

03. 확률 분포 추정

분류기 매개변수 θ

학습 (훈련)

훈련 집합 X

사전 확률 $P(W_i)$

우도 $P(x|W_i)$

최대우도법, 비오누 추정법

훈련 집합 X

일반적인 학습과정

비이시언 분류에서의 학습 과정

- 사전 확률 $P(W_i)$ 의 추정

$$P(W_i) = N_i / N$$

N 은 X 의 크기 (샘플의 수), N_i 는 W_i 에 속하는 샘플의 수
 N_i 가 충분히 크면 실제값에 근접

- 우도 $P(x|W_i)$ 추정

우도가 정규분포 같은 분포를 한다고 알고 있던지, 어떤 다른 이유로

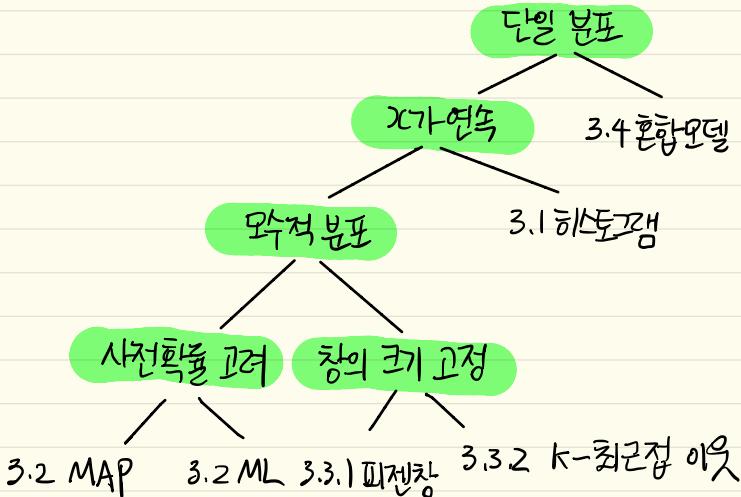
그러한 분포를 가정할 수 있다면 문제는 크게 쉬워짐. 이제 확률

분포를 위한 함수 추정 문제가 정규분포의 매개변수 (평균벡터와 공분산 행렬) 추정 문제로 줄어듬.

but. 임의의 모양을 가진 확률 분포를 하는 경우 \rightarrow 히스토그램
추정법

다른 부류에 속한 샘플은 서로 영향 미치지 X

따라서 우도는 독립적으로 수행할 수 있다.



3.1 히스토그램 추정

히스토그램: dynamic range (변수가 갖는 값의 범위) 가 $[0, 1]$ 이라 하고, 이 범위를 각각 10개와 20개 구간으로 나누고 구간 별로 그 안의 샘플의 수를 세어 만든 것. 하나의 구간을 bin이라 부른다.

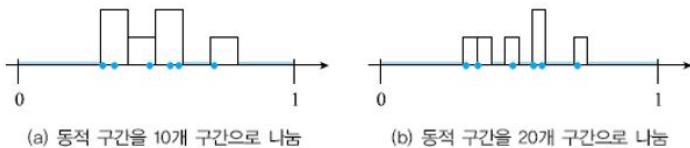


그림 3.3 1 차원에서 히스토그램 추정 사례

장점: 표현과 연산이 단순하면서 어떤 분포의 특성을 잘 표현하므로 유용

단점: 상황에 따라 그 쓰임새가 크게 제한
차원의 저주가 생김

특정 벡터가 차원이라 하고 각 차원을 S 개 구간으로 나누다면 총 S^d 개의 빈이 생긴다. 그러므로 공간의 차원이 낮아야 한다.

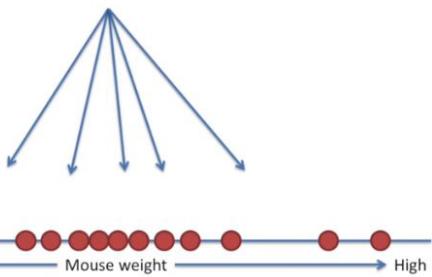
또, 샘플의 크기는 충분히 커야한다. 예를 들어, X 의 크기가 6이고, 구간이 1000개이면, 6개의 구간만 $1/6$ 의 확률값을 가지고, 나머지는 모두 0을 가지게 된다. 별 의미 없는 히스토그램이 된다.

3.2 최대우도

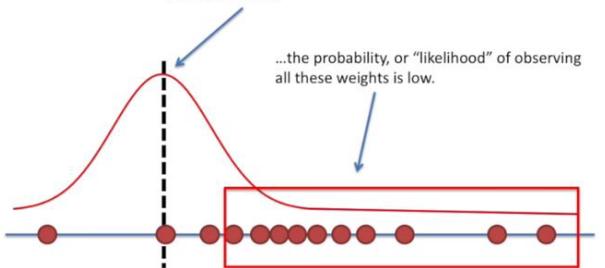
현실적으로 정규분포와 같이 매개 변수로 표현되는 경우만 적용이 가능하다. 그런 경우만 실제 계산이 가능하기 때문

- 문제 정의
- "주어진 X 를 발생시켰을 가능성이 가장 높은 매개 변수 θ 를 찾으라."
- 주어진 X 에 대해서 가장 큰 우도를 갖는 θ 를 찾으라

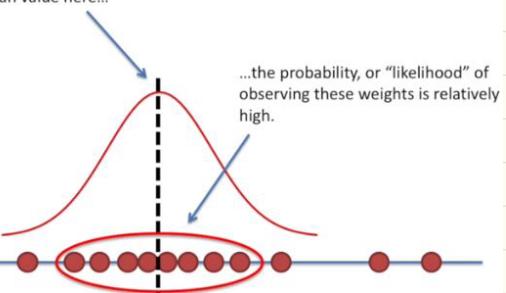
Let's say we weighed a bunch of mice...



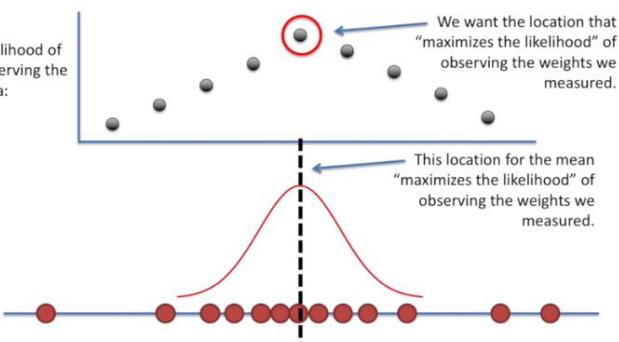
According to a normal distribution with a mean value over here...



According to a normal distribution with a mean value here...



Likelihood of observing the data:



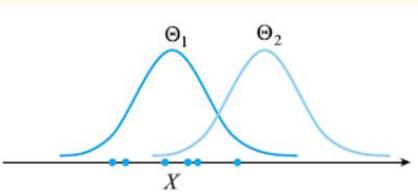


그림 3.4 최대 우도를 갖는 Θ 를 찾는 문제

→ X에서의 최대 우도를 갖는
 Θ 는 Θ_1

최대 우도 법칙 (maximum likelihood) ML

$$\hat{\Theta} = \arg_{\Theta} \max P(X|\Theta)$$

x가 연속값을 가정하여?

$$P(X|\Theta) = P(x_1|\Theta) P(x_2|\Theta) \cdots P(x_N|\Theta) = \prod_{i=1}^N P(x_i|\Theta)$$

$f(\cdot)$ 가 단조 증가 함수라면 $P(X|\Theta)$ 를 최대화하는 것과 $f(P(X|\Theta))$ 를 최대화 하는 것은 같은 답을 만들어 준다.

$$\hat{\Theta} = \arg_{\Theta} \max \sum_{i=1}^N \ln P(x_i|\Theta)$$

이는 최적화 문제 (Optimization Problem) 이다.

$$\frac{\partial L(\Theta)}{\partial \Theta} = 0, \text{ 이 때, } L(\Theta) = \sum_{i=1}^N \ln P(x_i|\Theta)$$

예제 3.1: 정규 분포를 위한 최대 우도

- ML에 의한 평균 벡터 μ 의 추정 (공분산 행렬은 있다고 가정)

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

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

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

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

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

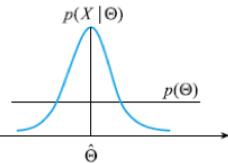
$$\begin{aligned}\sum_{i=1}^N \Sigma^{-1} (\mathbf{x}_i - \mu) &= 0 \\ \sum_{i=1}^N \mathbf{x}_i - N\mu &= 0\end{aligned}$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \tag{3.6}$$

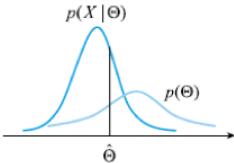
3.2 최대우도

$-P(\theta)$ 가 균일하지 않은 경우

$$\hat{\theta} = \arg \max_{\theta} P(\theta) \stackrel{?}{=} \ln P(X_i | \theta)$$



(a) ML 방법



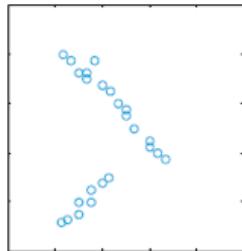
(b) MAP 방법

그림 3.5 ML과 MAP의 비교

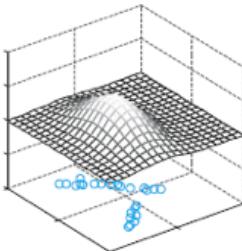
3.4.2 EM 알고리즘

• 예제 3.1과의 비교

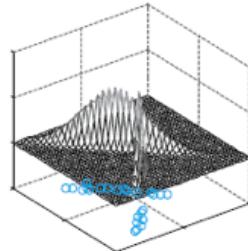
- 예제 3.1은 한 쌍의 M 와 \leq 을 추정 \rightarrow 미분 한번 적용으로 해결
- 지금은 k 개의 M 와 \leq , 그리고 그들의 혼합을 위한 혼합계수 π 를 추정
- 게다가 샘플이 어느 가우시안에 속하는지 정보가 없음



(a) 샘플 분포



(b) 한 개의 가우시안 사용



(c) 가우시안 혼합
(두 개의 가우시안) 사용

그림 3.11 가우시안 모델링

• 새로운 알고리즘

- 두 단계를 반복

- 샘플이 어느 가우시안에 속하는지 결정 (연성 소속, soft membership)
- 매개변수 추정 $\Theta = \{\pi = (\pi_1, \dots, \pi_K), (M_1, \leq_1), \dots, (M_K, \leq_K)\}$

알고리즘 [3.2]

가우시안 혼합 추정을 위한 EM 알고리즘의 골격

1. 매개 변수 집합 Θ 를 초기화 한다.

2. **repeat** {

3. E 단계: Θ 를 이용하여, 샘플 별로 K 개의 가우시안에 속할 확률을 추정한다.

4. M 단계: E 단계에서 구한 소속 확률을 이용하여 Θ 를 추정한다.

5. } **until** (멈춤 조건이 만족);



• EM 알고리즘의 구체화

- 샘플의 가우시언 소속을 어떻게 표현할 것인가?

- $\mathbf{z} = (z_1, z_2, \dots, z_K)^T$ 로 표현 (이런 종류의 변수를 **隱적 변수, latent variable**라고 부름), 샘플이 j 번째 가우시언에서 발생했다면 $z_j=1$ 이고 나머지는 0

• j 번째 가우시언에서 샘플 x_i 가 발생한 확률 ('우도'로 간주할 수 있음)

• 샘플 x_i 가 관찰되었는데 그것이 j 번째 가우시언에서 발생했을 확률 ('사후 확률'로 간주할 수 있음)

$$P(z_j=1 | x_i) = \frac{P(z_j=1) P(x_i | z_j=1)}{P(x_i)}$$

$$= \frac{P(z_j=1) P(x_i | z_j=1)}{\sum_{k=1}^K P(z_k=1) P(x_i | z_k=1)}$$

$$= \frac{\prod_j N(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \prod_k N(x_i | \mu_k, \Sigma_k)}$$

각 N 개의 샘플들의 K 개의 가우시언 소속 확률을 추정하였다.

이제 우리가 풀어야 하는 문제는 함수 $\ln P(X|\theta)$ 를 최대로 하는 매개 변수 θ 의 값을 찾는 것이다. θ 는 μ 와 Σ , 그리고 혼합계수 벡터 π 로 구성되어 있었다. 함수 $\ln P(X|\theta)$ 는 최대 점에서 μ_j 로 미분한 값이 0이 되어야 한다.

$$\begin{aligned}
& 1) \mu_j \text{에 대해 풀면, } \\
& \frac{\partial \ln p(x|\theta)}{\partial \mu_j} = \frac{\partial \left(\sum_{i=1}^N \ln \sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)}{\partial \mu_j} \\
& = \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \cdot \frac{\partial \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)}{\partial \mu_j} \\
& = \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \cdot \frac{\partial (\pi_j N(x_i | \mu_j, \Sigma_j))}{\partial \mu_j} \\
& = \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \cdot \frac{\pi_j \frac{1}{2} \times \Sigma_j^{-1} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}{\partial \mu_j} \\
& = \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \cdot (\pi_j N(x_i | \mu_j, \Sigma_j)) (\Sigma_j^{-1} (x_i - \mu_j)) \\
& = \sum_{i=1}^N \left(\frac{(\pi_j N(x_i | \mu_j, \Sigma_j))}{\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k)} \right) (\Sigma_j^{-1} (x_i - \mu_j)) \\
& = \sum_{i=1}^N P(z_j=1 | x_i) (\Sigma_j^{-1} (x_i - \mu_j))
\end{aligned}$$

$$\begin{aligned}
& \sum_{i=1}^N P(z_j=1 | x_i) (\Sigma_j^{-1} (x_i - \mu_j)) = 0 \text{이 되어야 하므로,} \\
& \sum_{i=1}^N P(z_j=1 | x_i) (x_i - \mu_j) = 0 \\
& \sum_{i=1}^N P(z_j=1 | x_i) x_i - \sum_{i=1}^N P(z_j=1 | x_i) \mu_j = 0
\end{aligned}$$

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^N P(z_j=1 | x_i) x_i \quad (3.26)$$

가중치

$$N_j = \sum_{i=1}^N P(z_j=1 | x_i) \quad (3.27)$$

(3.27)의 N_j 는 훈련집합에 있는 N 개 샘플 각각에 대해
 j 번째 가우시안에 속한 확률 $P(z_j=1 | x_i)$ 를 구하고 이들을
모두 더한 값이므로 'j번째 가우시안에 소속된 샘플의 개수'
 μ_j 는 모든 샘플의 평균 벡터, 즉, j번째 가우시안에 소속된
샘플의 평균벡터.

2) Ξ_j 에 대해 풀면,

위와 비슷한 과정을 거치면,

$$\Xi_j = \frac{1}{N_j} \sum_{i=1}^{N_j} P(z_j=1 | x_i) (x_i - \mu_j)(x_i - \mu_j)^T$$

3) π_i 에 대해 풀면,

π_i 는 단순히 미분하여 0으로 두고 해를 구하면 안됨.

$0 \leq \pi_k \leq 1$ 과 $\sum_{k=1}^K \pi_k = 1$ 의 두 가지 조건을 만족해야 함.

조건부 최적화(Constrained optimization) 문제!!

$$\ln p(x|\theta) + \alpha \left(\sum_{k=1}^K \pi_k - 1 \right)$$

$$\frac{\partial (\ln P(X|\theta) + \alpha \left(\sum_{k=1}^K \pi_k - 1 \right))}{\partial \pi_j}$$

$$= \frac{\partial \left(\sum_{i=1}^N \left(\ln \sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right) + \alpha \left(\sum_{k=1}^K \pi_k - 1 \right) \right)}{\partial \pi_j}$$

$$= \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \frac{\partial \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)}{\partial \pi_j} + \frac{\partial (\alpha(\pi_1 + \dots + \pi_j + \dots + \pi_K - 1))}{\partial \pi_j}$$

$$= \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \cdot x$$

$$\frac{\partial (\pi_1 N(x_i | \mu_1, \Sigma_1) + \dots + \pi_j N(x_i | \mu_j, \Sigma_j) + \dots)}{\partial \pi_j} + \alpha$$

$$= \sum_{i=1}^N \left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)^{-1} \frac{\partial (\pi_j N(x_i | \mu_j, \Sigma_j))}{\partial \pi_j} + \alpha$$

$$= \sum_{i=1}^N \frac{N(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k)} + \alpha$$

$$\sum_{i=1}^N \frac{\pi_j N(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k)} + \pi_j \alpha = 0$$

$$\sum_{i=1}^N P(z_j = 1 | x_i) + \pi_j \alpha = 0$$

$$N_j + \pi_j \alpha = 0$$

$$\pi_j = -\frac{N_j}{\alpha}$$

이에 $\sum_{k=1}^K \pi_k = 1$ 조건을 적용하면,

$$\pi_1 + \dots + \pi_K = 1 = \frac{-N_1 - \dots - N_K}{\alpha} = \frac{-N}{\alpha}$$

$$\alpha = -N$$

$$\therefore \pi_j = \frac{N_j}{N}$$

그 경우 시언에 소속된 샘플의 개수를 전체 샘플의 개수로 나누어 구하라는 것이다.

EM 알고리즘을 구하기 위한 재료를 모두 준비한 셈이다.

입력: 훈련 집합 X , 가우시안 개수 K

출력: (μ_j, Σ_j) , $1 \leq j \leq K$, 그리고 π

알고리즘:

1. μ_j 와 Σ_j , $1 \leq j \leq K$, 그리고 π 를 초기화 한다.

2. repeat {

// E 단계 (샘플의 가우시안 소속 확률 추정)

3. for $(i = 1 \text{ to } N)$

4. for $(j = 1 \text{ to } K)$

$$P(z_j = 1 | \mathbf{x}_i) = \frac{\pi_j N(\mathbf{x}_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \pi_k N(\mathbf{x}_i | \mu_k, \Sigma_k)} \quad // (3.25)$$

// M 단계 (Θ 추정)

6. for $(j = 1 \text{ to } K)$ {

$$7. N_j = \sum_{i=1}^N P(z_j = 1 | \mathbf{x}_i); \quad // (3.27)$$

$$8. \mu_j = \frac{1}{N_j} \sum_{i=1}^N P(z_j = 1 | \mathbf{x}_i) \mathbf{x}_i; \quad // (3.26)$$

$$9. \Sigma_j = \frac{1}{N_j} \sum_{i=1}^N P(z_j = 1 | \mathbf{x}_i) (\mathbf{x}_i - \mu_j)(\mathbf{x}_i - \mu_j)^T; \quad // (3.28)$$

$$10. \pi_j = \frac{N_j}{N}; \quad // (3.29)$$

}

11. } until (멈춤 조건 만족);



* EM 알고리즘에 대한 부연 설명

1. 군집화(Clustering)을 위한 k-평균(k-means) 알고리즘은 EM 알고리즘의 일종이다. 첫 단계는 샘플 각각을 가장 가까운 프로토타입에 할당한다. (E단계). 다음 단계에서는 각 프로토타입을 자신에게 배정된 샘플의 평균으로 대체한다. (M단계) 이 두 단계를 수렴할 때까지 반복한다.
2. EM 알고리즘은 속도가 느리다. 그래서 k-평균 알고리즘을 먼저 수행시키고, 그것으로 찾은 해를 초기값으로 삼아 수렴 속도를 빠르게 한다.
3. 멈춤 조건은? M단계 후, 그 우도 $\ln p(X|\theta)$ 가 이전 것보다 좋아지지 않을 때, 임계값 설정하여 차이가 임계값보다 작을 때
4. EM은 최적해로 수렴함(욕심 알고리즘이므로 전역 최적 해는 보장 못함)
5. EM은 불완전한 데이터가 주어진 경우를 위한 최대 우도(ML) 추정법의 일종