기계학습

8강 그래프모델(2), 혼합모델

장필훈 교수

학습목차

1 그래프모델(2)

2 혼합모델

01
그래프모델(2)

1-1 조건부독립 전시간 요약

- tail-to-tail, head-to-tail 노드는
  - 관측되지 않은 경우 경로를 뚫린 채로 두고
  - 관측되면 경로를 막는다(조건부 독립이다)
- head-to-head는 위의 반대
  - 관측되지 않은 경우 조건부 독립
  - 관측되면 경로가 뚫림.

1-2 조건부 독립 예제

- 배터리B가 충전되어 있을 때 1, 방전되어 있을 때 0

  연료F가 가득 차 있을 때 1, 비어 있을 때 0

  측정기 G는 B,F가 모두 가득할 때 1, 모두 비었을 때 0

- 관련 사전확률과 그에 따른 G의 조건부 확률

  $p(B=1)=0.9, \qquad p(F=1)=0.9$

  $p(G=1|B=1,F=1)=0.8, \qquad p(G=1|B=1,F=0)=0.2$

  $p(G=1|B=0,F=1)=0.2, \qquad p(G=1|B=0,F=0)=0.1$

1-2 조건부 독립 예제

- G=0일 때, F=0일 확률은?

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}\cong 0.257$$

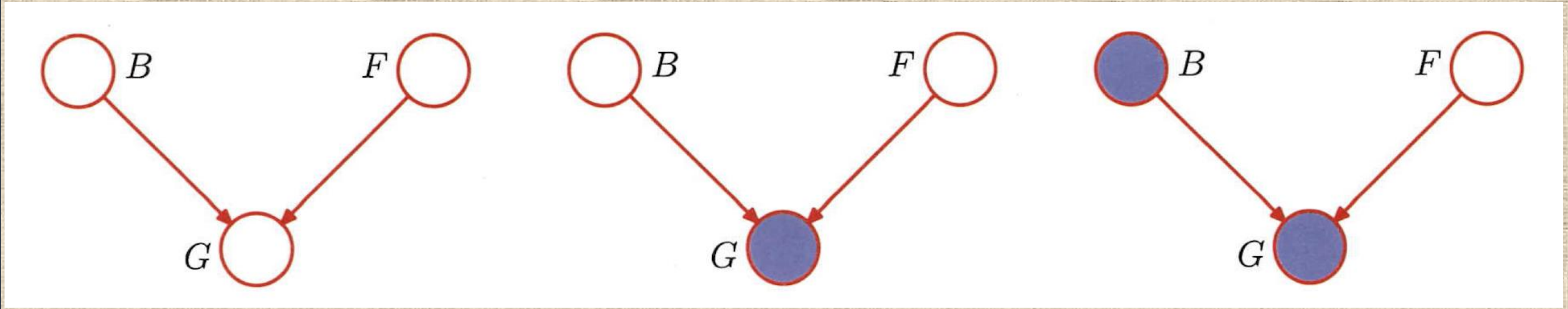
$$p(G=0|F=0)=\sum_{B\in\{0,1\}} p(G=0|B,F=0)p(B)=0.81$$

$$p(G=0)=\sum_{b\in\{0,1\}}\sum_{F\in\{0,1\}} p(G=0|B,F)p(B)p(F)=0.315$$

1-2 조건부 독립 예제

- $G = B = 0$이라면?

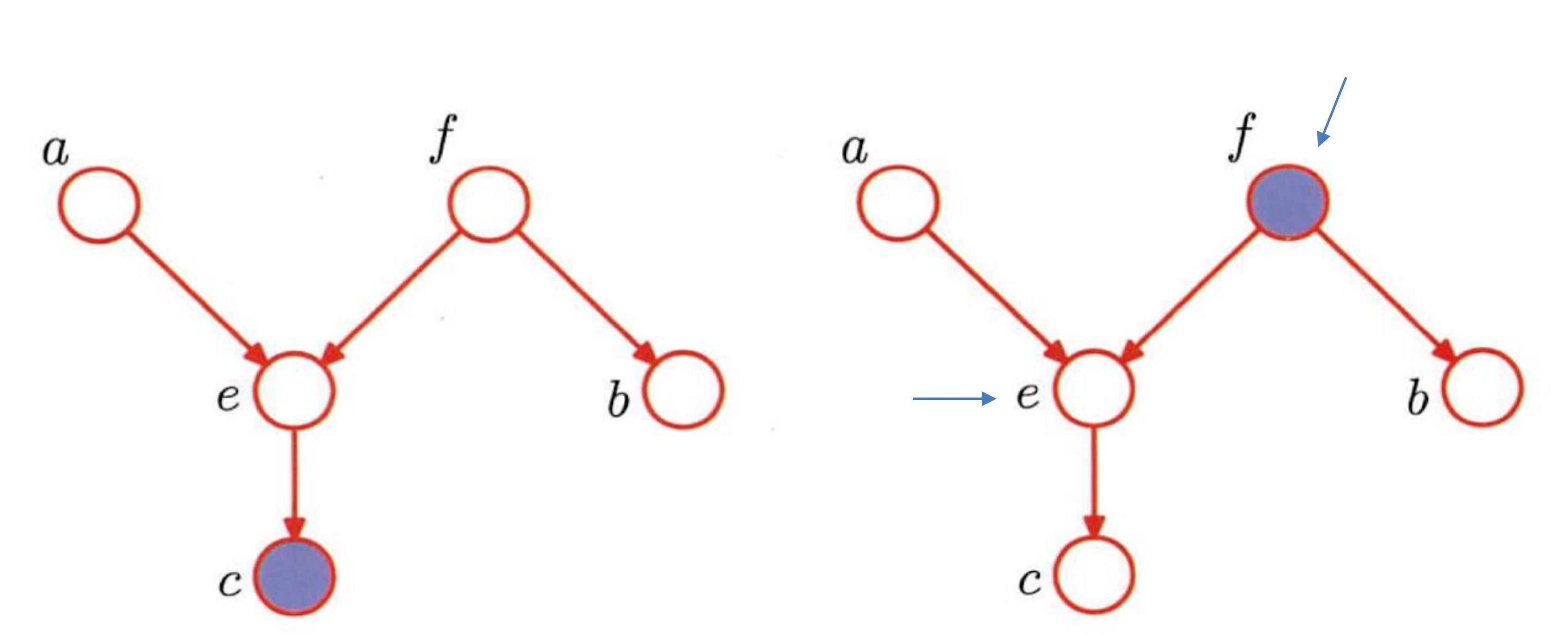
$$p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F\in\{0,1\}} p(G = 0|B = 0, F)p(F)} \cong 0.111$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.21

1-3 d분리

- a에서 b로 가는 경로를 보면,



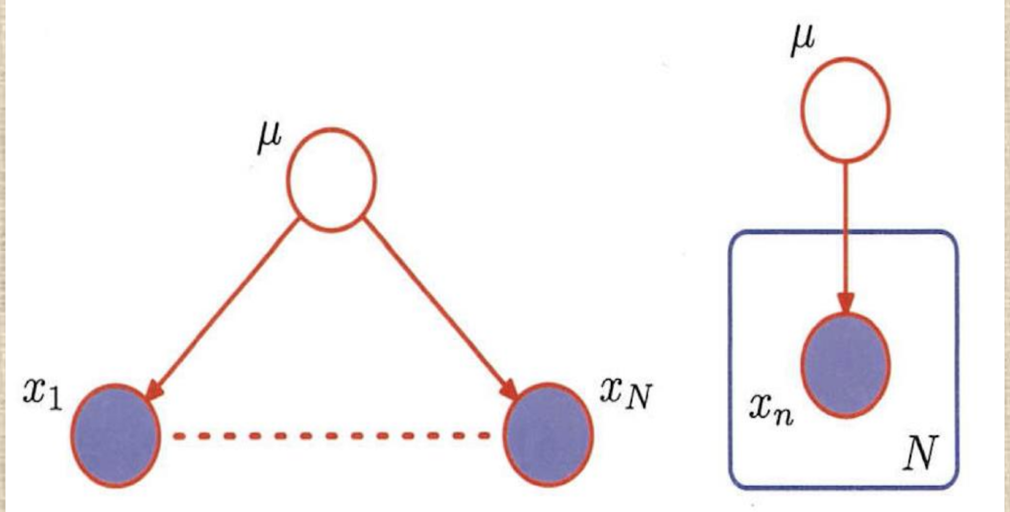
조건부 독립 아님　　　　조건부 독립(막혀있다)

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.22

1-3 d분리

- 가우시안분포의 평균에 대한 사후분포문제

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$



관측값(x_n)들은 μ가

주어졌을 때 독립적이다.

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

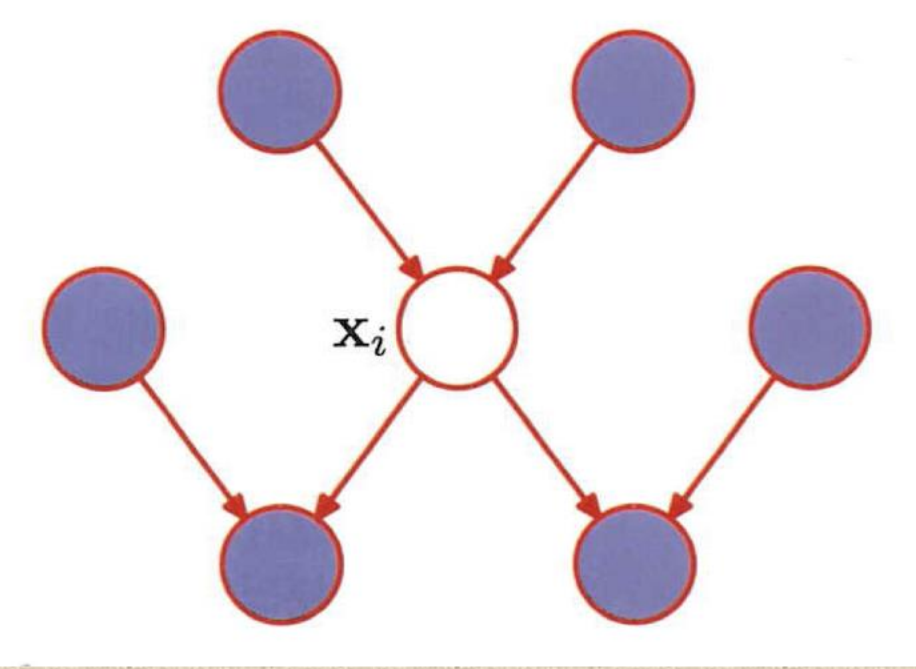
1-3 d분리

- μ는 관측되지 않았기 때문에, x_n들은 서로 독립이 아니다.

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

1-3 d분리

- 마르코프 블랭킷 : $\mathbf{x}_i$를 나머지 그래프로부터 분리시키는 최소한의 노드집합

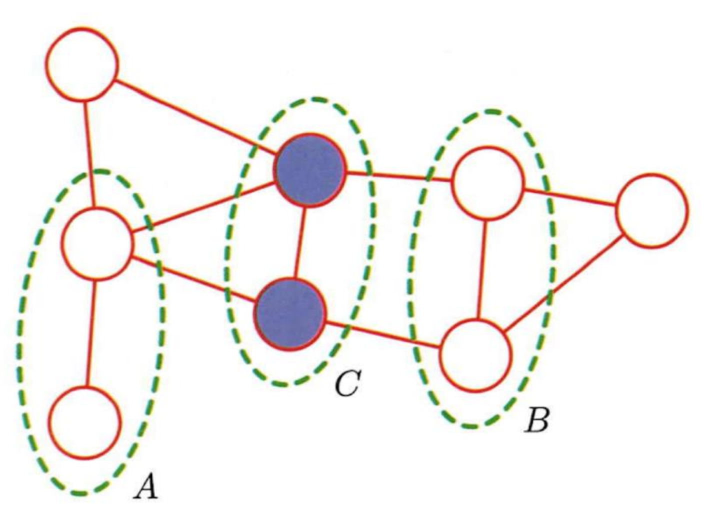


Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.26

1-4 마르코프 무작위장

- Markov random field

  = Markov network = undirected graphical model
- 앞서 다루었던 그래프와 ‘방향이 없다’는 것이 차이.
- 조건부 독립 성질:



1-4 마르코프 무작위장

- 조건부 독립 성질(계속)

  A의 노드들로부터 B의 노드들로 가는 모든 가능한 경로가

  집합 C의 하나 또는 여러 노드를 거쳐가야 한다면

  ‘전부 막힌 것’이고 따라서 조건부 독립성이 유효

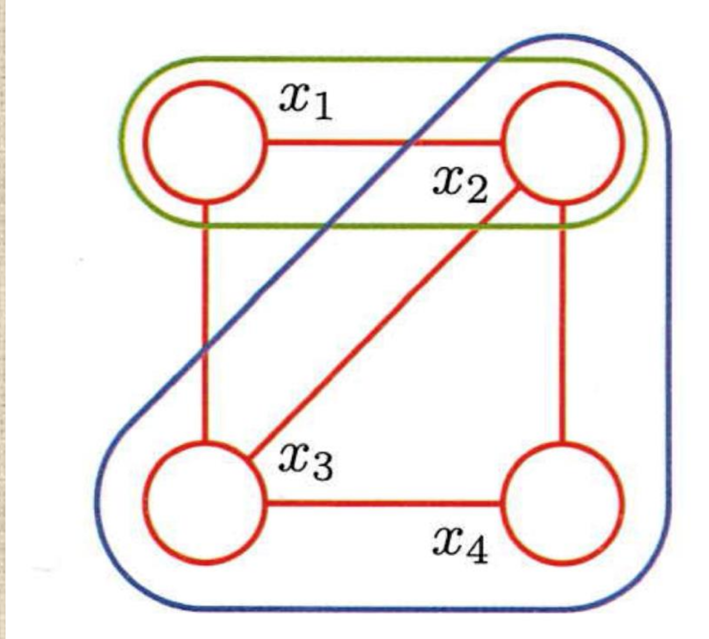
  (C가 주어졌을 때 A와 B는 독립)

1-4 마르코프 무작위장

- 인수분해성질
  - ‘지역성’을 어떻게 정의할 것인가?
- 클리크
  - 그래프의 부분집합
  - 그 부분집합에 속하는 모든 노드들간에 링크 존재
    - 완전연결

1-4 마르코프 무작위장

- 최대클리크
  - 임의의 노드를 추가하면 클리크가 망가진다



(파란색이 최대클리크,

나머지는 클리크)

- 결합분포를 최대클리크에 대한 포텐셜 함수로 적는다.

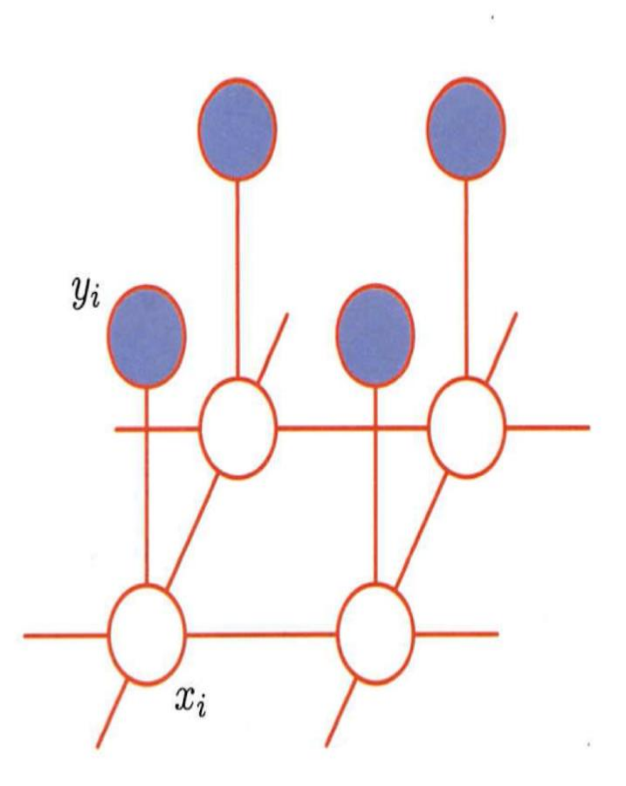
Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.29

1-4 마르코프 무작위장 예시

- 이미지 노이즈 제거
  - 관측 이미지가 y_i, 노이즈 없는 원 이미지가 x_i
  - 원래 이미지를 복원하는 것이 목표
  - 노이즈가 적으면 x_i, y_i간 상관관계가 크다
  - 근처 픽셀끼리 상관관계도 크다(x_i끼리)

1-4 마르코프 무작위장

- 문제풀이에 사용되는 마르코프 무작위장 비방향성 모델 →
- $\{x_i, y_i\}$형태 클리크와 $\{y_i, y_j\}$형태 클리크들로 이루어짐
  - 각 클리크가 같은 부호를 가지는 것을 선호하도록 디자인



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.31

1-4 마르코프 무작위장

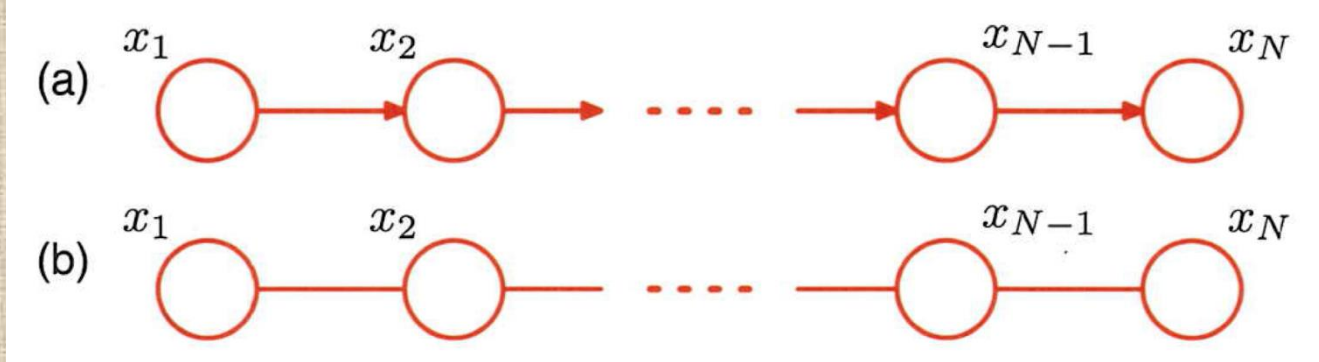
- 에너지함수를 정의한 후, 총 합이 적은 에너지를 가지도록 모든 픽셀에 대해 iteration.

$$E(x,y) = h\sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

  - 모든 i에 대해 $x_i = y_i$로 설정하고 시작, x_i마다 모두 순회하며 0, 1일때 에너지 계산후 변환.
  - $\beta = 0$이면 인접 픽셀간 연결이 사라짐($\therefore x_i = y_i$)

1-5 방향성 그래프의 변환

- 방향성 그래프 → 비방향성 그래프 문제



(a) $p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_N|x_{N-1})$

(b) $p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\ldots$

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.32

1-5 방향성 그래프의 변환

- 다음으로 놓을 수 있다.

$$\psi_{1,2}(x_1, x_2) = p(x_1)p(x_2|x_1)$$

$$\psi_{2,3}(x_2, x_3) = p(x_3|x_2)$$

$$\vdots$$

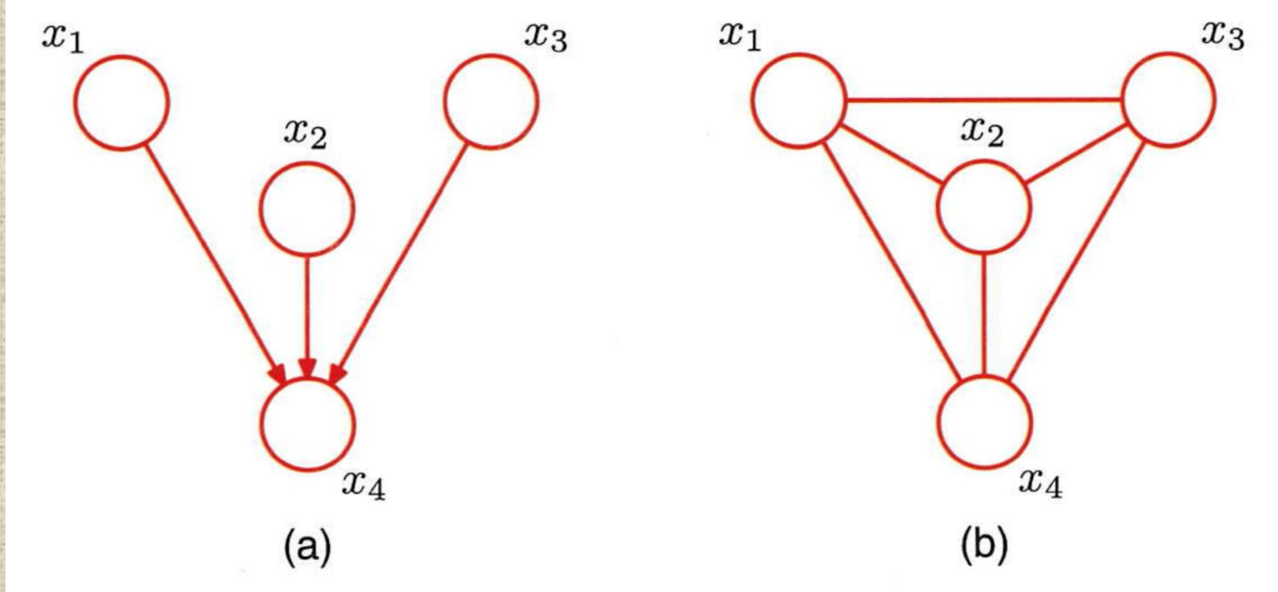
$$\psi_{N-1,N}(x_{N-1}, x_N) = p(x_N|x_{N-1})$$

- 더 복잡하게 생긴것을 고려해보자.

1-5 방향성 그래프의 변환

- 다음과 같다.

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.33

1-6 추론

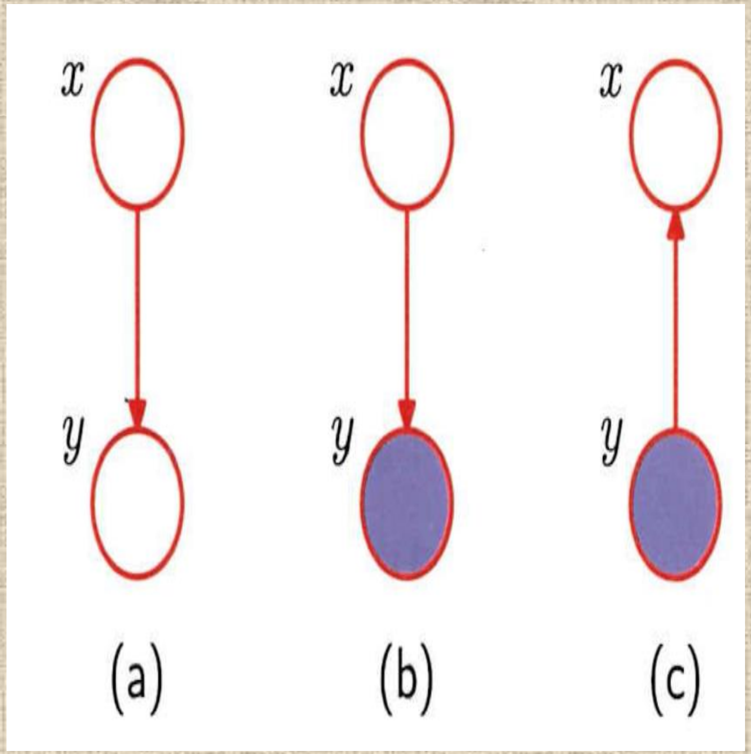
- 노드의 사후 분포를 계산해내는 것이 목표
- 예) 베이지안

(a) $p(x,y) = p(x)p(y|x)$

(b) y를 관측했다면,

(c) $p(x|y) = \dfrac{p(y|x)p(x)}{p(y)}$

화살표 방향이 반대로 바뀜



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig8.37

1-7 트리

- (비방향의 경우)

  모든 노드 쌍 사이에 하나만의 링크만 존재

  - 순환할 수 없다.

- (방향) 부모가 없는 하나의 노드(루트)와

  오직 하나의 부모만을 가지는 다른 노드로 이루어진 그래프.

- 익히 알고 있는 트리와 다르지 않음.

02
혼합모델

2-1 K-means

- 혼합모델: 더 복잡한 확률분포를 구성하기 위한 방법론
  - 데이터 집단화(clustering)에도 사용 가능
- K-means
  - ○ D차원 유클리드 확률변수 $\mathbf{x}$에 대한 N개의 관측값을 K개의 집단으로 나누는 것이 목표.
  - ○ K는 상수로 주어졌다고 가정.

2-1 K-means

- 각각의 데이터포인트로부터 가장 가까운 점까지 거리의 제곱합들이 최소가 되도록 한다.
- x_n이 집단k에 할당되면 $r_{nk} = 1, j \neq k$인 나머지 j에 대해서는 $r_{nj} = 0$ 이라고 하면, 목표 함수는,

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$$

2-1 K-means

- r_{nk}와 μ_k를 찾는것이 목표.
- 과정

1. μ_k 에 대한 초기값 설정
2. μ_k를 고정한 채로 J를 최소화하는 r_{nk}찾음 : Expectation
3. r_{nk}를 고정한 채로 J값을 최소화하는 μ_k찾음 : Maximization
4. 2~3 반복.

2-1 K-means

- 수식으로,

$$r_{nk} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \left\| \mathbf{x}_n - \mu_j \right\|^2 \\ 0 & \text{아닌 경우} \end{cases}$$

r_{nk}고정하고 μ_k최소화. (미분 $= 0$)

$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

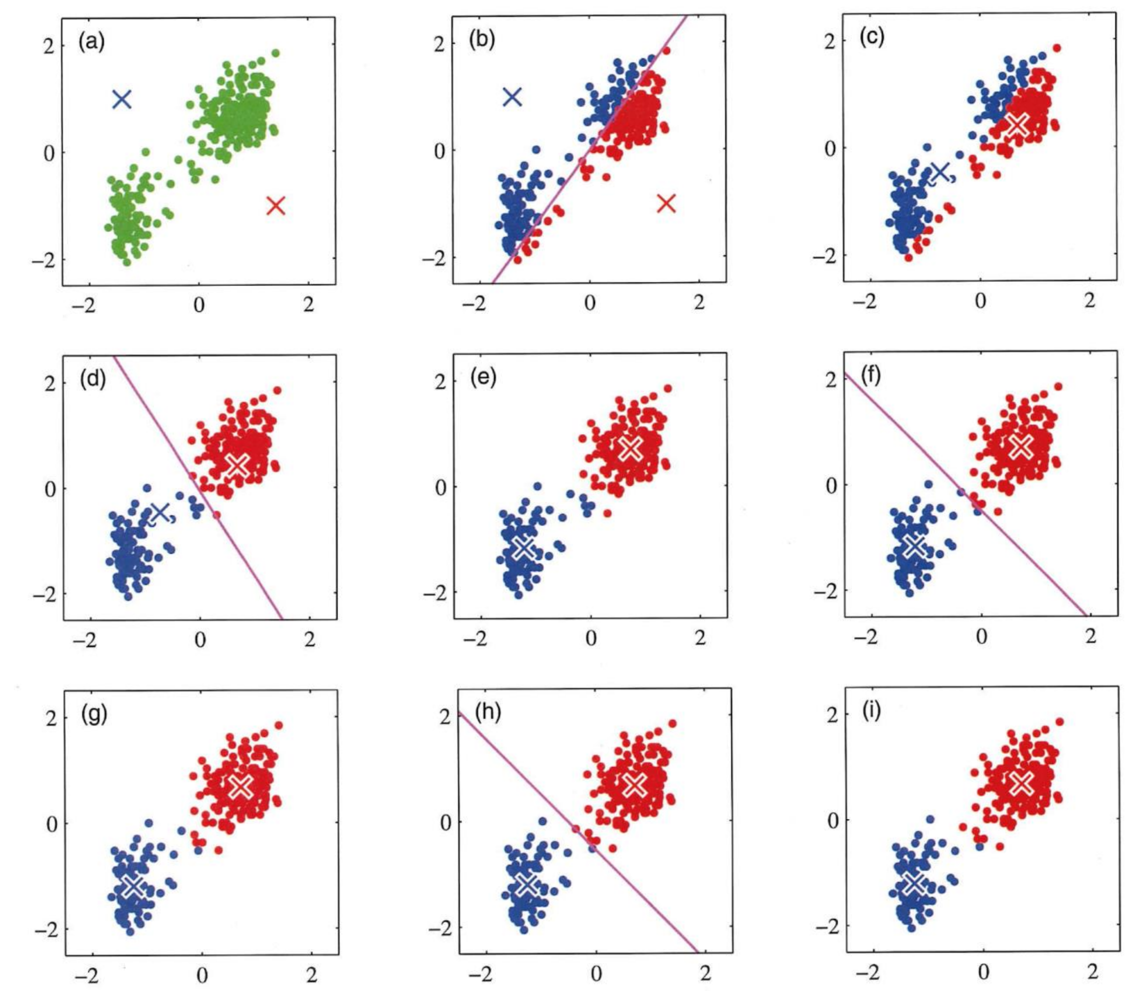
2-1 K-means

- 따라서,

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

- r_{nk}가 1-hot
- 집단 k에 할당된 모든 데이터포인트의 평균이 μ_k
- 더이상 새로운 할당이 없을때까지 반복
- 수렴은 보장되어 있음.

2-1 K-means



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig9.1

2-1 K-means

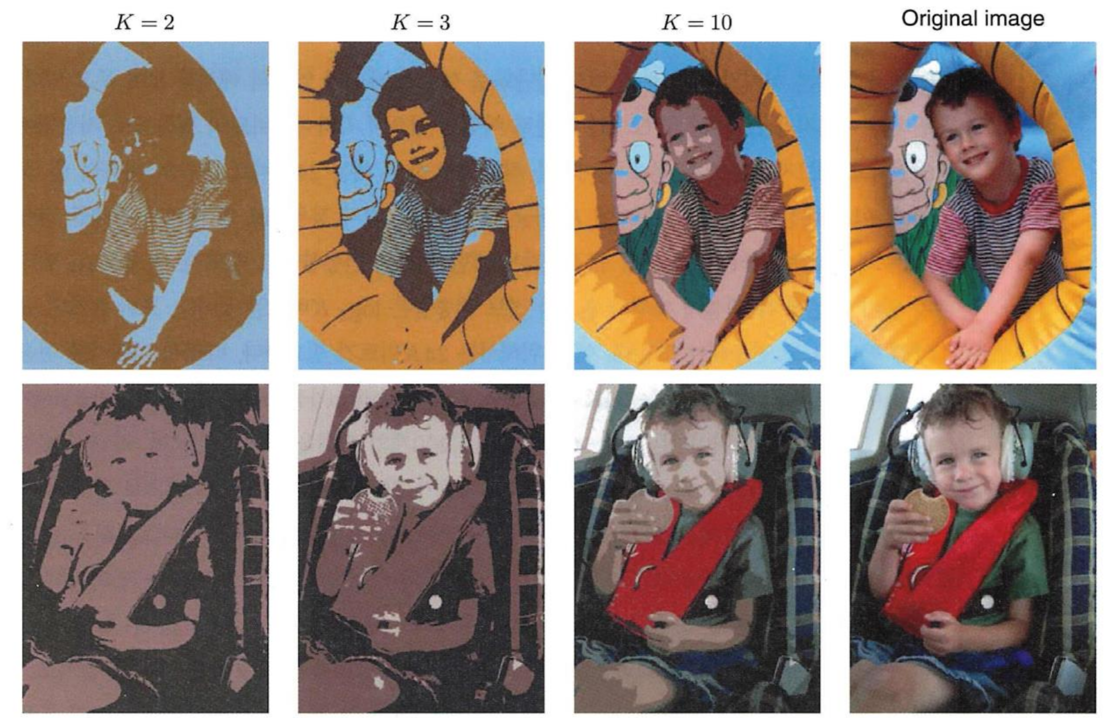
- 모든 데이터에 대해 매번 평균을 계산해야 한다
- 계산량을 줄이는 쪽으로 연구가 많음.
- 온라인방식도 가능

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{old} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{old})$$

- '거리'를 재정의하는 방법으로 확장 가능
- K-memoids : 평균이 아니라, 데이터포인트 내의 중앙값.

2-1 K-means

- 정확히 하나의 집단에만 매칭한다 ↔ 확률적 해석
- 응용의 예: 이미지 분할



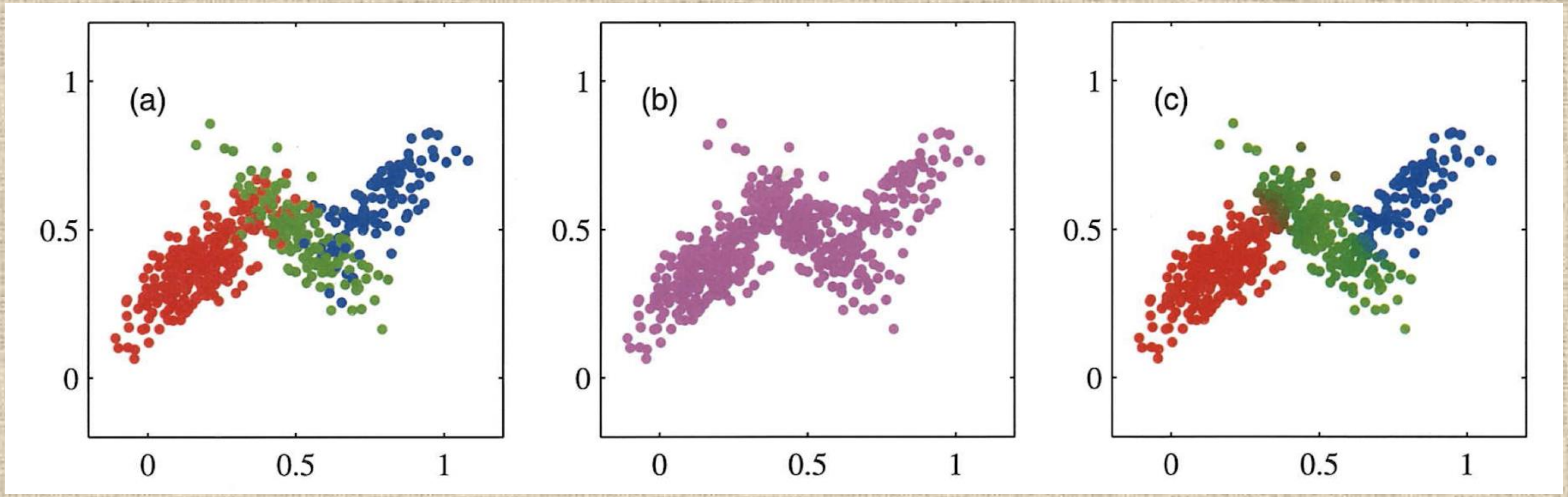
- 공간근접고려없음
- 의미반영 없음
- 활발한 연구주제.

Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig9.3

2-2 혼합 가우시안(Gaussian Mixture)

- 가우시안 혼합분포

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig9.5

2-2 혼합 가우시안(Gaussian Mixture)

- 최대가능도법으로 해결하기 어렵다

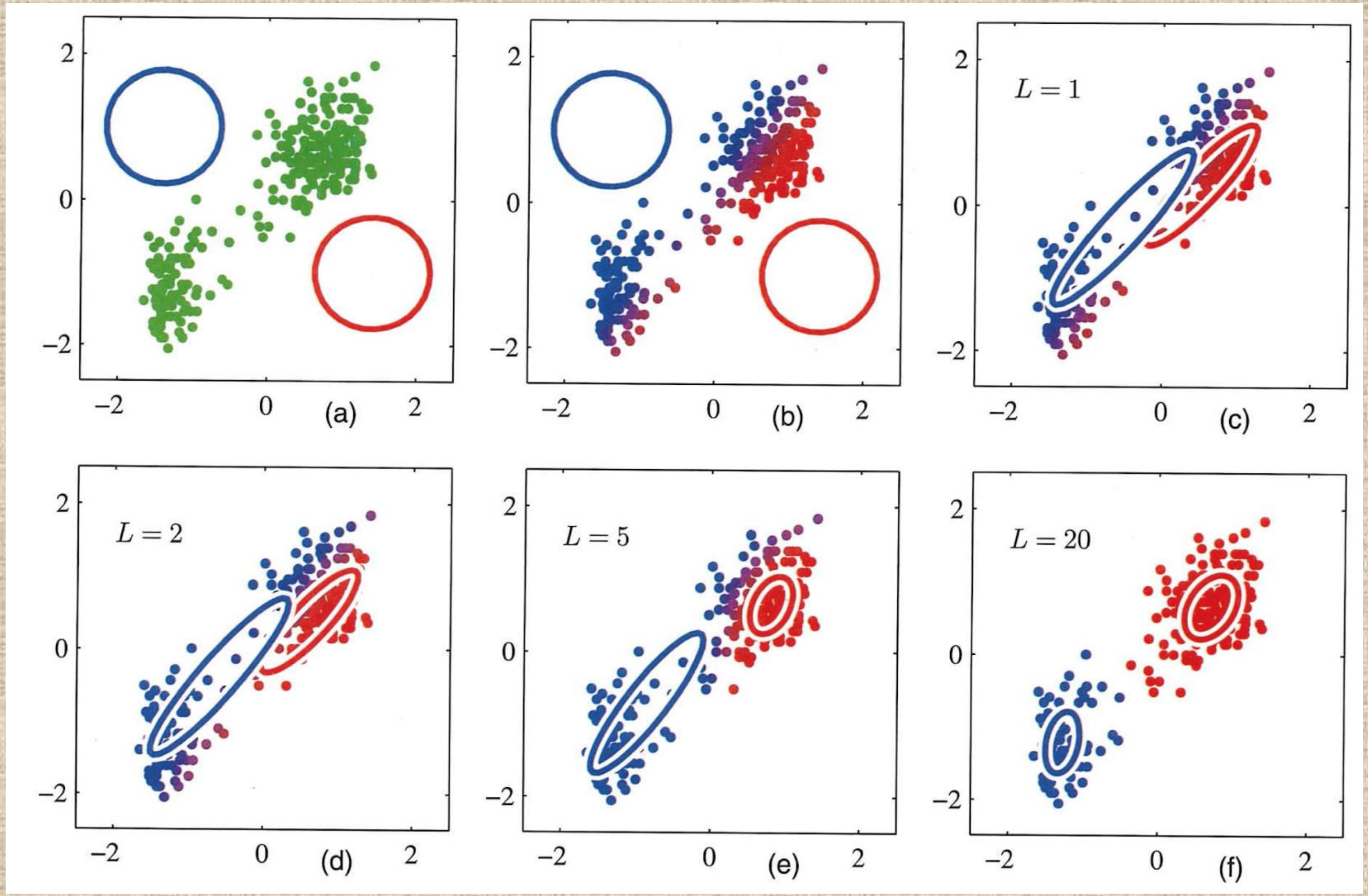
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

  - 로그 안의 합산항 때문에도 어렵고
  - 분산이 0으로 가면 로그값이 발산한다.

2-3 EM알고리즘

- 닫힌형태의 해를 제공하지는 않음
- 과정
  1. 평균,공분산,혼합계수의 초기값을 정한다
  2. E단계: 현재 매개변수로 사후확률값을 구한다
  3. M단계: 확률, 공분산, 혼합계수를 다시 계산한다
  4. 2,3을 반복(계수변화량등으로 종료조건 정해준다)

2-3 EM알고리즘



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig9.8

2-3 EM알고리즘

- 수렴을 위한 반복횟수가 K-means보다 훨씬 많다.
- 단계마다 계산량도 많음
- K평균을 미리 한번 사용하고, EM을 적용하는 경우가 많다
- 일반적으로 확장 가능
  - 매개변수 고정 → 사후확률계산으로 매개변수 재계산
  - 반복

2-3 EM알고리즘

- K평균은 가우시안 혼합분포에 대한 EM에 특정한 한계를 준 것으로 이해할 수 있다('오직 하나의 클래스에만 해당')
- 베이지안 선형회귀에도 적용할수 있다
- 잠재변수를 가지고 있는 확률모델의 최대가능도해를 찾기 위해 여기저기 다 쓸 수 있다.

03
KL divergence

3-1 KL Divergence

- 두 확률분포의 '차이'

$$KL(P\|Q) = \sum_i p_i(x)\log\frac{p_i(x)}{q_i(x)} = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx$$

- 차이지만, 대칭이 아니다$(KL(P\|Q) \neq KL(Q\|P))$
- KL divergece를 최소화 하는것이,

  log likelihood를 최대화 하는것과 같다.

다음시간

9강

- 표집법
- 연속잠재변수