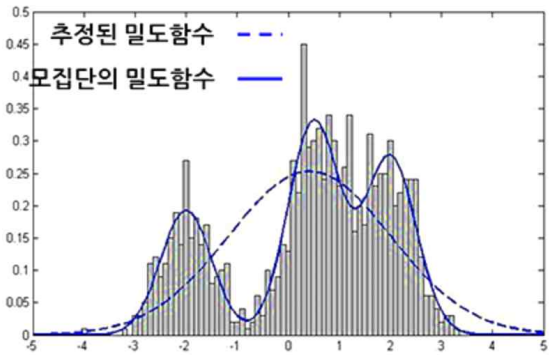
11강. 가우시안 혼합 모델

※ 점검하기

Q1. 다음 그림(교재 [그림 10-1])에서 나타난 것 같은 데이터 집합에 대하여 EM 알고리즘으로 가우시안 혼합 모델을 추정하는 프로그램을 직접 매트랩으로 구현해 본다. 이를 위해 다음과 같은 절차를 따르시오.



(1) 평균이 2.0이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 350개를 생성하시오.

(2) 평균이 0.5이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 400개를 생성하시오.

(3) 평균이 -2.0이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 250개를 생성하시오.

(4) (1)~(3)에서 생성된 데이터를 합쳐서 모두 1000개의 데이터를 가지는 집합 X를 만들고, 이 분포를 히스토그램으로 나타내 보시오.

(5) 1000개의 데이터 각각이 세 그룹 중 어디에 속하는지 알고 있다는 전제하에서, 각 그룹의 표본평균과 표본분산을 계산하고, 각 그룹이 전체 데이터에 대해서 차지하는 비율을 계산하여 3개의 성분을 가지는 가우시안 혼합 모델을 찾아보시오.

(6) (4)에서 찾아진 가우시안 혼합 모델 함수를 그래프로 그려서 (3)에서 그린 히스토그램과 형태가 일치하는지 확인하시오.

(7) 데이터들이 어떤 그룹에서 생성되었는지를 모른다는 전제하에, EM 알고리즘을 이용하여 적절한 가우시안 혼합 모델을 찾아보시오. 이때, 성분의 수는 3개로 하고, 파라미터의 초기값은 임의로 정한다.

(8) (7)에서 찾아진 가우시안 혼합 모델 함수를 그래프로 그려서 (4)의 히스토그램과 (5)에서 찾은 함수와 비교해 보시오.

해설

(1)~(3) 1차원 데이터 생성,

(4) 교재 3.3.1항(히스토그램법)

(5) 각 그룹에 대한 파라미터를 계산하여 가우시안 혼합 모델의 3개의 성분을 찾는다.

(6) [그림 10-1] 참고

(7)~(8) [프로그램 10-1]과 [그림 10-1] 참고

※ 정리하기

1. 가우시안 혼합 모델

1) 가우시안 혼합 모델을 사용하면 하나의 가우시안 분포함수로 나타낼 수 없었던 분포 특성뿐만 아니라,
아무리 복잡한 형태의 함수라도 충분한 개수의 가우시안 함수를 사용하면 원하는 만큼 정확하게 근사해 낼 수 있음

2) 가우시안 혼합 모델의 최우추정량은 각 식의 우변에 추정해야 하는 파라미터 추정치를 여전히 포함하고 있어
이 자체로서는 계산이 불가능함

→ 따라서 EM 알고리즘과 같은 반복 알고리즘을 통해 파라미터를 추정함

2. EM 알고리즘

1) 은닉변수에 대한 추정("E-step")과 파라미터의 추정("M-step")을 반복적으로 수행하는 알고리즘

① E-step : 파라미터 값이 주어졌다고 가정하고 은닉변수의 기대치를 계산함
- 초기에는 실제로 파라미터의 값이 주어지지 않았으므로 임의의 값으로 정함
- 이후에는 M-step을 수행하여 얻어지는 파라미터를 사용

② M-step : 은닉변수의 기대치를 이용하여 최우추정법에 따라 확률밀도함수의 로그우도를 최대화하는 파라미터를 추정

2) 기본 확률변수의 값이 관찰되면서 숨겨진 확률변수의 값도 함께 관찰된다면 EM과 같은 반복적인 학습 알고리즘에 의한 추정은 필요하지 않음

→ 즉 EM 알고리즘은 숨겨진 확률변수를 가지고 있는 확률 모델의 파라미터를 추정하기 위해 사용되는 방법

3) 가우시안 혼합 모델에 국한된 학습 알고리즘이 아니라 다양한 확률 모델에서 파라미터를 추정하기 위해 사용되는 일반적인 학습법

4) 관찰된 데이터로부터 정의되는 불완전 데이터 로그우도를 직접적으로 최대화하는 것이 불가능한 경우를 해결하기 위해
은닉변수를 도입하여 새로운 완전 데이터 로그우도를 정의함

→ 이것의 기대치로 정의되는 새로운 Q함수를 이용하여 파라미터를 최대화하는 알고리즘