열 $\chi[1 \dots T][1 \dots n]$ 와 $\tau[1 \dots T][1 \dots n]$ 을 생성하라.
2. **for** ($i = 1$ to n) $\chi[1][i] = \pi_i * b_i(o_1); // (7.18)$
3. **for** ($t = 2$ to T)
 4. **for** ($i = 1$ to n) {
 5. $\chi[t-1][j] * a_{ji}, 1 \leq j \leq n$ 중에 가장 큰 것의 첨자를 k 라 둔다.
 6. $\chi[t][i] = \chi[t-1][k] * a_{ki} * b_i(o_t); // (7.19)$
 7. $\tau[t][i] = k;$ $// (7.19)$
 8. }
- // 지금부터 (7.20)에 의한 경로 역 추적
9. $\chi[T][j], 1 \leq j \leq n$ 중에 가장 큰 것의 첨자를 \hat{q}_T 로 한다.
10. **for** ($t = T-1$ to 1) $\hat{q}_t = \tau[t+1](\hat{q}_{t+1}); // (7.20)$

시간복잡도:
 N^T

예제 7.7 Viterbi 알고리즘에 의한 디코딩

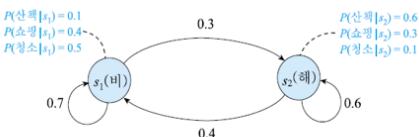


그림 7.9 그녀의 삶을 HMM으로 모델링

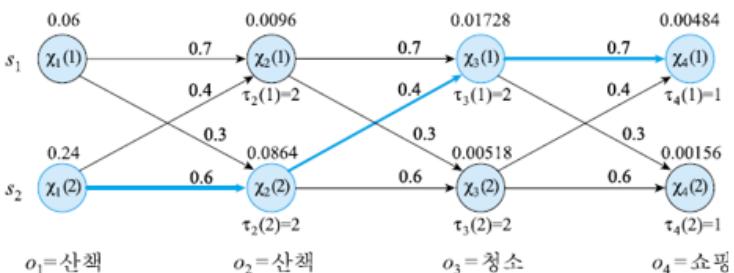


그림 7.17 Viterbi 알고리즘의 전방 계산과 최적 경로 역 추적 (파란색이 최적 경로)

$$t=1 \text{ 일 때, } \chi[1][1] = \pi_1 b_1(\text{산책}) = 0.6 \times 0.1 = 0.06$$

$$\chi[1][2] = \pi_2 b_2(\text{산책}) = 0.4 \times 0.6 = 0.24$$

$$\begin{aligned} t=2, \quad \chi[2][1] &= \max(\chi[1][1] a_{11}, \chi[1][2] a_{21}) b_1(\text{산책}) \\ &= (0.24 \times 0.4) \times 0.1 = 0.0096 \end{aligned}$$

$$\tau[2][1] = 2$$

$$\begin{aligned} \chi[2][2] &= \max(\chi[1][1] a_{12}, \chi[1][2] a_{22}) b_2(\text{산책}) \\ &= (0.24 \times 0.6) \times 0.6 = 0.0864 \end{aligned}$$

$$\tau[2][2] = 2$$

$$t=3, 4 \quad \chi[3][1] = \max(\chi[2][1] * a_{11}, \chi[2][2] * a_{21}) * b_1(\text{청소}) = (0.0864 * 0.4) * 0.5 = 0.01728$$

$$\tau[3][1] = 2$$

$$\chi[3][2] = \max(\chi[2][1] * a_{12}, \chi[2][2] * a_{22}) * b_2(\text{청소}) = (0.0864 * 0.6) * 0.1 = 0.00518$$

$$\tau[3][2] = 2$$

$$\chi[4][1] = \max(\chi[3][1] * a_{11}, \chi[3][2] * a_{21}) * b_1(\text{쇼핑}) = (0.01728 * 0.7) * 0.4 = 0.00484$$

$$\tau[4][1] = 1$$

$$\chi[4][2] = \max(\chi[3][1] * a_{12}, \chi[3][2] * a_{22}) * b_2(\text{쇼핑}) = (0.01728 * 0.3) * 0.3 = 0.00156$$

$$\tau[4][2] = 1$$

7.4.3 학습

· 학습 문제란?

- 관측 벡터 O 가 주어져 있을 때 HMM의 매개변수 Θ 는?
- 즉, O 의 발생 확률을 최대로 하는 Θ 를 찾는 문제
- 평가와 디코딩의 반대 작업
- 평가와 디코딩 같은 분석적 방법 없음
- 수치적 방법 필요
- 학습의 목적을 적어보면,

$$\hat{\Theta} = \arg \max_{\Theta} P(O|\Theta)$$

알고리즘 [7.3]

HMM 학습 알고리즘 스케치

입력: HMM 아키텍처 (n, m , 그래프 형태 등), 관측 벡터 $O = o_1 o_2 \cdots o_T$

출력: HMM $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \pi)$

알고리즘:

1. Θ 를 초기화 하라.
2. 적절한 방법으로 $P(\mathbf{O} | \Theta^{\text{new}}) > P(\mathbf{O} | \Theta)$ 인 개선된 Θ^{new} 를 찾아라.
3. 만족스러우면 $\hat{\Theta} = \Theta^{\text{new}}$ 로 하고 멈추고 그렇지 않으면 $\Theta = \Theta^{\text{new}}$ 로 하고 2로 가라.



• 가진 것과 찾아야 하는 것

- 가진 것 O ,

찾아야 하는 것 $\Theta = (\mathbf{A}, \mathbf{B}, \pi)$

- O 로 직접 Θ 를 찾을 수 없다.

→ 은닉 변수(latent variable)

γ 와 k 등장

→ 가우시언 혼합 추정과 유사성 있음

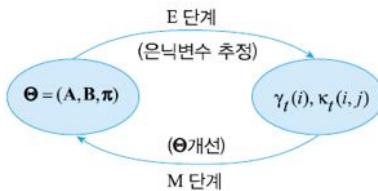


그림 7.18 Baum-Welch 학습 알고리즘은 EM 알고리즘의 일종임

입력: HMM 아키텍처 (n, m , 그래프 형태 등), 관측 벡터 $\mathbf{O} = o_1 o_2 \cdots o_T$

출력: HMM $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \pi)$

알고리즘:

1. Θ 를 초기화 하라.
2. // 개선된 Θ^{new} 찾기
 - 2.1 (E 단계) Θ 로 $\gamma_t(i), 1 \leq t \leq T, 1 \leq i \leq n$ 과 $\kappa_t(i, j), 1 \leq t \leq T-1, 1 \leq i, j \leq n$ 을 추정하라.
 - 2.2 (M 단계) $\gamma_t(i)$ 와 $\kappa_t(i, j)$ 를 가지고 Θ^{new} 를 찾아라.
3. 만족스러우면 $\hat{\Theta} = \Theta^{\text{new}}$ 로 하고 멈추고 그렇지 않으면 $\Theta = \Theta^{\text{new}}$ 로 하고 2로 가라.



• E단계 ($\gamma_t(i)$ 와 $\kappa_t(i, j)$ 의 추정)

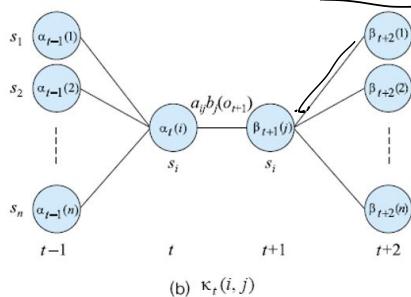
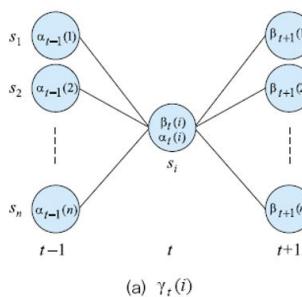


그림 7.19 은닉 변수 $\gamma_t(i)$ 와 $\kappa_t(i, j)$ 의 역할

• $\gamma_t(i)$ 는 시간 t 에서 상태 s_i 에 있을 확률

• $\kappa_t(i, j)$ 는 시간 t 에 s_i , $t+1$ 에 s_j 에 있을 확률

$$\gamma_t(i) = P(q_t = s_i | \mathcal{O}, \theta) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^n \alpha_t(j) \beta_t(j)}$$

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, q_t = s_i | \theta) = \left[\sum_{j=1}^n \alpha_{t-1}(j) \alpha_{ji} \right] b_i(o_t)$$

0의 일부 o_1, o_2, \dots, o_t 을 관측하고 t 일 때 상태 s_i 에 있을 확률

$$\beta_t(i) = p(o_{t+1}, o_{t+2}, \dots, o_T | q_t = s_i, \theta) = \left[\sum_{j=1}^n a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \right]$$

시간 t 에 상태 s_i 에 있다는 조건 하에, $t+1$ 부터 시작하여 나머지 기호 $o_{t+1}, o_{t+2}, \dots, o_T$ 를 관측할 확률



$$\alpha_t(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq n \quad (7.13)$$

$$\alpha_t(i) = \left[\sum_{j=1}^n a_{t-1}(j) a_{ji} \right] b_i(o_t), \quad 2 \leq t \leq T, \quad 1 \leq i \leq n \quad (7.14)$$

$$\beta_T(i) = 1, \quad 1 \leq i \leq n$$

$$\beta_t(i) = \sum_{j=1}^n a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad 1 \leq t \leq T-1, \quad 1 \leq i \leq n$$

$k_t(i, j)$ 는 모델 θ 와 관측 벡터 o 가 주어진 조건 하에서 시간 t 에서는 s_i , 그리고 $t+1$ 에서는 s_j 에 있을 확률이다.

$$k_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^n \alpha_t(k) a_{kj} b_j(o_{t+1}) \beta_{t+1}(j)}, \quad 1 \leq t \leq T-1, \quad 1 \leq i, j \leq n$$

→ 시간 t 와 $t+1$ 사이에는 총 n^2 개의 길이 있다. 따라서 $k_t(i, j)$ 는 $s_i \rightarrow s_j$ 를 거쳐갈 확률 $\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$ 을 n^2 개 길 각각에 대해 그것을 거쳐갈 확률을 구하고, 그것들을 모두 더한 값으로 나누어 주면 된다.

• M단계

- E단계에서 구한 γ 와 k 로 θ^{new} 의 재추정
- $P(O|\theta^{\text{new}}) > P(O|\theta)$ 여야 함. θ^{new} 는 θ 보다 O 를 '잘 설명해야' 함

$$\alpha_{ij}^{\text{new}} = \frac{S_i \text{에서 } S_j \text{로 이전할 기대값}}{S_i \text{에서 이전할 기대값}}$$

$$= \frac{t=1 \text{ 일 때 } S_i \text{에서 } S_j \text{로 이전할 확률} + \dots + t=T-1 \text{ 일 때 } S_i \text{에서 } S_j \text{로 이전할 확률}}{t=1 \text{ 일 때 } S_i \text{에서 이전할 확률} + \dots + t=T-1 \text{ 일 때 } S_i \text{에서 이전할 확률}}$$

$$= \frac{\sum_{t=1}^{T-1} k_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad 1 \leq i, j \leq n$$

$$b_i(U_k)^{\text{new}} = \frac{S_i \text{에서 } U_k \text{를 관측할 기대값}}{S_i \text{에 있을 기대값}}$$

$$= \frac{\theta_t = U_k \text{인 모든 } t \text{에 대해 } S_i \text{에 있을 확률의 합}}{t=1 \text{ 일 때 } S_i \text{에 있을 확률} + \dots + t=T \text{ 일 때 } S_i \text{에 있을 확률}}$$

$$= \frac{\sum_{t=1}^T \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)} \quad , \quad 1 \leq i \leq n, 1 \leq k \leq m$$

$$\pi_i^{\text{new}} = t \text{ 가 } 1 \text{ 일 때 } S_i \text{에 있을 확률} = \gamma_1(i), 1 \leq i \leq n$$

입력: HMM 아키텍처 (n, m , 그래프 형태 등), 관측 벡터 $\mathbf{O} = o_1 o_2 \cdots o_T$

출력: HMM $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \pi)$

알고리즘:

난수 발생 시킴

1. Θ 를 초기화 하라.

2. repeat {

// E 단계

// 라인 3의 α 는 (7.13)~(7.14), β 는 (7.24)~(7.25)

3. \mathbf{O} 와 Θ 를 가지고 $\alpha_t(i), 1 \leq t \leq T, 1 \leq i \leq n$ 과 $\beta_t(i), 1 \leq t \leq T, 1 \leq i \leq n$ 을 구한다.

// 라인 4의 γ 는 (7.22), κ 는 (7.26)

4. α 와 β 로 $\gamma_t(i), 1 \leq t \leq T, 1 \leq i \leq n$ 과 $\kappa_t(i, j), 1 \leq t \leq T-1, 1 \leq i, j \leq n$ 을 계산한다.

// M 단계

5. γ 와 κ 를 가지고 $a_i^{\text{new}}, 1 \leq i, j \leq n$ 을 추정한다. // (7.27)

6. γ 를 가지고 $b_i(v_k)^{\text{new}}, 1 \leq i \leq n, 1 \leq k \leq m$ 을 추정한다. // (7.28)

7. γ 를 가지고 $\pi_i^{\text{new}}, 1 \leq i \leq n$ 을 추정한다. // (7.29)

8. $\Theta = (\mathbf{A}^{\text{new}}, \mathbf{B}^{\text{new}}, \pi^{\text{new}});$

9. } until (멈춤 조건): new 추정값과 이전값의 차이

10. $\hat{\Theta} = \Theta;$

새로운 Θ 에서의 모든 확률이 어떤 임계값 이상

→ 반드시 수렴한다, 욕심 알고리즘이므로 전역 최적점 보장 못한다.

* 어고딕과 좌우모델

어고딕 모델: 허가에 따라 α 값의 크기 빠르게 작아짐. 1보다 훨씬 작은 값을 계속 곱해나가기 때문. 따라서 벡터의 길이 T 가 커지면 컴퓨터 메모리의 한계를 넘어 수치 오류 발생할 수 있음 \rightarrow scaling problem

좌우모델

- 어떤 상태에서 오른쪽 이후 상태로 전이해 버리면 다시는 자신으로 돌아올 수 없음. 따라서 T가 충분히 크더라도 학습 도중 자신이 방문될 기회가 충분치 않을 수 있음
- 이 모델은 주로 온라인 팬 인식, 음성인식에서 사용, T가 작음 대신 하나의 관측(샘플)이 주어지는 것이 아니라 여러개가 주어짐

7.5 부연설명

1. HMM의 출력은 확률

→ 다중 분류기 결합 등의 사후 작업에 출력 결과가 유용하게 쓰임

2. 샘플 생성 능력

알고리즘 [7.6]

HMM에 의한 관측 벡터 생성

입력: HMM $\Theta = (\mathbf{A}, \mathbf{B}, \pi)$, 관측 벡터의 길이 T

출력: 관측 벡터 $\mathbf{O} = o_1 o_2 \dots o_T$

알고리즘:

1. π 에 따라 초기 상태 q_1 을 결정한다.
2. $t = 1;$
3. **while** ($t \neq T$) {
 4. **B**에 따라 상태 q_t 에서 관측 값을 결정하고 그것을 o_t 에 기록한다.
 5. **A**에 따라 다음 상태 q_{t+1} 을 결정한다.
 6. $t++;$}
7. }



3. 예측 목적으로 사용
4. 분류기로 사용할 수 있음
 - 부류 별로 독립적으로 HMM 구축
$$O \in q = \arg \max_j P(O | \theta_{w_j}) P(w_j) \text{ 일 때 } w_q \text{ 로 분류하라.}$$
5. 적절한 아키텍처를 선택해야 한다.
 - 그녀의 삶 → 고딕 모델
 - 음성인식 → 좌우 모델
6. 상태의 개수를 적절히 해야 함
7. 공개 소프트웨어 있음