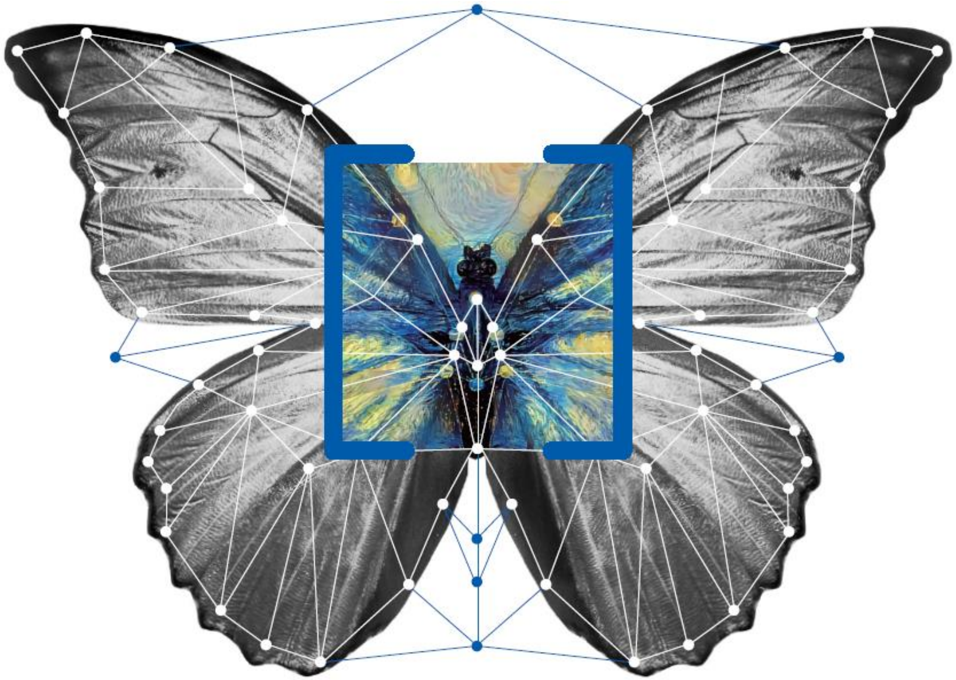
한빛아카데미
Hanbit Academy, Inc.



MACHINE LEARNING 기계 학습

오일석 지음

10. 확률 그래피컬 모델

PREVIEW

확률 추론 문제

- 흡연 환자의 엑스레이 진단에서 양성이 나타났을 때 폐암일 확률은?
- 한국은행이 기준금리를 올렸고 S전자의 1분기 실적이 양호일 때 S전자의 주식이 오를 확률은?

→답을 구할 수 있다면 매우 유용

확률 그래피컬 모델probabilistic graphical model

- 엑스레이, 흡연, 폐암을 확률변수로 정의하고 이들의 상호작용을 그래프로 표현하고, 그래프에서 확률 추론 수행
- 대표적 모델 3가지
  - 베이지안 네트워크
  - 마르코프 랜덤필드
  - RBM과 DBN

각 절에서 다루는 내용

10.1절_ 확률 그래피컬 모델의 유형을 구분하고 원리를 소개한다.

10.2절_ 베이지안 네트워크가 독립성을 이용하여 확률을 추론하는 방법을 설명한다.

10.3절_ 마르코프 네트워크가 에너지함수를 이용하여 확률을 추론하는 방법을 설명한다.

10.4절_ 딥러닝을 촉발시켰다고 평가되는 RBM과 RBM을 깊게 쌓은 DBN을 설명한다.

10.1 확률과 그래프의 만남

- 10.1.1 그래프 표현
- 10.1.2 그래프 분해와 확률 표현

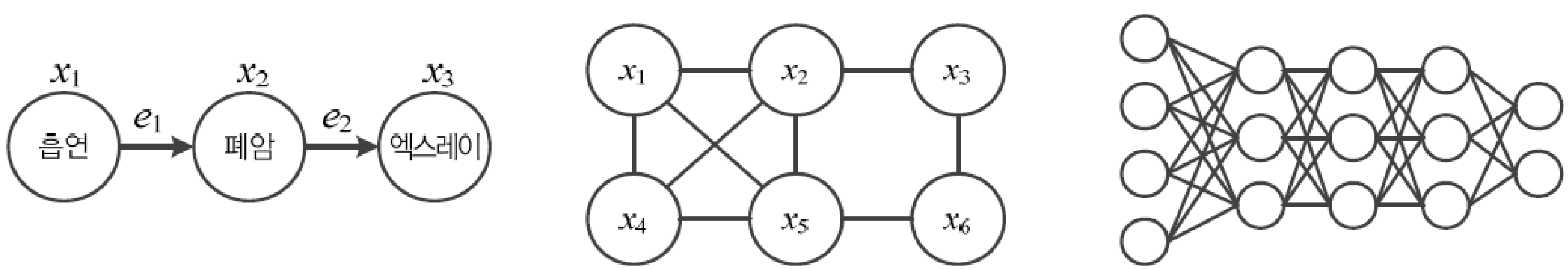
10.1.1 그래프 표현

- 그래프 표현
  - 예, 흡연 환자의 엑스레이 진단에서 양성이 나타났을 때 폐암일 확률은?
    - 주요 요인을 확률변수로 뽑음(엑스레이, 흡연, 폐암) - 노드
    - 인과관계 설정 -에지
  - 그래프는 확률변수의 상호작용을 표현하는 뼈대
  - 뼈대에 확률을 부여하면 확률 그래피컬 모델이 됨

10.1.1 그래프 표현

대표적인 3가지 모델

- 베이지안 네트워크Bayesian network
  - 인과관계를 나타내기 위해 방향 에지를 사용
- 마르코프 랜덤필드
  - 인과관계가 없어 무방향 에지를 사용
- DBN
  - RBM을 여러 층으로 쌓아 만듦



(a) 베이지안 네트워크(방향 그래프)　(b) 마르코프 랜덤필드(무방향 그래프)　(c) DBN(깊은 신경망)

그림 10-1 확률 그래피컬 모델

10.1.2 그래프 분해와 확률 표현

- 그래프 $G = \{X, E\}$
  - 노드의 집합 $X = \{x_1, x_2, \cdots, x_n\}$
  - 에지의 집합 $E = \{e_1, e_2, \cdots, e_m\}$
- 완전 그래프 예, [그림 10-2]
  - 결합확률 $P(\mathbf{x}) = P(x_1, x_2, x_3)$을 부여해야 함
  - 확률변수가 가질 수 있는 값이 $x_1 \in \{smoking, non\_smoking\}$, $x_2 \in \{lung\_cancer, not\_lung\_cancer\}$, $x_3 \in \{positive, negative\}$라면 8개($2^3$) 확률값을 지정해야 함

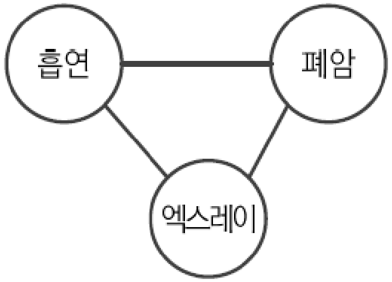


그림 10-2 세 확률변수의 완전 그래프(결합 확률 필요)

$$P(smoking, lung\_cancer, positive) = 0.02,\ P(non\_smoking, lung\_cancer, positive) = 0.01,$$
$$P(smoking, lung\_cancer, negative) = 0.01,\ P(non\_smoking, lung\_cancer, negative) = 0.01,$$
$$P(smoking, not\_lung\_cancer, positive) = 0.03,\ P(non\_smoking, not\_lung\_cancer, positive) = 0.02,$$
$$P(smoking, not\_lung\_cancer, negative) = 0.20,\ P(non\_smoking, not\_lung\_cancer, negative) = 0.70$$

10.1.2 그래프 분해와 확률 표현

- 결합확률
  - 완벽한 확률 정보로서 모든 확률 추론 가능
    - 예, 엑스레이가 양성일 때 폐암 확률은?
    - 예, 흡연자의 폐암 확률은?
  - 결합확률을 알아내는 일은 차원의 저주
    - n개 확률변수가 있고, 각각 q개의 값을 가진다면 총 $q^n - 1$개의 확률을 알아내야 함

- 그래프 분해
  - 직접 상호작용하는 확률변수만 에지로 연결함
  - 확률 그래피컬 모델의 핵심 아이디어
    - 베이지안 네트워크는 직접적인 인과관계가 있는 변수만 방향 에지로 연결
    - 마르코프 랜덤필드는 이웃한 변수만 무방향 에지로 연결
    - DBN은 이웃한 층 사이에만 무방향 에지로 연결
  - 결합확률을 알아낼 필요가 없어지고, 분해된 그래프에서 부분집합의 확률분포만 알아내면 됨
  - 에지 연결이 없는 노드는 중 간 노드를 통해 상호작용을 함(예, 흡연과 엑스레이는 폐암을 통해 상호작용)

10.2 베이지안 네트워크

- 10.2.1 간단한 예제
- 10.2.2 그래프 분해
- 10.2.3 d-분리
- 10.2.4 확률 추론

- 베이지안 네트워크의 장점
  - 마르코프 랜덤필드나 DBN보다 더 엄격한 확률 모델(다른 두 모델은 에너지 함수를 통해 간접적으로 확률 연산을 수행하는 반면, 베이지안 네트워크는 데이터로부터 확률 추정)
  - 확률변수 사이의 인과관계를 조건부 확률로 표현하므로 불완전 데이터를 처리할 수 있음(일부 확률변수를 관찰했을 때 나머지 변수 중 관심 있는 것의 확률을 계산할 수 있음)
  - 데이터와 전문가 지식을 혼용해 사용할 수 있음

10.2 베이지안 네트워크

세 가지 주요 문제

- 구조 학습structure learning: [그림 10-1(a)]와 같은 그래프 구조를 만드는 작업이다. 확률변수는 사람이 결정해야 하며, 확률변수가 정해지면 이들 간의 인과관계는 사람이 지정하거나 데이터로부터 자동으로 알아낼 수 있다.
- 확률 학습probability learning: 노드 또는 노드와 노드 사이에 확률을 부여하는 작업이다. 부모가 없는 루트 노드는 사전 확률, 부모가 있는 노드는 조건부 확률을 알아낸다.
- 확률 추론probabilistic inference: 구조 학습과 확률 학습을 마친 후 테스트 단계 또는 현장 설치 후 수행하는 작업이다. 흡연자의 폐암 확률을 알아내는 것과 같은 각종 질문에 대한 확률을 추정하는 일이다.

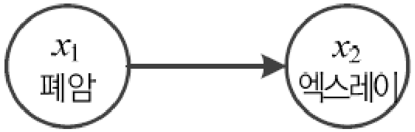
10.2.1 간단한 예제

예제 10-1 베이지안 네트워크로 폐암 진단

병원에서는 폐암 진단을 위해 엑스레이 사진을 찍는다. 이때 폐암과 엑스레이를 확률변수로 뽑고, 각각을 x_1과 x_2로 표기하자. 자칫 엑스레이 결과를 보고 폐암을 진단하므로 엑스레이→폐암이라는 인과관계를 맺으려 할 수 있는데, 자연계 현상에서는 폐암 여부가 엑스레이 결과를 좌우하므로 [그림 10-3]과 같이 폐암→엑스레이라고 표현해야 한다.

청정지역으로 유명한 마을에 사는 길동은 정기건강검진을 하던 중 엑스레이에서 양성 반응이 나타났다. 공황상태에 빠진 길동은 문득 기계 학습 과목에서 배운 베이지안 네트워크 이론이 떠올랐다. [그림 10-3]과 같이 노드 2개를

가진 베이지안 네트워크의 구조를 설계한 다음, 통계청의 의료 데이터를 열람하여 자신이 사는 지역의 폐암 환자 비율이 0.001, 즉 1천 명당 1명꼴이라는 사실을 알아낸다. 또한, 의사로부터 엑스레이는 완벽하지 않다는 말을 듣고, 폐암 환자 중 60%만 양성 반응이 나타나며 간혹 폐암이 아닌 정상인 100명에 2명꼴로 양성 반응이 나타난다는 설명을 들었다. 즉, 참 긍정률[true positive rate]이 0.6이고 참 부정률[false negative rate]이 0.98이다.[1] 길동은 이 통계 데이터를 그래프에 추가하여 [그림 10-3]의 베이지안 네트워크를 완성하였다.



$P(lung\_cancer) = 0.001$
$P(not\_lung\_cancer) = 0.999$

$P(positive|lung\_cancer) = 0.6$
$P(negative\ lung\_cancer) = 0.4$
$P(positive|not\_lung\_cancer) = 0.02$
$P(negative\ not\_lung\_cancer) = 0.98$

그림 10-3 노드가 2개인 베이지안 네트워크

10.2.1 간단한 예제

길동은 자신이 알고 싶어하는 확률, 즉 엑스레이가 양성인 조건에서 폐암일 확률을 수식 $P(x_1 = lung_cancer|x_2 = positive)$로 표현하였다. 그리고 2장의 식 (2.26)의 베이즈 정리를 이용하여 다음과 같이 계산하였다.

$$P(lung\_cancer|positive) = \frac{P(positive|lung\_cancer)P(lung\_cancer)}{P(positive)}$$

$$= \frac{P(positive|lung\_cancer)P(lung\_cancer)}{P(positive|lung\_cancer)P(lung\_cancer) + P(positive|not\_lung\_cancer)P(not\_lung\_cancer)}$$

$$= \frac{0.6 * 0.001}{0.6 * 0.001 + 0.02 * 0.999} = 0.029$$

길동은 엑스레이에서 양성 반응이 나타났지만 폐암일 확률은 불과 2.9%에 불과하다는 확률 추론 결과를 보고 안도하였다. 그리고 정밀 검사를 받기로 하였다.

10.2.1 간단한 예제

- 세 가지 주요 문제
  - [그림 10-3]의 그래프 구조를 만드는 일은 구조 학습
  - 통계청과 병원에서 확률을 수집한 일은 확률 학습
  - 엑스레이에서 양상이 나타난 자신이 폐암일 확률을 계산한 일은 확률 추론

- 미세먼지에 뒤덮인 탄광 마을 주민에 적용하면,
  - 탄광의 폐암 환자 비율이 0.5%라고 가정하면, 양성 반응인 사람의 폐암 확률은 13.1%

$$P(lung\_cancer|positive) = \frac{0.6 * 0.005}{0.6 * 0.005 + 0.02 * 0.995} = 0.131$$

10.2.1 간단한 예제

예제 10-2 **젖은 잔디**

비가 오거나 스프링클러를 틀면 잔디는 젖은 상태가 된다. 비가 오면 스프링클러를 틀 필요가 없다. 이 상황을 베이지안 네트워크로 표현해 보자. 먼저 비, 스프링클러, 잔디라는 3개의 확률변수를 뽑았다고 하자. 그리고 모든 변수가 두 가지 상태만 가진다고, 즉 비 $\in \{rain, not\_rain\}$, 스프링클러 $\in \{on, off\}$, 잔디 $\in \{wet, dry\}$라 가정하자. 어느 정도 관찰한 결과, [그림 10-4]와 같은 확률을 얻었다.

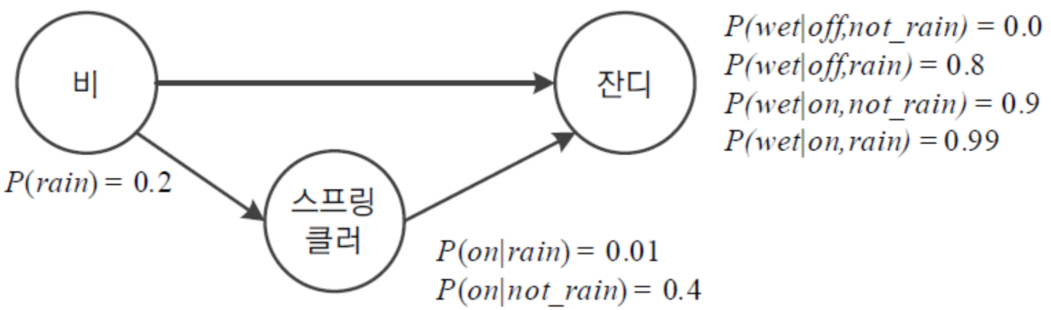


그림 10-4 **노드가 3개인 베이지안 네트워크**

확률은 아래와 같이 표로 표현할 수도 있다. 확률의 성질에 따라 행 방향으로 값을 더하면 항상 1이 되어야 한다. 따라서, 두 번째 열을 생략하여도 된다. [그림 10-4]에서는 이 성질을 이용하여 전체 경우 중 반만 제시하였다.

비

rain	*not-rain*
0.2	0.8

스프링쿨러

비	*on*	*off*
not_rain	0.4	0.6
rain	0.01	0.99

잔디

스프링쿨러	비	*wet*	*dry*
off	*not_rain*	0.0	1.0
off	*rain*	0.8	0.2
on	*not_rain*	0.9	0.1
on	*rain*	0.99	0.01

10.2.1 간단한 예제

이제 완성된 베이지안 네트워크로부터 다양한 확률을 추론할 수 있다. 예를 들어, 비가 오지 않았는데 스프링클러가 켜져 있을 확률을 구하고자 하면 $P(\text{스프링클러} = on|\text{비} = non_rain)$을 구하는 셈이므로 두 번째 표를 참조하여 40%라고 답하면 된다.

식 (10.1)을 적용하여 다음과 같이 결합확률을 구할 수도 있다. 총 8가지가 가능한데, 4가지만 보인다.

$$P(wet, on, rain) = P(rain)P(on|rain)P(wet|on, rain) = 0.2 * 0.01 * 0.99 = 0.00198$$
$$P(wet, on, not\_rain) = P(not\_rain)P(on|not\_rain)P(wet|on, not\_rain) = 0.8 * 0.4 * 0.9 = 0.288$$
$$P(wet, off, rain) = P(rain)P(off|rain)P(wet|off, rain) = 0.2 * 0.99 * 0.8 = 0.1584$$

$$P(wet, off, not\_rain) = P(not\_rain)P(off|not\_rain)P(wet|off, not\_rain) = 0.8 * 0.6 * 0.0 = 0.0$$

잔디가 젖어 있는 것을 관찰했을 때 비가 왔을 확률 $P(rain|wet)$을 구해 보자. 다음과 같이 계산하며, 35.77%라는 것을 알 수 있다.

$$\begin{aligned}
P(rain|wet) &= \frac{P(rain, wet)}{P(wet)} \\
&= \frac{\sum_{S\in\{on,off\}} P(wet, S, rain)}{\sum_{S\in\{on,off\}} \sum_{R\in\{rain,not_rain\}} P(wet, S, R)} \\
&= \frac{P(wet, on, rain) + P(wet, off, rain)}{P(wet, on, rain) + P(wet, on, not_rain) + P(wet, off, rain) + P(wet, off, not_rain)} \\
&= \frac{0.00198 + 0.1584}{0.00198 + 0.288 + 0.1584 + 0.0} = 0.3577
\end{aligned}$$

10.2.1 간단한 예제

- 확률 부여에서 중요한 점
  - 루트 노드는 사전 확률
  - 루트가 아닌 노드는 조건부 확률을 가짐

← 10.2.2절에서 마르코프 성질을 이용하여 엄밀하게 정의

10.2.2 그래프 분해

- 결합확률을 전개하면 식 (10.1)이 성립

$$
\begin{aligned}
P(\mathbf{x}) &= P(x_1, x_2, \cdots, x_n) \\
&= P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)P(x_4|x_3, x_2, x_1) \cdots P(x_n|x_{n-1}, x_{n-2}, \cdots, x_1)
\end{aligned}
\tag{10.1}
$$

- [그림 10-5]의 예제에서
  - 위상 정렬한 후, 식 (10.1)을 적용하면

$$
\begin{aligned}
&P(\text{흡연, 폐렴, 폐암, 피로, 엑스레이}) \\
&= P(\text{흡연})P(\text{폐렴}|\text{흡연})P(\text{폐암}|\text{폐렴, 흡연})P(\text{피로}|\text{폐암, 폐렴, 흡연}) \\
&\quad P(\text{엑스레이}|\text{피로, 폐암, 폐렴, 흡연})
\end{aligned}
\tag{10.2}
$$

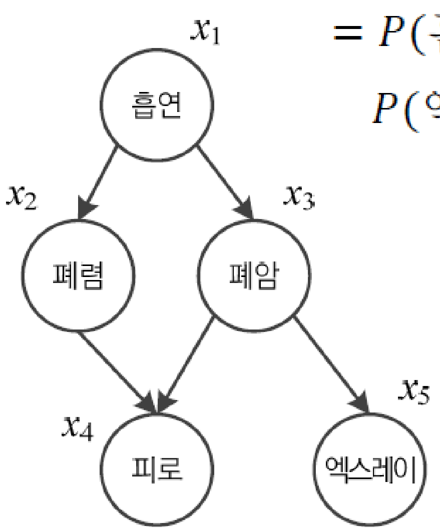


그림 10-5 노드가 5개인 인과관계 그래프

10.2.2 그래프 분해

- 마르코프 조건
  - 노드 x의 부모를 y라 할 때, y의 값이 주어지면 x는 비후손(후손을 제외한 모든 노드)와 조건부 독립
  - 예, $P(\text{엑스레이} \mid \text{피로}, \text{폐암}, \text{폐렴}, \text{흡연}) = P(\text{엑스레이} \mid \text{폐암})$

- 마르코프 조건을 식 (10.2)에 적용하면,

$$P(\text{흡연}, \text{폐렴}, \text{폐암}, \text{피로}, \text{엑스레이})$$
$$= P(\text{흡연})P(\text{폐렴} \mid \text{흡연})P(\text{폐암} \mid \text{흡연})P(\text{피로} \mid \text{폐암}, \text{폐렴})P(\text{엑스레이} \mid \text{폐암})$$

10.2.2 그래프 분해

- 이렇게 유도된 식에 따라 확률을 부여하면
  - [그림 10-6]의 베이지안 네트워크가 됨

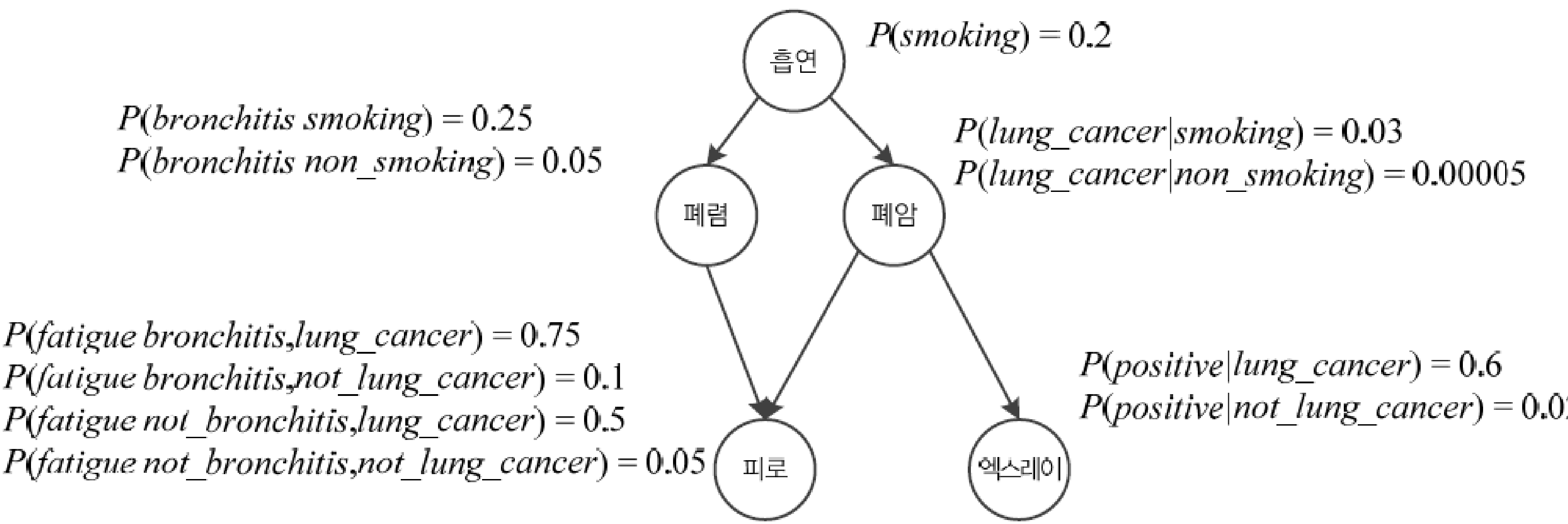


그림 10-6 베이지안 네트워크

10.2.2 그래프 분해

- 확률 학습
  - 부모 자식 사이에만 확률을 부여하면 됨

베이지안 네트워크에 확률을 부여하는 방법: 부모가 없는 루트 노드에는 사전 확률을 부여하고, 부모가 있는 노드에는 부모와 자식 사이로 한정하여 조건부 확률을 부여한다.

- 확률 학습 수행
  - 해당 분야 전문가가 경험이나 보유한 데이터를 기반으로 부여
  - 훈련집합을 가지고 자동 학습

10.2.3 d-분리

조건부 독립

- 세 가지 연결 패턴: 선형, 분기, 합류

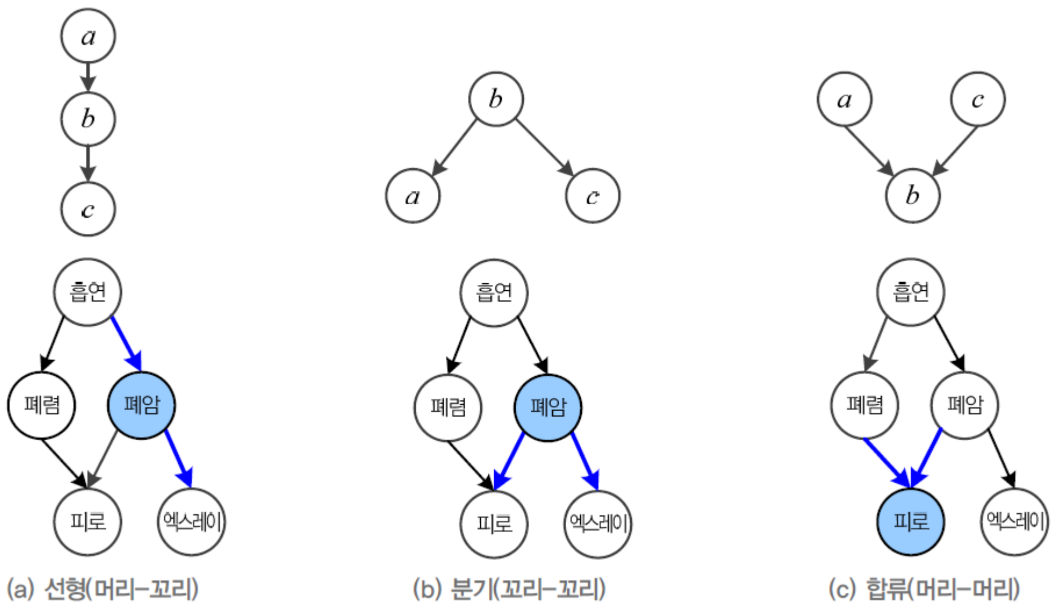


그림 10-7 세 가지 인과관계와 예제 체인

- 체인: 에지로 연결된 길
  - 예, 흡연→피로→폐렴
  - 예, 엑스레이←폐암→피로←폐렴(조상이 자식에게 영향을 미치기 때문에 역방향도 인정)

10.2.3 d-분리

- 선형
  - [그림 10-7(a)]에서 파란색은 값이 주어진 노드
  - 폐암 여부가 알려졌으므로, 마르코프 조건에 따라 엑스레이는 흡연과 조건부 독립임
    - 즉, $P(\text{엑스레이}|\text{폐암}, \text{흡연}) = P(\text{엑스레이}|\text{폐암})$
    - 독립 표기를 사용하면 $I(\text{엑스레이}, \text{흡연}|\text{폐암})$
- 분기
  - b를 모르면 a와 c는 독립이 아니고, b를 알면 a와 c는 독립임
  - 즉, $I(\text{피로}, \text{엑스레이}|\text{폐암})$
  - b를 알면 $a \leftarrow b \rightarrow c$ 체인은 폐쇄된다,라고 표현
- 합류
  - b를 모르면 a와 c는 독립, b를 알면 a와 c는 독립이 아님(분기와 반대)
  - b를 알면 a와 c는 독립이 아님을 설명됨explaining away 현상(할인 현상):
    - 예, 피로를 느낀다는 사실을 알게 되면, 폐암과 폐렴은 서로 영향을 줌(폐암이면 폐렴 가능성은 낮아짐)
    - b를 알면 $a \rightarrow b \leftarrow c$ 체인은 열린다,라고 표현

10.2.3 d-분리

예제 10-3 **조건부 독립**

[그림 10-8(a)]는 [그림 10-5]의 그래프에서 일부를 자르고 짜증이라는 노드를 추가한 것이다. 이 상황에서 I(피로, 흡연|폐렴)은 [그림 10-7(a)]의 선형 구조에 따라 성립한다. 그런데 짜증도 흡연과 조건부 독립일까? 즉 I(짜증, 흡연|폐렴)일까? 답은 '예'이다. 이러한 조건부 독립은 체인이 아주 길어도 성립한다. 또한, 값이 알려진 변수를 중심으로 위쪽에 있는 변수집합과 아래쪽에 있는 변수집합 사이에도 조건부 독립이 성립한다. 예를 들어 I({피로, 짜증}, 흡연|폐렴)이다. 폐렴이 체인을 폐쇄하였다.

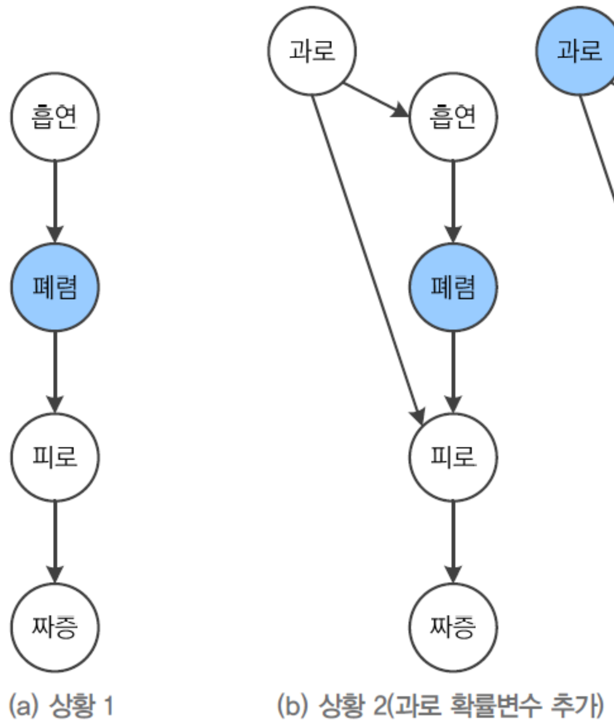


그림 10-8 조건부 독립을 설명하는 예제

10.2.3 d-분리

[그림 10-8(a)]에 과로라는 확률변수를 추가하여 약간 더 복잡한 [그림 10-8(b)]를 만들었다. 과로는 흡연과 피로에 직접적인 영향을 준다고 가정하여 에지로 연결하였다. 이 상황에서도 I(짜증, 흡연|폐렴)일까? 즉, 폐렴 여부를 알게 되었을 때 짜증과 흡연은 독립일까? 답은 '아니다'이다. 왜냐하면 흡연한다는 사실은 과로일 가능성을 키우고, 과로는 피로일 가능성, 피로는 짜증일 가능성을 키우기 때문이다.

[그림 10-8(b)]의 두 번째 그림에 해당하는 것으로, 폐렴 여부도 알고 과로 여부도 안다고 가정하자. 짜증과 흡연이 조건부 독립일까? 즉, I(짜증, 흡연|과로, 폐렴)일까? 답은 '그렇다'이다. 과로가 값을 가져 더는 흡연이 피로에 영향을 미치지 못하기 때문이다. 과로 확률변수가 흡연←과로→피로→짜증 체인을 폐쇄하였다.

흡연과 피로 여부를 아는 [그림 10-8(b)]의 세 번째 상황을 생각해 보자. 이때 과로와 폐렴은 조건부 독립, 즉 I(과로, 폐렴|흡연, 피로)일까? 답은 '아니다'이다. 왜냐하면 이 상황에서 환자가 과로한다는 사실을 알게 되면 폐렴일 가능성을 작게 보고, 반대로 과로하지 않는다는 사실을 알게 되면 폐렴을 의심, 즉 폐렴 가능성을 크게 보므로 과로와 폐렴은 조건부 독립이 아니다. 설명됨explaining away이라는 현상이 발생하였다.

10.2.3 d-분리

체인의 폐쇄

정의 10-1 체인의 폐쇄

a와 c를 잇는 체인 $a \rightsquigarrow c$와 노드 집합 $\mathcal{W}$가 주어졌을 때, 다음 조건 중 하나라도 성립하면 체인은 $\mathcal{W}$에 의해 폐쇄되었다고 말한다.

(1) (선형) $\mathcal{W}$에 속한 노드가 체인에 머리-꼬리 형태로 나타난다.

(2) (분기) $\mathcal{W}$에 속한 노드가 체인에 꼬리-꼬리 형태로 나타난다.

(3) (합류) 체인에 머리-머리 형태의 노드가 있을 때, 이 노드와 노드의 자손이 모두 $\mathcal{W}$에 속하지 않는다.

10.2.3 d-분리

예제 10-4 체인의 폐쇄

[그림 10-9]는 12개 확률변수를 가진 베이지안 네트워크이다. 여기에서 체인 $d \rightsquigarrow h$에 대한 폐쇄 여부를 확인하자.

- $d \rightarrow e \rightarrow f \rightarrow g \rightarrow h$는 $W = \{f\}$에 의해 폐쇄된다. 이 체인에 W에 속하는 머리-꼬리 노드 f가 있기 때문이다.

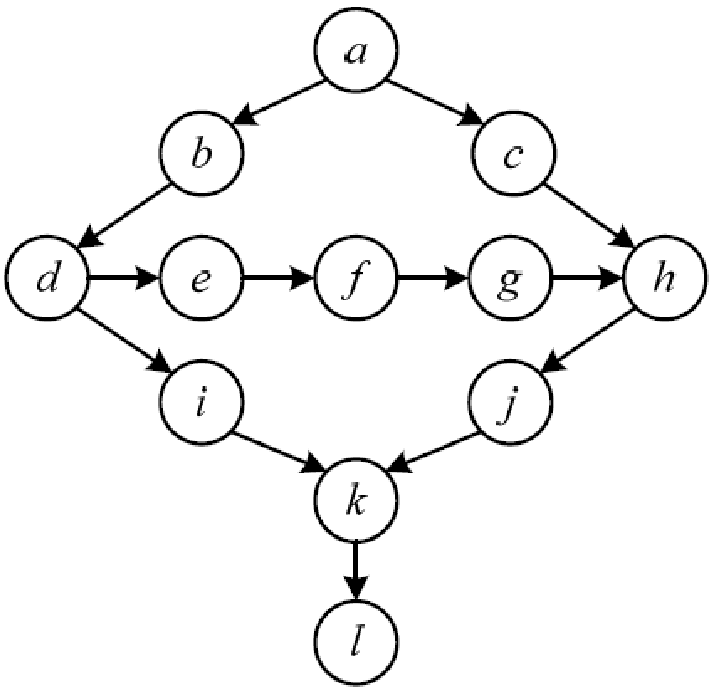


그림 10-9 체인의 폐쇄를 설명하는 예제

10.2.3 d-분리

- $d \rightarrow e \rightarrow f \rightarrow g \rightarrow h$는 $\mathcal{W} = \{f, a\}$에 의해 폐쇄된다. 이 체인에 $\mathcal{W}$에 속하는 머리-꼬리 노드 f가 있기 때문이다.
- $d \leftarrow b \leftarrow a \rightarrow c \rightarrow h$는 $\mathcal{W} = \{f, a\}$에 의해 폐쇄된다. 이 체인에 $\mathcal{W}$에 속하는 꼬리-꼬리 노드 a가 있기 때문이다.
- $d \rightarrow i \rightarrow k \leftarrow j \leftarrow h$는 $\mathcal{W} = \{f, a\}$에 의해 폐쇄된다. 이 체인에 머리-머리 노드 k가 있는데, k와 k의 자손 l이 $\mathcal{W}$에 속하지 않기 때문이다.
- $d \rightarrow i \rightarrow k \leftarrow j \leftarrow h$는 $\mathcal{W} = \{f, a, k\}$에 의해 폐쇄되지 않는다. 이 체인에 머리-머리 노드 k가 있는데, k가 $\mathcal{W}$에 속하기 때문이다.
- $d \rightarrow i \rightarrow k \leftarrow j \leftarrow h$는 $\mathcal{W} = \{f, a, k, j\}$에 의해 폐쇄된다. 이 체인에 머리-꼬리 노드 j가 있는데, j가 $\mathcal{W}$에 속하기 때문이다.

10.2.3 d-분리

- d-분리
  - 앞에서는 두 노드를 잇는 체인 하나의 폐쇄 여부를 따짐
  - 여기서는 두 노드를 잇는 체인에 여럿일 수 있으므로, 두 노드가 완전히 폐쇄되어 조건부 독립을 이루는지 확인

정의 10-2 d-분리

두 노드 a와 c 사이에 있는 모든 체인이 노드 집합 $\mathcal{W}$에 의해 폐쇄되면 두 노드는 d-분리[d-separated] 되었다고 하고, $d-sep(a, c|\mathcal{W})$라고 표기한다. 이 정의를 노드 집합으로 확장할 수 있다. 두 노드 집합 $\mathcal{A}$와 $\mathcal{C}$ 사이에 있는 모든 체인이 노드 집합 $\mathcal{W}$에 의해 폐쇄되면 두 노드 집합은 d-분리되었다고 하고, $d-sep(\mathcal{A}, \mathcal{C}|\mathcal{W})$라고 표기한다.

10.2.3 d-분리

예제 10-5 **d-분리**

[그림 10-5]의 예제 베이지안 네트워크를 사용하여 d-분리를 따져 보자. 몇 가지 질문에 대한 답과 이유를 제시한다.

- $d-sep$(흡연, 피로|폐렴)이 아니다. 폐쇄되지 않은 흡연→폐암→피로라는 체인이 있기 때문이다.
- $d-sep$(흡연, 피로|{폐렴, 폐암})이다. 2개의 체인, 흡연→폐렴→피로와 흡연→폐암→피로가 모두 폐쇄되었다.
- $d-sep$(폐렴, 엑스레이|피로)가 아니다. 폐렴→피로←폐암→엑스레이라는 체인이 피로에 의해 열렸기 때문이다.

[그림 10-9] 베이지안 네트워크의 d-분리를 생각해 보자.

- $d-sep(d, h|\{b, f\})$이다. d와 h를 잇는 체인 3개가 모두 폐쇄되었다.
- $d-sep(d, h|f)$가 아니다. d와 h를 잇는 $d \leftarrow b \leftarrow a \rightarrow c \rightarrow h$가 열려 있다.
- $d-sep(d, h|\{a, f, j, k\})$이다. d와 h를 잇는 체인 3개가 모두 폐쇄되었다.
- $d-sep(d, h|\{a, f, l\})$가 아니다. d와 h를 잇는 $d \rightarrow i \rightarrow k \leftarrow j \leftarrow h$가 열려 있다.

10.2.4 확률 추론

- d-분리와 확률 추론
  - 부모의 값이 지정된 경우 자식의 확률은 표 읽기로 알 수 있음
    - 예, 어떤 환자가 폐렴인데 폐암은 아니라는 사실을 알게 되면 피로를 느낄 확률은 확률 표에서 $P(fatigue|bronchitis, not_lung_cancer)$를 읽으면 됨 → 답은 0.1
  - 부모가 아닌 노드의 값까지 알게 되었다면,
    - 예, 비흡연자라는 사실까지 알게 되면 피로를 느낄 확률은? 즉 $P(fatigue|brochitis, not_lung_cancer, non_smoking)$은? ← 같을까?
    - *d-sep*(흡연,피로|{폐렴,폐암})이기 때문에 같다.

  ← 이렇게 확률을 알아내는 과정을 확률 추론이라 부름

10.2.4 확률 추론

- 현실 세계의 확률 추론
  - 주로 역방향 확률을 알아내고자 함
    - 예, 엑스레이가 $positive$일 때, 폐암일 확률은?
  - 아래에 있는 노드일수록 관찰 가능하고(정보 확률변수), 위쪽으로 갈수록 관찰 결과의 원인에 해당함(가설 확률변수)
    - 관찰을 통해 정보 확률변수를 알게 되었을 때, 가설 확률변수의 값을 알고자 함
    - 예, $P(lung_cancer|positive,not_fatigue)$를 추론

- 확률 추론에서 d-분리의 역할
  - 노드가 수십~수천 개인 베이지안 네트워크에서 확률 추론을 할 때, d-sep을 활용하면 계산량을 획기적으로 줄일 수 있음
    - 왜냐하면, $d_sep(a,c|\mathcal{W})$이면 $I(a,c|\mathcal{W})$이고, $d_sep(\mathcal{A},\mathcal{C}|\mathcal{W})$이면 $I(\mathcal{A},\mathcal{C}|\mathcal{W})$이기 때문
    - 예, $P(fatigue,positive|not\_lung\_cancer)$ 를 추정하는 경우, $d_sep(\text{피로},\text{엑스레이}|\text{폐암})$이므로 $I(\text{피로},\text{엑스레이}|\text{폐암})$임 → $P(fatigue,positive|not_lung_cancer)$를 $P(fatigue|not\_lung\_cancer)*P(positive|not\_lung\_cancer)$ 처럼 분해하여 계산 가능함

10.2.4 확률 추론

- 정확한 해 구하기
  - 값이 알려진 확률변수에서 출발하여 이웃 확률변수로 정보를 파급하는 메시지 전달 방식을 사용

예제 10-6 **메시지 전달을 통한 확률 추론**

선형 구조를 가진 [그림 10-10]의 베이지안 네트워크에서 확률 추론을 해 보자. 이 네트워크에는 확률변수 4개가 있는데 각각 두 가지 값을 가지는 이진변수라고 가정하자. 예를 들어, 확률변수 $x \in \{\mathscr{x}1, \mathscr{x}2\}$이다. 이때 확률변수의 값은 특별한 글씨체 $\mathscr{x}$를 사용하여 확률변수 자체와 구분한다. 확률변수 x가 $\mathscr{x}1$ 값을 가진 상황을 가정하고, 순방향의 확률 추론으로서 $P(\mathscr{w}1|\mathscr{x}1)$을 계산한다고 하자. 먼저 x의 정보가 y로 전달된다. 이때는 베이지안 네트워크가 가진 조건부 확률을 사용한다.

$$P(\mathscr{y}1|\mathscr{x}1) = 0.9$$

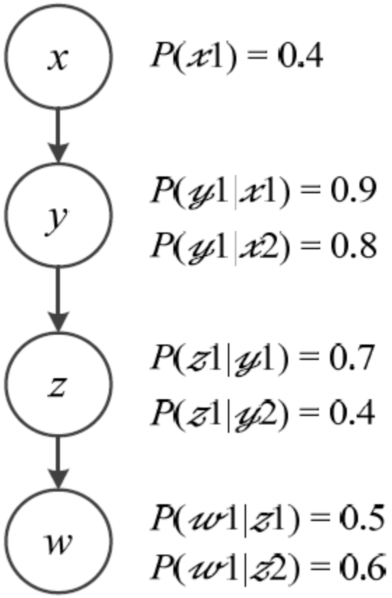


그림 10-10 확률 추론 예제(선형 구조)

10.2.4 확률 추론

이제 y의 정보가 z로 전달된다. $P(z1|x1)$은 다음과 같이 계산할 수 있다. 첫 번째 줄이 두 번째 줄로 바뀔 때는 마르코프 조건을 적용한다. 다음 단계의 전달을 위해 $P(z2|x1)$도 계산해야 하는데, $P(z1|x1)$처럼 수식을 전개하여 계산할 수도 있고, z가 두 가지 값만 가지므로 $P(z1|x1) + P(z2|x1) = 1$을 이용하여 0.330이라고 바로 결정할 수도 있다.

$$\begin{aligned}
P(z1|x1) &= P(z1|y1,x1)P(y1|x1) + P(z1|y2,x1)P(y2|x1) \\
&= P(z1|y1)P(y1|x1) + P(z1|y2)P(y2|x1) \\
&= 0.7 * 0.9 + 0.4 * 0.1 = 0.67
\end{aligned}$$

이제 다음 식을 적용하여 z의 정보를 w로 전달한다.

$$\begin{aligned}
P(w1|x1) &= P(w1|z1,x1)P(z1|x1) + P(w1|z2,x1)P(z2|x1) \\
&= P(w1|z1)P(z1|x1) + P(w1|z2)P(z2|x1) \\
&= 0.5 * 0.67 + 0.6 * 0.33 = 0.533
\end{aligned}$$

결국, 확률 추론을 통해 x 확률변수에서 $x1$을 관찰했을 때 w 확률변수가 $w1$을 가질 확률은 0.533이라는 사실을 알게 되었다.

10.2.4 확률 추론

이제 확률변수 w가 $w1$ 값을 가진 상황을 가정하고, 역방향의 확률 추론으로서 $P(x1|w1)$을 계산하자. 먼저 2장의 식 (2.26) 베이즈 정리를 적용하고 앞에서 계산한 $P(w1|x1) = 0.533$과 x의 사전확률 $P(x1) = 0.4$를 대입하면 다음과 같다.

$$P(x1|w1) = \frac{P(w1|x1)P(x1)}{P(w1)} = \frac{0.533 * 0.4}{P(w1)}$$

$P(w1)$을 계산하려면 메시지 전달을 적용해야 한다. 먼저 x의 사전확률로부터 y의 사전확률을 계산한다. $P(y1) + P(y2) = 1$이므로 $P(y2) = 0.16$이다.

$$P(y1) = P(y1|x1)P(x1) + P(y1|x2)P(x2) = 0.9 * 0.4 + 0.8 * 0.6 = 0.84$$

이어 y를 z로 전달하면 다음과 같다. $P(z1) + P(z2) = 1$이므로 $P(z2) = 0.348$이다.

$$P(z1) = P(z1|y1)P(y1) + P(z1|y2)P(y2) = 0.7 * 0.84 + 0.4 * 0.16 = 0.652$$

z를 w로 전달하면 다음과 같다.

$$P(w1) = P(w1|z1)P(z1) + P(w1|z2)P(z2) = 0.5 * 0.652 + 0.6 * 0.348 = 0.5348$$

$P(x1|w1)$ 식에 대입하여 $P(w1)$을 계산하면 다음과 같다. 결국 w 확률변수에서 $w1$을 관찰했을 때 x 확률변수가 $x1$을 가질 확률은 0.3987이라는 사실을 추론하였다.

$$P(x1|w1) = \frac{0.533 * 0.4}{P(w1)} = \frac{0.533 * 0.4}{0.5348} = 0.3987$$

10.2.4 확률 추론

근사 추론

- 정확한 해 알고리즘은 NP-hard → 대안은 정확성을 포기하고 근사해
- 근사 접근방법은 샘플링 사용

논리 샘플링 기법

- 예, [그림 10-11]

[그림 10−11]의 베이지안 네트워크에서 $P(y = y1)$을 구하는 코드:

h=0
for (k=1 to m)
    $P(x)$에 따라 x의 값 $\tilde{x}$을 생성한다.
    $P(y|\tilde{x})$에 따라 y의 값 $\tilde{y}$을 생성한다.
    if ($\tilde{y} = y1$) h++
$P(y = y1) = \dfrac{h}{m}$

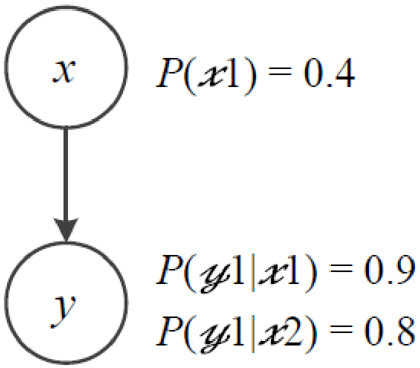


그림 10−11 확률 추론 예제

10.2.4 확률 추론

예제 10-7 논리 샘플링

$m=10$으로 설정하고 앞의 코드를 실행해 보자. 라인 3에서는 [0,1] 사이의 난수를 생성하여 0.4보다 작으면 $x1$, 그렇지 않으면 $x2$로 결정한다. 0.7이 나와 $\tilde{x}=x2$가 되었다고 가정하자. 라인 4에서는 난수를 생성하여 0.8보다 작으면 $y1$, 그렇지 않으면 $y2$로 결정한다. 0.2가 나와 $\tilde{y}=y1$이 되었다고 가정하자. 이렇게 하여 첫 번째 샘플 $(x2, y1)$을 얻었다. h를 증가시킨다.

두 번째 루프에서는 0.3이 나와 $\tilde{x}=x1$이 되었다고 하자. 라인 4에서는 0.9보다 작으면 $y1$, 그렇지 않으면 $y2$로 결정해야 하는데, 0.92가 나와 $\tilde{y}=y2$가 되었다고 가정하자. 이렇게 하면 두 번째 샘플 $(x1, y2)$를 얻는다. h를 증가시키지 않는다.

이렇게 10번 반복하여 다음과 같은 샘플을 얻었다고 하자. 샘플 10개 중 $y1$인 것이 8개이므로 $P(y=y1)=\frac{8}{10}=0.8$이 된다.

$$(x2, y1)\ (x1, y2)\ (x2, y1)\ (x2, y2)\ (x1, y1)\ (x2, y1)\ (x2, y1)\ (x1, y1)\ (x2, y1)\ (x1, y1)$$

[예제 10-6]에서 추정한 정확한 값 0.84와 논리 샘플링으로 추정한 근사해 0.8은 0.04만큼 오차가 있다.

10.2.4 확률 추론

- 예, [그림 10-11]

[그림 10-11]의 베이지안 네트워크에서 $P(x1|y2)$를 구하는 코드:

```
h=0
for (k=1 to m)
    repeat
        P(x)에 따라 x의 값 x̃을 생성한다.
        P(y|x̃)에 따라 y의 값 ỹ을 생성한다.
    until (ỹ = y2)        // y2인 샘플만 취함
    if (x̃ = x1) h++
```

$$P(x1|y2) = \frac{h}{m}$$

10.2.4 확률 추론

예제 10-8 **논리 샘플링**

$m = 10$으로 설정하고 앞의 코드를 실행해 보자. 라인 4에서 0.1이 나와 $\tilde{x} = x1$이 되었다. 라인 5에서 0.4가 나와, 이번 샘플을 버리고 repeat 루프를 다시 시작한다. 라인 4에서 0.3이 나와 $\tilde{x} = x1$이 되고 라인 5에서 0.95가 나와 $\tilde{y} = y2$가 되었다. 이렇게 첫 번째 샘플 $(x1, y2)$를 얻었다. h를 증가시킨다.

10번 반복하여 다음과 같은 샘플을 얻었다고 하자. 샘플 10개 중 $x1$인 것이 3개이므로 $P(x1|y2) = \frac{3}{10} = 0.3$이 된다.

$$(x1, y2)\ (x2, y2)\ (x2, y2)\ (x2, y2)\ (x1, y2)\ (x2, y2)\ (x2, y2)\ (x1, y2)\ (x2, y2)\ (x2, y2)$$

베이즈 정리를 적용하여 정확한 해를 구하면 0.25이다. 논리 샘플링으로 추정한 근사해 0.3은 0.05만큼 오차가 있다.

- [예제 10-8]에서는 샘플을 10개만 사용하여 오차가 큰데, 샘플 수를 늘리면 정확도 증가
- [예제 10-7]과 [예제 10-8]은 [그림 10-11]의 특정 베이지안 네트워크에 대한 알고리즘

10.2.4 확률 추론

- 일반적인 베이지안 네트워크에서 논리 샘플링 알고리즘

알고리즘 10-1 베이지안 네트워크의 확률 추론을 위한 확률 논리 샘플링

입력: 베이지안 네트워크 $G = (V, E)$, 값이 알려진 노드집합 $\mathcal{A}$

// $V = \{x_1, x_2, \cdots, x_n\}$은 자식이 부모보다 큰 번호를 가지도록 위상 정렬됨(그림 10-5)

출력: $V - \mathcal{A}$에 속한 노드 x_i 각각에 대한 조건부 확률 $P(x_i|\mathcal{A})$

for ($x_i \in V - \mathcal{A}$)
    x_i가 l개의 값을 가진다면 $h_{i1} \sim h_{il}$을 0으로 초기화한다.
for (k=1 to m)
    j=1 // 루트 노드에서 시작
    while (j<=n)
        $P(x_j|\wp(x_j))$에 따라 x_j의 값 $\tilde{x}_j$를 생성한다. // $\wp(x_j)$는 x_j의 부모의 값
        if($x_j \in \mathcal{A}$ and $\tilde{x}_j \neq$입력으로 주어진 x_j의 값) j=1 // 처음부터 다시 시작
        else j++ // 다음 변수로 넘어감
    for($x_i \in V - \mathcal{A}$)
        $\tilde{x}_i$에 해당하는 인덱스 q에 대해 h_{iq}++
for($x_i \in V - \mathcal{A}$)
    $P(x_i\text{가 }q\text{번 째 값}\,|\,\mathcal{A}) = \dfrac{h_{iq}}{m},\ \forall q$

10.2.4 확률 추론

- 지금은 마르코프 체인 몬테카를로MCMC(Markov chain Monte Carlo) 기법을 주로 사용
  - 논리 샘플링은 이전 샘플과 현재 샘플 사이에 아무런 연관이 없음
  - MCMC는 현재 샘플링이 직전에 생성된 샘플의 정보를 활용함
  - 연관성은 주로 Metropolis-Hastings 알고리즘이나 깁스 샘플링을 사용
- 근사해 알고리즘조차 작은 오류를 보장해야 하는 경우는 NP-hard
  - 중요한 연구 주제

10.3 마르코프 랜덤필드

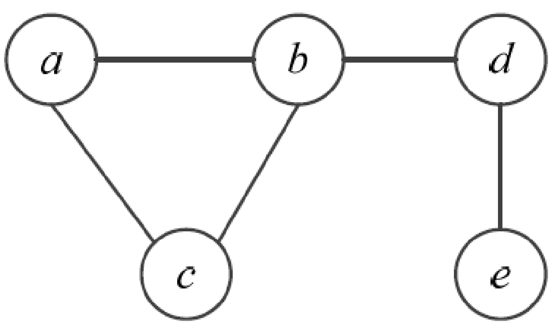
- 10.3.1 동작 원리
- 10.3.2 사례 연구: 영상 잡음 제거

- 마르코프 랜덤필드
  - 노드 사이에 인과관계가 없는 문제를 다루므로 무방향 그래프 사용
  - 이웃한 노드 사이에만 직접적인 상호작용 허용 → 마르코프
  - 확률변수가 동일한 자격으로 영향을 주고받으며 필드를 형성 → 랜덤필드

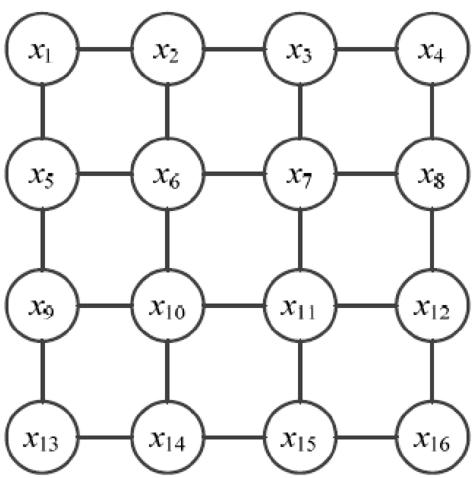
10.3.1 동작 원리

그래프 분해

- 클릭을 이용하여 분해(베이지안 네트워크는 부모-자식 사이에만 확률 부여하여 분해)
- 클릭은 완전 부분그래프, 극대 클릭은maximal clique 노드를 추가하면 더 이상 완전 그래프를 유지하지 못하는 클릭
  - 예, G_1은 $\{a,b,c\}$, $\{b,d\}$, $\{d,e\}$의 극대 클릭 3개를 가짐
  - 예, G_2는 $\{x_1,x_2\}$, $\{x_1,x_5\}$, $\{x_2,x_3\}$, $\{x_2,x_6\}$, …의 극대 클릭을 가짐



(a) G_1



(b) G_2 (4*4 영상)

그림 10-12 마르코프 랜덤필드 예제

10.3.1 동작 원리

- 깁스 확률분포Gibbs distribution
  - 클릭 q는 퍼텐셜 $\psi(q)$를 가짐
  - 식 (10.4)는 퍼텐셜로 정의되는 확률분포($\sum_{\mathbf{x}} \tilde{P}(\mathbf{x}) = 1$를 만족하지 못해 확률로서 결함)

$$\tilde{P}(\mathbf{x}) = \prod_{q \in \mathrm{G}} \psi(q) \qquad (10.4)$$

  - 분할함수 Z로 나누어 정규화하면,

$$P(\mathbf{x}) = \frac{1}{Z} \tilde{P}(\mathbf{x}) \qquad (10.5)$$

$$Z = \sum_{\mathbf{x}} \tilde{P}(\mathbf{x}) \qquad (10.6)$$

  - 퍼텐셜함수는 에너지함수로 정의됨
    - 에너지함수 $energy(q)$는 응용과 목적에 맞게 정의해야 함

$$\psi(q) = \exp\big(-energy(q)\big) \qquad (10.7)$$

10.3.1 동작 원리

예제 10-9 **마르코프 랜덤필드의 확률분포**

[그림 10-12(a)]의 마르코프 랜덤필드의 에너지함수를 다음과 같이 정의하자. 확률변수 5개는 모두 0 또는 1을 가지는 이진변수라 가정한다.

$$energy(q) = \begin{cases} 0, & q\text{의 모든 노드가 같은 값을 가짐} \\ 1, & \text{그렇지 않음} \end{cases}$$

[그림 10-13]이 보여 주는 상태의 에너지와 퍼텐셜을 계산하자. 클릭 $\{a,b,c\},\{b,d\},\{d,e\}$의 $energy$와 ψ는 다음과 같다.

$$energy(\{a,b,c\}) = 0,\quad energy(\{b,d\}) = 1,\quad energy(\{d,e\}) = 0$$
$$\psi(\{a,b,c\}) = 1,\quad \psi(\{b,d\}) = 0.3679,\quad \psi(\{d,e\}) = 1$$

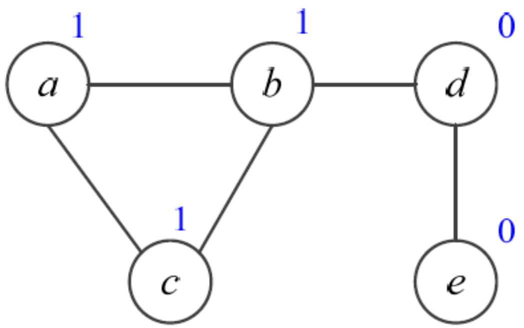


그림 10-13 이진 확률변수의 예제 상태(총 2^5=32개의 상태가 있음)

이 퍼텐셜값을 식 (10.4)에 대입하면 다음 결과를 얻는다.

$$\tilde{P}(\mathbf{x} = (1,1,1,0,0)^{\mathrm{T}}) = \psi(\{a,b,c\})\psi(\{b,d\})\psi(\{d,e\}) = 0.3679$$

지금까지 (1,1,1,0,0)이라는 한 상태에 대해 계산하였는데 [표 10-1]은 32개의 상태를 모두 나열한다.

10.3.1 동작 원리

표 10-1 [그림 10-13]의 마르코프 랜덤필드의 확률분포를 계산하는 과정

상태	$\psi(\{a,b,c\})$	$\psi(\{b,d\})$	$\psi(\{d,e\})$	$\tilde{P}$	상태	$\psi(\{a,b,c\})$	$\psi(\{b,d\})$	$\psi(\{d,e\})$	$\tilde{P}$
(0,0,0,0,0)	1	1	1	1.000	(1,0,0,0,0)	0.3679	1	1	0.3679
(0,0,0,0,1)	1	1	0.3679	0.3679	(1,0,0,0,1)	0.3679	1	0.3679	0.1353
(0,0,0,1,0)	1	0.3679	0.3679	0.1353	(1,0,0,1,0)	0.3679	0.3679	0.3679	0.0498
(0,0,0,1,1)	1	0.3679	1	0.3679	(1,0,0,1,1)	0.3679	0.3679	1	0.1353
(0,0,1,0,0)	0.3679	0.3679	1	0.1353	(1,0,1,0,0)	0.3679	0.3679	1	0.1353
(0,0,1,0,1)	0.3679	0.3679	0.3679	0.0498	(1,0,1,0,1)	0.3679	0.3679	0.3679	0.0498
(0,0,1,1,0)	0.3679	1	0.3679	0.1353	(1,0,1,1,0)	0.3679	1	0.3679	0.1353
(0,0,1,1,1)	0.3679	1	1	0.3679	(1,0,1,1,1)	0.3679	1	1	0.3679
(0,1,0,0,0)	0.3679	1	1	0.3679	(1,1,0,0,0)	0.3679	1	1	0.3679
(0,1,0,0,1)	0.3679	1	0.3679	0.1353	(1,1,0,0,1)	0.3679	1	0.3679	0.1353
(0,1,0,1,0)	0.3679	0.3679	0.3679	0.0498	(1,1,0,1,0)	0.3679	0.3679	0.3679	0.0498
(0,1,0,1,1)	0.3679	0.3679	1	0.1353	(1,1,0,1,1)	0.3679	0.3679	1	0.1353
(0,1,1,0,0)	0.3679	0.3679	1	0.1353	(1,1,1,0,0)	1	0.3679	1	0.3679
(0,1,1,0,1)	0.3679	0.3679	0.3679	0.0498	(1,1,1,0,1)	1	0.3679	0.3679	0.1353
(0,1,1,1,0)	0.3679	1	0.3679	0.1353	(1,1,1,1,0)	1	1	0.3679	0.3679
(0,1,1,1,1)	0.3679	1	1	0.3679	(1,1,1,1,1)	1	1	1	1.000

10.3.1 동작 원리

식 (10.6)의 분할함수 Z는 32개의 값을 모두 더하여 $Z = 7.872$가 된다. 이제 식 (10.5)를 적용하여 확률분포를 구하자. [표 10-1]에 있는 $\tilde{P}$ 을 Z 로 나누면 된다. 예를 들어, $P(\mathbf{x} = (1,1,1,0,0)^{\mathrm{T}}) = \frac{0.3679}{7.872} = 0.0467$, $P(\mathbf{x} = (0,0,1,0,1)^{\mathrm{T}}) = \frac{0.0498}{7.872} = 0.0063$, $P(\mathbf{x} = (0,0,0,0,0)^{\mathrm{T}}) = \frac{1.0}{7.872} = 0.127$이다.

10.3.1 동작 원리

- 마르코프 랜덤필드의 특성
  - 식 (10.4)의 $\tilde{P}(\mathbf{x})$ 계산은 빠름
  - 하지만 식 (10.6)의 분할함수 계산은 차원의 저주
    - n개의 확률변수(노드)가 있고 변수마다 k개의 값을 가진다면 k^n번의 계산 필요
    - 따라서 추정치를 사용

- 마르코프 랜덤필드의 학습
  - 확률을 최대로 하는 상태를 찾는 과정

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} P(\mathbf{x}) \tag{10.8}$$

10.3.2 사례 연구: 영상 잡음 제거

컴퓨터비전에 응용

- 잡음 제거, 영상 복원, 에지 검출, 텍스쳐 분석, 스테레오 비전, 영상 분할 등
- 여기서는 잡음 제거 응용을 살펴봄([그림 10-14]에서 오른쪽 영상이 주어지면 왼쪽 영상을 찾아내는 일)

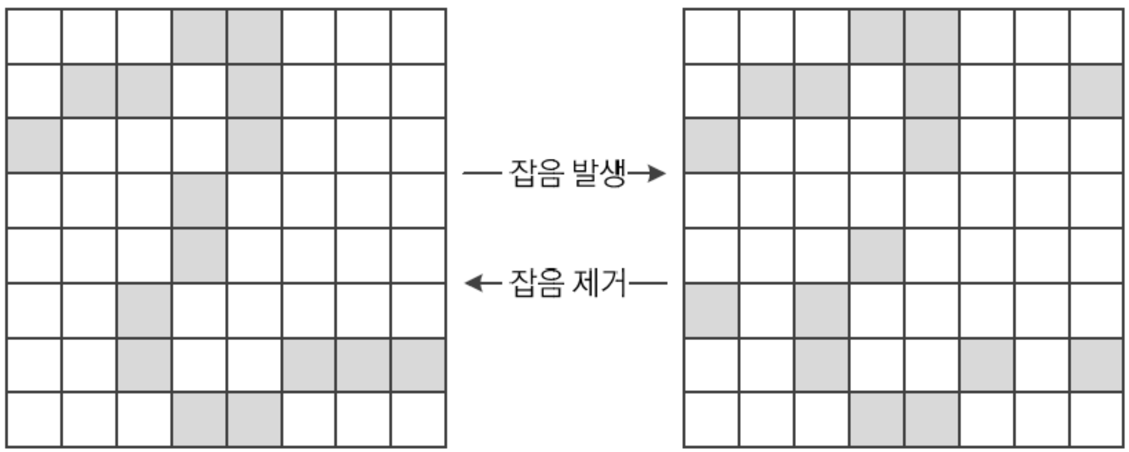


그림 10-14 잡음 제거를 통한 원본 복원

- 오른쪽 영상만 가지고 왼쪽에 가까운 영상을 어떻게 찾나? → 에너지 함수에 매끄러움 성질을 반영하면 됨

10.3.2 사례 연구: 영상 잡음 제거

잡음 제거 문제의 공식화

- 컴퓨터비전은 [그림 10-15]의 틀을 사용
  - $\mathbf{y}$는 입력 영상이고 $\mathbf{x}$는 출력 영상

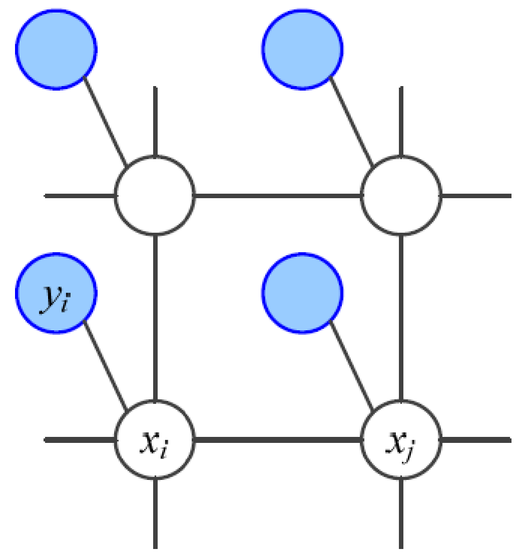


그림 10-15 마르코프 랜덤필드를 이용한 영상 잡음 제거

- 잡음 제거는 식 (10.9)로 공식화

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} P(\mathbf{x}|\mathbf{y}) \quad (10.9)$$

10.3.2 사례 연구: 영상 잡음 제거

에너지함수 공식화

- 두 종류의 클릭에 대한 에너지함수

$$energy(\{x_i, x_j\}) = -\alpha x_i x_j \qquad (10.10)$$

$$energy(\{x_i, y_i\}) = -\beta x_i y_i \qquad (10.11)$$

- 타당성
  - x_i와 x_j가 같으면 $\psi(\{x_i, x_j\}) = \exp(\alpha) = 2.718$이고, 다르면 $\psi(\{x_i, x_j\}) = \exp(-\alpha) = 0.3679$ ($\alpha = 1$일 때)
  - 즉 같은 상태를 선호함 → 영상을 매끄럽게 유지하려는 힘
- 식 (10.11)도 비슷하게 작동함(즉 영상 $\mathbf{x}$와 $\mathbf{y}$를 같게 만들려는 힘)
- 식 (10.10)의 힘과 식 (10.11)의 힘이 균형을 이루어 잡음이 제거된 최적 영상을 찾아냄

10.3.2 사례 연구: 영상 잡음 제거

- 영상 전체를 위한 식의 유도

$$
\begin{aligned}
P(\mathbf{x}|\mathbf{y}) &= \frac{1}{Z}\prod_{q\in G}\psi(q) \\
&= \frac{1}{Z}\exp(-energy(\{x_1,x_2\}))\exp(-energy(\{x_1,x_3\}))\cdots \\
&\quad \exp(-energy(\{x_1,y_1\}))\exp(-energy(\{x_2,y_2\}))\cdots \\
&= \frac{1}{Z}\exp(-energy(\{x_1,x_2\}) - energy(\{x_1,x_3\})\cdots \\
&\quad -energy(\{x_1,y_1\}) - energy(\{x_2,y_2\})\cdots) \\
&= \frac{1}{Z}\exp\big(-energy(\mathbf{x},\mathbf{y})\big)
\end{aligned}
\qquad (10.12)
$$

10.3.2 사례 연구: 영상 잡음 제거

최적화 알고리즘

- 가장 간단한 ICM(iterated conditional modes) 알고리즘

알고리즘 10-2 영상 잡음 제거를 위한 ICM

입력: 잡음이 섞인 영상 $\mathbf{y}$

출력: 잡음이 줄어든 영상 $\mathbf{x}$

1 $\mathbf{x} = \mathbf{y}$
2 $energy(\mathbf{x}, \mathbf{y})$를 계산하여 $e_{current}$라 한다.
3 while (true)
4   for ($\mathbf{x}$의 화소 x_i 각각에 대해)
5   x_i를 반전시킨 영상을 $\mathbf{x}'$라 하고 $energy(\mathbf{x}', \mathbf{y})$를 계산한 결과를 e_{new}라 한다.
6   if ($e_{new} < e_{current}$) $\mathbf{x} = \mathbf{x}', e_{current} = e_{new}$
7   if (라인 4~6)에서 부호가 바뀐 화소가 없으면 break

ICM 알고리즘은 한 화소씩 처리하므로 탐욕 알고리즘greedy algorithm이다. 따라서 전역 최적해가 아닌 지역 최적해를 찾는다. 전역 최적해에 더 근접한 해를 찾아 주는 개선된 알고리즘은 여러 가지 있으며, 그래프 절단 알고리즘이 대표적이다[Kolmogorov2004].

10.4 RBM과 DBN

- 10.4.1 RBM의 구조와 원리
- 10.4.2 RBM 학습
- 10.4.3 DBN

- 통계 역학
  - 볼츠만과 깁스의 연구
  - 홉필드는 통계 역학 이론을 신경망에 도입

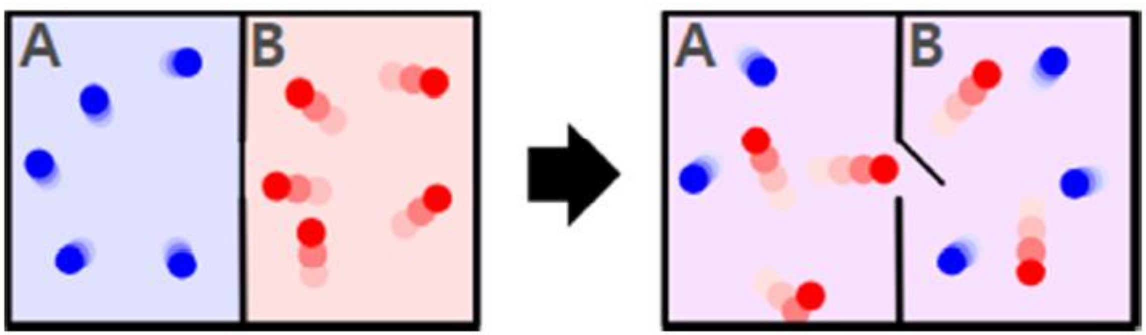


그림 10-16 통계 역학

10.4 RBM과 DBN

- 제한 볼츠만 기계RBM(restricted Boltzmann machine)([그림 10-17(c)])
  - 가시 노드는 관찰된 특징을 입력, 은닉 노드는 중간 계산 결과 저장
  - 같은 종류의 노드 사이에는 에지가 없는 볼츠만 기계

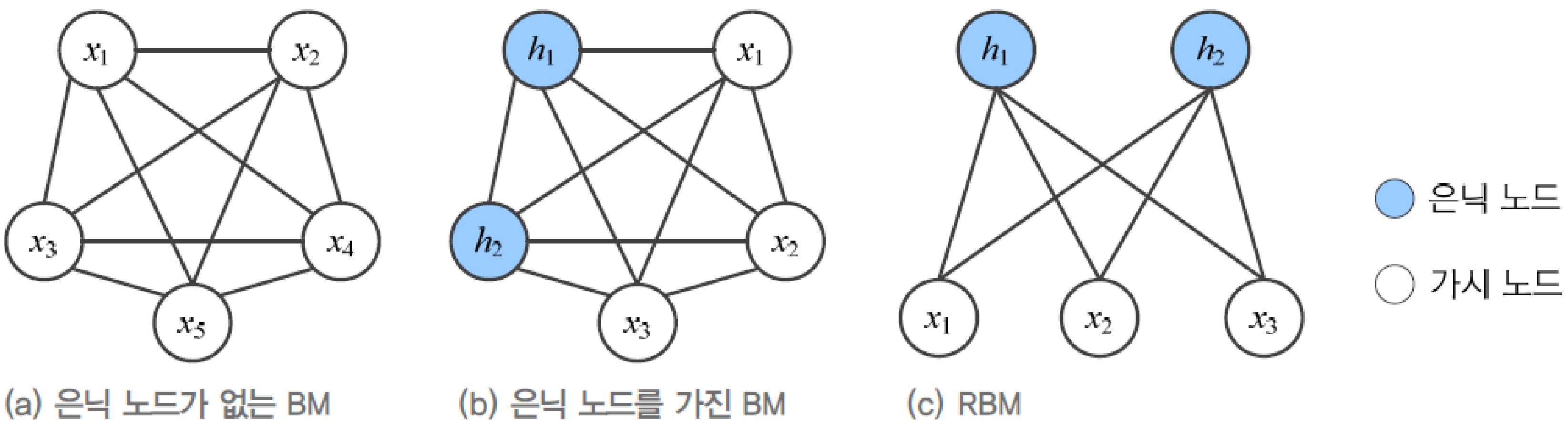


그림 10-17 볼츠만 기계(BM)와 제한 볼츠만 기계(RBM)

  - RBM조차 마땅한 학습 알고리즘이 없었는데, 2002년 대조 발산 알고리즘 탄생
  - 2006년에 층별 사전 학습 알고리즘으로 RBM을 여러 층 쌓아 만든 DBN 탄생 ← 딥러닝을 널리 확산하는 기폭제

10.4.1 RBM의 구조와 원리

RBM은 에너지 모델

- 노드 값에 따라 에너지가 정해지는데, 에너지가 낮을수록 발생 확률이 높음
- 특정 패턴을 높은 확률로 발생시킬 수 있는 능력 → 생성 모델과 분별 모델로 쓸 수 있음

RBM 구조

- 바이파타이트 그래프
  - x_i로 표기되는 가시 노드와 h_j로 표기되는 은닉 노드
  - 벡터로 표기하면, $\mathbf{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}}$, $\mathbf{h} = (h_1, h_2, \cdots, h_m)^{\mathrm{T}}$
- 학습이 알아내야 하는 매개변수는 $\Theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$
- 모든 노드는 이진 값을 가진다고 가정

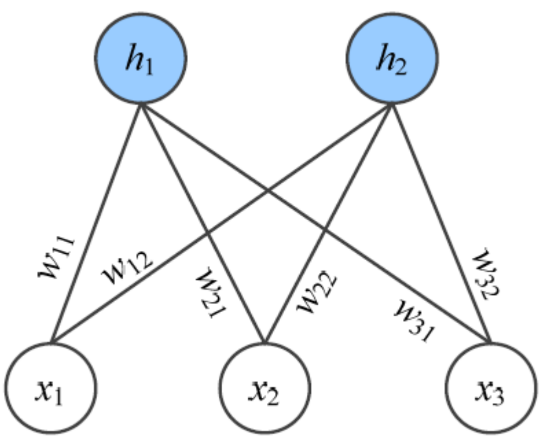


그림 10-18 RBM의 구조와 매개변수

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}$$

10.4.1 RBM의 구조와 원리

에너지와 확률분포

- RBM의 $\mathbf{x}$와 $\mathbf{h}$ 값이 지정되면,

$$\left.\begin{aligned} energy(\mathbf{x},\mathbf{h}) &= -\sum_{i=1}^{d} a_i x_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{d}\sum_{j=1}^{m} x_i w_{ij} h_j \\ &= -\mathbf{a}^{\mathrm{T}}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{h} - \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h} \end{aligned}\right\} \qquad (10.14)$$

- 발생 확률은,

$$P(\mathbf{x},\mathbf{h}) = \frac{1}{Z}\exp\big(-energy(\mathbf{x},\mathbf{h})\big) \qquad (10.15)$$

$$Z = \sum_{\mathbf{x}}\sum_{\mathbf{h}} \exp\big(-energy(\mathbf{x},\mathbf{h})\big) \qquad (10.16)$$

- $\mathbf{x}$의 발생 확률은,

$$P(\mathbf{x}) = \sum_{\mathbf{h}} P(\mathbf{x},\mathbf{h}) = \frac{1}{Z}\sum_{\mathbf{h}} \exp\big(-energy(\mathbf{x},\mathbf{h})\big) \qquad (10.17)$$

10.4.1 RBM의 구조와 원리

예제 10-10 **x와 h의 발생 확률**

[그림 10-19]의 예제 RBM을 가지고 수식 (10.14)~(10.17)을 명확히 이해하자. RBM의 매개변수는 다음과 같다. 그림에서 수평선은 바이어스에 해당한다.

$$\mathbf{W} = \begin{pmatrix} 0.1 & -0.2 \\ 0.0 & -0.1 \\ 0.2 & 0.1 \end{pmatrix},\ \mathbf{a} = \begin{pmatrix} 0.2 \\ -0.1 \\ 0.0 \end{pmatrix},\ \mathbf{b} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}$$

먼저 $\mathbf{x} = (1,0,1)^{\mathrm{T}}$이고 $\mathbf{h} = (1,1)^{\mathrm{T}}$일 때 에너지를 계산하면 다음과 같이 -0.7이다.

$$\begin{aligned} energy(\mathbf{x} = (1,0,1)^{\mathrm{T}}, \mathbf{h} = (1,1)^{\mathrm{T}}) \\ &= -(1 \times 0.2 + 0 \times (-0.1) + 1 \times 0) - (1 \times 0.1 + 1 \times 0.2) - (1 \times 1 \times 0.1 + 1 \times 1 \\ &\quad \times (-0.2) + 0 \times 1 \times 0.0 + 0 \times 1 \times (-0.1) + 1 \times 1 \times 0.2 + 1 \times 1 \times 0.1) = -0.7 \end{aligned}$$

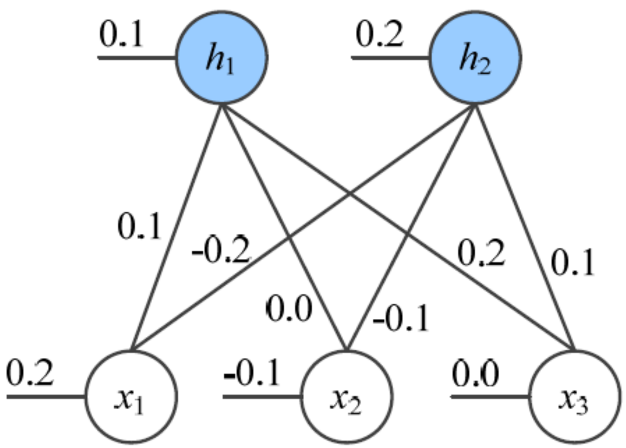


그림 10-19 RBM의 예제

10.4.1 RBM의 구조와 원리

확률을 계산하려면 먼저 Z값을 구해야만 한다. Z는 $\mathbf{x}$와 $\mathbf{h}$가 가질 수 있는 모든 상태의 값을 더한 것이므로, 총 2^5=32가지의 에너지를 계산해야 한다. $\mathbf{x} = (1,0,1)^{\mathrm{T}}$, $\mathbf{h} = (1,1)^{\mathrm{T}}$ 를 계산한 것처럼, 모든 경우를 계산하여 식 (10.16)에 대입하면 $Z = 40.9493$이 된다. $\mathbf{x} = (1,0,1)^{\mathrm{T}}$, $\mathbf{h} = (1,1)^{\mathrm{T}}$가 발생할 확률은 다음과 같이 계산한다.

$$P(\mathbf{x} = (1,0,1)^{\mathrm{T}},\ \mathbf{h} = (1,1)^{\mathrm{T}}) = \frac{1}{40.9493}\exp(0.7) = 0.04918$$

식 (10.17)을 이용하여 $\mathbf{x} = (1,0,1)^{\mathrm{T}}$가 발생할 확률을 계산하면, 0.156465이다.

$$\begin{aligned} P(\mathbf{x} = (1,0,1)^{\mathrm{T}}) &= P((1,0,1)^{\mathrm{T}},(0,0)^{\mathrm{T}}) + P((1,0,1)^{\mathrm{T}},(0,1)^{\mathrm{T}}) + P((1,0,1)^{\mathrm{T}},(1,0)^{\mathrm{T}}) + P((1,0,1)^{\mathrm{T}},(1,1)^{\mathrm{T}}) \\ &= 0.0298272 + 0.0329641 + 0.0444969 + 0.0491767 = 0.156465 \end{aligned}$$

$\mathbf{x}$가 가질 수 있는 모든 경우의 발생 확률을 계산하면 다음과 같다.

$$\begin{aligned} &P((0,0,0)^{\mathrm{T}}) = 0.114200, \quad P((1,0,0)^{\mathrm{T}}) = 0.132516 \\ &P((0,0,1)^{\mathrm{T}}) = 0.134846, \quad P((1,0,1)^{\mathrm{T}}) = 0.156465 \\ &P((0,1,0)^{\mathrm{T}}) = 0.097926, \quad P((1,1,0)^{\mathrm{T}}) = 0.114200 \\ &P((0,1,1)^{\mathrm{T}}) = 0.115343, \quad P((1,1,1)^{\mathrm{T}}) = 0.134502 \end{aligned}$$

10.4.2 RBM 학습

- 학습의 목적
  - 훈련집합 $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$에 속한 샘플은 높은 확률로 발생시키고, 속하지 않은 샘플은 낮은 확률로 발생시키는 것
  - 예, [그림 10-19]의 RBM은 모든 샘플이 비슷한 확률임 → 쓸모가 없음
  - 예, [그림 10-20]의 RBM은 $\mathbb{X} = \{(1,0,0)^\mathrm{T}, (1,0,1)^\mathrm{T}, (1,1,0)^\mathrm{T}, (1,1,1)^\mathrm{T}\}$이라면, $\mathbb{X}$를 높은 확률로 발생시킴 → 유용한 RBM

- RBM이 사용하는 목적함수

$$J(\Theta) = \prod_{\mathbf{x} \in \mathbb{X}} P(\mathbf{x}) \quad \text{또는} \quad J(\Theta) = \sum_{\mathbf{x} \in \mathbb{X}} \log P(\mathbf{x}) \qquad (10.18)$$

- RBM 학습 알고리즘이 할 일

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \prod_{\mathbf{x} \in \mathbb{X}} P(\mathbf{x}) \quad \text{또는} \quad \widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{\mathbf{x} \in \mathbb{X}} \log P(\mathbf{x}) \qquad (10.19)$$