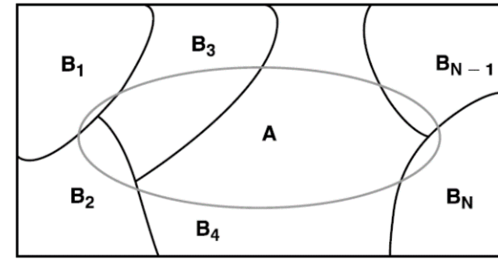
# 확률 이론

## ▪ 베이즈의 정리

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] \cdot P[B_j]}{\sum_{k=1}^{N} P[A \mid B_k] \cdot P[B_k]}$$




$$P[\omega_j \mid \mathrm{x}] = \frac{P[\mathrm{x} \mid \omega_j] \cdot P[\omega_j]}{\sum_{k=1}^{N} P[\mathrm{x} \mid \omega_k] \cdot P[\omega_k]} = \frac{P[\mathrm{x} \mid \omega_j] \cdot P[\omega_j]}{P[\mathrm{x}]}$$

※ $\omega_j$ : $j$번째 클래스

x : 특징 벡터

- $P[\omega_j]$ : 클래스 $\omega_j$ 의 사전 확률(prior probability)
- $P[\omega_j \mid \mathrm{x}]$ : 관측 x가 주어질 경우 클래스 $\omega_j$ 에 대한 사후 확률(posterior probability)
- $P[\mathrm{x} \mid \omega_j]$ : 우도(likelihood: 클래스 $\omega_j$ 가 주어질 경우 관측 x가 일어날 조건부 확률)
- $P[\mathrm{x}]$ : x 가 일어날 확률로 결정에 영향을 미치지 않은 정규화 상수

# 우도비 검증 (Likelihood Ratio Test: LRT)

## ❖ 데이터의 확률밀도함수를 알 경우의 클래스의 분류

"만약 $\mathbf{P(\omega_1 \mid x) > P(\omega_2 \mid x)}$라면 $\boldsymbol{\omega_1}$을 선택하고, 그렇지 않으면 $\boldsymbol{\omega_2}$를 선택한다."

$$\mathbf{P(\omega_1 \mid x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2 \mid x)}$$

베이즈의 정리 ⇒

$$\mathbf{\frac{P(x \mid \omega_1)P(\omega_1)}{P(x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(x \mid \omega_2)P(\omega_2)}{P(x)}}$$

$$\mathbf{\Lambda(x) = \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}}$$

## ❖ 데이터의 확률밀도함수를 알 경우의 클래스의 분류

**예제 5-1**

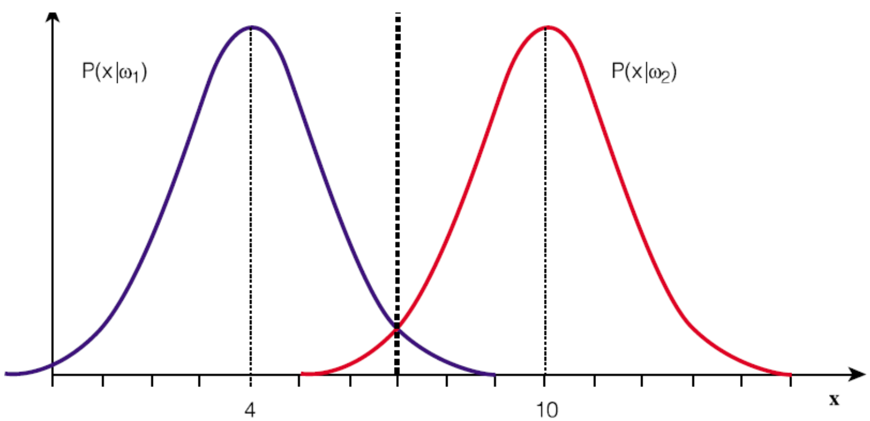
다음과 같이 2-클래스 $\omega_1, \omega_2$가 주어진 경우(조건부) 확률밀도함수(혹은 우도함수)가 주어질 경우, LRT 결정규칙을 유도해 보자. 단 사전 확률(a prior probability)은 같다고 가정한다.

$$P(x|\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}, \quad P(x|\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}$$

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{1}$$

2. LRT 표현식을 단순화하면 다음과 같다.

$$\Lambda(x) = \frac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$



# 우도비 검증

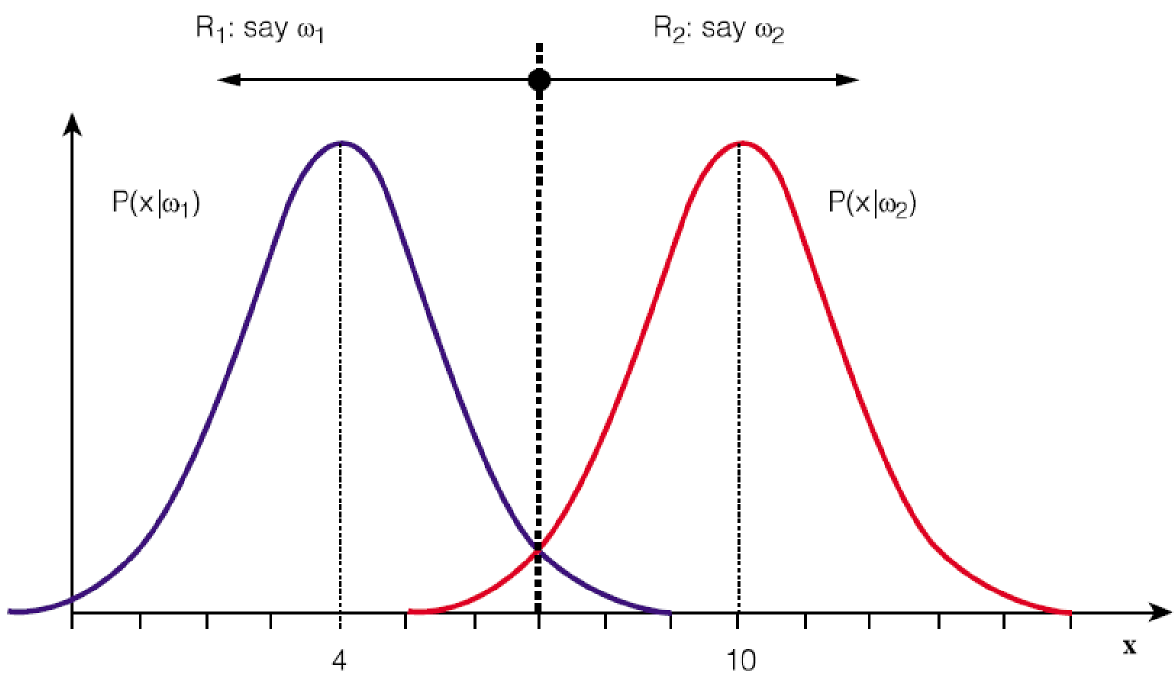
## ❖ 데이터의 확률밀도함수를 알 경우의 클래스의 분류

3. 양변에 로그를 취하고 부호를 바꾸면 다음과 같다.

$$(x-4)^2-(x-10)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$$

4. 이를 풀면 다음과 같다.

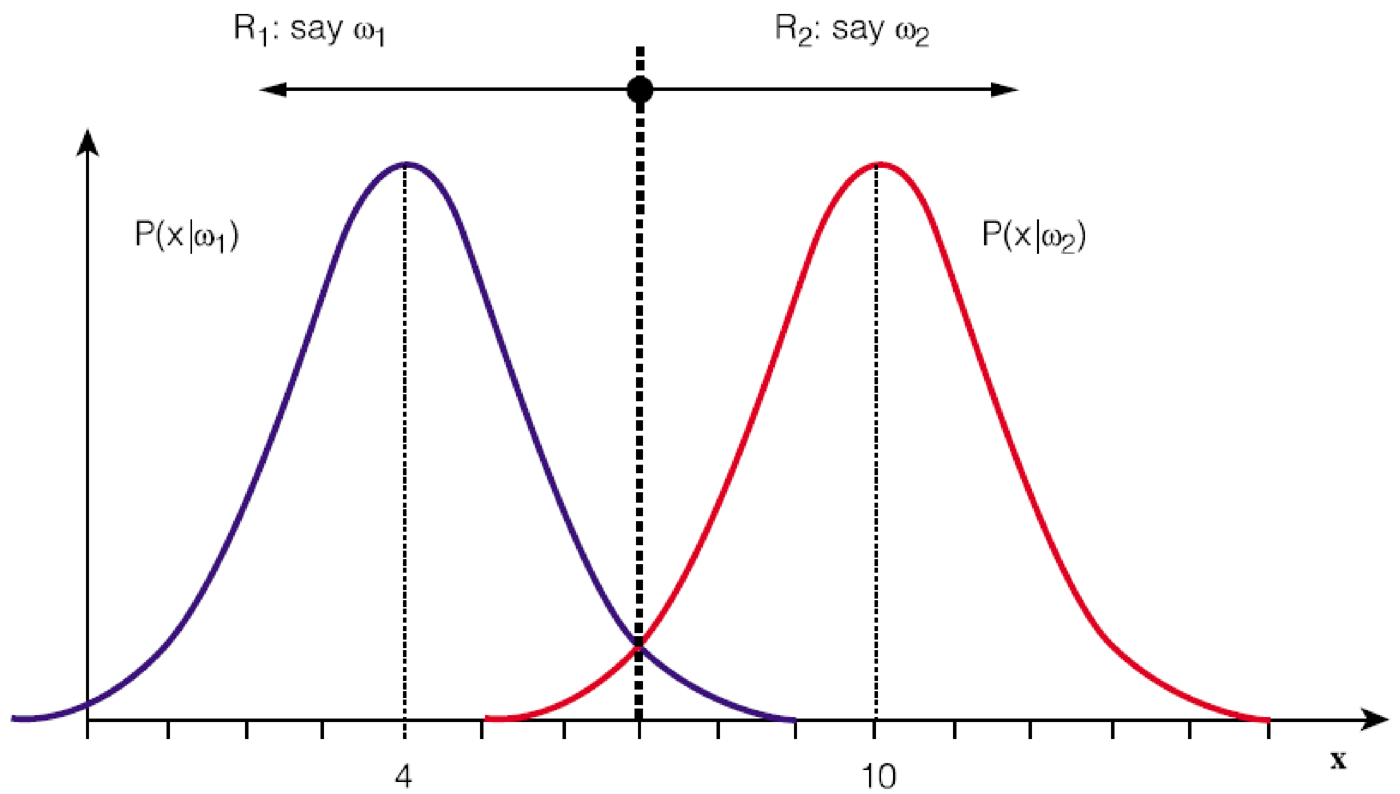
$$x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 7$$




[그림 5-1] LRT에 의한 결정 경계의 결정

# 우도비 검증

## ❖ 데이터의 확률밀도함수를 알 경우의 클래스의 분류

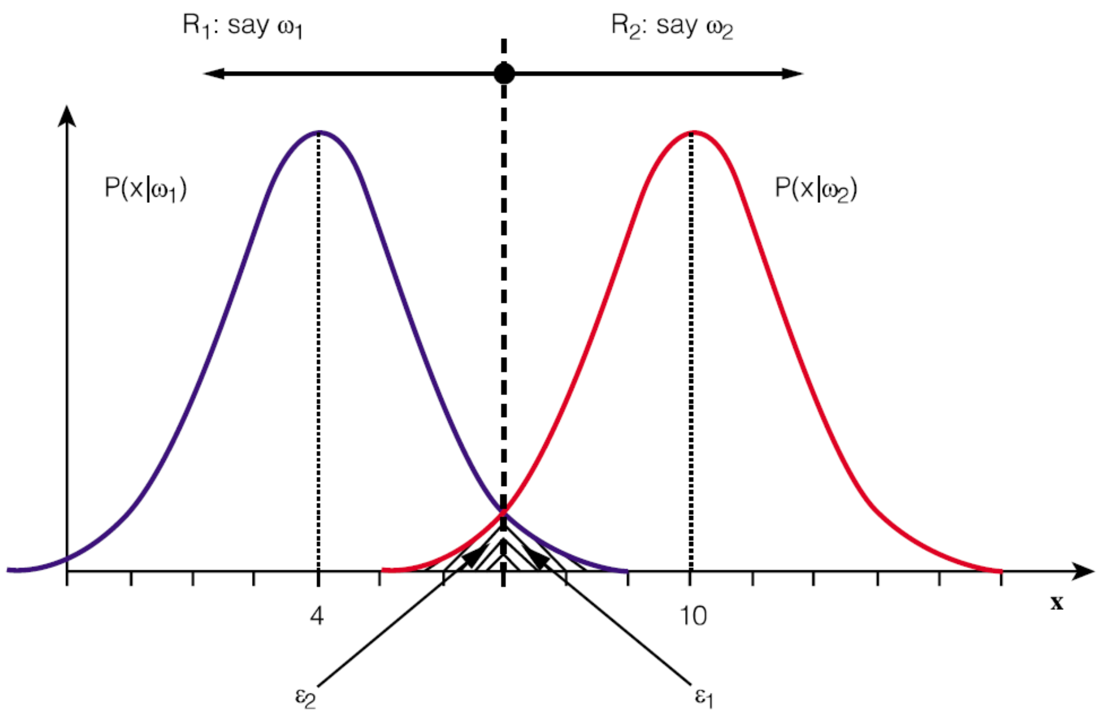



[그림 5-1] LRT에 의한 결정 경계의 결정

각 클래스에 대한 우도의 모양은 동일하고 각 클래스의 평균값만 달라지는 모양이다. 만약 사전 확률이 $\mathbf{P(\omega_1) = 2P(\omega_2)}$와 같다면, LRT 결정규칙은 어떻게 바뀔까?

# 오류확률

$$P[\text{error}] = \underbrace{P[\omega_1]\int_{R_2} P(x\mid\omega_1)dx}_{\varepsilon_1} + \underbrace{P[\omega_2]\int_{R_1} p(x\mid\omega_2)dx}_{\varepsilon_2}$$




[그림 5-2] 오류확률

- 평균(mean)
  - 자료의 총합을 자료의 개수로 나눈 것을 말한다.
  - 자료의 분포의 무게 중심에 해당한다.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 분산(variance)
  - 자료로부터 평균값의 차이에 대한 제곱 값의 평균을 말함
  - 자료의 흩어진 정도를 나타낸다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 표준 편차(standard deviation)
  - 분산의 제곱근을 취하여 자료의 단위와 일치시킨 것을 말한다.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

# 2.2.4 평균과 분산

## 데이터의 요약 정보로서 평균과 분산

$$\left.\begin{aligned} \text{평균} \quad \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \text{분산} \quad \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}\right\} \qquad (2.36)$$

## 평균 벡터와 공분산 행렬

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad (2.37)$$

*두 개 이상의 변량 데이터가 주어질 경우에 각 변량간의 변화하는 양상을 나타내는 통계적 척도*

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \qquad (2.39)$$

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

## 평균 벡터와 공분산 행렬 예제

**예제 2-7**

Iris 데이터베이스의 샘플 중 8개만 가지고 공분산 행렬을 계산하자.

$$\mathbb{X} = \left\{\mathbf{x}_1 = \begin{pmatrix}5.1\\3.5\\1.4\\0.2\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}4.9\\3.0\\1.4\\0.2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}4.7\\3.2\\1.3\\0.2\end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix}4.6\\3.1\\1.5\\0.2\end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix}5.0\\3.6\\1.4\\0.2\end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix}5.4\\3.9\\1.7\\0.4\end{pmatrix}, \mathbf{x}_7 = \begin{pmatrix}4.6\\3.4\\1.4\\0.3\end{pmatrix}, \mathbf{x}_8 = \begin{pmatrix}5.0\\3.4\\1.5\\0.2\end{pmatrix}\right\}$$

먼저 평균벡터를 구하면 $\boldsymbol{\mu} = (4.9125,\ 3.3875,\ 1.45,\ 0.2375)^{\mathrm{T}}$이다. 첫 번째 샘플 $\mathbf{x}_1$을 식 (2.39)에 적용하면 다음과 같다.

$$(\mathbf{x}_1 - \boldsymbol{\mu})(\mathbf{x}_1 - \boldsymbol{\mu})^{\mathrm{T}} = \begin{pmatrix}0.1875\\0.1125\\-0.05\\-0.0375\end{pmatrix}(0.1875 \quad 0.1125 \quad -0.05 \quad -0.0375)$$

$$= \begin{pmatrix}0.0325 & 0.0211 & -0.0094 & -0.0070\\0.0211 & 0.0127 & -0.0056 & -0.0042\\-0.0094 & -0.0056 & 0.0025 & 0.0019\\-0.0070 & -0.0042 & 0.0019 & 0.0014\end{pmatrix}$$

나머지 7개 샘플도 같은 계산을 한 다음, 결과를 모두 더하고 8로 나누면 다음과 같은 공분산 행렬을 얻는다.

$$\boldsymbol{\Sigma} = \begin{pmatrix}0.0661 & 0.0527 & 0.0181 & 0.0083\\0.0527 & 0.0736 & 0.0181 & 0.0130\\0.0181 & 0.0181 & 0.0125 & 0.0056\\0.0083 & 0.0130 & 0.0056 & 0.0048\end{pmatrix}$$

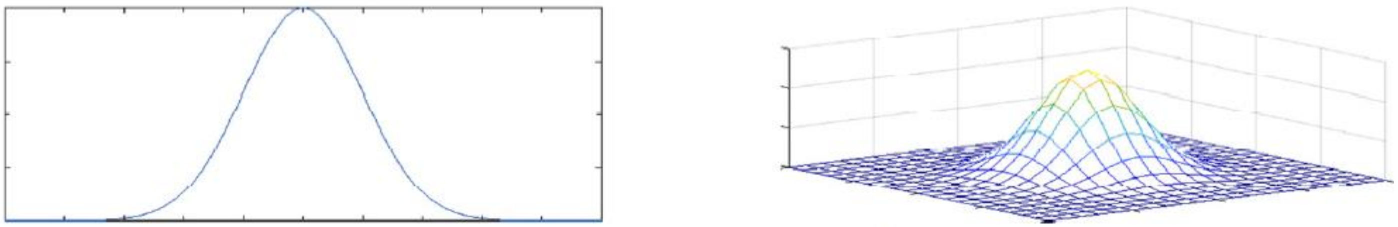
# 2.2.5 유용한 확률분포

## ■ 가우시안 분포

- 평균 $\mu$와 분산 $\sigma^2$으로 정의

확률변수 매개변수

$$N(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



(a) 1차원 (b) 2차원

그림 2-19 가우시안 분포

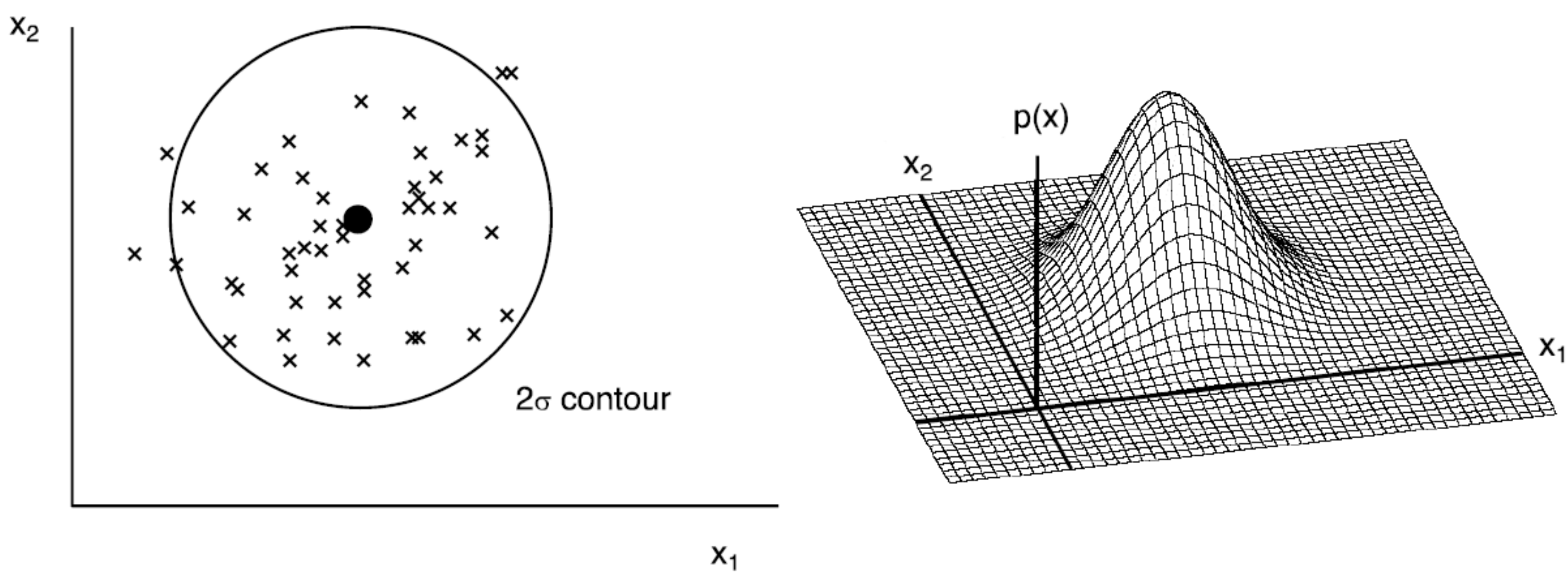
- 다차원 가우시안 분포: 평균벡터 $\boldsymbol{\mu}$와 공분산행렬 $\Sigma$로 정의

$$N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

공분산 행렬식 차원수

# 가우시안 분포

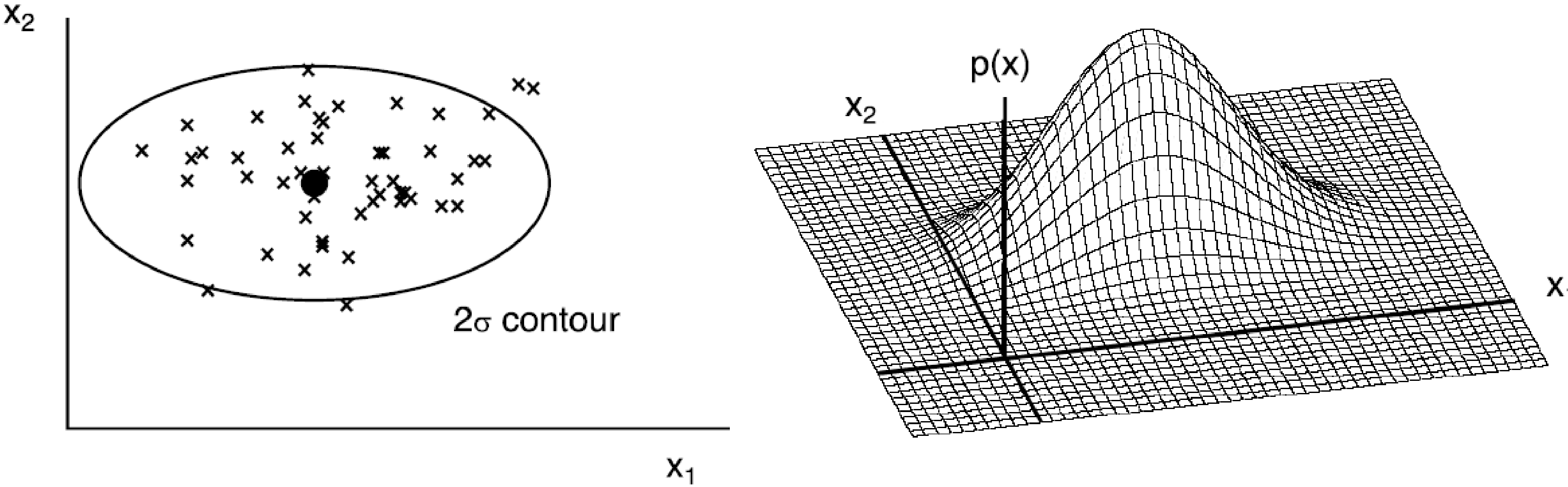
## ❖ 구형 공분산 가우시안 형태




(a) 구형 공분산 2차원 표현    (b) 구형 공분산 3차원 표현

[그림 4-13] 구형 공분산 가우시안 형태

### ❖ 대각 공분산 가우시안 형태

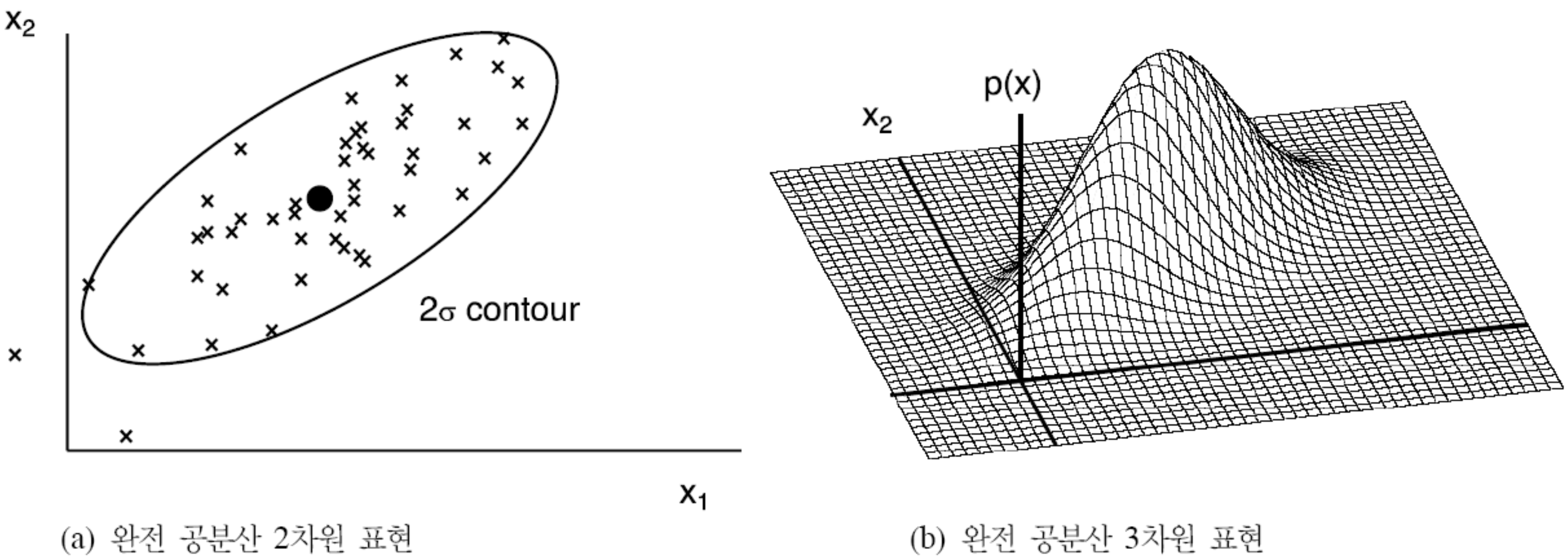



(a) 대각 공분산 2차원 표현

(b) 대각 공분산 3차원 표현

[그림 4-12] 대각 공분산 가우시안 형태

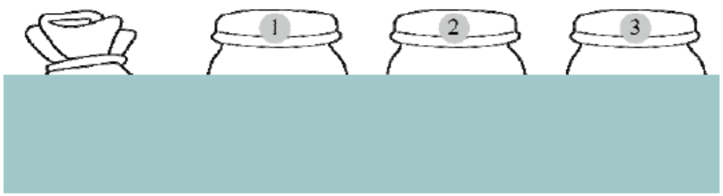
## ❖ 완전 공분산 가우시안 형태




[그림 4-11] 완전 공분산 가우시안 형태

## ■ 매개변수 Θ를 모르는 상황에서 매개변수를 추정하는 문제

(a) $\Theta = \{p_1, p_2\}$ (b) $\Theta = \{q_3\}$



(c) $\Theta = \{p_1, p_2, q_1, q_2, q_3\}$

**그림 2-17** 매개변수가 감추어진 여러 가지 상황

- **예) [그림 2-17(b)] 상황**

*(a)각 병이 뽑힐 확률추정 필요*

*(b)3번 병의 공 색깍의 확률분포 추정 필요*

*(c)모든 매개변수 추정 필요*

데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}

*반복실험에 의해 관찰된 결과*

"데이터 $\mathbb{X}$가 주어졌을 때, $\mathbb{X}$를 발생시켰을 가능성을 최대로 하는 매개변수 $\Theta = \{q_3\}$의 값을 찾아라."

*위의 (b) 경우 파란공이 뽑힐 확률, $q_3$를 추정하는 예제*

# 2.2.3 최대 우도

## 최대 우도법

- [그림 2-17(b)] 문제를 수식으로 쓰면,

$$\hat{q}_3 = \underset{q_3}{\operatorname{argmax}}\, P(\mathbb{X}|q_3) \qquad (2.31)$$

- 일반화 하면,

최대 우도 추정: $$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, P(\mathbb{X}|\Theta) \qquad (2.32)$$

- 수치 문제를 피하기 위해 로그 표현으로 바꾸면,

데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}

최대 로그우도 추정: $$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, \log P(\mathbb{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta) \qquad (2.34)$$

*Maximum Log Likelihood*

X에서 관찰된 공 색깔들을 이용하여
첫 번째 매개변수의 확률분포로 확률계산 후
확률을 높일 수 있는 방향으로 매개변수를
반복적으로 추정하여 최대의 확률을 보이는
매개변수를 최종적으로 추정함.

# 2.2.6 정보이론

- **메시지가 지닌 정보를 수량화할 수 있나?**
  - "고비 사막에 눈이 왔다"와 "대관령에 눈이 왔다"라는 두 메시지 중 어느 것이 더 많은 정보의 가치를 가지나?
  - 정보이론의 기본 원리 → 확률이 작을수록 가치있는(또는 많은) 정보

- **자기 정보self information**
  - 사건(메시지) $e_i$의 정보량 (단위: 비트 또는 나츠)

$$h(e_i) = -\log_2 P(e_i) \quad \text{또는} \quad h(e_i) = -\log_e P(e_i) \tag{2.44}$$

- **엔트로피**
  - 확률변수 $x$의 확률분포가 가지고 있는 불확실성을 나타내는 엔트로피

$$\text{이산 확률분포} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i) \quad \text{또는} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_e P(e_i) \tag{2.45}$$

$$\text{연속 확률분포} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x) \quad \text{또는} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_e P(x) \tag{2.46}$$

## ■ 자기 정보와 엔트로피 예제

**예제 2-8**

윷을 나타내는 확률변수를 $x$라 할 때 $x$의 엔트로피는 다음과 같다.

$$H(x) = -\left(\frac{4}{16}\log_2\frac{4}{16} + \frac{6}{16}\log_2\frac{6}{16} + \frac{4}{16}\log_2\frac{4}{16} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16}\right) = 2.0306\text{비트}$$

주사위는 눈이 6개인데 모두 1/6이라는 균일한 확률을 가진다. 이 경우 엔트로피를 계산하면 다음과 같다.

$$H(x) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 2.585\text{ 비트}$$

- **주사위가 윷보다 엔트로피가 높은 이유는?**

주사위의 숫자는 모든 같은 확률을 가지고 있어서 어떤
숫자가 나올지 조금의 예측도 불가능.

**즉 모든 사건이 발생할 확률이 동일하면 엔트로피가 MAX**

*혹시 주사위가 윷놀이보다*
*경우의 수가 많아서*
*그렇지도?*
***2.3219로 여전히 높다.***

## ■ 교차 엔트로피cross entropy

- 두 확률분포 $P$와 $Q$ 사이의 교차 엔트로피

$$H(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,k} P(e_i)\log_2 Q(e_i) \qquad (2.47)$$

- 식을 전개하면,

$$\begin{aligned} H(P,Q) &= -\sum_{x} P(x)\log_2 Q(x) \\ &= -\sum_{x} P(x)\log_2 P(x) + \sum_{x} P(x)\log_2 P(x) - \sum_{x} P(x)\log_2 Q(x) \\ &= H(P) + \underbrace{\sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)}}_{\text{KL 다이버전스}} \end{aligned}$$

$$H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i)$$

## KL 다이버전스

- 식 (2.48)은 $P$와 $Q$ 사이의 KL 다이버전스
- 두 확률분포 사이의 거리를 계산할 때 주로 사용

두 구름 속의 입자들의 거리가 아닌
구름들간의 거리는 어떻게 측정?

구름의 중심부간의 거리계산으로 충분?

$$KL(P \parallel Q) = \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (2.48)$$

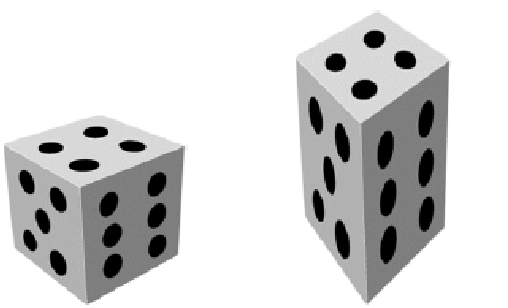
## 교차 엔트로피와 KL 다이버전스의 관계

$$\begin{aligned} P\text{와 } Q\text{의 교차 엔트로피 } H(P,Q) &= H(P) + \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (2.49) \\ &= P\text{의 엔트로피} + P\text{와 } Q \text{ 간의 } KL \text{ 다이버전스} \end{aligned}$$

**예제 2-9**

[그림 2-21]과 같이 정상적인 주사위와 찌그러진 주사위가 있는데, 정상적인 주사위의 확률분포는 $P$, 찌그러진 주사위의 확률분포는 $Q$를 따르며, $P$와 $Q$가 다음과 같이 분포한다고 가정하자.

$$P(1)=\frac{1}{6},\ P(2)=\frac{1}{6},\ P(3)=\frac{1}{6},\ P(4)=\frac{1}{6},\ P(5)=\frac{1}{6},\ P(6)=\frac{1}{6}$$
$$Q(1)=\frac{3}{12},\ Q(2)=\frac{1}{12},\ Q(3)=\frac{1}{12},\ Q(4)=\frac{1}{12},\ Q(5)=\frac{3}{12},\ Q(6)=\frac{3}{12}$$



(a) 정상 주사위 (b) 찌그러진 주사위

**그림 2-21 확률분포가 다른 두 주사위**

$$H(P,Q) = -\sum_x P(x)\log_2 Q(x)$$

확률분포 $P$와 $Q$ 사이의 교차 엔트로피와 KL 다이버전스는 다음과 같다.

$$H(P,Q) = -\left(\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{3}{12}\right) = 2.7925$$
$$KL(P \parallel Q) = \frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2\frac{2}{3} = 0.2075$$

# 과제 1 Feedback

## 좋은 관찰 예시 (하지만 기계학습을 위해서는 부족함.)

**- 귀 모양과 위치**

고양이는 뾰족한 귀를 가지고 있으며, 대체로 머리의 상단에 위치한다. 반면, 개의 귀는 종류에 따라 달라지지만, 일반적으로 고양이보다는 크고 때로는 처져 있다.

**- 코의 구조**

고양이의 코는 끝이 뾰족하고 작은 편이며, 개의 코는 보통 더 크고, 끝이 둥글다.

**- 수염의 길이와 분포**

고양이의 수염은 매우 길고, 얼굴 주변에 집중되어 있다. 개는 수염이 있긴 하지만, 고양이만큼 길지 않을 수 있다.

**- 눈의 모양과 크기**

고양이의 눈은 크고, 수직으로 긴 동공을 가지며, 개는 동공이 둥글고 눈이 고양이보다 상대적으로 작을 수 있다.

**- 턱과 입의 구조**

고양이는 작고 덜 발달된 아래턱을 가지고 있는 반면, 개는 보통 더 발달된 아래턱과 큰 입을 가지고 있다.

# 과제 1 Feedback

## 학습하기 좋은 예시

- 얼굴 면적에 비해 눈과 동공의 크기가 크다.
- 얼굴 면적에 비해 수염이 긴 편이다.
- 얼굴 면적에 비해 코가 작은 편이다.
- 얼굴면적에 비해 눈과 동공의 크기가 크다.
- 눈의 폭이 얼굴 폭의 1/3 이상을 차지한다.
- 코의 크기가 얼굴의 1/4 미만을 차지한다.
- 얼굴의 가로 폭이 세로 폭보다 길다.

# 과제 1 Feedback

- 동공의 형태: 모든 고양이들의 동공이 타원의 형태
- 귀의 형태: 고양이들의 귀가 삼각형 형태의 뾰족
- 입 모양의 형태: 고양이들의 입의 모양이 코로부터 이어져 'ㅅ' 형태
- 코와 입 사이의 털: 고양이들의 코와 입 사이에 양 쪽으로 뻗은 털
- 코의 형태: 고양이들의 코가 역삼각형 형태
- 안구의 위치: 고양이는 안구의 위치가 두개골의 중앙에 가깝게 위치.
- 눈동자와 코 크기의 비율: 코가 눈동자보다 항상 작다.
- 귀 끝이 향하는 방향: 개는 시계 12시 방향을, 고양이는 양쪽 끝이 10시와 2시 방향
- 주둥이가 튀어나온 정도: 주둥이가 얼굴로부터 튀어나온 정도. 고양이보다 개가 길다
- 눈꼬리 방향: 고양이는 위로 올라가 있는 반면, 개는 쳐져 있다.