워크북

교과목명 : 머신 러닝

차시명: 8차시

◈ 담당교수: **장 필 훈**

◉ 세부목차

- 조건부독립 예제
- d분리
- 마르코프 무작위장
- 트리
- 혼합모델
  - K-means
  - 혼합 가우시안
  - EM

학습에 앞서

■ 학습개요

이전시간에 이어 그래프 모델에 관해 조건부독립을 복습하고 예제를 통해 이해한다. 조건부독립을 바탕으로, d분리 개념에 대해 새로 배우고 이해한다. 분리집합인 마르코프 블랭킷에 대해서도 배운다. 방향이 있는 그래프에 이어 방향이 없는 그래프인 마르코프 무작위장에 대해서도 배우고 '클리크'의 개념을 익힌다. 이 개념을 이용하여 이미지의 노이즈 제거 문제를 다루어본다. 마지막으로 방향성그래프와 비방향성 그래프를 변환하는 방법에 대해 알아본다.

혼합모델의 경우, 대표적으로 K-means를 자세히 살펴보고, 그 특성을 수식 전개를 통해 알아본다. K-means는 EM알고리즘의 일종으로도 볼 수 있고, 해당 관점에서 알고리즘을 이해한다. 더 나아가 가우시안 혼합분포를 EM을 이용하여 푸는 예제를 살펴본다.

■ 학습목표

1	그래프모형에서 조건부독립을 이해하고 예제를 통해 개념을 숙지한다.
2	조건부독립을 이용해서 d분리를 유도해낸다.
3	마르코프 무작위장에서 클리크의 개념을 이용해 독립을 설명해내는 방법을 이해한다.
4	이미지의 노이즈 제거 문제에서 그래프모델이 어떻게 쓰이는지 본다.
5	혼합모델의 대표격인K-means 알고리즘을 이해하고, EM의 특수한 형태임을 안다.
6	가우시안 혼합모델에서 EM이 어떻게 쓰이는지 보고, EM의 활용방법을 숙지한다.

■ 주요용어

용어	해설
d분리	그래프상의 두 노드(확률변수)가 서로 조건부 독립이면 두 노드는 d분리조건을 충족하는 것이다. 그래프 위에서 노드의 집합을 정의할 수 있는데, 각 집합간에도 d분리조건을 만족하는지 알아볼 수 있다. 마르코프블랭킷은 방향그래프에서 특정 노드를 다른 노드들로부터 분리시키는 최소한의 노드집합을 말하는데, 관점을 바꾸면, 마르코프 블랭킷이 특정노드와 다른 노드 사이에 d분리를 가능하게 한다고 볼 수 있다.
마르코프무작위장	방향성 그래프 모델의 대표적인 예가 베이지안 네트워크이듯, 비 방향성 그래프의 대표적인 예로 마르코프 무작위장을 들 수 있다. 다만, 마르코프 무작위장에서는 순환이 가능하다. 마르코프 무작위장에서도 조건부독립 성질을 논할 수 있고, 인수분해 성질도 수식화할 수 있다. 베이지안 네트워크로 변환도 가능하다.
클리크	그래프의 부분집합중에, 그에 속하는 모든 노드간 연결이 존재하는 집합(완전연결)을 클리크라고 한다.
EM알고리즘	제약조건 있는 최적화 문제에서 해석적 해를 구하기 어렵거나 불가능할 때, 근사하는 알고리즘 중에 대표적인 것. 더 형식적으로는 '잠재변수를 가지고 있는 확률적 모델의 최대가능도 해를 찾기 위한 일반적인 테크닉'이다. E(Expectation)단계와 M(Maximization)단계의 iteration으로 최적해에 점점 근사해간다.

	학습하기

조건부독립의 예제문제를 하나 풀어보겠습니다. 배터리B가 충전 되었을 때 1, 방전 되었을 때 0으로 나타냅니다. (중간은 없습니다). 연료F가 가득 차 있을 때 1, 비어 있을 때 0입니다. 측정기G는 B, F가 모두 가득할 때 1, 모두 비어 있을 때 0을 가리킵니다. 관련 사전 확률과 그에 따른 G 의 조건부 확률은 다음과 같습니다.

$\mathrm{p}(B=1)=0.9,\ p(F=1)=0.9,$

$p(G=1|B=1,F=1)=0.8,$

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

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

$p(G=1|B=0,F=0)=0.1.$

G=0일때 F=0일 확률은 다음과 같습니다.

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

F=0일때 G=0일 확률은 다음과 같습니다.

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

G=0일 확률은 다음과 같습니다.

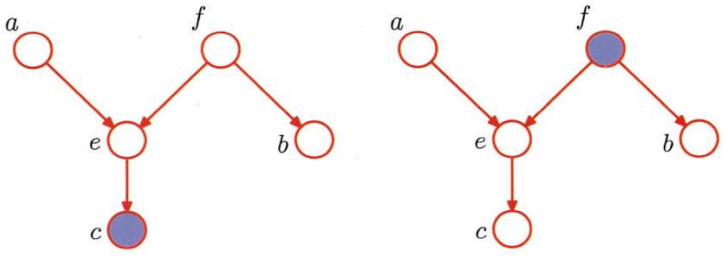
$$p(G=0)=\sum_{b\in\{0,1\}}\sum_{F\in\{0,1\}}p(G=0|B,F)p(B)p(F)=0.315$$

만약 G=B=0이면 F=0일 확률은 얼마일까요. 다음과 같이 구합니다.

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

<d분리>

앞서 우리는 조건부 독립이 그래프에서 어떻게 나타나는지 관찰해 보았습니다. 그래프의 구조에 따라 특정 노드가 관찰 되면 다른 두 노드가 조건부 독립이 되거나 조건부 독립이었던 두 노드가 의존하게 되었습니다. 앞서 배운 간단한 구조 세가지만으로 복잡한 그래프상의 조건부독립여부를 알아낼 수 있습니다. 다음의 두 예를 봅시다.

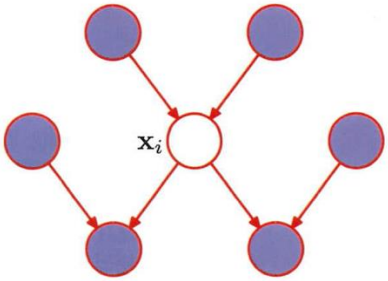


Bishop 8.22

왼쪽에서 c 가 관찰되었을 때 a와 b가 독립일까요. 우선 f는 이 경로를 '막지' 않습니다. 우리는 앞서, 꼬리 대 꼬리노드는 중간 노드가 관측되지 않았다면 조건부독립이 아님을 이미 보았습니다. e노드도 경로를 막지 않습니다. 관측되지 않은 머리 대 머리노드이지만 e의 자손 c가 관측되었으므로 a와 f를 의존성있게 만듭니다. 결국 a-b는 독립이 아니게 됩니다.

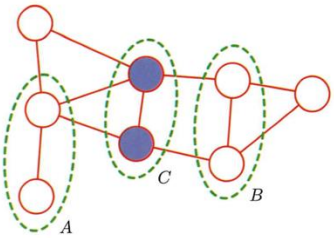
오른쪽 그림은 그 반대입니다. 노드 f와 노드 e가 모두 경로를 막고 있고 결국 a와 b는 조건부 독립이게 됩니다.

복잡한 그래프에서, 노드를 다른 모든 그래프상의 노드로부터 분리시키는 최소한의 노드집합을 마르코프 블랭킷이라고 하고 다음과 같이 생겼습니다.



Bishop 8.26

이제 방향성없는 그래프를 다루어 보겠습니다. 방향성 없는 그래프의 경우 조건부 독립이 되려면 두 노드 사이에 가능한 모든 경로가 '막혀'있어야 합니다. 즉 관측되어야 합니다. 그림으로 나타내면 아래와 같이, C의 모든 노드가 관측되었을 때 A의 모든 노드와 B의 모든 노드가 조건부독립입니다.



마르코프 무작위장이 적용되는 예로 이미지의 노이즈 제거 문제를 들 수 있습니다. 구체적인 과정과 수식을 강의시간에 다루었으니 참고바랍니다.

트리는 그래프의 특별한 형태입니다. 모든 노드 쌍 사이에 하나만의 링크가 존재하고 순환도 없어야 합니다. 트리도 방향을 가질 수 있는데, 이럴때는 부모가 없는 단 하나의 노드(루트)와 오직 하나의 부모만을 가지는 다른 노드들로 이루어진 그래프를 말합니다.

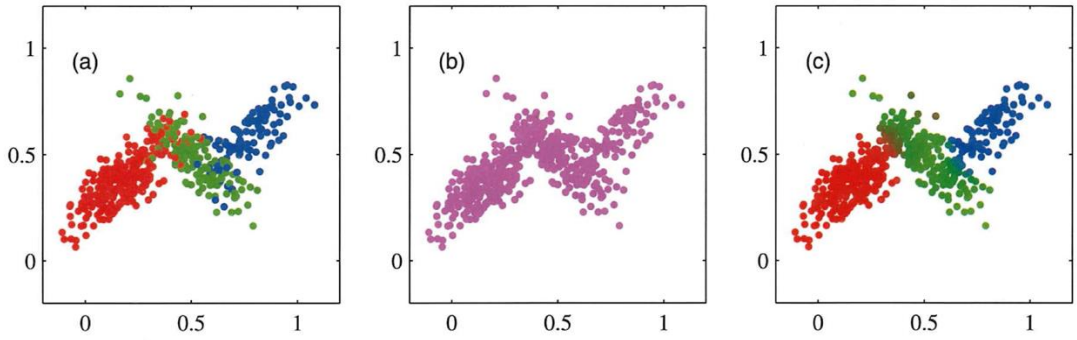
<혼합모델>

우리는 지금까지 잘 알려진 분포를 가정하고 문제를 풀었습니다. 만약 분포가 복잡한 형태라면 이렇게 단순한 형태 하나를 가정하고 추정하는 것보다, 단순한 분포 여러개의 합이라고 가정하고 추

정해 내는것이 데이터포인트의 원래 분포에 대한 좀 더 정확한 추정일 것입니다.

가장 단순한 예로 우리에게 익숙한 k-means를 살펴보겠습니다. k-means는 d차원 유클리드 확률변수 x에 대한 n개의 관측값을 k개의 집단으로 나누는 것이 목표입니다. 각 집단이 가지는 구체적인 분포의 형태가 아니라, 어느 집단에 속하는지 알아내는것만을 목표로 한다는 점에 주의하세요. 각 집단이 어떤 분포를 가지든 그 결합분포상에서 주어진 데이터포인트가 어떤 집단에 속하는 것이 가장 그럴듯한지 찾아내는 문제라고 생각하면 됩니다. 집단의 개수 k는 주어졌다고 가정합니다. 구체적으로는, 각각의 데이터포인트로부터 가장 가까운 점(가장 가까운 집합의 센터)까지 거리의 제곱합들이 최소가 되도록 합니다. 이것은 E-M으로 이해할 수도 있습니다. E-M은 뒤에 더 구체적으로 보겠지만, Expectation-Maximization과정을 이르는 말로, 우리의 가정에 따른 기댓값을 최대화 하는 과정을 반복함으로써 답에 접근해가는 과정을 말합니다. k-means의 경우 우리가 일단 각 집합의 센터를 가정하고서, 모든 데이터포인트를 가장 가까운 센터로 배정하는 과정이 E과정입니다. 그리고 배정된 데이터포인트들의 센터를 다시 계산하는 과정이 M에 해당합니다. k-means는 온라인 방식도 가능한데, 데이터 포인트를 센터에 모두 배정한 뒤 새로운 센터를 계산하는 것이 아니라, 하나 배정할때마다 해당 클러스터의 센터를 다시 계산하는 방식으로 동작합니다. 평균이 아니라 중간값을 사용하는 방법(k-medoids)도 있습니다. k-means의 자세한 과정을 녹화강의에 다루었으니 참고 바랍니다.

혼합 가우시안도 대표적인 혼합모델의 예입니다. 다음 그림을 보면 직관적으로 이해할 수 있습니다.



Bishop.9.5

맨 왼쪽이 원 분포, 두번째가 주어진 데이터 포인트, 세번째가 두 번째 데이터 포인트를 이용해 추정해 낸 분포입니다. 가우시안 혼합분포는 다음과 같이 주어지기 때문에 최대가능도법으로 해결하기 어렵습니다.

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

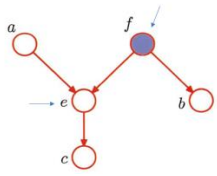
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

가능도함수를 보면 로그 안에 합산이 있는 것을 볼 수 있습니다. 이것이 계산을 어렵게 만듭니다. (닫힌 형태의 해를 제공하지 않습니다.) 그래서 이 경우 EM알고리즘을 씁니다. 우선 평균,공분산,혼합

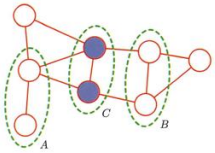
계수의 초깃값을 정하고, 현재 매개변수들을 사용해서 사후확률을 구합니다. 그리고 이 사후분포를 바탕으로 각 매개변수를 다시 추정해냅니다(이 단계는 최대가능도법과 같습니다). 이 두 단계를 값이 어느정도까지 수렴할때까지 반복합니다. 수렴을 위한 반복횟수가 k-means보다 많고, 단계마다 계산량도 매우 많습니다. k-means는 가우시안 혼합분포에 대한 EM에 특정한 한계를 준 것으로도 이해할 수 있습니다(한계: 오직 하나의 클래스에만 할당한다). 베이지안 선형회귀에도 EM을 적용할 수 있습니다. 나아가, 잠재변수를 가지고 있는 확률모델의 최대가능도해를 찾기 위해 거의 일반적으로 적용할 수 있는 방법입니다. 해를 한번에 계산해낼 수 없기 때문에 임의의 초깃값을 주고서 한번 계산한 뒤(초깃값은 틀린 답일 것이므로, 사후확률로 다시 가능도를 최대화 했을 때 값이 바뀝니다) 최대가능도법을 사용하고, 이 과정을 반복한다고 기억해두세요.

연습문제

1. head-to-head노드는 들어오는 엣지만 있고 나가는 엣지는 없다.
    a. O
    b. head-to-head노드의 정의
2. head-to-head노드의 경우, 관측이 되면 들어오는 두 노드를 서로 연결한다(조건부독립이 아니게 된다)
    a. O
    b. 본문참고. head-to-tail, tail-to-tail과 다름.
3. 아래의 경우 a와 b는 조건부독립이다.



    a. O
    b. e노드는 자손노드c까지 관측되지 않은 h2h노드이므로 연결을 끊고,
    f노드는 t2t 노드인데 관측되었으므로 연결을 끊는다. 따라서 a와 b는 조건부독립이 된다.
4. 비방향성 그래프에서도 조건부독립 성질을 가지도록 만들 수 있다.
    a. O
    b. 다음과 같을때 A와 B는 조건부독립이다.



5. 엣지가 하나 존재하는 두 노드는 무조건 클리크이다.

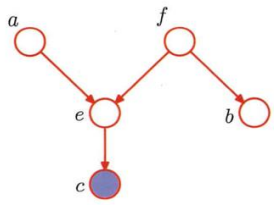
a. O

b. 클리크의 정의(집합 내 존재하는 모든 노드간 연결이 존재하기만 하면 클리크)에 따라 무조건 클리크이다. 최대 클리크는 아닐 수 있다.

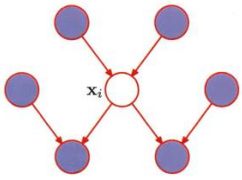
6. K평균은 모든 데이터에 대해, K개의 평균까지 거리를 계산해야 한다.
   a. O
   b. 가장 가까운 점을 알아내기 위해 매번 모든 점에 대해 연산해야 한다. 연산량을 줄이는 방향으로 여러가지 알고리즘이 연구되어 있다.
7. K평균은 혼합가우시안의 EM알고리즘 해결에 특별한 제약을 가한 것이다.
   a. O
   b. EM알고리즘에 '오직 한 클래스에만 할당'이라는 제약을 부여한 것이다. 출력값이 확률적으로 해석되지 않게 한다(혹은 확률100%나 0% 둘중 하나로만 해석되게 한다)
8. 두 확률분포의 차이를 나타내는 KL-divergence는 KL(p,q)=KL(q,p)이다.
   a. X
   b. KL-divergence 는 '차이'지만 대칭은 아니다. 기준이 되는 분포가 존재한다. 그래서 쓸 때도 순서에 유의하여 쓴다. KL-divergence를 최소화하는것이 log likelihood를 최대화하는 것과 같다.

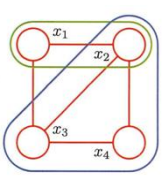
정리하기

1. head-to-tail, head-to-tail 노드는 관측이 되었을 경우 양쪽을 막는다.
2. head-to-head노드는 반대(관측이 되면 연결한다)
   a. tail이 들어가면, 관측된경우 양쪽을 막는다고 외우면 쉽다.
3. 연결이 하나만 있는 자손 노드도 해당 노드와 동일하게 생각한다, 예를들어 아래에서 e노드는 관측되지 않았지만, c노드(e노드(h2h)의 자손)는 관측되었으므로 a와 f는 조건부독립이 아니다.



4. x_i 를 나머지 그래프로부터 분리시키는 최소한의 노드집합을 마르코프 블랭킷이라고 한다.





5. 비방향성 그래프에서 모든 노드 사이에 연결이 존재하는 부분집합을 클리크라고 하고, 더이상의 노드를 추가하면 클리크가 되지 않는 상태를 최대클리크라고 한다. 오른쪽에서 파란색이 최대클리크, 나머지는 클리크.
6. 방향성 그래프를 비방향성 그래프로, 비방향성 그래프를 방향성 그래프로 변환하는 방법이 존재한다.
7. K-means방법
    a. 초기값은 K로 주어진다.
    b. 수렴성이 보장된다.
    c. 계산량이 많다.
    d. 온라인 방식으로도 가능하다
    e. '거리'를 재정의하는 방법으로 일반화도 가능하다.
    f. 정확히 하나의 집단에만 매칭한다
8. 혼합가우시안의 경우
    a. 하나의 집단에만 매칭하지 않고 확률값을 준다.
    b. EM알고리즘으로 근사가 가능하다.
    c. 반복횟수가 K평균 에 비해 훨씬 많다.
    d. K평균은 가우시안 혼합분포에 대한 EM에 특정한 한계를 준 것과 같다.(한 class에만 배정)

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 표집법
    - 거부표집법
    - 중요도표집법
    - MCMC
    - 기브스 표집법
    - 조각표집법
- PCA