3 장. 확률 분포 추정

오일석, 패턴인식, 교보문고, 2008.

들어가는 말

- 베이시언 분류에서의 학습은 사전 확률과 우도의 추정

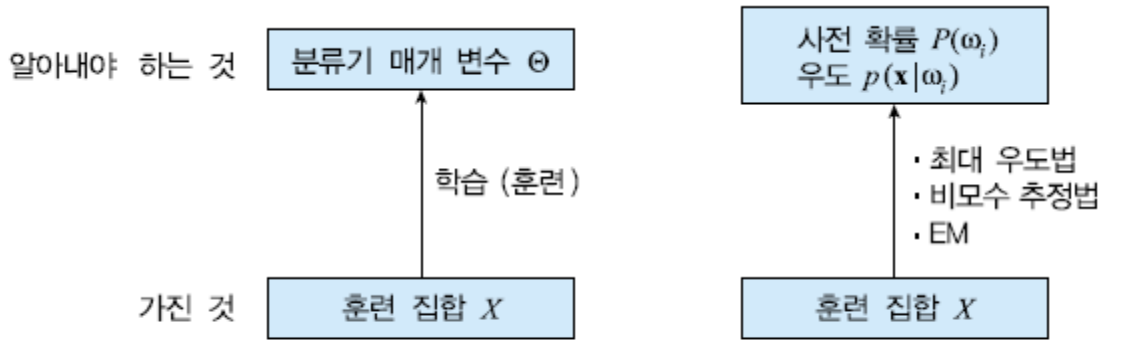


그림 3.1 일반적인 학습 과정과 베이시언 분류에서의 학습

들어가는 말

- 사전 확률 $P(\omega_i)$의 추정

$$P(\omega_i) = N_i / N \qquad (3.1)$$

  - N은 X의 크기이고 N_i는 ω_i에 속하는 샘플 수
  - N이 충분히 크면 (3.1)은 실제 값에 근접

- 우도 $P(\mathbf{x}|\omega_i)$ 추정

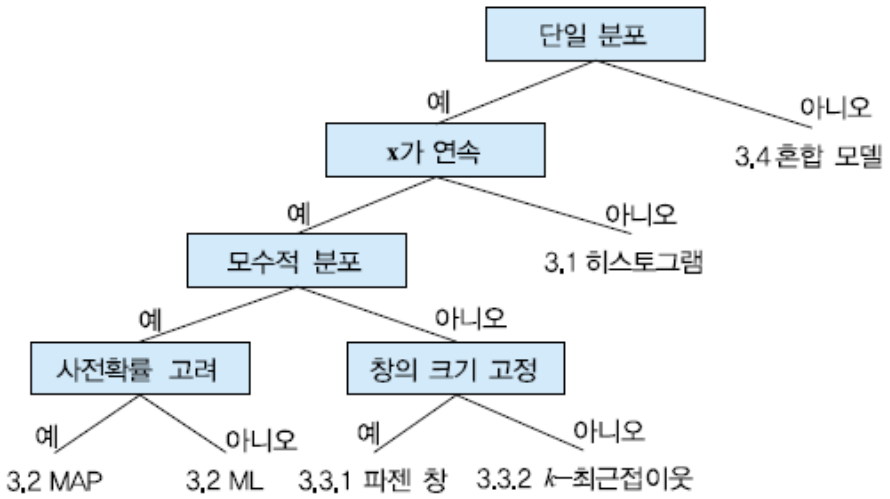
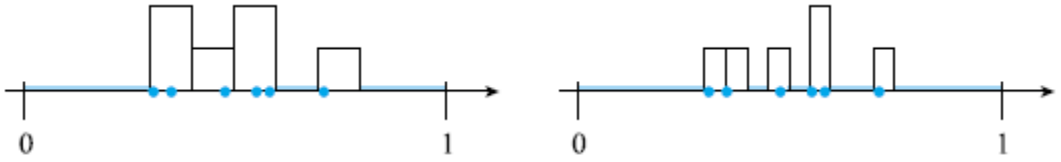


그림 3.2 여러 방법을 구별하는 트리

3.1 히스토그램

- 히스토그램



(a) 동적 구간을 10개 구간으로 나눔 (b) 동적 구간을 20개 구간으로 나눔

그림 3.3 1 차원에서 히스토그램 추정 사례

  - 총 s^d개의 빈이 발생 (각 차원을 s 개 구간으로 나눈다 했을 때)
    - 전형적인 차원의 저주
    - N은 충분히 크고 d는 작아야 함

3.2 최대 우도

- 문제 정의
  - "주어진 X를 발생시켰을 가능성이 가장 높은 매개 변수 Θ를 찾아라."
  - 주어진 X에 대해 가장 큰 우도를 갖는 Θ를 찾아라.
  - 아래 예에서
    - $P(X|\Theta_1) > P(X|\Theta_2)$
    - 최대 우도를 갖는 Θ는?

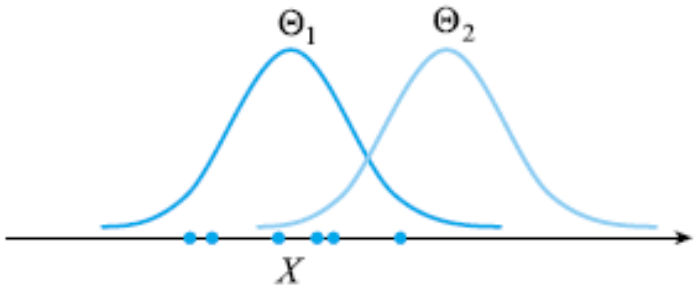


그림 3.4 최대 우도를 갖는 Θ를 찾는 문제

3.2 최대 우도

- 최대 우도ML 방법
  - 아래 최적화 문제를 풀어 답을 구하는 방법

$$\hat{\Theta} = \arg\max_{\Theta} p(X \mid \Theta) \tag{3.2}$$

$$p(X \mid \Theta) = p(\mathbf{x}_1 \mid \Theta) p(\mathbf{x}_2 \mid \Theta) \cdots p(\mathbf{x}_N \mid \Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \Theta) \tag{3.3}$$

- 로그 우도로 바꾸면

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{N} \ln p(\mathbf{x}_i \mid \Theta) \tag{3.4}$$

  - 미분을 이용한 최적화 문제 풀이
    - $L(\Theta)$의 도함수를 0으로 두고 풀어 구한 답이 $\hat{\Theta}$

$$\left.\begin{aligned} &\frac{\partial L(\Theta)}{\partial \Theta} = \mathbf{0} \\ &\text{이 때 } L(\Theta) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i \mid \Theta) \end{aligned}\right\} \tag{3.5}$$

3.2 최대 우도

- 예제 3.1: 정규 분포를 위한 최대 우도
  - ML에 의한 평균 벡터 μ의 추정 (공분산 행렬은 안다고 가정)

$$p(\mathbf{x}_i \mid \Theta) = p(\mathbf{x}_i \mid \boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)$$

$$\ln p(\mathbf{x}_i \mid \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}|$$

$$L(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - N\left(\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}|\right)$$

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

이제 $\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ 를 0으로 두고 식을 정리해 보자.

$$\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$$

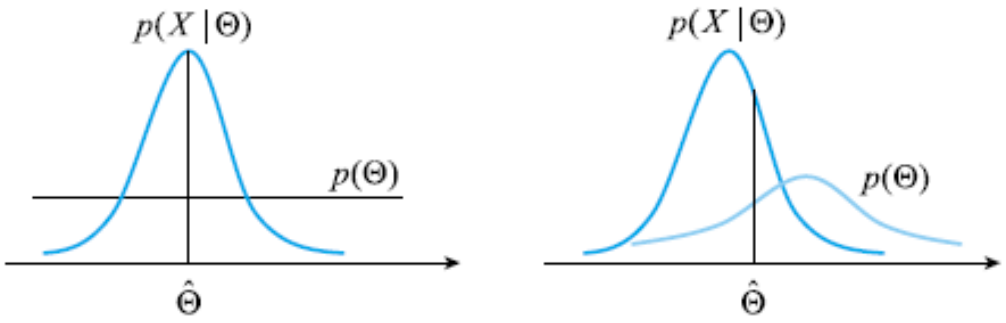
$$\sum_{i=1}^{N} \mathbf{x}_i - N\boldsymbol{\mu} = \mathbf{0}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \qquad (3.6)$$

3.2 최대 우도

- MAP 방법
  - $P(\Theta)$가 균일하지 않은 경우

$$\hat{\Theta} = \arg\max_{\Theta} p(\Theta) \sum_{i=1}^{N} \ln p(\mathbf{x}_i \mid \Theta) \tag{3.7}$$



(a) ML 방법 (b) MAP 방법

그림 3.5 ML과 MAP의 비교

Ronald Aylmer Fisher

(1890년 2월 17일 ~ 1962년 7월 29일) 영국

패턴 인식에서 Fisher가 등장하는 곳은 크게 두 군데다. 하나는 최대 우도이고 다른 하나는 Fisher의 선형 분별이다. 그가 정립한 최대 우도 이론을 역사적으로 조망한 흥미로운 논문이 있다 [Aldrich97]. 테스트용 데이터베이스로 널리 쓰이는 Iris 데이터도 그가 만들었다. 이런 이유로 그가 통계학자인 것으로 알려져 있는데 사실은 통계학뿐 아니라 유전학에도 큰 공헌을 하였다. '이기적 유전자'라는 The selfish gene 책으로 세계적인 주목을 받은 Richard Dawkins는 Fisher를 '다윈의 후계자중에 최고'라고 평할 정도이다. Fisher는 우생학에 eugenics 많은 관심을 가졌으며 Cambridge 대학 시절 우생학 그룹을 결성하여 Charles Darwin의 아들인 Horace Darwin과 같이 활동하기도 하였다. 그는 우생학을 유전학과 통계학의 접점으로 간주하였다. 그는 다윈의 자연 선택은 natural selection 알려진 것보다 강한 힘으로 작용한다는 연구 결과를 제시하기도 하였다. Fisher는 인도 통계 연구원을 방문하여 Mahalanobis와 교류하기도 하였으며 1957년에 Cambridge 대학을 은퇴한 후에는 호주 Adelaide 대학에서 말년을 보냈다. 그는 인종이 선천적으로 능력 차이를 지닌다는 주장을 펴기도 하였으며, 담배와 폐암의 상관 관계를 부정하는 주장을 펴기도 하였다. Fisher에 대한 보다 자세한 내용은 Adelaide 대학의 웹 [Adelaide(웹)] 또는 [Yates63]을 참고하라.

3.3 비모수적 방법

- 확률 분포 추정 방법
  - 모수적 방법
    - 확률 분포가 매개 변수 (모수)로 표현되는 형태
    - ML,MAP 방법등
  - 비모수적 방법
    - 확률 분포가 임의의 형태
    - 파젠 창, k-최근접 이웃 추정 방법 등

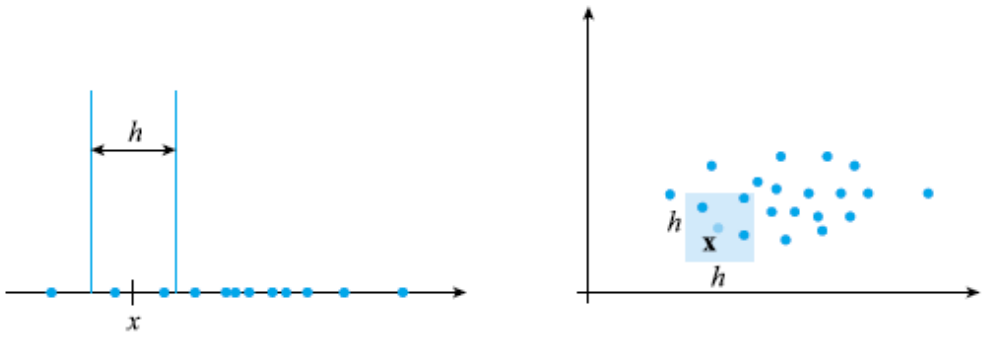
3.3.1 파젠 창

- 히스토그램 방법을 확장하여 확률 밀도 함수pdf 추정
  - 그림 3.6에서 임의의 점 x에서 확률 값 추정
  - 크기 h인 창을 씌우고 그 안의 샘플의 개수를 k라 하면,

$$p(x) = \frac{1}{h}\frac{k}{N}$$

  - d 차원으로 확대하면,

$$p(\mathbf{x}) = \frac{1}{h^d}\frac{k_{\mathbf{x}}}{N} \qquad (3.8)$$



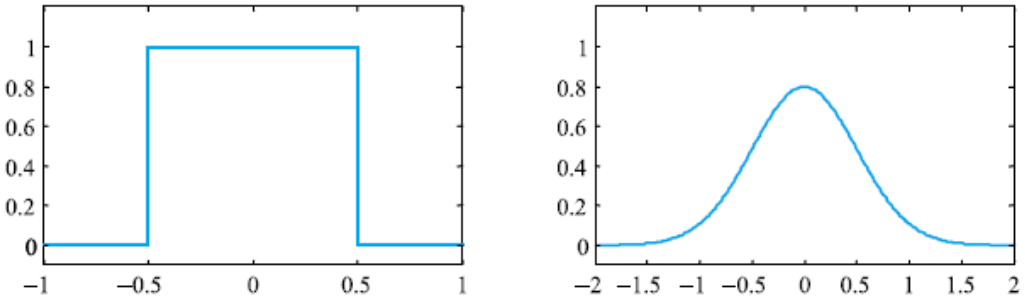
(a) 1 차원 특징 공간 (b) 2 차원 특징 공간

그림 3.6 파젠 창에 의한 확률 밀도 추정

3.3.1 파젠 창

- 여전히 매끄럽지 않은 함수
  - 예를 들어 그림 3.6(a)에서 x를 오른쪽 옮기면 계속 두 개이다가 어느 순간에 3으로 바뀜. 따라서 불연속인 pdf
- 매끄러운 pdf
  - 창 안의 샘플에 가중치를 준다. (중앙에 가까운 샘플이 더 높은 가중치)
  - 어떻게 이러한 아이디어를 구현할까?
- 커널 함수

$$\kappa(\mathbf{x}) = \begin{cases} 1, & |x_i| \le 0.5, 1 \le i \le d \\ 0, & \text{그 외} \end{cases} \qquad (3.9)$$



(a) 계단 함수 (b) 가우시언 함수

그림 3.7 파젠 창이 사용하는 대표적인 커널 함수

3.3.1 파젠 창

- 커널 함수를 사용하여 수식을 다시 쓰면,

$$k_{\mathbf{x}} = \sum_{i=1}^{N} \kappa\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \tag{3.10}$$

$$p(\mathbf{x}) = \frac{1}{h^d}\frac{1}{N}\sum_{i=1}^{N} \kappa\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \tag{3.11}$$

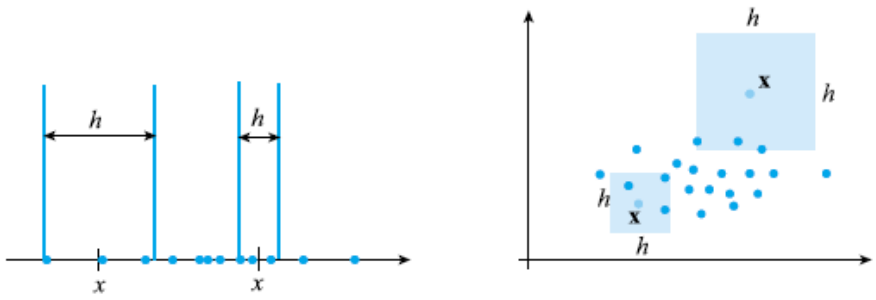
  - 커널 함수로 가우시언을 채택하면 매끄러운 pdf를 얻게 된다.

- 파젠 창의 특성
  - 차원의 저주에서 자유로운가?
  - 추정한 pdf가 실제에 가까우려면 N과 h는 어떻게 되어야 하나?

3.3.2 k-최근접 이웃 추정

- k-최근접 이웃 추정
  - $\mathbf{x}$를 중심으로 창을 씌우고 k 개 샘플이 안에 들어올 때까지 확장하고 그 순간의 창의 크기를 h라 한다. 즉 k가 고정되고 h가 가변이다.
  - 파젠 창에서는 h가 고정되고 k가 가변이다.

$$p(\mathbf{x}) = \frac{1}{h_{\mathbf{x}}{}^{d}} \frac{k}{N} \qquad (3.12)$$



(a) 1 차원 특징 공간 (b) 2 차원 특징 공간

그림 3.8 k-최근접 이웃 방법으로 확률 밀도 함수의 추정 (k=3일 때)

  - 시간 복잡도: $\Theta(kdN)$
    - 보로노이 도형으로 복잡도 줄일 수 있음

3.3.3 k-최근접 이웃 분류기

- k-NN 분류기
  - 확률 분포 추정이 아니라 분류기인데, k-NN 추정과 동작이 흡사하여 여기에서 설명함
  - $\mathbf{x}$를 중심으로 창을 씌우고, k 개 샘플이 안에 들어올 때까지 확장. 이때의 창의 크기를 $h_{\mathbf{x}}$라 하면 창의 부피는 $h_{\mathbf{x}}{}^{d}$
  - 창 안의 샘플 중에 ω_i에 속하는 것의 개수를 k_i라 하면,

$$p(\mathbf{x}\,|\,\omega_i) = \frac{k_i}{h_{\mathbf{x}}{}^{d} N_i} \tag{3.13}$$

$$p(\omega_i) = \frac{N_i}{N} \tag{3.14}$$

$$p(\mathbf{x}) = \frac{k}{h_{\mathbf{x}}{}^{d} N} \tag{3.15}$$

3.3.3 k-최근접 이웃 분류기

- 베이스 정리를 적용하면

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x})} = \frac{k_i}{k} \tag{3.16}$$

- k-NN 분류기

$$\left.\begin{array}{l} k-\text{NN 분류기: } \mathbf{x}\text{를 } \omega_q\text{로 분류하라.} \\ \text{이때 } q = \arg\max_i k_i \end{array}\right\} \tag{3.17}$$

알고리즘 [3.1] k-NN 분류기

입력: 훈련 집합 $X = \{(\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_N, t_N)\}$, 미지의 샘플 $\mathbf{x}$

출력: 부류 ω_q

알고리즘:

1. 훈련 샘플 중에 $\mathbf{x}$에 가장 가까운 k 개를 찾는다.
2. k 개가 속한 부류를 조사하여 가장 빈도가 높은 부류를 ω_q라 한다.

3.3.3 k-최근접 이웃 분류기

- k-NN 분류기의 오류율 특성

$N \to \infty$인 경우의 1-NN 분류기의 오류 확률 $E_{1\text{-NN}}$은 베이시언 분류기의 오류 확률 E_B에 대하여 (3.18)과 같은 범위에 있다.[7] 다시 말해 1-NN 분류기의 오류 확률은 베이시언 분류기의 오류 확률의 두 배를 넘지 않는다. 베이시언 분류기가 최적인 점을 감안하면 아주 바람직한 성능이라 할 수 있다. 물론 이론적으로 $N \to \infty$이어야 하며 실제적으로는 충분히 큰 N인 경우에 기대할 수 있는 오류 확률이다. $N \to \infty$라는 말은 (3.13) ~ (3.16)으로 추정한 확률이 실제 확률과 같아진다는 뜻이다. (3.19)는 $k > 1$인 경우의 오류 확률이다.

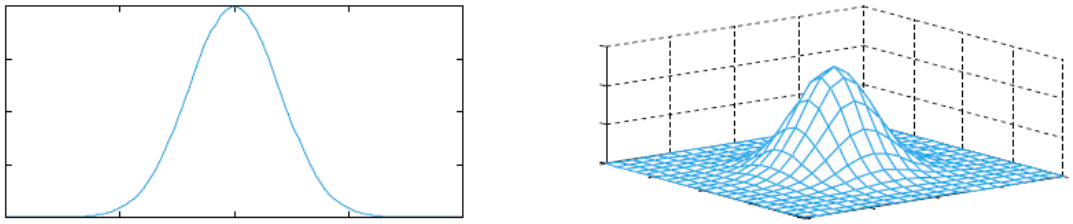
$$E_B \le E_{1\text{-NN}} \le 2E_B \tag{3.18}$$

$$E_B \le E_{k\text{-NN}} \le E_B + \sqrt{\frac{2E_{1\text{-NN}}}{k}} \tag{3.19}$$

3.4 혼합 모델

- 두 개 이상의 서로 다른 확률 분포의 혼합으로 X를 모델링함
  - 보통 요소 확률 분포로는 가우시언을 사용함

$$\left.\begin{aligned} N(\mu,\sigma^2) &= \frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ N(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \end{aligned}\right\} \quad (2.32)$$

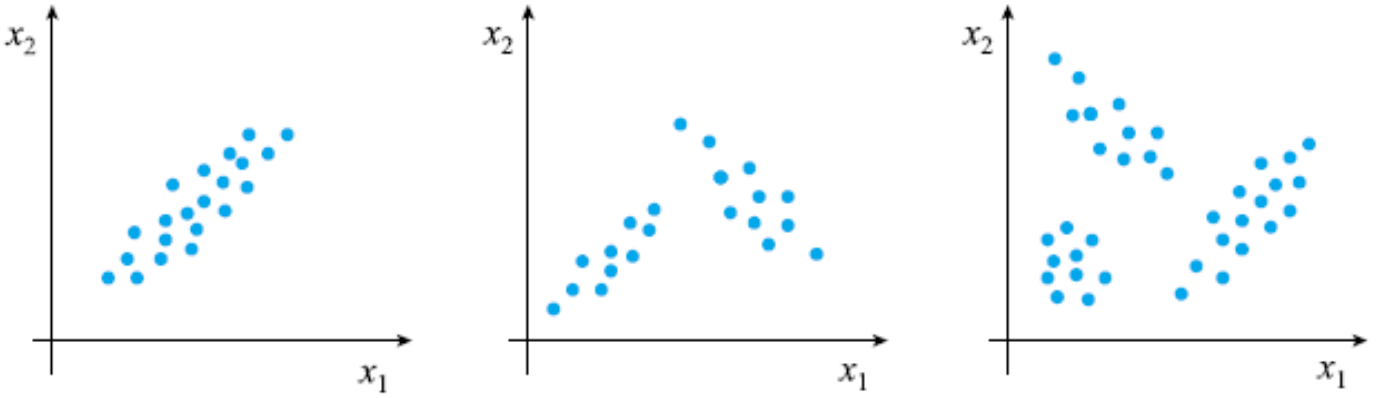


(a) 1 차원 정규 분포 (b) 2 차원 정규 분포

그림 2.9 정규 분포의 예

3.4.1 가우시언 혼합

- 다양한 분포들



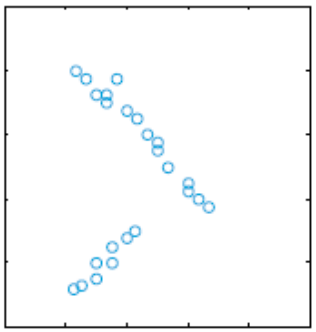
(a) 한 개의 모드 (b) 두 개의 모드 (c) 세 개의 모드

그림 3.10 다양한 분포들

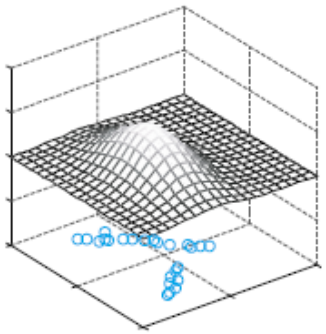
  - 어떻게 다중 모드 분포를 정확히 모델링 할 수 있을까?

3.4.1 가우시언 혼합

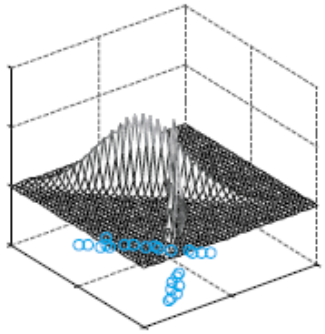
- 가우시언 혼합Gaussian mixture



(a) 샘플 분포



(b) 한 개의 가우시언 사용



(c) 가우시언 혼합
(두 개의 가우시언) 사용

그림 3.11 가우시언 모델링

- 추정해야 할 매개 변수

  가우시언의 개수 K

  k 번째 가우시언의 매개 변수 $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \cdots, K$

  k 번째 가우시언의 가중치 π_k, $k = 1, \cdots, K$

$$p(\mathbf{x}) = \frac{1}{3} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{2}{3} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

3.4.1 가우시언 혼합

- 최적화 문제로 공식화 해 보자.
  - 가우시언 혼합의 일반 공식

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (3.20)$$

    - π_k는 혼합 계수, $N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$는 요소 분포

  - 주어진 것과 추정해야 할 것

주어진 값 $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$

추정할 값 $\Theta = \{\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K), (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$

3.4.1 가우시언 혼합

- 최대 우도 문제로 공식화
  - Θ에 대한 $\mathbf{x}$의 우도와 로그 우도

$$p(X \mid \Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \Theta) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{3.21}$$

$$\ln p(X \mid \Theta) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{3.22}$$

  - X에 대해 최대 우도를 갖는 Θ를 찾는 문제

$$\hat{\Theta} = \arg\max_{\Theta} \ln p(X \mid \Theta) \tag{3.23}$$

    - 이 최적화 문제를 어떻게 풀 것인가?

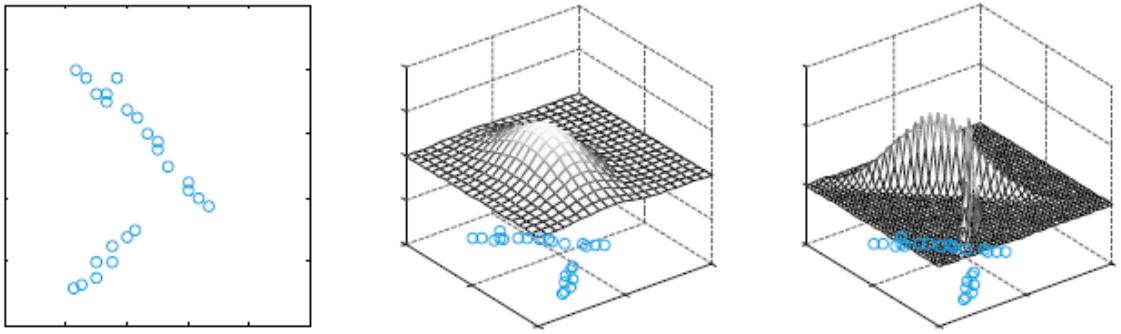
Johann Carl Friedrich Gauss

(1777년 4월 30일 ~ 1855년 2월 23일) 독일

Gauss는 독일의 수학자이다. 그의 모토는 '*pauca sed matura* (*few, but ripe*)'였다고 전해 진다. 이를 증명하듯 그는 지독할 정도의 완벽주의자였고 발표한 논문과 책이 많은 편은 아닌데 하나하나 수학과 과학에 커다란 영향을 미치는 것들이다. Gauss는 여섯 자녀를 두었는데 '가문의 이름을 더럽힐까 봐' 그들이 과학이나 수학을 전공하는 것을 극구 반대했다고 한다. 그의 이름이 형용사화되어 있는 가우시언 분포(정규 분포라고도 부름)는 Hanover 지역의 측량 작업에서 발생하는 오차를 표현하기 위해 개발하였다. 가우시언 분포는 특징 벡터의 확률 분포를 표현하는데 많이 사용한다. 그는 수학뿐 아니라 지리학, 천문학, 그리고 전자기학에도 큰 족적을 남겼다. 전자기학에서는 그의 공헌을 기리기 위해 그의 이름을 딴 가우스라는 단위를 사용하고 있다. Gauss는 신동이었다고 한다. 초등학교에서 Büttner 선생님이 학생들을 집중시키기 위해 $1 + 2 + \cdots + 100$을 시켰는데 Gauss가 선생님을 깜짝 놀래 켰다고 한다. 모든 학생이 선생님의 바램대로 끙끙거리고 있는데 그는 5050이라는 정답을 순식간에 내놓았다고 한다. $1 + 100 = 101$, $2 + 99 = 101$, $3 + 98 = 101$, … 이런 식으로 101이 50개 있으니 $101 \times 50 = 5050$이라는 규칙을 발견한 것이다. Gauss에 대한 보다 상세한 내용은 그의 탄생 150년을 기리는 뜻에서 쓴 논문을 참고하기 바란다 [Dunnington27].

3.4.2 EM 알고리즘

- 문제에 대한 통찰 (예제 3.1과의 비교)
  - 예제 3.1은 한 쌍의 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$를 추정 → 미분 한번 적용으로 해결
  - 지금은 K 개의 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$ 그리고 그들의 혼합을 위한 혼합 계수 $\boldsymbol{\pi}$를 추정
  - 게다가 샘플이 어느 가우시언에 속하는지에 정보가 없음 (손실 정보)



(a) 샘플 분포 (b) 한 개의 가우시언 사용 (c) 가우시언 혼합 (두 개의 가우시언) 사용

그림 3.11 가우시언 모델링

3.4.2 EM 알고리즘

- 새로운 알고리즘
  - 두 단계를 반복
    - 샘플이 어느 가우시언에 속하는지 결정 (연성 소속soft membership)
    - 매개 변수 추정 $\Theta = \{\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K), (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$

알고리즘 [3.2] 가우시언 혼합 추정을 위한 EM 알고리즘의 골격

1. 매개 변수 집합 Θ를 초기화 한다.
2. **repeat** {
3. E 단계: Θ를 이용하여, 샘플 별로 K 개의 가우시언에 속할 확률을 추정한다.
4. M 단계: E 단계에서 구한 소속 확률을 이용하여 Θ를 추정한다.
5. } **until** (멈춤 조건이 만족);

3.4.2 EM 알고리즘

- EM 알고리즘의 구체화
  - 샘플의 가우시언 소속을 어떻게 표현할 것인가?
    - $\mathbf{z}=(z_1, z_2, \ldots, z_K)^T$로 표현 (이런 종류의 변수를 은닉 변수라latent variable 부름) 샘플이 j 번째 가우시언에서 발생했다면 z_j=1이고 나머지는 0
  - j 번째 가우시언에서 샘플 $\mathbf{x}_i$가 발생할 확률 ('우도'로 간주할 수 있음)

$$p(\mathbf{x}_i \mid z_j = 1) = N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \qquad (3.24)$$

  - 샘플 $\mathbf{x}_i$가 관찰되었는데 그것이 j 번째 가우시언에서 발생했을 확률 ('사후 확률'로 간주할 수 있음)

$$\left.\begin{aligned} P(z_j = 1 \mid \mathbf{x}_i) &= \frac{P(z_j = 1)p(\mathbf{x}_i \mid z_j = 1)}{p(\mathbf{x}_i)} \\ &= \frac{P(z_j = 1)p(\mathbf{x}_i \mid z_j = 1)}{\sum_{k=1}^{K} P(z_k = 1)p(\mathbf{x}_i \mid z_k = 1)} \\ &= \frac{\pi_j N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \end{aligned}\right\} \qquad (3.25)$$

3.4.2 EM 알고리즘

- EM 알고리즘의 구체화
  - (3.23)의 최적화 문제를 풀기 위해, (3.22)의 $\ln P(X|\Theta)$을 미분하여 얻은 도함수를 0으로 두고 그것의 해를 구한다.

$$\ln p(X \mid \Theta) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \qquad (3.22)$$

$$\hat{\Theta} = \arg\max_{\Theta} \ln p(X \mid \Theta) \qquad (3.23)$$

  - 먼저 $\boldsymbol{\mu}_j$에 대해 풀면,

$$\boldsymbol{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i) \mathbf{x}_i \qquad (3.26)$$

$$N_j = \sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i) \qquad (3.27)$$

    - N_j는 'j 번째 가우시언에 소속된' 샘플의 개수로 해석할 수 있음
    - $\boldsymbol{\mu}_j$는 j 번째 가우시언에 소속된 샘플의 가중치 평균으로 해석

3.4.2 EM 알고리즘

- EM 알고리즘의 구체화
  - $\boldsymbol{\Sigma}_j$에 대해 풀면,

$$\boldsymbol{\Sigma}_j = \frac{1}{N_j}\sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \qquad (3.28)$$

    - $\boldsymbol{\Sigma}_j$는 j 번째 가우시언에 소속된 샘플의 가중치 공분산 행렬로 해석 가능

  - 혼합 계수 π_j에 대해 풀면,
    - 조건부 최적화 문제이므로 라그랑제 승수를 도입하여 해결

$$\pi_j = \frac{N_j}{N} \qquad (3.29)$$

3.4.2 EM 알고리즘

알고리즘 [3.3] 가우시언 혼합 추정을 위한 EM 알고리즘

입력: 훈련 집합 X, 가우시언 개수 K

출력: $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $1 \le j \le K$, 그리고 $\boldsymbol{\pi}$

알고리즘:

1. $\boldsymbol{\mu}_j$와 $\boldsymbol{\Sigma}_j$, $1 \le j \le K$, 그리고 $\boldsymbol{\pi}$를 초기화 한다.
2. **repeat** {
   // E 단계 (샘플의 가우시언 소속 확률 추정)
3. **for** ($i = 1$ **to** N)
4. **for** ($j = 1$ **to** K)
5. $$P(z_j = 1 \mid \mathbf{x}_i) = \frac{\pi_j N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \quad // \ (3.25)$$
   // M 단계 (Θ 추정)
6. **for** ($j = 1$ **to** K) {
7. $$N_j = \sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i); \quad // \ (3.27)$$
8. $$\boldsymbol{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i) \mathbf{x}_i; \quad // \ (3.26)$$
9. $$\boldsymbol{\Sigma}_j = \frac{1}{N_j} \sum_{i=1}^{N} P(z_j = 1 \mid \mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}; \quad // \ (3.28)$$
10. $$\pi_j = \frac{N_j}{N}; \quad // \ (3.29)$$
   }
11. } **until** (멈춤 조건 만족);

3.4.2 EM 알고리즘

- EM 알고리즘에 대한 부연 설명
  - 군집화를 위한 *k*-means 알고리즘은 EM의 일종이다.
  - EM은 속도가 느리다.
  - 멈춤 조건은?
  - EM은 최적 해로 수렴함 (욕심 알고리즘이므로 전역 최적 해 보장 못함)
  - EM은 불완전 데이터에 대한 최대 우도 추정법으로 간주할 수 있다.