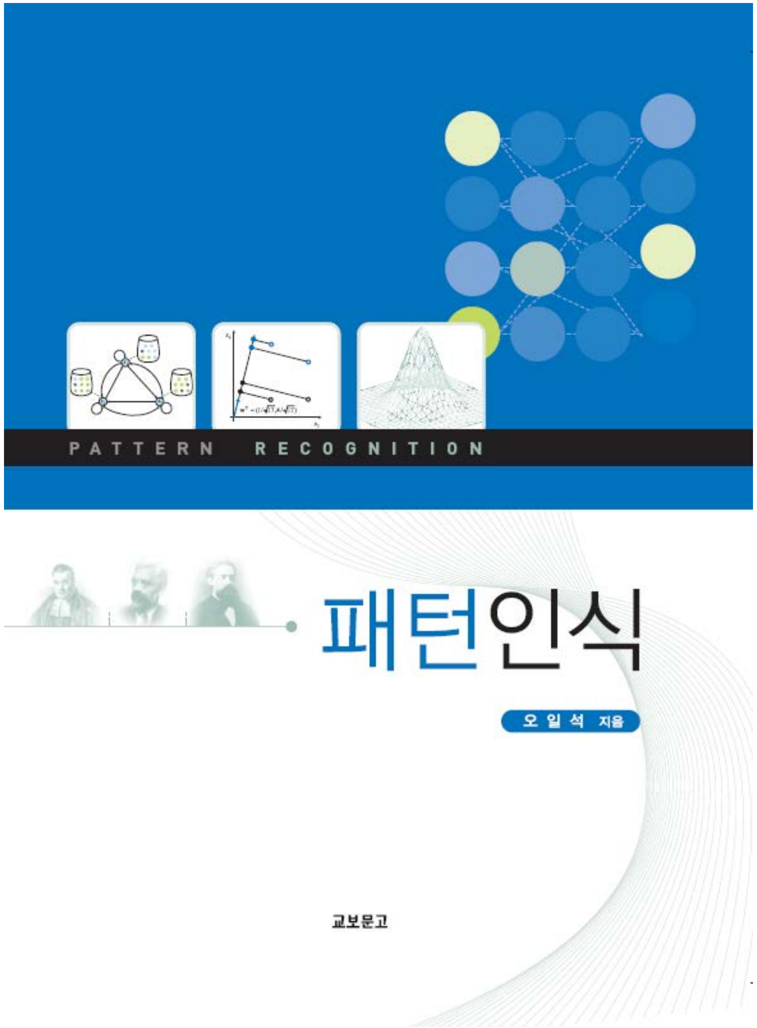
2 장. 베이시언 결정 이론

오일석, 패턴인식, 교보문고, 2008.



가장 그럴듯한 이라는 보편 법칙

- 보편적인 인식 법칙
  - '가장 그럴듯한' 부류로 분류
    - 라디오 목소리가 배철수인지 배칠수인지 헛갈리는데 배칠수 같다.
    - 도로 표지판이 전주인지 진주인지 확실치 않은데 전주인것 같다.
- 기계 (컴퓨터)의 인식
  - 수학 틀에 넣어야 프로그래밍이 가능해짐

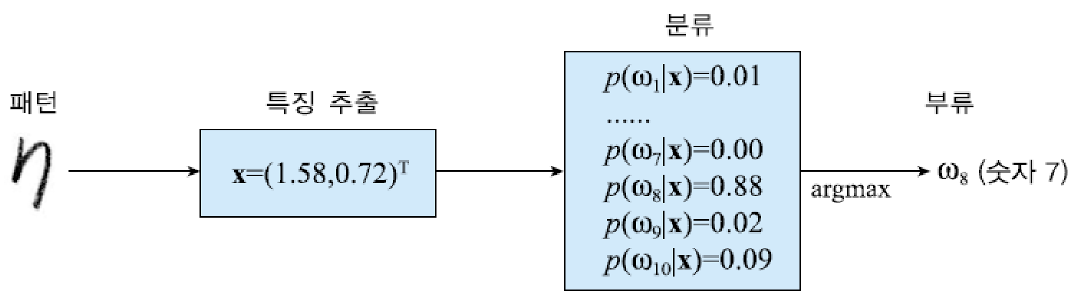


그림 2.1 '가장 그럴듯한' 이라는 철학을 수학 틀에 넣음

  - $P(\omega_i|\mathbf{x})$: $\mathbf{x}$가 주어졌을때 그것이 부류 ω_i에서 발생했을 확률 (사후 확률)

어렵고 중요한 문제

- 사후 확률 $P(\omega_i|\mathbf{x})$의 추정
  - 어려운가? (그림 1.6을 가지고 생각해 보자.)
  - 왜?
  - 어떻게 추정하나?
  - 2~3장의 핵심 주제

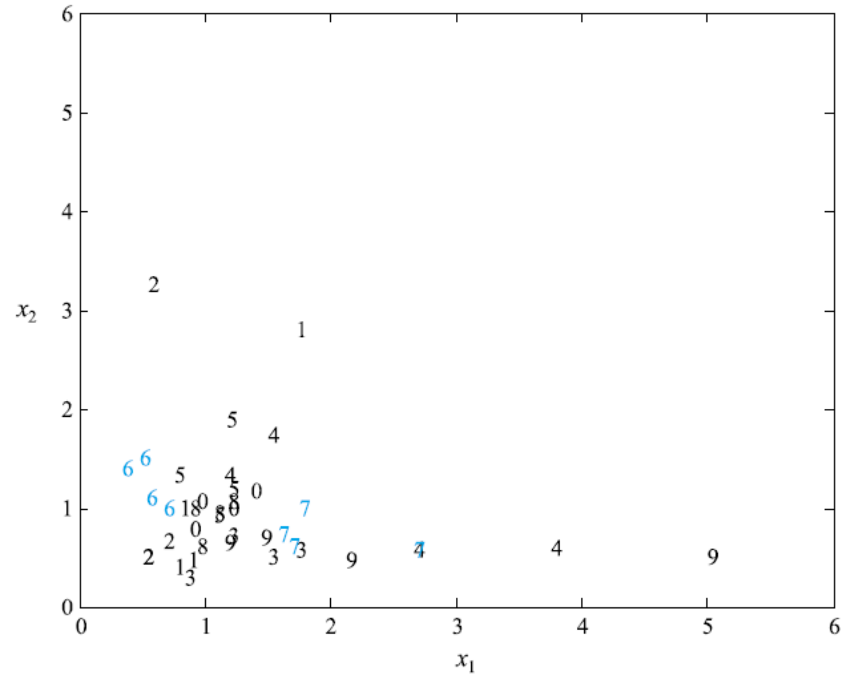


그림 1.6 특징 공간에서 샘플의 분포

2.1.1 확률 기초

- 주사위
  - 주사위 던졌을 때 3이 나올 확률 $P(X=3)=1/6$
  - X를 랜덤 변수라 부름
  - 이 경우 X는 이산 값을 가짐
- 사람 키
  - 연속 값
  - 확률 밀도 함수 $p(X)$

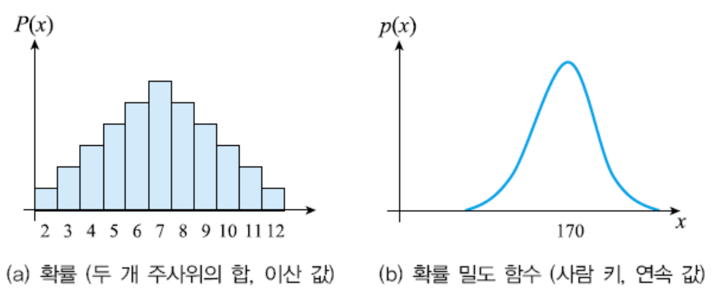


그림 2.2 확률 분포 함수

- 패턴 인식에서 특징 각각이 랜덤 변수에 해당

2.1.1 확률 기초

- 확률 실험 (사전 확률, 우도, 사후 확률을 설명할 목적의 시나리오)

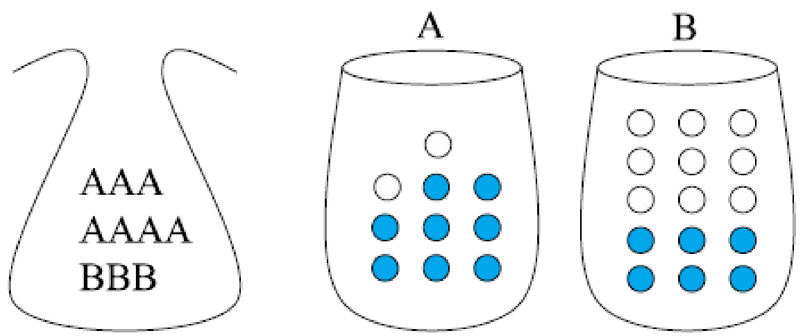


그림 2.3 확률 실험 장치

  - 주머니에서 카드를 뽑아 상자를 선택하고 선택된 상자에서 공을 뽑아 관찰
  - 랜덤 변수 $X \in \{A,B\}$, $Y=\{$파랑, 하양$\}$

2.1.1 확률 기초

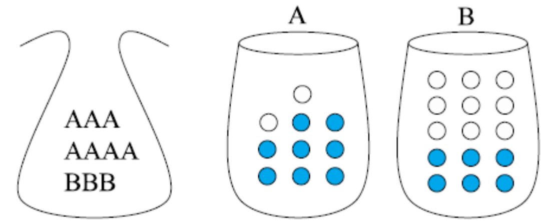


그림 2.3 확률 실험 장치

- 확률
  - 상자 A가 선택될 확률은?
    - $P(X\text{=A})=P(\text{A})=7/10$
  - 상자 A에서 하얀 공이 뽑힐 확률은?
    - 조건부 확률 $P(Y\text{=하양}|X\text{=A})=P(\text{하양}|\text{A})=2/10$
  - 상자는 A이고 공은 하양이 뽑힐 확률은?
    - 결합 확률 $P(\text{A, 하양})=P(\text{하양}|\text{A})P(\text{A})=(2/10)(7/10)=7/50$
  - 하얀 공이 나올 확률은?
    - 주변 확률 $P(\text{하양})=P(\text{하양}|\text{A})P(\text{A})+P(\text{하양}|\text{B})P(\text{B})$
      $=(2/10)(7/10)+(9/15)(3/10)=8/25$
  - $P(X,Y)=P(X)P(Y)$이면 X와 Y는 독립
- $P(X)$를 **사전 확률**이라 prior probability 부름

2.1.1 확률 기초

- 이런 문제를 생각해 보자.
  - 하얀 공이 뽑혔는데 어느 상자에서 나왔는지 맞추어라.
    - 기본 전략: 상자 A와 B에서 나왔을 가능성 각각을 구하고 큰 가능성을 보인 상자를 답으로 취한다.
    - 이렇게 해야 맞출 가능성이 최대 (오류 범할 가능성이 최소)가 됨
    - 가능성은 어떻게 계산?

- 생각 1
  - 상자 A의 하얀 공 확률과 상자 B의 하얀 공 확률을 비교하여 큰 쪽을 취한다.
  - P(하양|B)=9/15 > P(하양|A)=2/10 이므로 '상자 B에서 나왔다'고 말함
  - 조건부 확률 $P(Y|X)$를 사용한 셈이다. 타당한가?
  - 이 조건부 확률을 우도라고[likelihood] 부름

2.1.1 확률 기초

- 생각 2
  - 상자 A와 상자 B의 선택 가능성을 비교하여 큰 쪽을 취한다.
  - $P(\mathrm{A})=7/10 > P(\mathrm{B})=3/10$ 이므로 '상자 A에서 나왔다'고 말함
  - 사전 확률 $P(X)$를 사용한 셈이다. 타당한가?
- 올바른 생각
  - 생각 1과 생각 2의 한계
    - 극단적으로 $P(\mathrm{A})=0.999$라면 생각 1이 틀린 것이 확실하다.
    - 극단적으로 $P(\text{하양}|\mathrm{A})=0.999$라면 생각 2가 틀린 것이 확실하다.
    - 우도와 사전 확률을 모두 고려함이 타당해 보임
  - 문제에 충실하자.
    - 조건부 확률 $P(\mathrm{A}|\text{하양})$과 $P(\mathrm{B}|\text{하양})$ 을 비교하여 큰 쪽을 취함
    - 즉 $P(X|Y)$를 사용하겠다는 생각이 타당하다.
    - $P(X|Y)$를 사후 확률 이라posterior probability 함
    - 어떻게 계산할 것인가?

2.1.1 확률 기초

- 베이스 정리의 유도

$$P(X,Y) = P(Y,X)$$

$$P(X)P(Y \mid X) = P(Y)P(X \mid Y)$$

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} = \frac{\text{우도} * \text{사전확률}}{P(Y)} \qquad (2.1)$$

- 베이스 정리를 이용한 사후 확률 계산

$$P(\text{A} \mid \text{하양}) = \frac{P(\text{하양} \mid \text{A})P(\text{A})}{P(\text{하양})} = \frac{(2/10)(7/10)}{(8/25)} = 0.4375$$

$$P(\text{B} \mid \text{하양}) = \frac{P(\text{하양} \mid \text{B})P(\text{B})}{P(\text{하양})} = \frac{(9/15)(3/10)}{(8/25)} = 0.5625$$

2.1.2 평균과 분산

- 평균 벡터와 공분산 행렬

이산 확률 분포 $\boldsymbol{\mu} = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$ (2.9)

연속 확률 분포 $\boldsymbol{\mu} = \int_{R^d} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$ (2.10)

샘플 집합 $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ (2.11)

이산 확률 분포 $\boldsymbol{\Sigma} = \sum_{\mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} P(\mathbf{x})$ (2.13)

연속 확률 분포 $\boldsymbol{\Sigma} = \int_{R^d} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} p(\mathbf{x}) d\mathbf{x}$ (2.14)

샘플 집합 $\boldsymbol{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}$ (2.15)

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2d} \\ \vdots & & \ddots & \\ \sigma_{d1} & \sigma_{d2} & & \sigma_{dd} \end{pmatrix} = \begin{pmatrix} \sigma_1^{\ 2} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^{\ 2} & & \sigma_{2d} \\ \vdots & & \ddots & \\ \sigma_{d1} & \sigma_{d2} & & \sigma_d^{\ 2} \end{pmatrix}$$

- 예제 2.3
  - 8개 샘플이 주어진 상황에서 평균 벡터와 공분산 행렬 구함

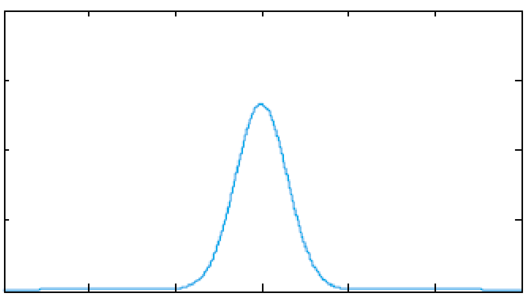
학생	$\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$ (x_1:키, x_2:몸무게, x_3:학점)
1	$\mathbf{x}_1 = (170, 60, 4.1)^{\mathrm{T}}$
2	$\mathbf{x}_2 = (165, 55, 3.0)^{\mathrm{T}}$
3	$\mathbf{x}_3 = (174, 75, 2.8)^{\mathrm{T}}$
4	$\mathbf{x}_4 = (169, 67, 2.9)^{\mathrm{T}}$
5	$\mathbf{x}_5 = (155, 49, 3.1)^{\mathrm{T}}$
6	$\mathbf{x}_6 = (172, 63, 3.6)^{\mathrm{T}}$
7	$\mathbf{x}_7 = (166, 58, 3.7)^{\mathrm{T}}$
8	$\mathbf{x}_8 = (168, 61, 4.0)^{\mathrm{T}}$

$$\boldsymbol{\mu} = (167.375, 61.0, 3.4)^{\mathrm{T}}$$

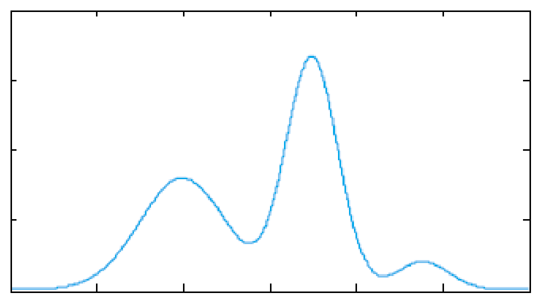
$$\boldsymbol{\Sigma} = \begin{pmatrix} 33.696 & 39.429 & 0.371 \\ 39.429 & 60.857 & -0.943 \\ 0.371 & -0.943 & 0.263 \end{pmatrix}$$

2.1.3 확률 분포의 표현과 추정

- 이산인 경우
  - 차원의 저주
  - 변수의 수가 d이고 각 변수가 q개의 구간을 가진다면 q^d에 비례하는 메모리 필요
- 연속인 경우
  - 일정한 형태를 가는 상황
  - 그렇지 않은 상황



(a) 정규 분포



(b) 세 개의 모드를 가진 분포

그림 2.4 다양한 확률 분포

2.2 베이시언 분류기

- 분류기 학습 (훈련)에 사용하는 정보는 '훈련 집합'
  - 훈련 집합 $X=\{(\mathbf{x}_1,t_1), (\mathbf{x}_2,t_2), \ldots, (\mathbf{x}_N,t_N)\}$
    - $\mathbf{x}_i=(x_1,x_2,\ldots,x_d)$는 특징 벡터
    - $t_i \in \{\omega_1,\omega_2,\ldots,\omega_M\}$ 부류 표지 (이진 분류기인 경우 M=2)

- 예) 필기 숫자

$\mathbf{x}_1=(13/11,12/12)^T$, $t_1=\omega_1$ (숫자 0)

$\mathbf{x}_2=(12/7,14/5)^T$, $t_2=\omega_2$ (숫자 1)

$\mathbf{x}_3=(6/11,13/4)^T$, $t_3=\omega_3$ (숫자 2)

$\mathbf{x}_4=(13/11,10/14)^T$, $t_4=\omega_4$ (숫자 3)

……

$\mathbf{x}_{40}=(15/13,11/17)^T$, $t_{40}=\omega_{10}$ (숫자 9)

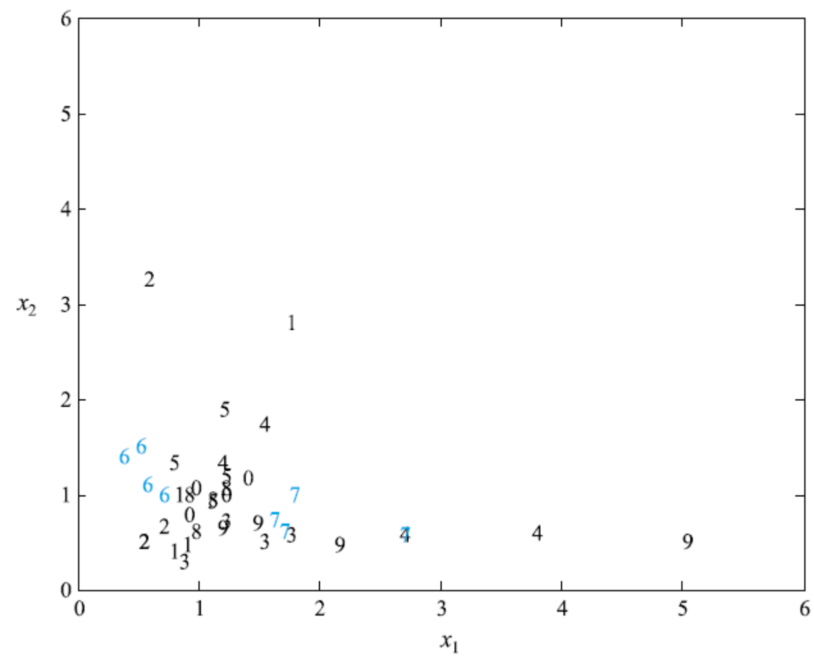


그림 1.6 특징 공간에서 샘플의 분포

2.2.1 최소 오류 베이시언 분류기

- 주어진 특징 벡터 **x**에 대해 '가장 그럴듯한' 부류로 분류

$$\left.\begin{array}{l} P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x}) \text{이면, } \mathbf{x}\text{를 } \omega_1\text{로 분류하고} \\ P(\omega_1 \mid \mathbf{x}) < P(\omega_2 \mid \mathbf{x}) \text{이면, } \mathbf{x}\text{를 } \omega_2\text{로 분류하라.} \end{array}\right\} \quad (2.16)$$

  - (2.16)에서 사후 확률은 직접 구할 수 없음. 왜?
  - 베이스 정리를 이용하여 사후 확률 계산을 사전 확률과 우도로 대치

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{\text{우도*사전확률}}{p(\mathbf{x})} \quad (2.17)$$

  - 분모는 무시해도 됨. 왜?
  - 우도와 사전 확률은 어떻게 계산?

2.2.1 최소 오류 베이시언 분류기

- 사전 확률 계산
  - $P(\omega_1)=n_1/N$, $P(\omega_2)=n_2/N$
  - 정확한 값이 아니라 추정 (N이 커짐에 따라 실제 값에 가까워짐)

- 우도 계산
  - 훈련 집합에서 ω_i에 속하는 샘들들을 가지고 $P(\mathbf{x}|\omega_i)$ 추정
  - 부류 조건부 확률 이라고도class-conditional probabilty 함
  - 3장의 주제

2.2.1 최소 오류 베이시언 분류기

- 최소 오류 베이시언 분류기
  - 결정 규칙

$$\left.\begin{array}{l} p(\mathbf{x} \mid \omega_1)P(\omega_1) > p(\mathbf{x} \mid \omega_2)P(\omega_2)\text{이면, } \mathbf{x}\text{를 } \omega_1\text{로 분류하고} \\ p(\mathbf{x} \mid \omega_1)P(\omega_1) < p(\mathbf{x} \mid \omega_2)P(\omega_2)\text{이면, } \mathbf{x}\text{를 } \omega_2\text{로 분류하라.} \end{array}\right\} \quad (2.18)$$

  - 특수한 경우로 (2.18)의 의미 해석하면,
    - 사전 확률이 0.5인 경우 우도만으로 분류
    - $P(\omega_1) >> P(\omega_2)$인 경우 사전 확률이 의사 결정 주도

2.2.1 최소 오류 베이시언 분류기

- 최소 오류 베이시언 분류기
  - 오류 확률

$$E = \frac{1}{2}\left(\int_{-\infty}^{t} p(x \mid \omega_2)dx + \int_{t}^{\infty} p(x \mid \omega_1)dx\right)$$

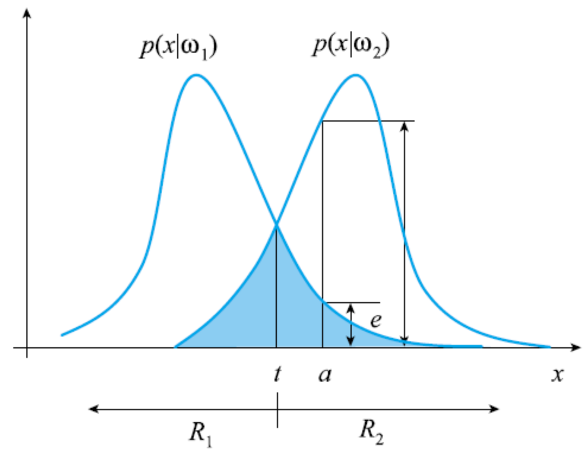


그림 2.5 베이시언 분류기의 오류 확률 (사전 확률은 같다고 가정)

  - 최적성

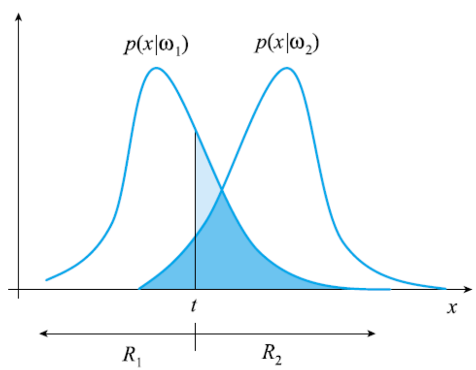


그림 2.6 베이시언 분류기를 사용하지 않을 때 증가하는 오류

2.2.2 최소 위험 베이시언 분류기

- 성능 기준으로 오류가 적절하지 못한 상황
  - 정상인과 암 환자 분류
  - 과일을 상품과 하품으로 분류

- 손실 행렬

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

2.2.2 최소 위험 베이시언 분류기

- 최소 위험 베이시언 분류기

$$\left.\begin{array}{l} \mathbf{x}\text{를 } q_2 > q_1\text{이면 } \omega_1\text{로 분류하고, } q_1 > q_2\text{이면 } \omega_2\text{로 분류하라.} \\ \text{이때 } q_1 = c_{11}p(\mathbf{x}\mid\omega_1)P(\omega_1) + c_{21}p(\mathbf{x}\mid\omega_2)P(\omega_2) \\ \qquad\; q_2 = c_{12}p(\mathbf{x}\mid\omega_1)P(\omega_1) + c_{22}p(\mathbf{x}\mid\omega_2)P(\omega_2) \end{array}\right\} \quad (2.23)$$

- 우도비로 다시 쓰면
  - 우도비 결정 규칙

$$\left.\begin{array}{l} \mathbf{x}\text{를 } \dfrac{p(\mathbf{x}\mid\omega_1)}{p(\mathbf{x}\mid\omega_2)} > T \text{ 이면 } \omega_1\text{로 분류하고, } \dfrac{p(\mathbf{x}\mid\omega_1)}{p(\mathbf{x}\mid\omega_2)} < T \text{ 이면 } \omega_2\text{로 분류하라.} \\ \text{이때 } T = \dfrac{(c_{21}-c_{22})P(\omega_2)}{(c_{12}-c_{11})P(\omega_1)} \end{array}\right\} \quad (2.24)$$

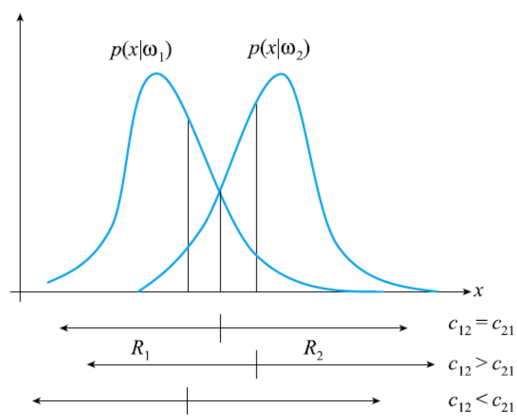


그림 2.7 최소 위험 베이시언 분류기 (사전 확률은 같다고 가정)

2.2.3 M 부류로 확장

- M 부류 최소 오류 베이시언 분류기
  - 사후 확률로 쓰면

$$\mathbf{x}\text{를 } k = \arg\max_i P(\omega_i \mid \mathbf{x}) \text{ 일때 } \omega_k\text{로 분류하라.} \tag{2.26}$$

  - 사전 확률과 우도로 쓰면

$$\mathbf{x}\text{를 } k = \arg\max_i p(\mathbf{x} \mid \omega_i) P(\omega_i) \text{ 일때 } \omega_k\text{로 분류하라.} \tag{2.27}$$

- M 부류 최소 위험 베이시언 분류기

$$\left.\begin{array}{l} \mathbf{x}\text{를 } k = \arg\min_i q_i \text{일때 } \omega_k\text{로 분류하라.} \\ \text{이때 } q_i = \sum_{j=1}^{M} c_{ji}\, p(\mathbf{x} \mid \omega_j) P(\omega_j) \end{array}\right\} \tag{2.29}$$

2.3 분별 함수

- 지금까지 분류기를 분별 함수로 다시 작성하면

$$\left.\begin{array}{l} \mathbf{x}\text{를 } k = \arg\max_i g_i(\mathbf{x}) \text{ 일때 } \omega_k \text{로 분류하라.} \\ \text{이때 } g_i(\mathbf{x}) = \begin{cases} p(\mathbf{x}|\omega_i)P(\omega_i) & \text{(최소 오류 베이시언 분류기)} \\ \dfrac{1}{\sum_{j=1}^{M} c_{ji}\, p(\mathbf{x}|\omega_j)P(\omega_j)} & \text{(최소 위험 베이시언 분류기)} \end{cases} \end{array}\right\} \quad (2.30)$$

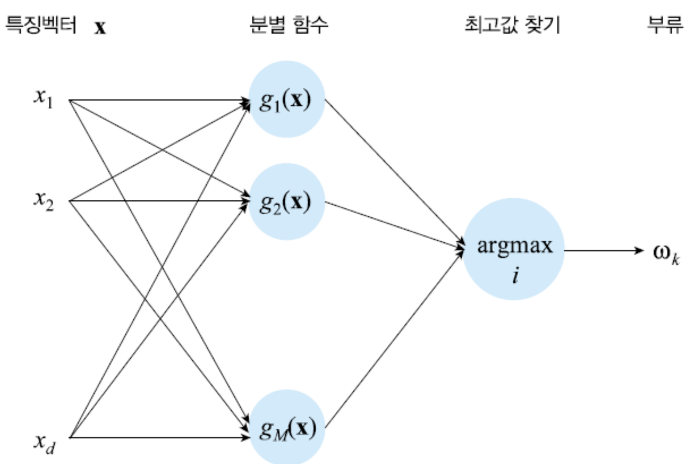


그림 2.8 분류기의 일반적인 틀

알고리즘 [2.1] *M* 부류 베이시언 분류기

입력: 특징 벡터 $\mathbf{x}$

출력: 부류

알고리즘:

1. **for** ($i = 1$ **to** M)
2. **if** (최소 오류) $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$;
3. **else if** (최소 위험) $g_i(\mathbf{x}) = 1 \Big/ \sum_{j=1}^{M} c_{ji}\, p(\mathbf{x}|\omega_j)P(\omega_j)$;
4. $g_i(\mathbf{x})$, $1 \le i \le M$ 중 가장 큰 것을 찾아 그것의 첨자를 k라 하자.
5. **return** ω_k;

2.3 분별 함수

- 분별 함수 표현의 장점
  - 여러 분류기를 하나의 틀로 표현
  - $f(.)$가 단조 증가라면 $p(\mathbf{x}|\omega_i)\ P(\omega_i)$대신 $g_i(\mathrm{x})=f(p(\mathbf{x}|\omega_i)\ P(\omega_i))$ 사용하여도 같은 결과
    - $f(.)$로 log 함수를 주로 사용
    - log는 곱셈을 덧셈으로 바꾸어 주므로 수식 전개에 유리하고 log 취하면 값의 규모가 커져 수치 오류에 둔감한 이점

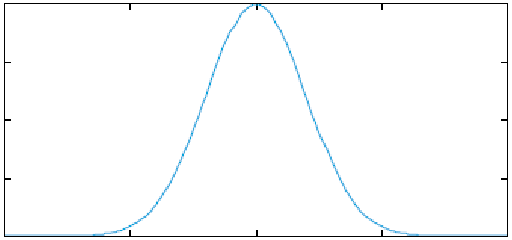
2.4 정규분포에서 베이시언 분류기

- 정규 분포 (가우시언 분포)
  - 현실 세계에 맞는 경우 있음
  - 평균과 분산이라는 두 종류의 매개 변수만으로 표현 가능
  - 수학적인 매력
- 우도가 정규 분포를 따른다는 가정 하에 베이시언 분류기의 특성을 해석해 보자.

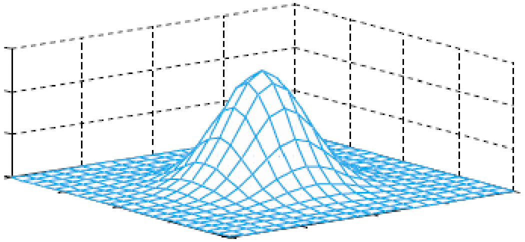
2.4.1 정규분포와 분별 함수

- 정규 분포

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



(a) 1 차원 정규 분포



(b) 2 차원 정규 분포

그림 2.9 정규 분포의 예

2.4.1 정규분포와 분별 함수

- 우도를 다시 쓰면,

$$p(\mathbf{x}\,|\,\omega_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)) \qquad (2.33)$$

- 로그를 취하여 분별 함수를 만들어 보면,

$$\left.\begin{aligned} g_i(\mathbf{x}) &= \ln\left(p(\mathbf{x}\,|\,\omega_i)P(\omega_i)\right) \\ &= \ln N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \ln P(\omega_i) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \\ &= -\frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \\ &= -\frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \end{aligned}\right\} \quad (2.34)$$

  - $g_i(\mathbf{x})$는 변수 $\mathbf{x}$에 대한 2차 식

2.4.1 정규분포와 분별 함수

- 예제 2.4
  - d=2이고 아래와 같다고 가정

    부류 ω_i의 정규 분포가 $\boldsymbol{\mu}_i = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$와 $\boldsymbol{\Sigma}_i = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

  - 분별 함수를 유도해 보면,

$$g_i(\mathbf{x}) = -\frac{1}{2}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)^{\mathrm{T}} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}^{-1} \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right) - \frac{2}{2}\ln 2\pi - \frac{1}{2}\ln\left|\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right| + \ln P(\omega_i)$$

$$= -\frac{1}{2}\begin{pmatrix} x_1 - 3 & x_2 - 1 \end{pmatrix} \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 - 3 \\ x_2 - 1 \end{pmatrix} - \ln 2\pi - \frac{1}{2}\ln 4 + \ln P(\omega_i)$$

$$= -\frac{1}{4}\underbrace{(x_1{}^2 + x_2{}^2)}_{2\text{차 식}} + \frac{1}{2}\underbrace{(3x_1 + x_2)}_{1\text{차 식}} - \frac{1}{2}\underbrace{(5 + 2\ln 2\pi + \ln 4 - 2\ln P(\omega_i))}_{\text{상수}}$$

2.4.1 정규분포와 분별 함수

- 결정 경계
  - 두 부류가 차지하는 영역의 경계
    - $g_i(\mathbf{x})=g_j(\mathbf{x})$인 점
    - 즉 $g_{ij}(\mathbf{x})=0$인 점

$$g_{ij}(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}) \tag{2.35}$$

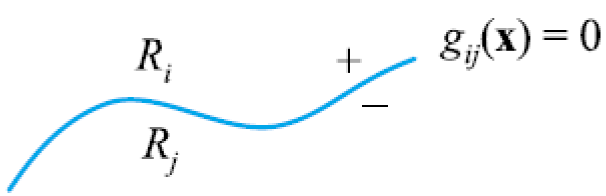


그림 2.10 두 부류 ω_i와 ω_j의 결정 경계

2.4.2 선형 분별

- 모든 부류의 공분산 행렬이 같은 상황, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$로 표기
  - 분별 함수를 다시 쓰면

$$g_i(\mathbf{x}) = \frac{1}{2}\underbrace{\left(2\boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + 2\ln P(\omega_i)\right)}_{i\text{에 따라 다름}} - \frac{1}{2}\underbrace{\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + d\ln 2\pi + \ln|\boldsymbol{\Sigma}|\right)}_{i\text{에 무관함}} \quad (2.36)$$

  - i에 무관한 항은 제거해도 됨. 따라서 2차 항 $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}$ 없어짐
  - 선형식이 됨

$$\left.\begin{aligned} g_i(\mathbf{x}) &= \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}\mathbf{x} + \left(\ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right) \\ &= \mathbf{w}_i{}^{\mathrm{T}}\mathbf{x} + b_i \end{aligned}\right\} \quad (2.37)$$

2.4.2 선형 분별

- 결정 경계

$$
\left.\begin{aligned}
g_{ij}(\mathbf{x}) &= g_i(\mathbf{x}) - g_j(\mathbf{x}) \\
&= \left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\right)^{\mathrm{T}} \mathbf{x} + \left(\ln P(\omega_i) - \ln P(\omega_j) - \frac{1}{2}\boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right) \\
&= \underbrace{\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\right)^{\mathrm{T}}}_{\mathbf{w}} \left(\mathbf{x} - \underbrace{\left(\frac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) - \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)} \ln\frac{P(\omega_i)}{P(\omega_j)}\right)}_{\mathbf{x}_0}\right) \\
&= \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0)
\end{aligned}\right\} \quad (2.38)
$$

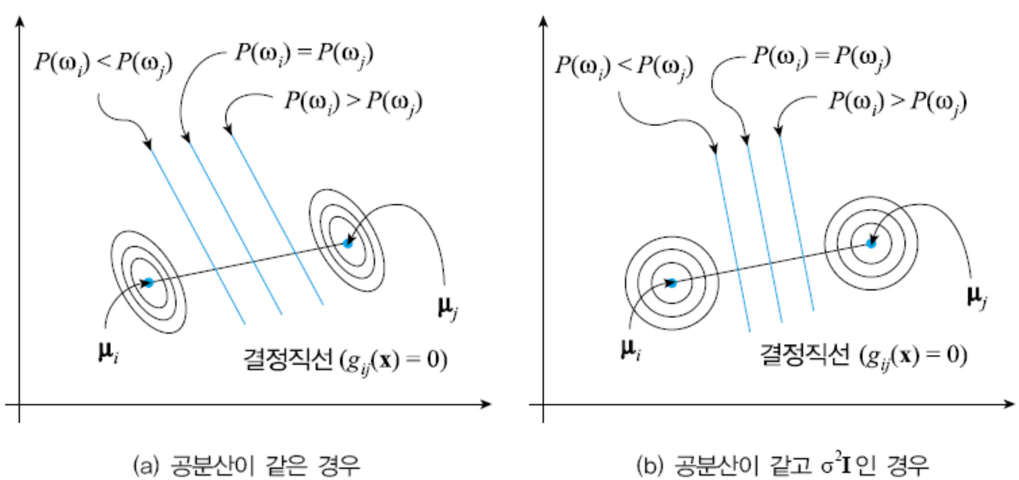


그림 2.11 선형 분류기

2.4.2 선형 분별

- 예제 2.5

ω_1: $(1,2)^T$, $(3,1)^T$, $(5,2)^T$, $(3,3)^T$

ω_2: $(6,6)^T$, $(8,5)^T$, $(10,6)^T$, $(8,7)^T$

$$\boldsymbol{\mu}_1 = \frac{1}{4}\left(\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}3\\1\end{pmatrix}+\begin{pmatrix}5\\2\end{pmatrix}+\begin{pmatrix}3\\3\end{pmatrix}\right)=\begin{pmatrix}3\\2\end{pmatrix},\quad \boldsymbol{\Sigma}_1=\begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{4}\left(\begin{pmatrix}6\\6\end{pmatrix}+\begin{pmatrix}8\\5\end{pmatrix}+\begin{pmatrix}10\\6\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}\right)=\begin{pmatrix}8\\6\end{pmatrix},\quad \boldsymbol{\Sigma}_2=\begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}$$

$$g_{12}(\mathbf{x}) = \left(\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}3-8\\2-6\end{pmatrix}\right)^T \mathbf{x}$$

$$+\left(\ln P(\omega_1)-\ln P(\omega_2)-\frac{1}{2}\begin{pmatrix}3 & 2\end{pmatrix}\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}3\\2\end{pmatrix}+\frac{1}{2}\begin{pmatrix}8 & 6\end{pmatrix}\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}8\\6\end{pmatrix}\right)$$

$$=\begin{pmatrix}-15/8 & -6\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}+\left(\ln P(\omega_1)-\ln P(\omega_2)+34.3125\right)$$

$$=-15/8x_1-6x_2+\left(\ln P(\omega_1)-\ln P(\omega_2)+34.3125\right)$$

$P(\omega_1)=0.5, P(\omega_2)=0.5$인 경우 : $5x_1+16x_2-91.5=0$

$P(\omega_1)=0.8, P(\omega_2)=0.2$인 경우 : $5x_1+16x_2-95.197=0$

$P(\omega_1)=0.2, P(\omega_2)=0.8$인 경우 : $5x_1+16x_2-87.803=0$

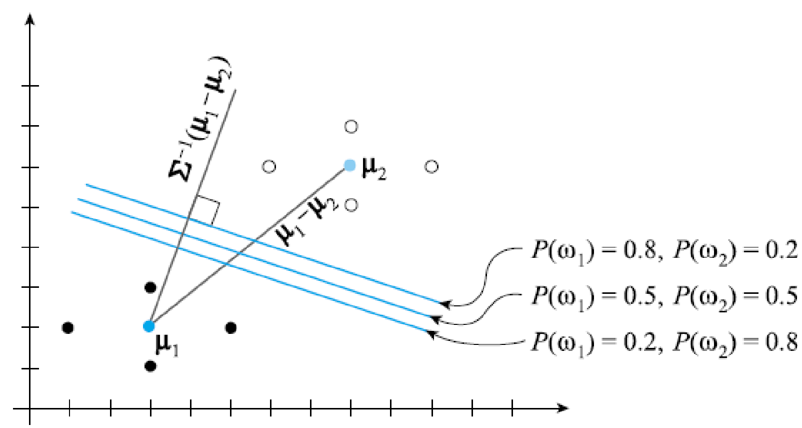


그림 2.12 선형 분별 분석 (LDA) 사례

Thomas Bayes
(1702 ~ 1761) 영국

Bayes는 목사의 아들로 영국 런던에서 태어났다. 그는 Edinburgh 대학에서 논리학과 신학을 전공하였다. 그 이후 목사로 부임하여 평생 신학의 길을 걷지만 그에 못지 않게 과학과 수학에 흥미와 뛰어난 재능을 보였다. Bayes를 유명하게 만든 논문은 사후에 그의 친구 Richard Price에 의해 출판되었다. 이 논문의 원고는 웹을 통해 쉽게 구할 수 있다 [Bayes1764]. 이 논문은 역 확률을inverse probability 다루고 있고 이것이 베이스 정리의 기초를 제공하였다. 1950년경에 그의 이론적인 공헌을 기리는 뜻에서 그의 이름을 형용사화한 Bayesian이라는 단어가 만들어 졌고 이제는 확률에 관련된 모든 분야에서 가장 널리 쓰이는 단어 중의 하나가 되었다. 물론 패턴 인식에서도 마찬가지이다. 패턴 인식을 다루는 거의 모든 교과서가 앞 부분에서 베이시언 분류를 소개하고 있다. Bayes에 대한 보다 상세한 내용을 원하면 그의 탄생 300 주년을 기리는 뜻에서 발표된 논문 [Bellhouse04]를 참고하기 바란다.

2.4.3 2차 분별

- 임의의 공분산 행렬

$$g_i(\mathbf{x}) = \underbrace{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x}}_{2\text{차 항}} + \underbrace{\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x}}_{1\text{차 항}} + \underbrace{\left(-\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)\right)}_{\text{상수 항}} \qquad (2.40)$$

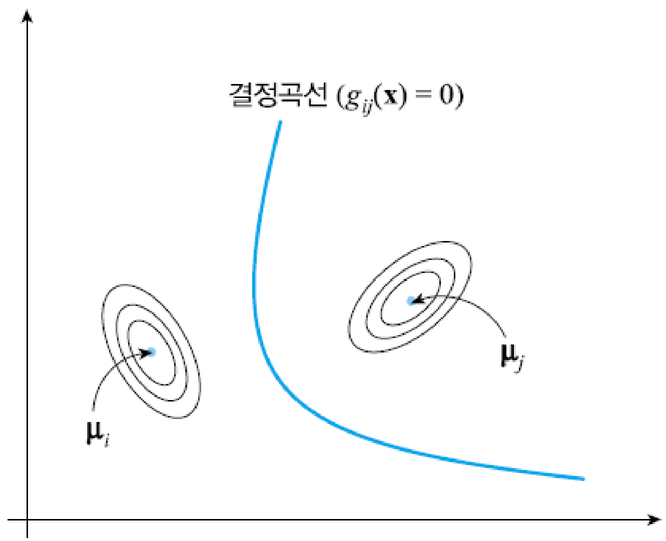


그림 2.13 2차 분별 분석으로 만든 2차 분류기

2.4.3 2차 분별

- 예제 2.6

ω_1: $(1,2)^T$, $(3,1)^T$, $(5,2)^T$, $(3,3)^T$

ω_2: $(7,6)^T$, $(8,4)^T$, $(9,6)^T$, $(8,8)^T$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 8/3 & 0 \\ 0 & 2/3 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 8 \\ 6 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2/3 & 0 \\ 0 & 8/3 \end{pmatrix}$$

$$g_{12}(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = \frac{9}{16}x_1^2 - \frac{9}{16}x_2^2 - \frac{87}{8}x_1 + \frac{3}{4}x_2 + \frac{801}{16} + \ln P(\omega_1) - \ln P(\omega_2)$$

$P(\omega_1) = 0.8, P(\omega_2) = 0.2$인 경우 : $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 + 274.3936 = 0$

$P(\omega_1) = 0.5, P(\omega_2) = 0.5$인 경우 : $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 - 267 = 0$

$P(\omega_1) = 0.2, P(\omega_2) = 0.8$인 경우 : $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 + 259.6064 = 0$

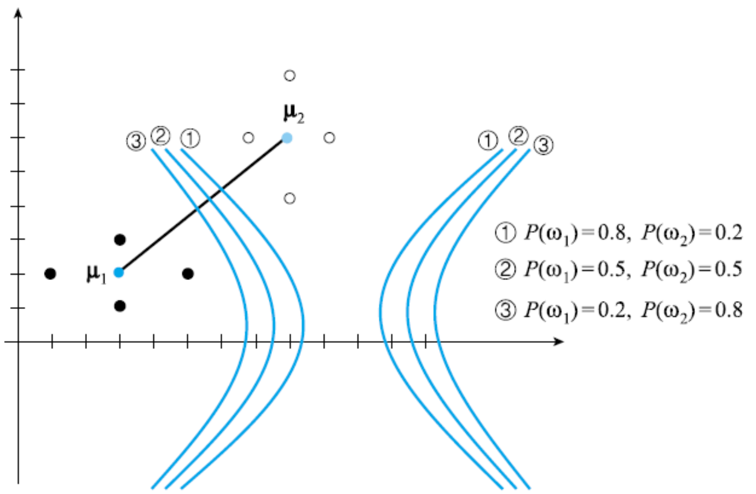


그림 2.14 2차 분별 분석 (QDA) 사례

2.4.4 최소 거리 분류기

- 최소 거리 분류기로 다시 해석해 보자.
  - 수식 유도 편의를 위해 두 부류의 사전 확률과 공분산 행렬 같다고 가정

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \qquad (2.41)$$

  - 최소 거리 분류기

$\mathbf{x}$를 $k = \arg\min_i \left((\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)^{1/2}$ 일 때 ω_k로 분류하라. (2.44)

- 거리 척도

마할라노비스 거리: $\left((\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)^{1/2}$ (2.42)

유클리디언 거리: $\left((\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}_i)\right)^{1/2} = \|\mathbf{x}-\boldsymbol{\mu}_i\|$ (2.43)

2.4.4 최소 거리 분류기

- 예제 2.7 ω_1: $(1,2)^T$, $(3,1)^T$, $(5,2)^T$, $(3,3)^T$

  ω_2: $(6,6)^T$, $(8,5)^T$, $(10,6)^T$, $(8,7)^T$

  $\mathbf{x} = (8,2)^T$

$$\boldsymbol{\mu}_1\text{까지의 마할라노비스 거리} = \left(\begin{pmatrix} 8-3 & 2-2 \end{pmatrix} \begin{pmatrix} 3/8 & 0 \\ 0 & 3/2 \end{pmatrix} \begin{pmatrix} 8-3 \\ 2-2 \end{pmatrix} \right)^{1/2} = 3.062$$

$$\boldsymbol{\mu}_2\text{까지의 마할라노비스 거리} = \left(\begin{pmatrix} 8-8 & 2-6 \end{pmatrix} \begin{pmatrix} 3/8 & 0 \\ 0 & 3/2 \end{pmatrix} \begin{pmatrix} 8-8 \\ 2-6 \end{pmatrix} \right)^{1/2} = 4.899$$

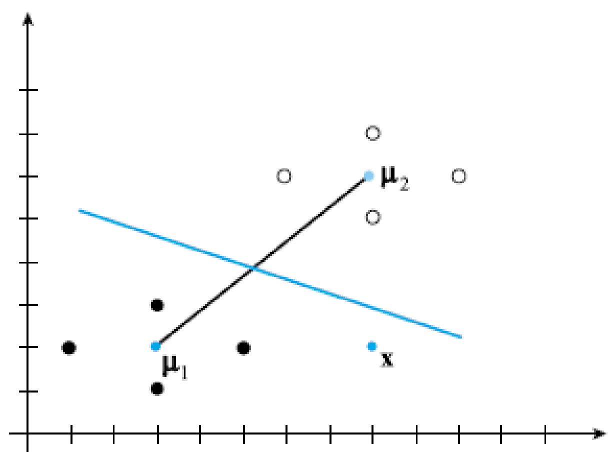
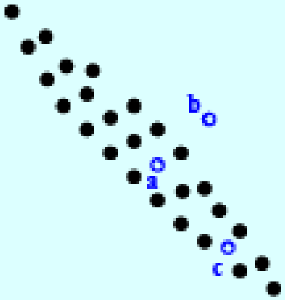


그림 2.15 두 통계 분포까지의 마할라노비스 거리

Prasanta Chandra Mahalanobis
(1893년 6월 29일 ~ 1972년 6월 28일) 인도

인도에 고등 교육과 연구를 같이 수행하는 우리나라 KAIST와 비슷한 연구원이 있다. 인도 통계 연구원이 Indian Statistical Institute (ISI) 바로 그것인데 통계학과 그것의 응용에 관련된 분야를 다룬다. ISI는 Mahalanobis가 1931년에 설립하였고 현재는 세계적인 명성을 누리고 있다. Mahalanobis는 영국 Cambridge 대학에서 수학과 물리를 공부하였으며 인도로 돌아와 본격적으로 통계학을 공부하였다. 그는 통계학 전반에 걸쳐 탁월한 업적을 남겼다. 특히 대규모 모집단 조사를 위한 샘플링 기법에 큰 공헌을 하였으며 인도 정부가 안고 있는 많은 문제를 통계를 이용하여 해결하여 실용적인 공헌에서도 독보적이다. 무엇보다 그를 세계적으로 유명하게 만든 것은 Mahalanobis 거리이다. 그에 대한 보다 상세한 내용은 [Rao73]을 참고하라.

Mahalanobis 거리의 발상은 간단하다. 오른쪽 그림에서 b와 c 중에 어느 것이 a에 더 가까운가? 확률 분포를 고려하지 않으면 당연히 b가 가깝다. 하지만 확률 분포를 고려하면?



- 베이시언 분류의 특성

1. 비현실성을 내포하고 있다. 이미 언급한 바와 같이 일반적인 확률 분포를 사용하려 하면 차원의 저주가 발생한다. 정규 분포를 가정하면 실제 확률 분포와 차이가 발생한다.
2. 1의 비현실성을 무시하고 실제 확률 분포를 안다고 가정하면 베이시언 분류기는 오류율 측면에서 최적이다. 이 최적성을 수학적으로 완벽하게 증명할 수 있다.
3. 베이시언 분류기는 M 개 부류 각각에 대해 그에 속할 확률을 출력한다. 즉 (2.30)의 베이시언 분류기에서 $g_j(\mathbf{x})$를 확률로 해석할 수 있다. 따라서 이 확률을 신뢰도 값으로 삼아 후처리 등에 유용하게 활용할 수 있다.

2.5 베이시언 분류의 특성

- 나이브naïve 베이시언 분류기
  - 특징들이 서로 독립이라는 가정

$$p(\mathbf{x} \mid \omega_i) = \prod_{j=1}^{d} p(x_j \mid \omega_i) \qquad (2.45)$$

  - 우도 계산을 (2.45)로 하는 분류기를 나이브 베이시언 분류기라 함
  - 얻은 것: 차원의 저주를 피함
  - 잃은 것: 성능 저하

2.6 기각 처리

- 기각
  - 신뢰도가 충분치 않은 경우는 의사 결정 포기
  - 그림 2.16에서 두 부류의 확률 차이가 Δ보다 작으면 기각

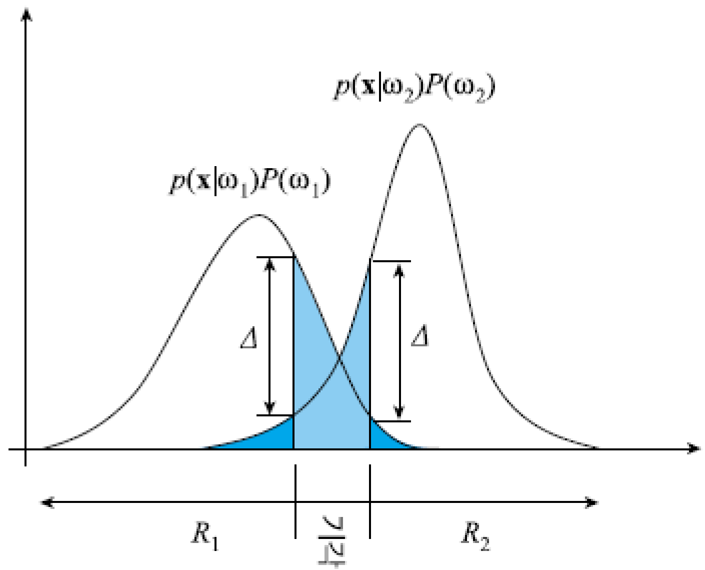


그림 2.16 기각 처리