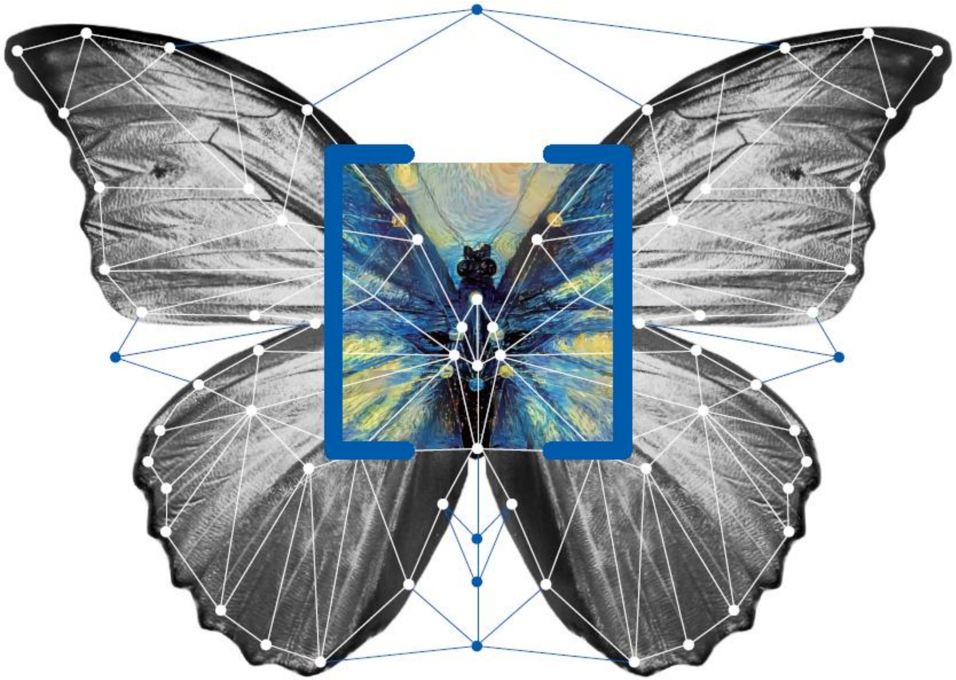
한빛아카데미
Hanbit Academy, Inc.



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

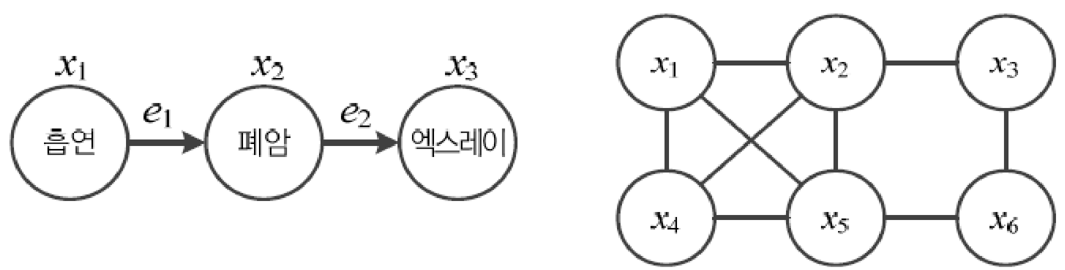
확률 그래피컬 모델probabilistic graphical model

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

10.1.1 그래프 표현

그래프 표현

- 예, 흡연 환자의 엑스레이 진단에서 양성이 나타났을 때 폐암일 확률은?
  - 주요 요인 엑스레이, 흡연, 폐암을 확률변수로 뽑아 노드로 취함 (x1,x2,x3)
  - 인과관계는 에지로 표현 (e1,e2)
- 그래프는 확률변수의 상호작용을 표현하는 뼈대
- 뼈대에 확률을 부여하면 확률 그래피컬 모델이 됨



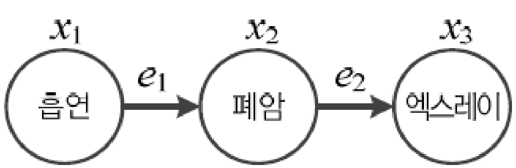
(a) 베이지안 네트워크(방향 그래프) (b) 마르코프 랜덤필드(무방향 그래프)

그림 10-1 확률 그래피컬 모델

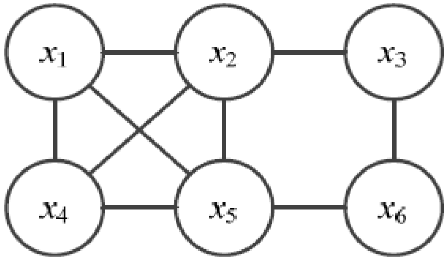
10.1.1 그래프 표현

대표적인 모델

- 베이지안 네트워크Bayesian network
  - **인과관계**를 나타내기 위해 방향 에지를 사용
- 마르코프 랜덤필드
  - 인과관계가 없는 문제를 다루므로 **무방향** 에지를 사용



(a) 베이지안 네트워크(방향 그래프)



(b) 마르코프 랜덤필드(무방향 그래프)

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($2^3$) 상태에 대해 확률값을 지정해야 함

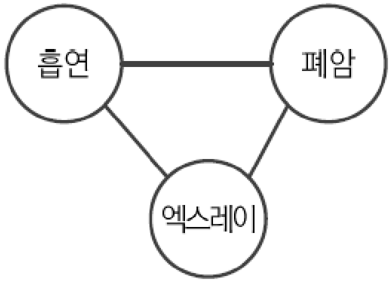


그림 10-2 세 확률변수의 완전 그래프(결합 확률 필요)

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

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

마지막 한 개는 확률 총합인 1에서 빼면 알 수 있기 때문

8개의 확률변수에 각각 옵션이 4씩이면 $4^8 - 1$개를 알아야하는데 65535개를 어느 세월에 !!

그래프 분해

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

10.1.2 그래프 분해와 확률 표현

– 결합 확률
  - A와 B 사건이 동시에 발생하는 확률, 조건부 확률의 수식으로 유도할 수 있음
  - 이를 '곱셈 법칙'이라고 한다.
  - 만약, 각 사건 A와 B가 독립이라면, $P[A|B]=P(A)$이 되므로 위의 곱셈 법칙에 대입하면, $P[B]\times P[A]=P(A\cap B)$가 성립한다.

$$P[A\,|\,B]=\frac{P[A\cap B]}{P[B]}$$

– 체인 규칙(chain rule)
  - 각 사건 이 일어날 확률은 조건부 확률과 결합 확률을 이용하여 연쇄적인(chain) 조건부 확률의 곱의 표현식으로 표현할 수 있다.
  - 이를 '체인 규칙'이라고 한다.

$$\begin{aligned}&P(A_1, A_2, A_3, A_4, \cdots, A_n)\\&=P(A_1\,|\,A_2, A_3, A_4, \cdots, A_n)\times P(A_2\,|\,A_3, A_4, \cdots, A_n)\times\cdots\times P(A_{n-1}\,|\,A_n)\times P(A_n)\end{aligned}$$

$$P(A_2|A_1)=\frac{P(A_1\cap A_2)}{P(A_1)} \Rightarrow P(A_1\cap A_2)=P(A_1)\cdot P(A_2|A_1)$$

$$\begin{aligned}&P(A_3|A_1\cap A_2)=\frac{P(A_1\cap A_2\cap A_3)}{P(A_1\cap A_2)}\\&\Rightarrow P(A_1\cap A_2\cap A_3)=P(A_1\cap A_2)\cdot P(A_3|A_1\cap A_2)\\&\Leftrightarrow P(A_1\cap A_2\cap A_3)=P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1\cap A_2)\end{aligned}$$

10.2 베이지안 네트워크

베이지안 네트워크의 장점

- 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)

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.

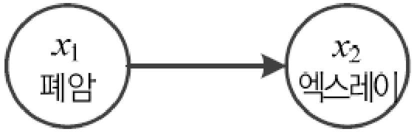
10.2.1 간단한 예제

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 긍정률true positive rate이 0.6이고 참 부정률false negative rate이 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

- [예제 10-1]에서 세 가지 주요 문제를 풀어본 셈
  - [그림 10-3]의 그래프 구조를 만드는 일은 구조 학습
  - 통계청과 병원에서 확률을 수집한 일은 확률 학습
  - 엑스레이에서 양성이 나타났을 때 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%

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

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.

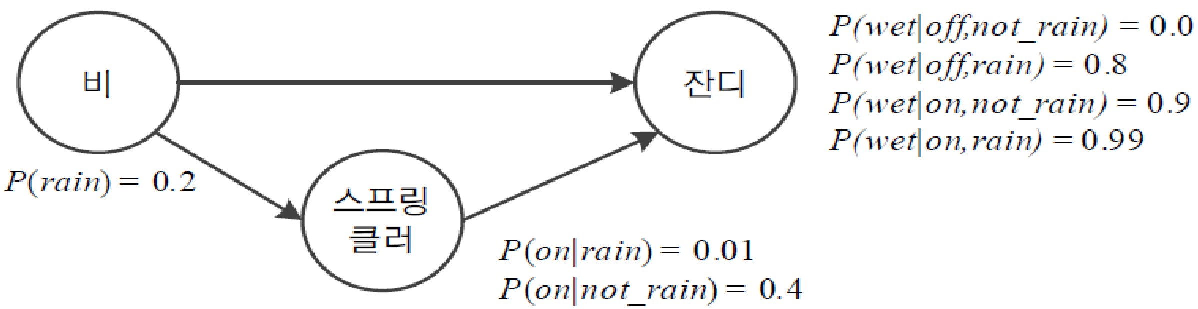


그림 10-4 노드가 3개인 베이지안 네트워크

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

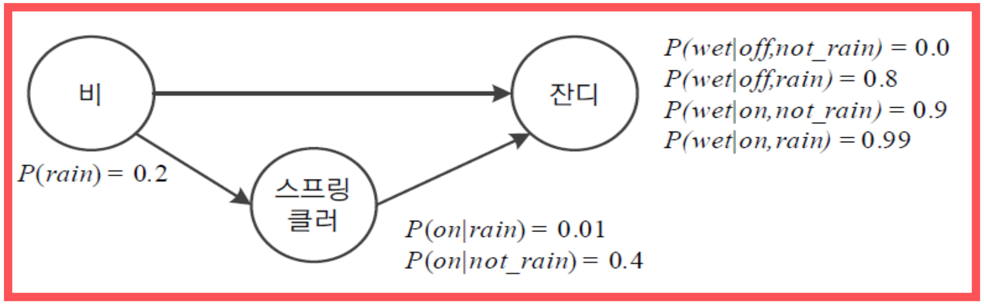
이제 완성된 베이지안 네트워크로부터 다양한 확률을 추론할 수 있다. 예를 들어, 비가 오지 않았는데 스프링클러가 켜져 있을 확률을 구하고자 하면 $P(\text{스프링클러} = on|\text{비} = non_rain)$을 구하는 셈이므로 두 번째 표를 참조하여 40%라고 답하면 된다.

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.



$$P(rain|wet) = \frac{P(rain, wet)}{P(wet)}$$
$$= \frac{\sum_{S\in\{on,off\}} P(wet, S, rain)}{\sum_{S\in\{on,off\}} \sum_{R\in\{rain,not\_rain\}} P(wet, S, R)}$$
$$= \frac{P(wet, on, rain) + P(wet, off, rain)}{P(wet, on, rain) + P(wet, on, not\_rain) + P(wet, off, rain) + P(wet, off, not\_rain)}$$
$$= \frac{0.00198 + 0.1584}{0.00198 + 0.288 + 0.1584 + 0.0} = 0.3577$$

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

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

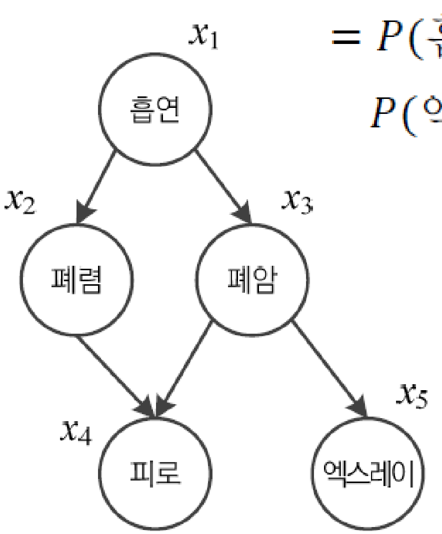


그림 10-5 노드가 5개인 인과관계 그래프

- 식 (10.1)과 식 (10.2)는 보편 수식 (그래프 분해를 적용하기 이전)

10.2.2 그래프 분해

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

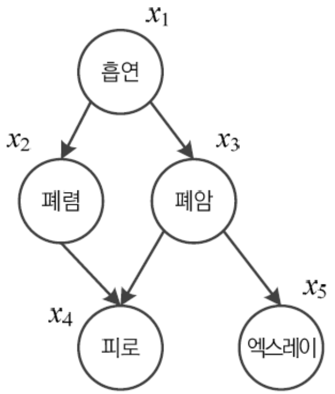


그림 10-5 노드가 5개인 인과관계 그래프

- 마르코프 조건을 식 (10.2)에 적용하면,

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

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

10.2.2 그래프 분해

- 이렇게 유도된 식에 따라 가상의 확률을 부여해 보면,
  - [그림 10-6]의 베이지안 네트워크가 됨

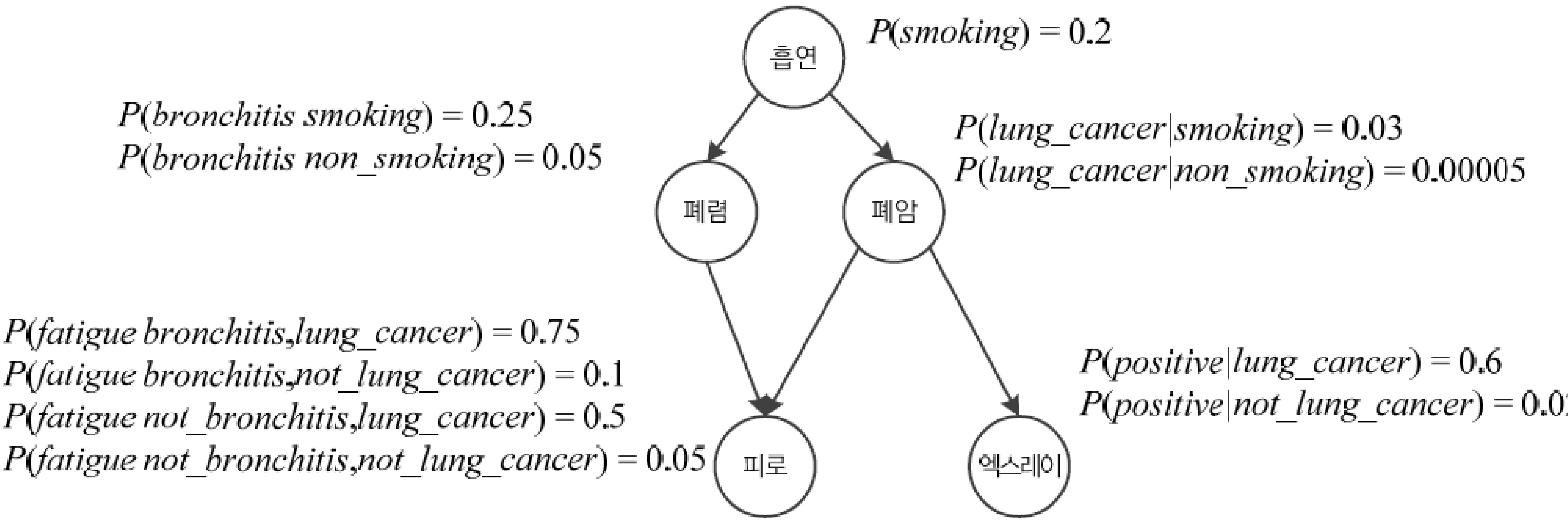


그림 10-6 베이지안 네트워크

10.2.2 그래프 분해

- 그래프 분해로 확률 학습이 쉬워짐
  - 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
  - 또는 훈련집합을 가지고 자동 학습

- 구조 학습과 확률 학습을 통해 베이지안 네트워크를 완성한 다음, 확률 추론을 수행

pgmpy

- pgmpy is a library for probabilistic graphical models in Python.
- Bayesian Network

(https://pgmpy.org/models/bayesiannetwork.html)