$$EI_{y^*}(x) = \int_R \max(y^* - y, 0) p_m(y|x) dy. \tag{22}$$

y^*: 기존 탐색값 중 임의의 임계값(보통 하위 15%)
y: 하이퍼파라미터 x에 대한 확률분포 p_m상의 성능측정값
x: 하이퍼파라미터 탐색값
p_m: 하이퍼파라미터 x에 대한 성능측정값 y의 확률분포(surrogate model)

- 즉, 성능 측정값의 기댓값이 크거나(p_m) 하이퍼파라미터가 x일 때 성능측정값의 불확실성이 큰 경우 EI 가 큰 값을 갖게 되므로 베이지안 최적화는 순차적 탐색과정에서 성능측정값이 컸던 지점 또는 확률분포상 불확실정이 큰 지점에 큰 비중을 두고 순차적 탐색과정을 진행한다는 것을 의미한다.
- 다수의 하이퍼파라미터를 지닌 딥러닝 모델의 특성상 하이퍼파라미터 값에 따라 도출되는 성능측정값의 함수 $f : X \rightarrow Y$는 구조를 알기 어렵고 알아내려는 과정자체가 비현실적이기 때문에 대안으로 확률모델인 Surrogate model(p_m)을 활용한다.
- 본 연구에서는 surrogate model로 TPE(Tree Parzen Estimator)를 사용하며, TPE에서는 $p_m(x|y)$을 다음과 같이 표현할 수 있고, 이때 y^*는 이전 탐색결과 중 임의의 값(주로 하위 15%)을 의미한다.
  - 대표적인 surrogate 모델로 gaussian process(GP)와 TPE가 있으며, 다양한 선행 연구 사례에서 TPE가 더 우수한 성능을 나타낸 바 있어 TPE를 사용하고자 한다.

$$p_m(x|y) = \begin{cases} l(x) & if \ \ y < y^* \\ g(x) & if \ \ y \geq y^* \end{cases} \tag{23}$$

$l(x)$: y^*이하일 때 TPE의 surrogate model
$g(x)$: y^*이상일 때 TPE의 surrogate model

- 따라서, 식(19)에 따라 식(18)은 아래와 같이 표현할 수 있다.

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(x|y) \frac{p(y)}{p(x)} dy \tag{24}$$

- 그리고, r과 $p(x)$는 아래와 같이 정의할 수 있다.

$$r = p(y < y^*) \tag{25}$$

$$p(x) = \int_{-\infty}^{+\infty} p(x|y)p(y)dy = rl(x) + (1-r)g(x) \tag{26}$$

- 정리하면, 아래와 같이 표현할 수 있다.

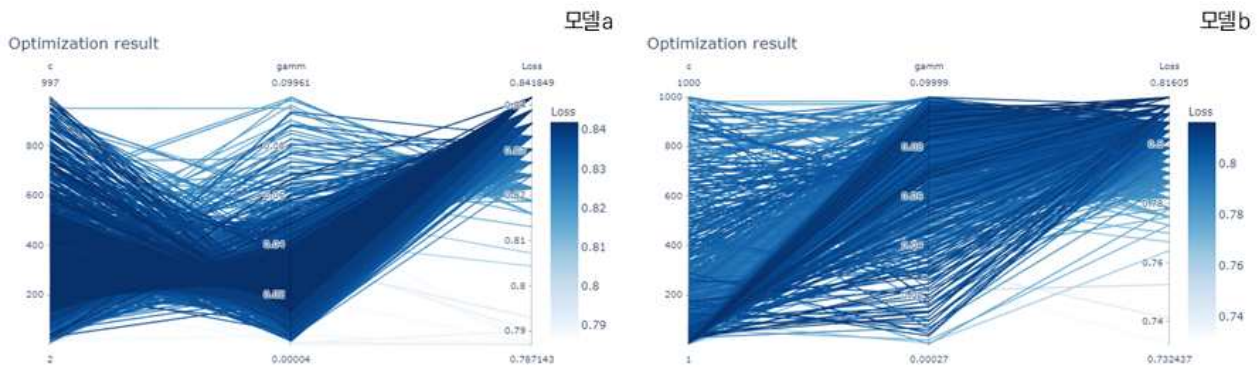
$$\int_{-\infty}^{y^*} (y^* - y)p(x|y)p(y)dy = l(x)\int_{-\infty}^{y^*} (y^* - y)p(y)dy \tag{27}$$

$$l(x)\int_{-\infty}^{y^*} (y^* - y)p(y)dy = ry^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy$$

- 최종적으로 식(23)은 아래와 같이 정리될 수 있다.

$$EI_{y^*}(x) = \frac{ry^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{rl(x) + (1-r)g(x)} \propto (r + \frac{g(x)}{l(x)}(1-r))^{-1} \tag{28}$$

- 즉, EI는 $l(x)$에 비례하고 $g(x)$에 반비례하며, 이는 TPE가 최적화 문제에서 특정 임계값 r보다 더 좋은 성능을 보인 탐색결과에 더 가중치를 두고 최적화를 진행하는 것을 의미한다 (Bergstra et al., 2011).

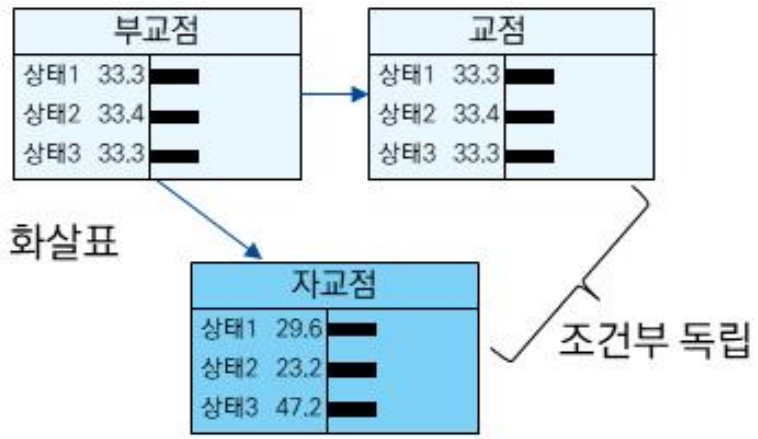


<그림 2-30> TPE를 활용한 하이퍼파라미터 최적화 예시

2). 조건부확률 모델 구축

✓ 베이지안 네트워크(Bayesian networks)

- 베이지안 네트워크는 해석적인 측면에서 우수할 뿐만 아니라, 변수 간의 상호 관계 혹은 예측에 대한 불확실성에 대해 조건적인 의존 관계를 확률적으로 표현하므로 환경 정책 등 의사결정을 지원에 있어 매우 적절한 방법이다.
- 베이지안 네트워크는 도식적 모델로 변수 사이의 상호 인과관계를 기반으로 해석 가능한 모델 구조를 제공한다.
  - 베이지안 네트워크는 변수의 상태를 나타내는 교점과 변수들 간 연결을 나타내는 화살표로 구성되는 도식적 모델이다.
  - 두 교점이 화살표로 이어져 있다면, 부교점과 자교점 간의 확률적 관계가 조건부 확률표(Conditional Probability Table, CPT)로 표현되는 의존성이 있음을 나타낸다.
  - 반면, 두 교점이 화살표로 이어져 있지 않다면, 두 변수는 조건부 독립성을 나타낸다.



<그림 2-31> 베이지안 네트워크의 구조

- 따라서, 특정 자교점 X_i가 부교점의 집합 $pa(X_i)$에 의해 확률적으로 표현된다고 할 때, 자교점의 조건부 확률 분포는 $P(X_i|pa(X_i))$로 표현이 된다.
- 이때, 전체 랜덤 변수에 대한 확률 분포는 결합 확률 분포(Joint probability distribution)라 하며, 베이지안 네트워크상 조건부 확률 분포를 고려한 결합 확률 분포 $P_X(X)$는 아래와 같이 표현할 수 있다.

$$P_X(X) = \prod_{i=1}^{p} P_{X_i}(X_i|pa(X_i)) \tag{29}$$

- 베이지안 네트워크의 구축 과정은 크게 데이터 전처리와 학습으로 구분할 수 있으며, 전처리 과정은 변수 이산화(discretization), 변수 선택(variable selection)을 포함하고, 학습은 구조 학습(structure learning) 및 파라미터 학습(parameter learning)을 포함한다.

✓ **자료 전처리**

- 변수 이산화과정은 연속적인 변수를 N개의 이산화된 범주로 나누는 것을 의미한다.
  - 이산화를 통해 생성된 범주의 개수가 너무 적거나 많은 경우, 분포의 의미를 손실하거나 모델의 복잡성이 급수 적으로 증가해 모델의 성능을 저하 시킬 수 있으으로 현실적인 범위가 필요하다.
  - 변수 이산화를 위한 방법 대표적인 방식은 범위법, 빈도법, 상호 정보법의 크게 세 가지 방식으로 구분되며, 전문가의 의견 수렴을 통한 방식도 사용된다.
  - 범위법(Interval-based method)은 연속변수를 동일한 범위로 이산화하는 방법으로 적용이 간단하나, 연속변수 분포의 왜도가 클 경우, 효과적이지 못하며, 관련된 변수 사이의 관계성이 고려되지 않는 한계가 있다.
  - 빈도법(Quantile-based method)은 연속변수를 빈도를 기반으로 이산화하는 방법으로 범위법 보다 변수의 분포 모양을 고려한 방법이나, 관련된 변수 사이의 관계성은 고려하지 못하는 한계가 있다.
  - 상호 정보법(Mutual information-based method)은 연속 변수 사이의 상호 정보를 기반으로 이산화하는 방법으로 Hartmink 방법이라고도 부르며(Hartmink, 2001), 범위법이나 빈도법보다 변수의 분포를 잘 표현하는 방법이지만, 복잡한 관계를 가지고 있는 다수의 변수에 대해 적용하기에 한계가 있다.
- 변수 선택 과정은 다양한 변수 중 모델의 종속변수와 상당한 관련이 있는 것으로 판단되는 주요변수들을 선택하는 과정이다.
  - 베이지안 네트워크에서 너무 많은 입력변수는 모델의 복잡도를 증가시키며, 과적합과 성능의 감소로 이어질 수 있으며, 너무 적은 독립변수는 입력변수를 설명하기에 충분한 정보가 되지 않을 수 있다.
  - 변수 선택 과정의 대표적인 방법은 필터 방법, 래퍼 방법, 임베디드 방법의 크게 세 가지로 구분할 수 있다.

- 필터 방법(Filter method)은 종속변수와 입력변수 간의 상관계수, 상호정보 등과 같은 단순 통계량을 바탕으로 변수를 선택하는 방법으로 단순하고 간단히 구현할 수 있는 방법이나, 비슷한 관계성을 나타내는 변수들을 구분하지 못해 다중공선성을 해결하지 못한다.
- 래퍼 방법(Wrapper method)은 방법으로 다양한 입력변수의 조합을 통해 모델의 성능을 평가한 뒤, 가장 좋은 모델 성능을 보이는 대입 변수의 조합을 선택한다.
- 모든 대입 변수 경우의 수 조합에 대해 수행하거나 전진선택법(Forward selection), 후진소거법(Backward elimination), 유전알고리즘(Genetic algorithm) 등의 방법이 사용되며, 다중공선성에 대한 문제는 해결할 수 있으나, 시간과 비용적인 측면에서 구현하기 상대적으로 어렵다.
- 임베디드 방법(Embedded method)은 다른 모델을 사용하여 특정 목적함수를 최적화하거나 모델 퍼포먼스의 증가를 통해 중요 변수를 추출하는 방식으로 주로 L1 회귀분석(Lasso regression)이나 의사결정 나무를 활용한다.
- 다중공선성과 시간 및 비용적인 측면을 동시에 해결하여 탁월한 방법이나, 실제 분석에 사용할 모델 방법과는 다른 모델을 통해 중요한 변수를 선택하기 때문에 해석적인 측면에서 주의가 필요한 방법이다.

✓ **베이지안 네트워크 학습**

- 베이지안 네트워크의 학습은 구조 학습과 파라미터 학습으로 이루어지며, 구조 학습은 입력변수 간 인과관계 및 그 방향성을 설정하는 과정이며, 파라미터 학습은 구조 학습을 통해 결정된 조건부 확률표를 최적화하는 과정이다.
- 즉, 베이지안 네트워크의 학습은 다음과 같은 구조로 표현할 수 있다.

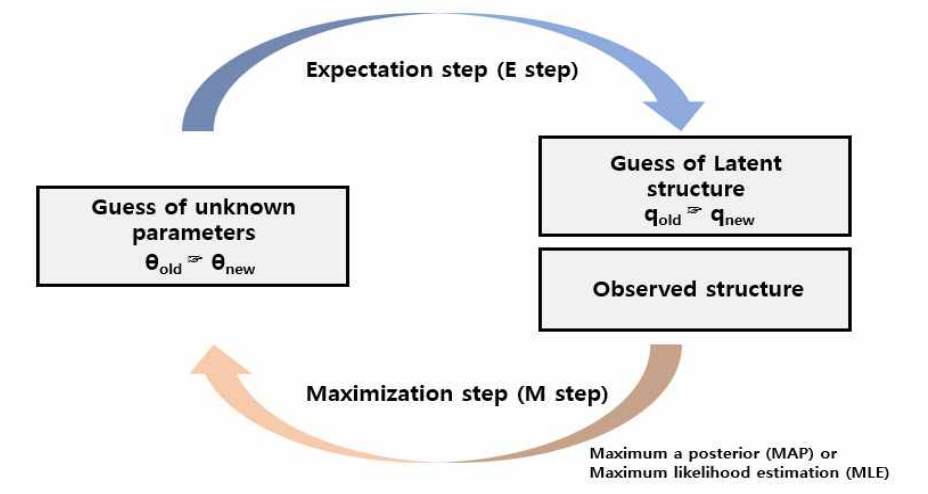
$$\underbrace{P(B|D) = P(G,\Theta|D)}_{learning\ \ BNs} = \underbrace{P(G|D)}_{Structure\,learning} \bullet \underbrace{P(\Theta|G,D)}_{Parameter\,learning} \tag{30}$$

D : 데이터

$B = (G, X)$: 베이지안 네트워크

Θ: 입력변수의 집합 X의 분포를 설명하는 모수의 집합

- 구조 학습은 선행 연구사례 및 전문가 자문 등을 활용하여 구조를 구성하는 지식 기반 구조 학습과 데이터에 대해 알고리즘을 적용하여 구조를 학습하는 데이터 기반 구조 학습 방식으로 구분할 수 있다.
  - 지식 기반 구조 학습은 지속적인 전문가들의 자문을 바탕으로 입력변수 사이의 인과관계를 결정하는 방법으로, 현재까지 인정된 지식 체계를 반영할 수 있으며, 데이터가 부족한 상황이거나 여러 오차에 의한 불확실성이 커 데이터 상 노이즈가 심할 경우 효과적으로 활용할 수 있다.
  - 데이터 기반 구조 학습은 다양한 알고리즘을 통해 자동적으로 랜덤 변수 사이의 인과 관계를 결정하는 방법이며, 새로운 인과 관계의 가능성에 대해 발견할 수 있어 확장적인 장점이 있다.
  - 그러나, 사용되는 알고리즘마다 결과로 나타나는 구조에 많은 차이가 있고, 입력변수의 분포에 대해 통계적 유의성을 판단 및 검증하여야 올바른 인과 구조의 생성이 가능하다.
  - 구조 학습 알고리즘은 크게 인과 관계의 통계적 가설 검정을 통해 구조를 형성하는 Constraint-based 알고리즘(PC, Grow-Shrink(GS), Incremental Association(IAMB) 등), 휴리스틱 한 방식으로 전체 모델 구조의 목적함수 수치를 최적화하는 Score-based 알고리즘(Greedy search, Genetic algorithm, Simulated annealing 등), 두 가지 방식을 동시에 사용하는 Hybrid 알고리즘(Sparse candidate algorithm(SC), Max-Min Hill-Climbing algorithm(MMHC)) 세 가지로 분류할 수 있다.
- 파라미터 학습은 데이터를 기반으로 랜덤 변수의 조건부 확률표를 최적화하는 과정으로 주로 Expectation-Maximization 알고리즘(EM 알고리즘)이나 경사 하강법(Gradient decent)으로 수행된다. 일반적으로 EM 알고리즘이 더 모델의 견고함이(robustness)이 높다고 알려져 있다.
  - EM 알고리즘은 관측되지 않은 잠재 변수에 의존하는 확률 모델에서 관측 변수의 분포에 대해 최대 가능도(Maximum likelihood)나 최대 사후 확률(Maximum a posteriori)을 갖는 매개변수를 반복적으로 찾는 알고리즘이다.

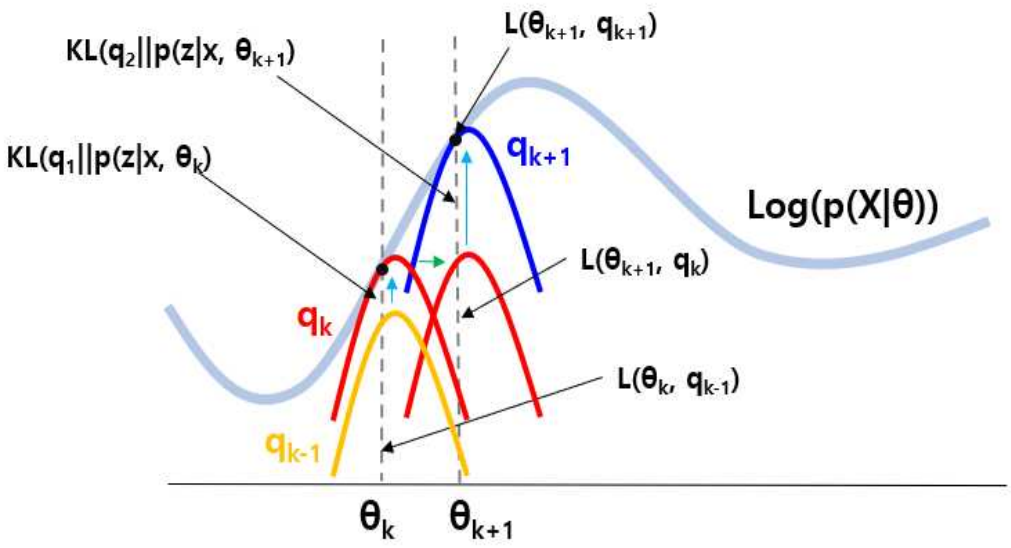


<그림 2-32> 간략화한 EM 알고리즘의 적용 과정

- EM 알고리즘은 파라미터로부터 잠재적 구조를 추론하는 Expectation step(E-step)과 관측된 구조를 통해 파라미터의 가능도나 사후확률을 최대화하여 갱신하는 Maximization step(M-step) 구조로 구성된다. EM 알고리즘의 최종 목적은 가능도를 최대화하는 잠재 변수에 대한 확률 분포와 모수를 구하는 것이 목표이다.

$$\max_{\theta} \log(p(X|\theta)) = \sum_{i=1}^{N} \log(p(x_i|\theta)) \tag{31}$$

θ : 모수, X : 무작위 변수

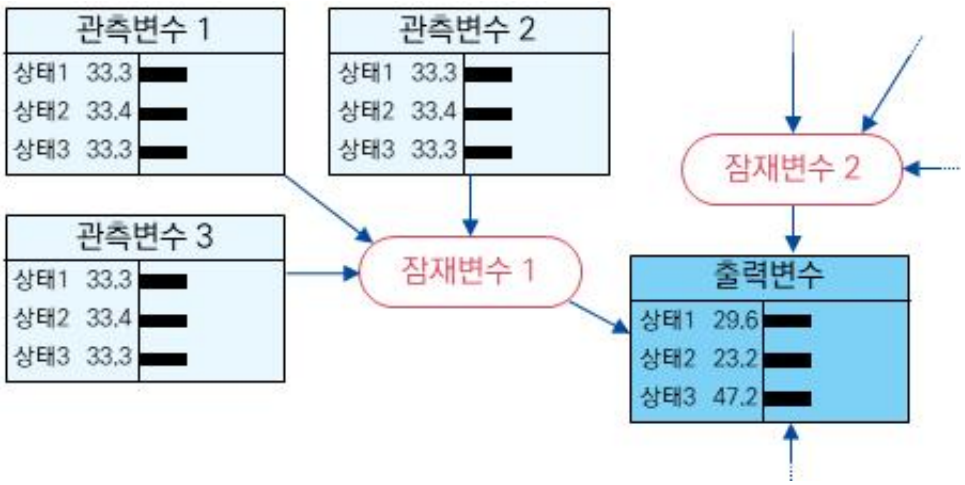


<그림 2-33> EM 알고리즘의 개념도

- EM 알고리즘의 수렴할 때까지 E와 M 단계를 반복적으로 수행되며 이러한 반복은 새로운 모델의 변수가 기존 모델보다 부적합할 때 멈추게 된다.
- 이때 두 가지 멈춤 조건(stopping criterion)을 사용할 수 있으며, 1) 가능도 함수의 차가 수렴하여 거의 차이가 없거나 2) 모수의 차이가 유의하지 않을 경우 EM 알고리즘은 종료된다.

✓ **계층적 베이지안 네트워크(Hierarchical Bayesian network)**

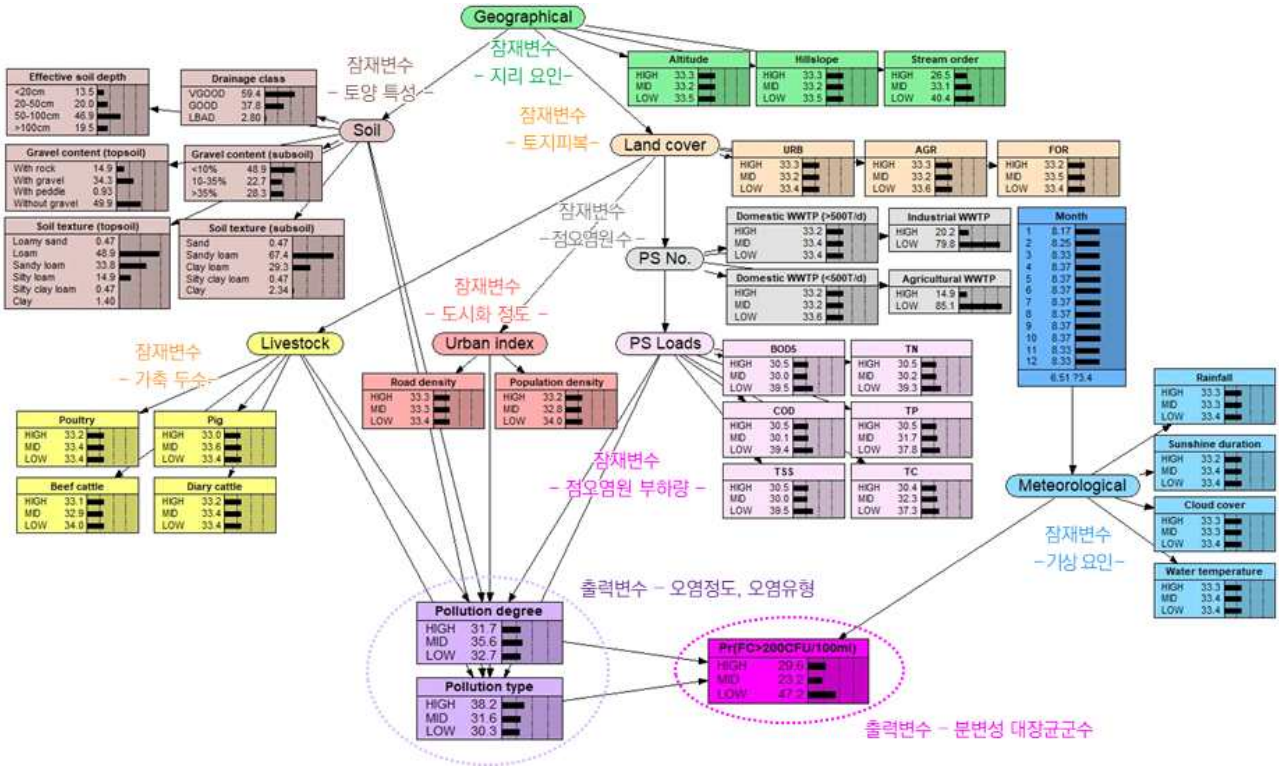
- 일반적인 베이지안 네트워크에서 입력변수 수의 증가는 중요 입력변수들의 변화에 대한 종속변수의 민감도를 감소시키며, 결과적으로 과적합(Overfitting)을 일으키고 모델 성능을 감소시킨다.
- 입력변수의 증가에 따른 문제는 변수 선택 등을 해결할 수 있으나, 이는 부득이하게 정보의 손실을 야기한다.
- 변수 선택에 대안적인 방법으로 베이지안 네트워크 내부에서 잠재적인 변수(Latent variable)를 활용하여 모델을 구조화하는 방법이 있으며, 이와 같이 잠재 변수를 활용해 구조화한 베이지안 네트워크를 계층적 베이지안 네트워크라 한다.



<그림 2-34> 계층적 베이지안 네트워크의 구조

- 계층적 베이지안 네트워크의 잠재변수는 유사한 관측변수(observed variable) 간의 정보를 압축, 공통적이며 대표적인 정보를 추출함으로써 특정 변수 그룹의 일반화된 정보를 제공한다.

- 따라서, 계층적 베이지안 네트워크는 수 많은 입력변수를 필요로 하는 복잡한 환경 현상을 표현함에 있어 변수 선택 및 제거 등으로 인한 정보손실을 방지하면서도 복잡하고 다중적인 계층 구조에 대해 효과적으로 표현할 수 있으며, 관측변수 및 잠재변수 간의 상호작용 효과 등을 예측 가능하고 과적합이나 교란 관계(spurious relationship)에 따른 잠재적인 함정을 회피할 수 있다.



<그림 2-35> 분변성 오염에 대한 계층적 베이지안 네트워크 구축 사례

2-2-2-4. 구축된 평가모델에 대한 예측성능 평가

- 최적화 과정 이후 구축이 완료된 평가모델들은 분류모델에 대해 다양한 평가지표를 활용하여 예측성능에 대한 검증을 수행한다.
- 분류모델의 예측결과는 주로 혼동행렬(confusion matrix)으로 표현되며, 예측결과에 대한 성능지표는 정확도(accuracy), 정밀도(precision) 및 재현율(recall), F1 score, AUC(Area under reciever operating characteristic curve) 등이 있다 <표 2-8>.

<표 2-8> 혼동행렬의 형태

Confusion matrix		Observed class	
		True	False
Predicted class	Positive	True positive (TP)	False positive (FP, 일종 오류)
	Negative	False negatives (FN, 이종 오류)	True negatives (TN)

- 정확도는 전체 샘플 수에 대한 예측을 통해 얻어진 분류 범주가 실제의 분류 범주와 같은 샘플 수의 비율을 의미하는 가장 직관적이고 간편한 성능 평가방법이다.
- 다만, 분류 문제에서 불균형 데이터(Imbalanced data)의 경우에 희소한 경우에 대한 모델 성능을 쉽게 평가하지 못한다.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{32}$$

- 정밀도는 모델이 positive라고 분류한 것 중에서 실제 positive인 샘플 수의 비율을 의미하며 재현율은 전체 positive 샘플 수에 대한 모델에서 positive이라고 예측한 샘플 수의 비율을 의미한다.
- 정밀도와 재현율은 공통적으로 TP(true positive)가 분자이며, 분모에서 TP를 공통으로 활용함에 더해 제 1종 오류와 제 2종 오류에 해당하는 FN(false negative)과 FP(false positive)를 가지고 있으므로 정밀도와 재현율은 서로 trade-off 관계에 있는 것을 확인할 수 있다.