기계학습

7강 SVM(2), 그래프모델(1)

장필훈 교수

학습목차

1. SVM(2)
2. 그래프모델(1)

01

SVM(2)

1-1 클래스 분포간 중첩(cont.)

- softmargin 제약조건
  - hardmargin의 경우와 마찬가지로,

    $a_n(t_n y(\mathbf{x}_n) - 1 + \xi_n) = 0$ 이고,

    따라서,

    $t_n y(\mathbf{x}_n) = 1 - \xi_n$

1-1 클래스 분포간 중첩(cont.)

- hardmargin의 경우와 마찬가지로 b를 구할 수 있다.
- hardmargin, softmargin모두

  $\mathbf{a}$를 찾는 것은 2차 계획법을 푸는 것.

  →계산복잡도가 높다. 메모리도 많이 소요.

  여러방법이 연구됨.

1-2 로지스틱 회귀와 관계

- soft margin의 목표는 다음을 최소화 하는것.

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2$$

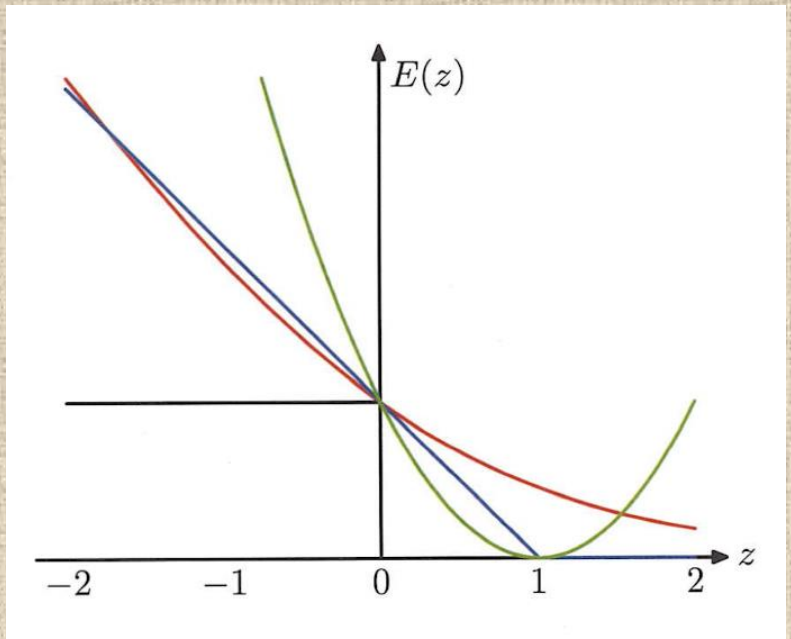
- $y_n t_n \geq 1$인 마진경계에서 올바른 쪽은 $\xi = 0$, 나머지는 $\xi_n = 1 - y_n t_n$이므로, 위 식은, (상수배차이무시)

$$\sum_{n=1}^{N}[1 - y_n t_n]_+ + \lambda\|\mathbf{w}\|^2$$

1-2 로지스틱 회귀와 관계

- 힌지오류함수

$$[1 - y_n t_n]_+ = \max(1 - y_n t_n, 0)$$



제곱오류는 결정경계로부터

멀리 떨어진 포인트를 더 강조

: 오분류율을 낮추는것이 목표라면

부적절

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig7.5

1-3 multiclass SVM

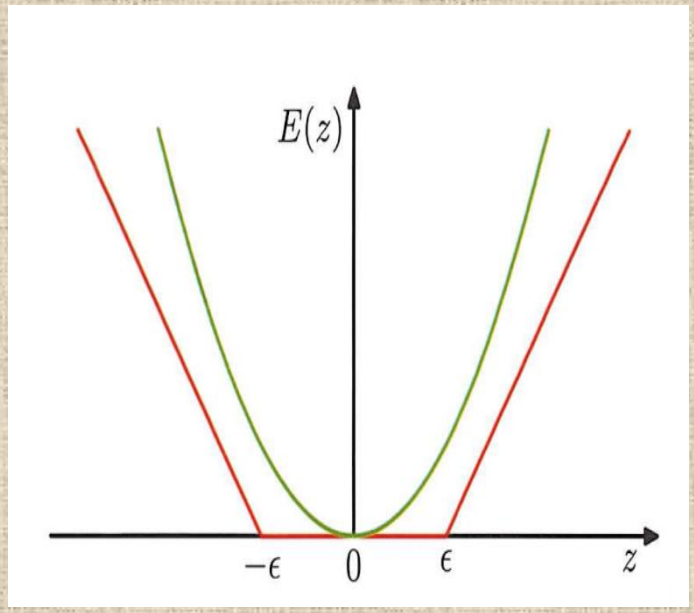
- $y(\mathbf{x}) = \max_k y_k(\mathbf{x})$: 앞서 나왔던 것.
  - 한계점: 분류기들이 서로 다르게 훈련되었기 때문에, $y_k(\mathbf{x})$들이 비교 가능하다는 보장이 없다. (절대치 차이가 클 수 있다)
  - 훈련집합들의 n수가 불균형하면 더 큰 문제가 됨
  - 여러 종류가 있음.

1-4 SVM을 이용한 회귀

- 선형회귀의 정규화된 오류함수

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 + \frac{1}{2}\|\mathbf{w}\|^2$$

- 오류함수를 교체해보자



$$E_\epsilon(y(\mathbf{x}) - t) = \begin{cases} 0, & \text{if } |y(\mathbf{x}) - t| < \epsilon; \\ |y(\mathbf{x}) - t| - \epsilon, & \text{나머지 경우} \end{cases}$$

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig7.6

1-4 SVM을 이용한 회귀

- slack variable($\xi_n \geq 0, \hat{\xi}_n \geq 0$) 도입해서

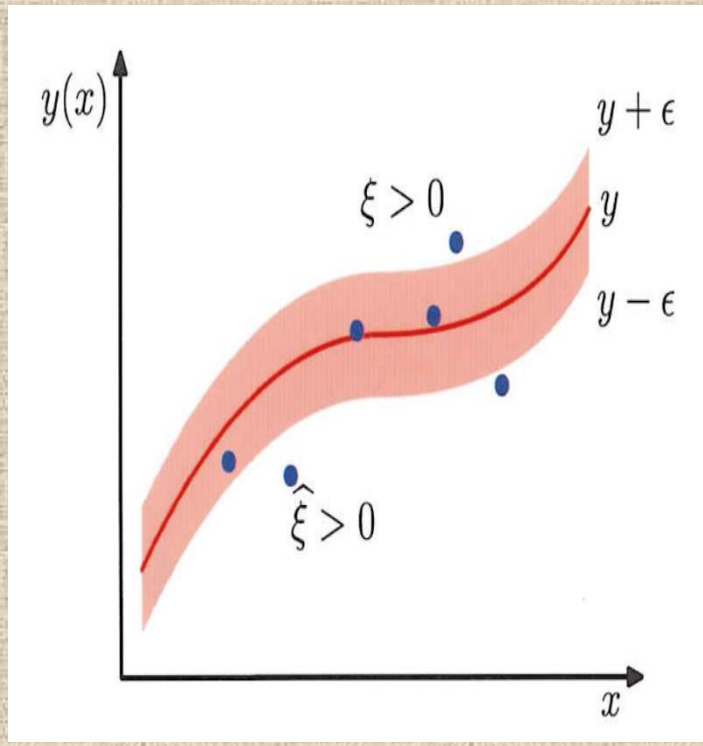
  표적포인트 범위를 적으면 다음과 같고,

$$t_n \leqslant y(\mathbf{x}_n) + \epsilon + \xi_n$$

$$t_n \geqslant y(\mathbf{x}_n) - \epsilon - \hat{\xi}_n$$

  오류함수를 다음과 같이 적을 수 있다.

$$C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig7.7

1-4 SVM을 이용한 회귀

- 오류함수의 최소화가 목표이므로,

  라그랑주 함수를 만들고 최적화한다.

$$L = C\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n)$$
$$-\sum_{n=1}^{N}a_n(\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N}\widehat{a}_n(\epsilon + \widehat{\xi}_n - y_n + t_n)$$

  $\mathbf{w}, b, \xi_n, \hat{\xi}_n$에 대한 라그랑주 함수 미분값을 0으로 설정.

  다음을 얻음. 예측에 사용. $y(\mathbf{x}) = \sum_{n=1}^{N}(a_n - \widehat{a}_n)k(\mathbf{x}, \mathbf{x}_n) + b$

1-5 상관벡터머신

- SVM의 한계점
  - 출력값이 사후확률이 아니라 결정값
  - K>2클래스에 관해 확장이 어려움
  - 커널의 조건에 제약이 있음
- 대안 : 상관벡터머신 relevance vector machine
  - 많은 성질 공유. 주로 더욱 희박한 모델을 결과로 줌.

1-5 상관벡터머신

- RVM을 이용한 회귀
  - 입력벡터 $\mathbf{x}$일 때 타겟변수 $\mathbf{t}$에 대한 조건부 분포가정

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}), \beta^{-1})$$

  - 일반식이 다음과 같은 형태가 됨(SVM과 동일)

$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n k(\mathbf{x}, \mathbf{x}_n) + b$$

1-5 상관벡터머신

- 커널이 양의 정부호 형태를 가져야 한다는 제약 없음
- 기저함수가 훈련집합의 위치나 숫자에 묶이는 제약 없음
- RVM과정 예시:

  입력 $\mathbf{x}$를 N개 입력받았고, 이때 타겟을 $\mathbf{t}$로 두면,

  가능도함수는

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \beta)$$

  ($\mathbf{X}$는 $\mathbf{x}$들의 행렬)

1-5 상관벡터머신

- RVM과정 예시(cont.)

  매개변수 벡터 $\mathbf{w}$에 대한 사전분포로 $\mu = 0$인 가우시언 가정.

  가중 매개변수 w_i각각에 대해 hyperparameter α_i를 각각 따로 쓴다. 따라서 가중치 사전분포는,

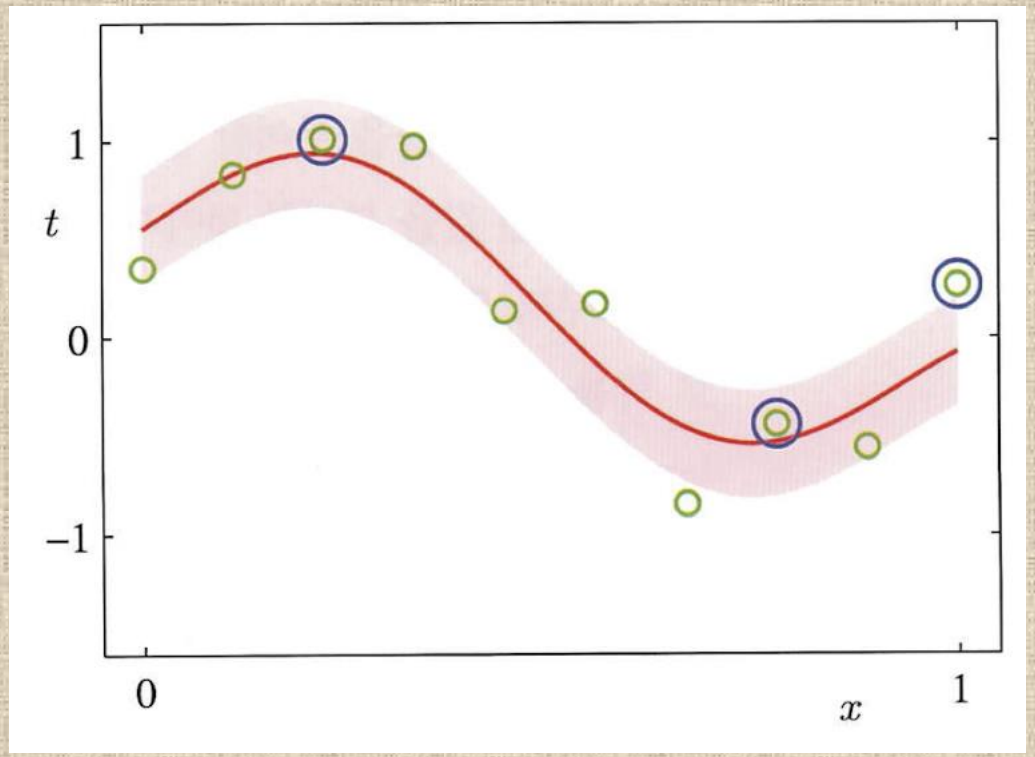
$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1}), \qquad \alpha_i\text{는 매개변수정밀도}$$

1-5 상관벡터머신

- α와 β를 근사한다(제약조건을 이용한 근사)
- α_i중 일부가 매우 큰 값을 가지고, 이에 해당하는 w_i들의 사후분포 평균,분산은 모두 0이 됨. 그러면 이에 해당하는 기저함수 $\phi_i(\mathbf{x})$는 모델에서 빠진다.
- 0이 아닌 나머지 가중치에 해당하는 입력값 $\mathbf{x}_n$들을 연관벡터(relevance vector)라고 한다. SVM의 서포트 벡터에 해당. 경계지역에 놓여있지 않는 성향이 있다.

1-5 상관벡터머신

- 사인곡선 회귀데이터 집합에 RVM적용결과



- 연관벡터 숫자가 서포트 벡터 숫자보다 확연히 적다
- SVM에 비해 훈련시간이 긴 단점. 다만 훈련이 한번. (SVM은 매개변수를 교차검증법을 통해 찾기 때문에 여러번)

02

그래프모델(1)

2-1 그래프모델

- 노드와 링크로 이루어짐(꼭짓점과 변/호)
  - 노드: 확률변수
  - 링크: 변수들간의 확률적 관계있음을 표현
- 방향 유무에 따라
  - 방향성 그래프 모델(베이지안 네트워크)
  - 비방향성 그래프 모델(마르코프 무작위장)

2-2 베이지안 네트워크

- directed graphical model
- 노드 a에서 b로 가는 링크가 있는 경우,
  - a가 부모노드
  - b가 자식노드
- 결합분포의 분해

$$p(a,b,c) = p(c|a,b)p(a,b) = p(c|a,b)p(b|a)p(a)$$

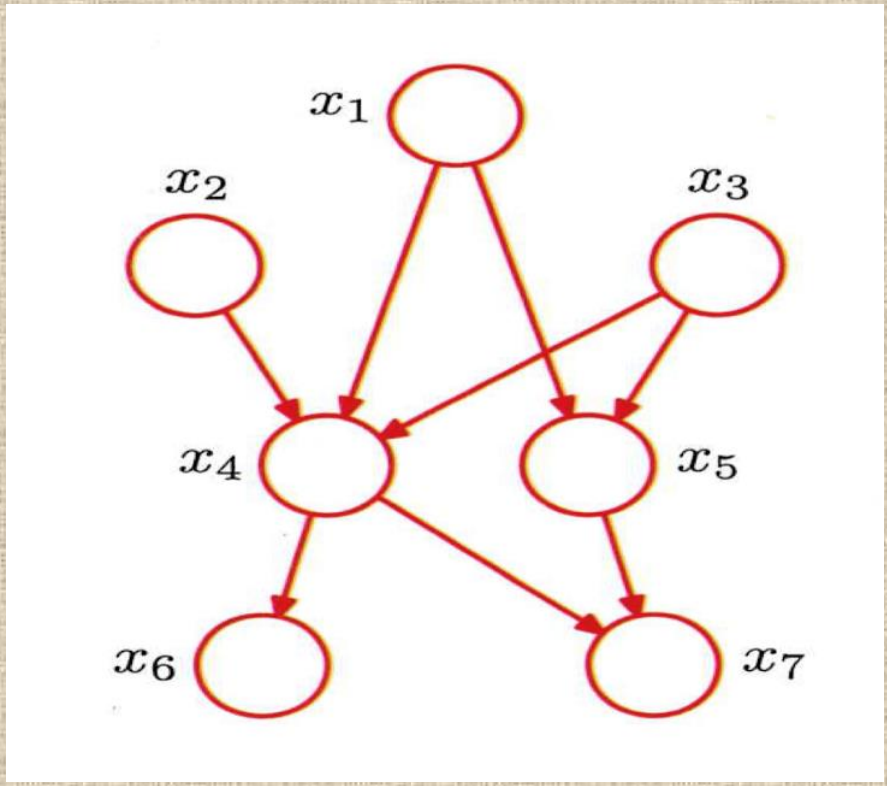
2-2 베이지안 네트워크

- 결합분포를 분해할 때 순서에 따라 다른 결과를 얻는다.
- 모든 노드쌍 사이에 연결이 있으면 '완전연결'
  - 예) K개 변수에 대한 (완전연결) 결합분포

$$p(x_1, \cdots, x_K) = p(x_K|x_1, \cdots, x_{K-1}) \cdots p(x_2|x_1)p(x_1)$$

2-2 베이지안 네트워크

- $p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$



2-2 베이지안 네트워크

- K개의 노드를 가지는 그래프의 경우 결합분포는,

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

  - 방향성그래프모델에서 '인수분해'를 나타냄
- 방향성 **순환**이 없어야 한다.

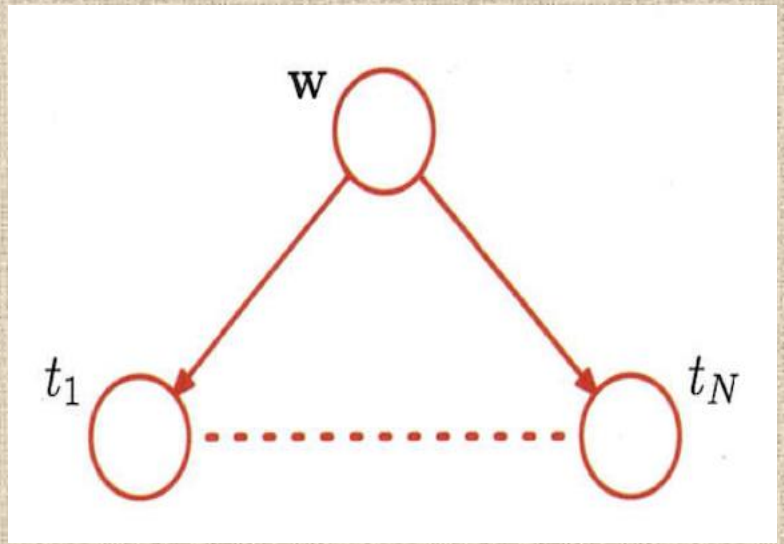
  = 방향성 비순환 그래프(directed acyclic graph: DAG)

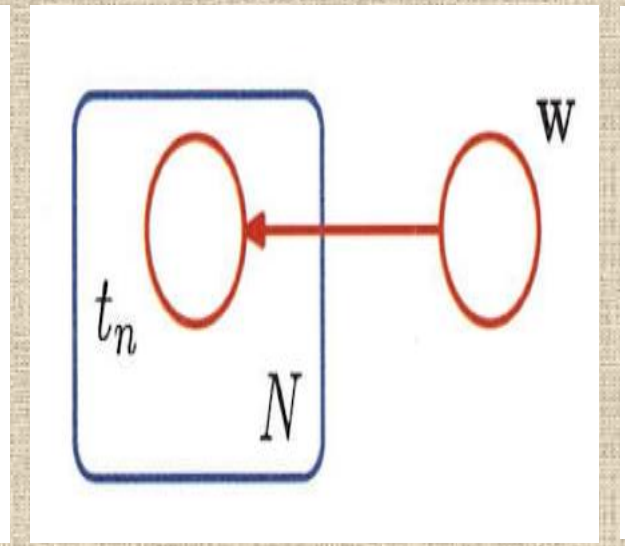
  = 모든 노드에 순서를 부여할 수 있다.

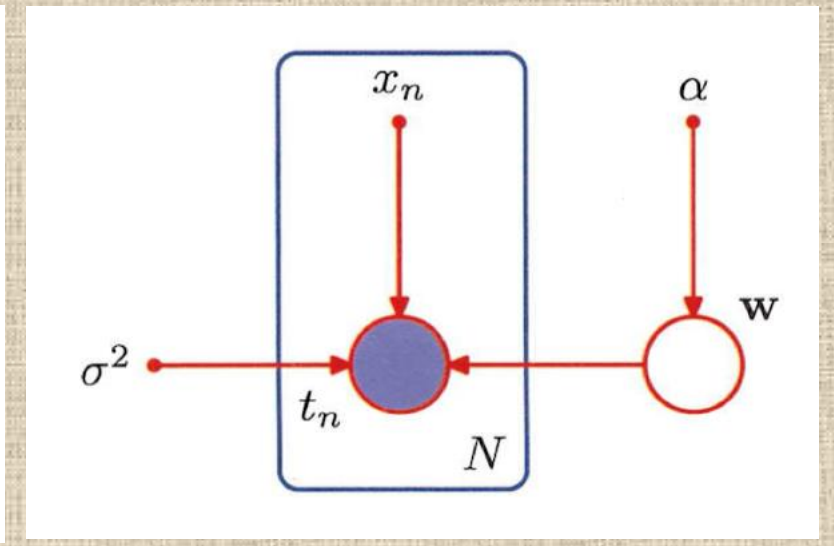
2-2 베이지안 네트워크 예시

1. 다항근사 - 앞서 나왔던 베이지안 다항회귀모델

$$p(\mathbf{t}|\mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$







Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.3, 8.4, 8.6

2-2 베이지안 네트워크

2. 생성적 모델 - 표본을 추출해야 할 상황: ancestral sampling

K개의 변수에 대한 결합분포 $p(x_1, \cdots, x_K)$를 가정.

그래프는 DAG. 목표는 이 분포로부터 표본을 추출하는 것.

- 가장 조상(ancestor)의 분포$p(x_1)$로부터 표본추출.
- 각각의 노드를 순서대로 따라가면서 추출.
- 마지막 변수 x_K로부터 표본 추출하면 끝.

2-2 베이지안 네트워크

- 보통 단말(terminal)노드들이 관측변수, ancestor쪽 노드들이 잠재(latent)변수에 해당
  - 잠재변수는 관측변수에 대한 복잡한 분포를 더 단순한 분포를 바탕으로 구성된 모델을 통해 표현 : 추상화
- 데이터를 만들어 낼 수 있으므로 generative model
  - 앞의 선형회귀는 아니다.

2-2 베이지안 네트워크

- 부모와 자식노드가 모두 이산변수일 경우
  - K개의 상태 가능한 단일이산변수 x(1-hot encoding)의 확률분포 = $$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$\sum_k \mu_k = 1$조건이 있기 때문에 K-1개의 μ_k값만 필요.

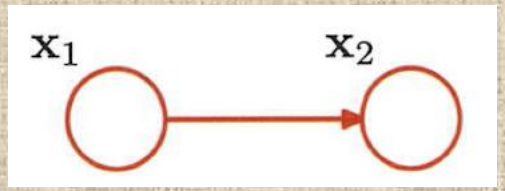
2-2 베이지안 네트워크

- 두개라면?

$$p(\mathbf{x_1}, \mathbf{x_2}|\boldsymbol{\mu}) = \prod_{k=1}^{K}\prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}$$

$\sum_k \sum_l \mu_{kl} = 1$을 가지기 때문에 매개변수 개수는 $K^2 - 1$

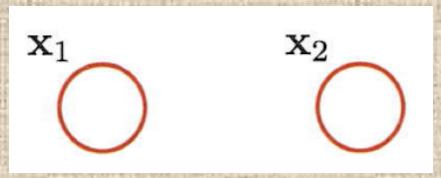
- M개라면? $K^M - 1$: 기하급수적 증가



$$p(\mathbf{x_1}, \mathbf{x_2}) = p(\mathbf{x_2}|\mathbf{x_1})p(\mathbf{x_1})$$

2-2 베이지안 네트워크

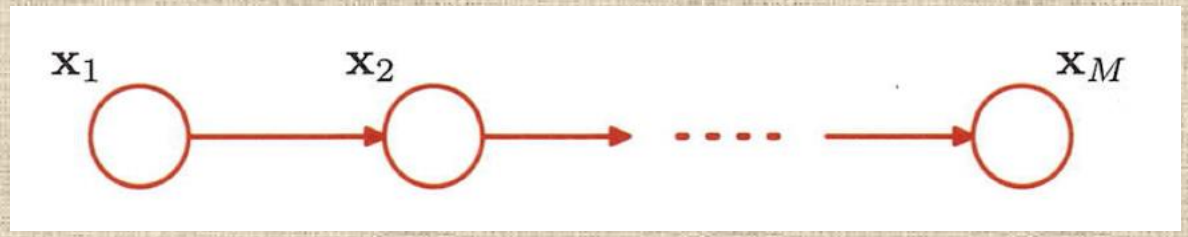
- 독립이면?



별개의 다항분포이므로 전체 매개변수 숫자는 $M(K-1)$

- 중간쯤이면 -

독립인 그래프보다 더 일반적이지만, complete graph 보다는 덜 하다. 예)

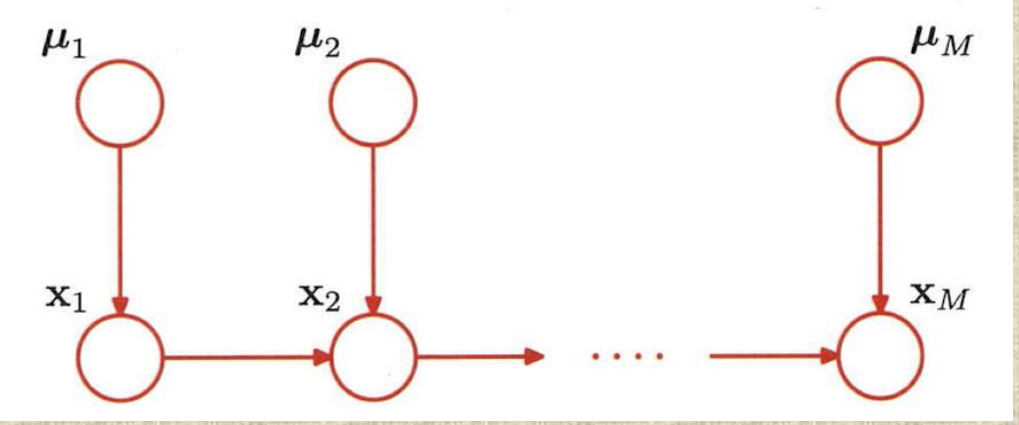


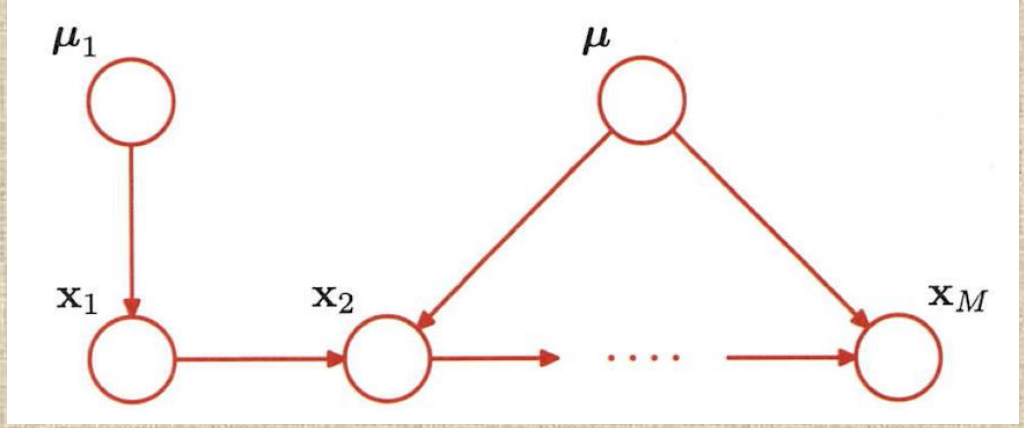
Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.10

2-2 베이지안 네트워크

- 매개변수 숫자의 기하급수적인 증가를 제어하기 위해 조건부 분포로 매개변수화된 *모델*을 사용할 수 있다.

  예) 매개변수에 대한 디리클레 사전분포를 도입.





Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.11, 8.12

2-2 베이지안 네트워크

- 부모와 자식노드가 모두 다변량 가우시안일 경우
  - 노드 i는 가우시안분포를 가지는 연속확률변수 x_i.

이 평균은 부모노드들의 상태 pa_i의 선형 *결합(sum)*

$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i|\sum_{j\in\mathrm{pa}_i} w_{ij}x_j + b_i\ , v_i\right)$$

w_{ji}와 b_i는 평균을 조정하는 매개변수, v_i는 분산

2-2 베이지안 네트워크

- (cont.) 이 경우 결합분포의 로그

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i|\mathrm{pa}_i) = -\sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j\in \mathrm{pa}_i} w_{ij}x_j - b_i\right)^2 + C$$

$\mathbf{x}$의 성분들에 대한 제곱식이므로

결합분포 $p(\mathbf{x})$는 다변량 가우시안.

- 결합분포의 평균과 공분산을 재귀적으로 구할 수 있다.

2-2 베이지안 네트워크

- (cont.) 그래프에 링크가 하나도 없어서 D개의 고립된 노드들로만 이루어져있으면, w_{ij}가 존재하지 않는다.
  (계산하면, 평균은 b_i에만 관계되고, 공분산행렬은 대각행렬.
  $\therefore$ D개의 독립적인 단변량 가우시안)
- Complete graph인 경우는 위 식대로 계산하면 된다.
  모든 i , j에 대해 w_{ij}가 존재.($D(D-1)/2$개)

2-3 조건부 독립

- $p(a|b,c) = p(a|c)$라면,

  a는 c가 주어진 상황에서 b로부터 조건부 독립

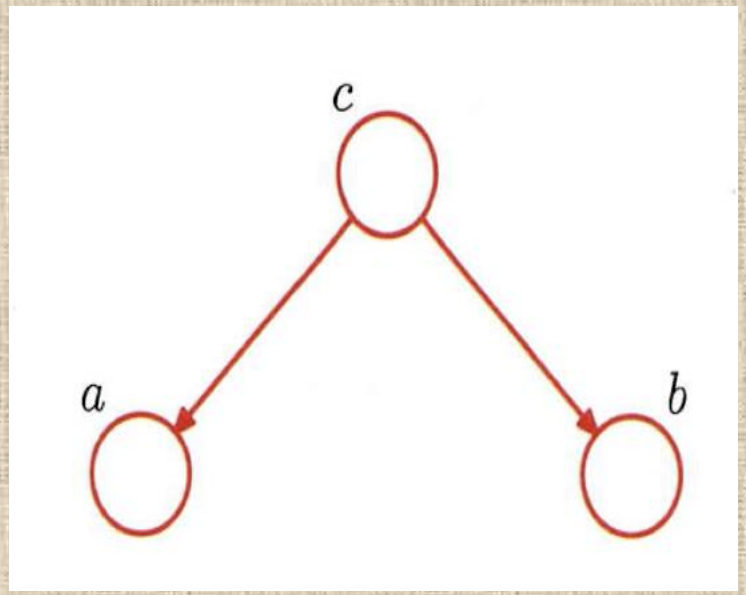
- $p(a,b|c) = p(a|b,c)p(b|c) = p(a|c)p(b|c)$

  c가 주어진 상황에서 a와 b는 통계적으로 독립적

- 둘은 같은 의미.

2-3 조건부 독립

- 아래 그림은 $p(a,b,c) = p(a|c)p(b|c)p(c)$

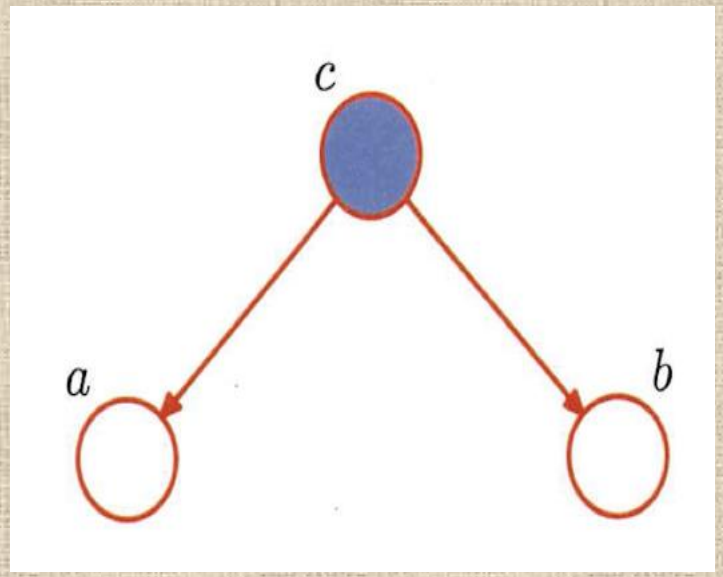


- 아무 변수도 관측되지 않았다면, $p(a,b) = \sum_c p(a|c)p(b|c)p(c)$이고, 이것은 $p(a)p(b)$로 인수분해되지 않는다. 따라서 조건부독립 아님.

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.15

2-3 조건부 독립

- 만일 같은 상태에서 c가 주어졌다면,



tail-to-tail 노드

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

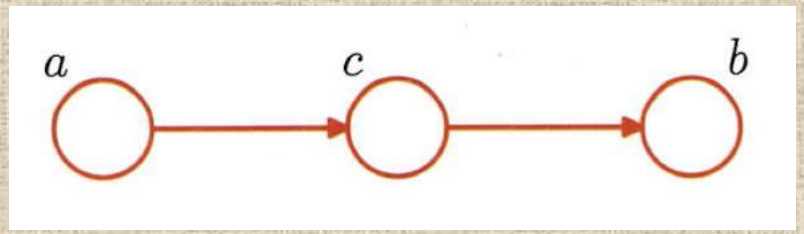
$$= p(a|c)p(b|c)$$

조건부 독립

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.16

2-3 조건부 독립

- 아래 그림은 $p(a,b,c) = p(a)p(c|a)p(b|c)$



head-to-tail

c에 대해 marginalize: $p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a)$

따라서 조건부 독립이 아님($p(a)p(b)$가 아니기 때문)

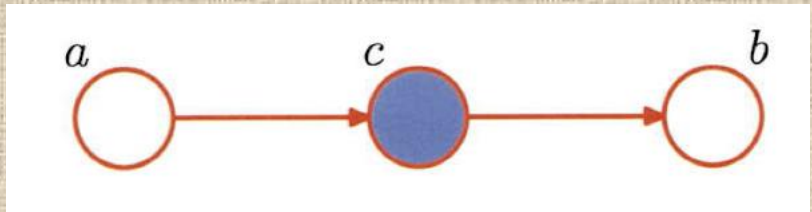
Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.17

2-3 조건부 독립

- c값이 주어지면?

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

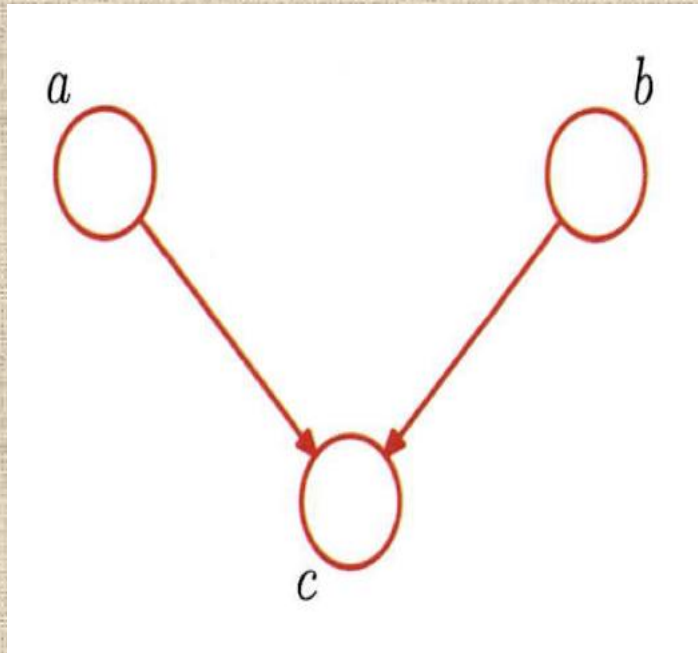
조건부 독립이다.



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.18

2-3 조건부 독립

- 아래 그림은 $p(a,b,c) = p(a)p(b)p(c|a,b)$



양변을 c에 대해 marginalize,

$= \sum_c p(a)p(b)p(c|a,b)$

$= p(a)p(b)$

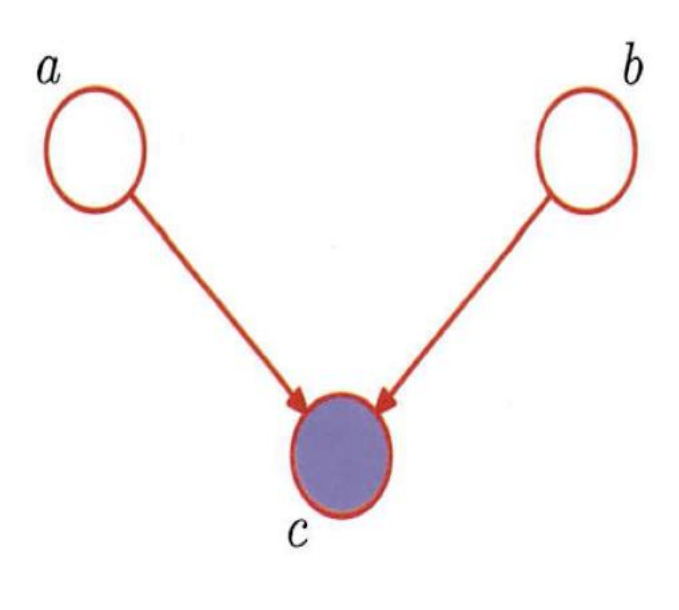
아무변수도 관측되지 않은 상태로

조건부 독립.

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.19

2-3 조건부 독립

- c값이 주어지면?



head-to-head

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

$$= \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

$\neq p(a|c)p(b|c)$ 조건부독립아님

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.20

다음시간

8강

- 그래프모델(2)
  - 조건부독립(계속)
  - d분리
  - 추론