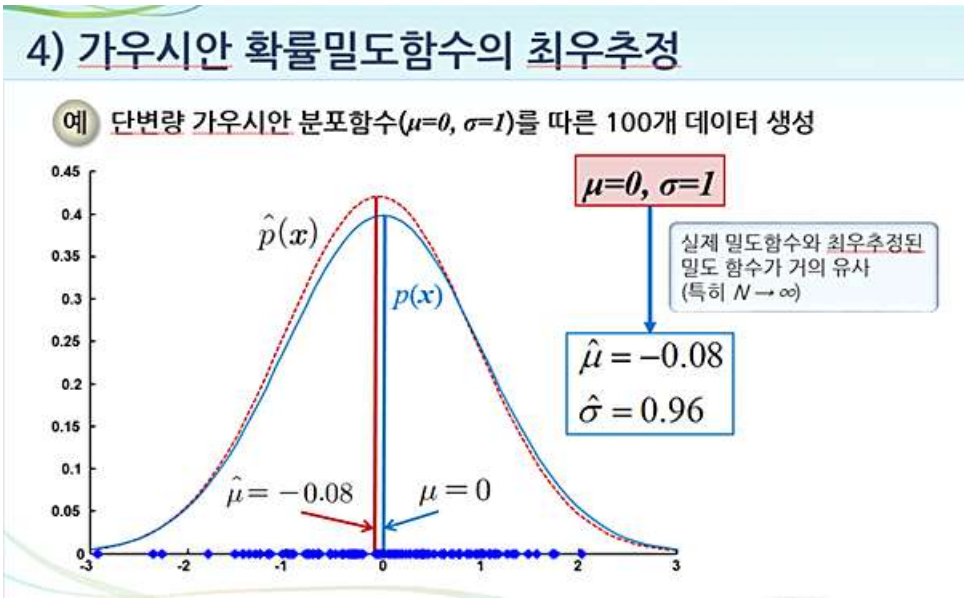
3강. 데이터의 통계적 분석

※ 점검하기

Q1. 두 클래스 집합 C1, C2는 각각 다음과 같은 평균과 공분산을 가지는 가우시안 분포를 따른다.
매트랩을 이용하여 각 클래스별로 데이터를 100개씩 생성하시오.
생성된 데이터를 2차원 평면상의 점의 위치에 대해 설명하시오.

$$\mu_1 = [0,0]^T,\ \mu_2 = [4,4]^T \qquad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

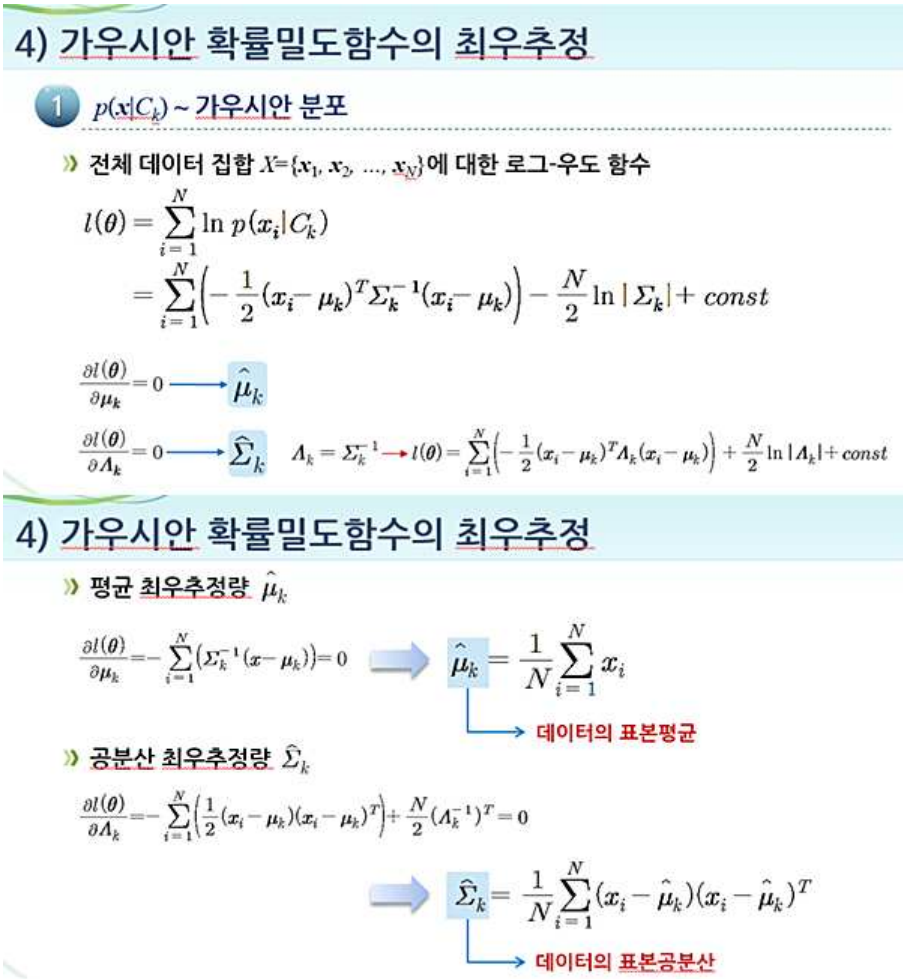
〈관련학습보기〉



교재 2장 [프로그램 2-1, 데이터 생성]에서 데이터 개수(N),
평균(m1,m2)와 공분산(s1,s2) 부분을 조정하면 된다.
[참조] 2. 모수적 확률밀도 추정의 「4) 가우시안 확률밀도함수의 최우추정」

Q2. 각 클래스의 데이터 분포가 가우시안 함수를 따른다는 가정 하에, 1번에서 생성한 데이터 집합을 이용하여 각 클래스의 확률밀도함수 $p(x|c_i)$의 μ_1 와 $\sum_1$를 추정하시오.

〈관련학습보기〉



4) 가우시안 확률밀도함수의 최우추정

1 $p(x|C_k)$ ~ 가우시안 분포

» 전체 데이터 집합 $X=\{x_1, x_2, \ldots, x_N\}$에 대한 로그-우도 함수

$$l(\theta) = \sum_{i=1}^{N} \ln p(x_i|C_k)$$
$$= \sum_{i=1}^{N}\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right) - \frac{N}{2}\ln|\Sigma_k| + const$$

$\frac{\partial l(\theta)}{\partial \mu_k}=0 \rightarrow \hat{\mu}_k$

$\frac{\partial l(\theta)}{\partial \Lambda_k}=0 \rightarrow \hat{\Sigma}_k$ $\quad \Lambda_k=\Sigma_k^{-1} \rightarrow l(\theta)=\sum_{i=1}^{N}\left(-\frac{1}{2}(x_i-\mu_k)^T\Lambda_k(x_i-\mu_k)\right)+\frac{N}{2}\ln|\Lambda_k|+const$

4) 가우시안 확률밀도함수의 최우추정

» 평균 최우추정량 $\hat{\mu}_k$

$$\frac{\partial l(\theta)}{\partial \mu_k} = -\sum_{i=1}^{N}(\Sigma_k^{-1}(x-\mu_k)) = 0 \Rightarrow \hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i$$

→ 데이터의 표본평균

» 공분산 최우추정량 $\hat{\Sigma}_k$

$$\frac{\partial l(\theta)}{\partial \Lambda_k} = -\sum_{i=1}^{N}\left\{\frac{1}{2}(x_i-\mu_k)(x_i-\mu_k)^T\right\} + \frac{N}{2}(\Lambda_k^{-1})^T = 0$$

$$\Rightarrow \hat{\Sigma}_k = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T$$

→ 데이터의 표본공분산

교재 2장 [프로그램 2-2, 데이터 분석]를 이용하여 집합의 표본평균과 표본공분산을 계산한다.

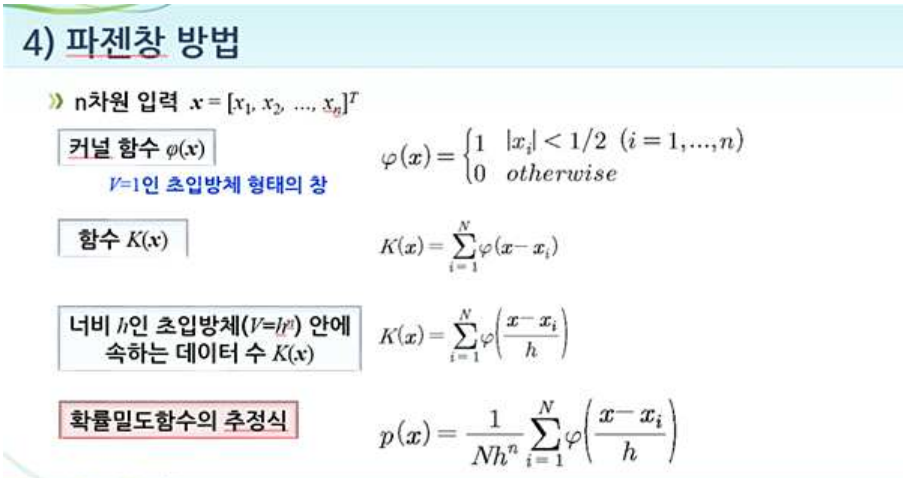
[참조] 2. 모수적 확률밀도 추정의 「4) 가우시안 확률밀도함수의 최우추정」

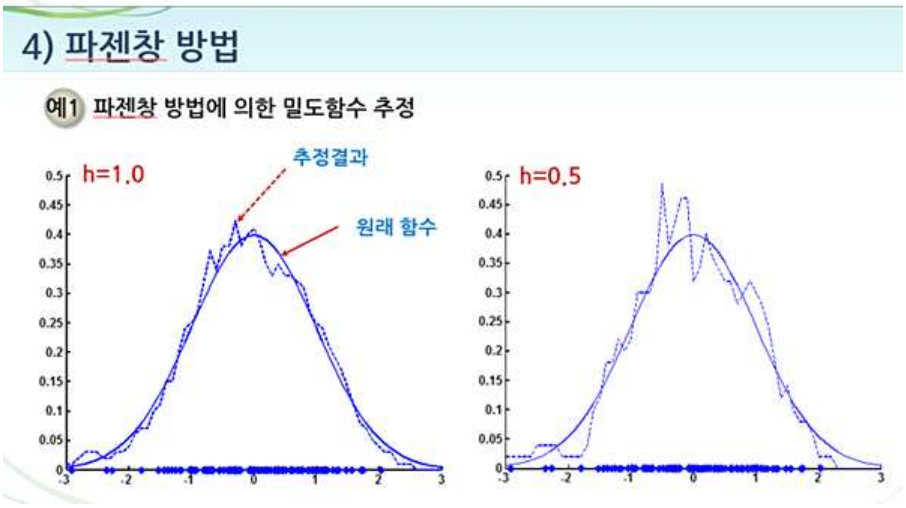
Q3. 1번에서 사용한 데이터에 대해, 커널 밀도함수 추정법에 의해 밀도함수를 추정하고자 한다. 다음을 각각 수행하시오.

(1) 파젠창 방법을 사용하여 밀도함수를 추정하되, 값을 여러 가지로 변형시켜 보면서 얻어지는 결과를 비교해 보시오.

(2) 가우시안 커널 방법을 사용하여 밀도함수를 추정하되, 값을 여러 가지로 변형시켜 보면서 얻어지는 결과를 비교해 보시오.

〈관련학습보기〉





[프로그램 3-2, 파젠창에 의한 밀도함수 추정] 활용
[참조] 3. 비모수적 확률밀도 추정 「4) 파젠창 방법」

5) 가우시안 커널 방법

1 보다 일반적인 커널 함수로의 확장

» 파젠창 방법의 불연속적인 문제 해결

2 새로운 커널 함수가 만족해야 하는 조건

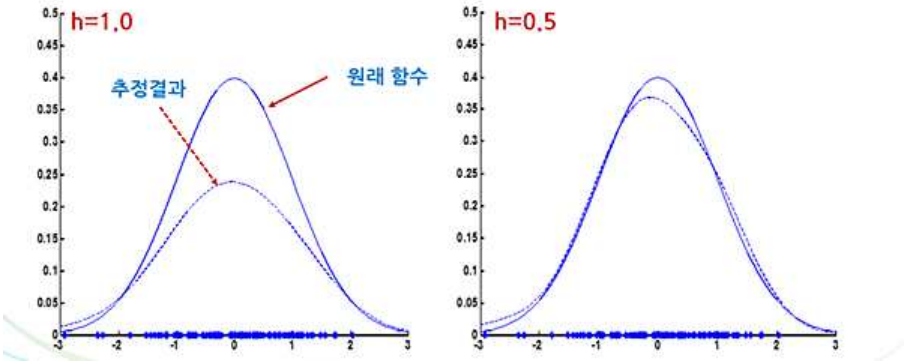
$$\varphi(x) \geq 0, \quad \int \varphi(x)dx = 1$$

» 가우시안 밀도함수 → 「가우시안 커널 방법」 $\varphi(x) = \frac{1}{\sqrt{2\pi^n}} \exp\left\{-\frac{1}{2}x^T x\right\}$

확률밀도함수의 추정식 $p(x) = \frac{1}{Nh^n}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi^n}} exp\left(-\frac{1}{2}\frac{(x-x_i)^T(x-x_i)}{h^2}\right)$

5) 가우시안 커널 방법

예1 가우시안 커널 방법에 의한 밀도함수 추정



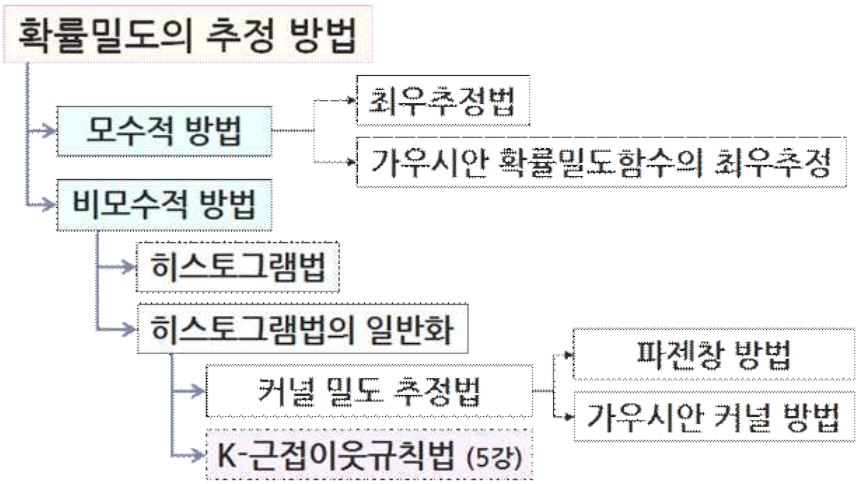
[프로그램 3-3, 가우시안 커널에 의한 밀도함수 추정] 활용
[참조] 3. 비모수적 확률밀도 추정 「5) 가우시안 커널 방법」

※ 정리하기

1. 데이터의 확률분포와 패턴인식

1) 통계적 패턴인식 처리 과정에서 「학습」이란 주어진 데이터로부터 각 클래스별 확률밀도함수 $p(x \mid C_k)$를 추정하는 것을 의미하며, 「인식/분류」라는 것은 새로운 데이터 x_{new}가 주어졌을 때 이것이 각 클래스 $C_k(k = 1, \ldots, M)$로부터 관찰될 확률 $P(C_k \mid x_{new})$을 계산하여 확률값이 가장 큰 클래스로 할당하는 것을 의미함

2. 모수적·비모수적 확률밀도 추정



1) 모수적 방법에서는 확률 모델을 미리 가정한 후, 데이터를 이용해서 이에 필요한 파라미터들을 추정하여 구체적인 확률밀도를 얻지만, 비모수적 방법에서는 특정 확률 모델을 가정하지 않기 때문에 추정해야 할 파라미터도 따로 존재하지 않고, 오로지 주어진 데이터 집합을 이용하여 직접 밀도함수를 추정하는 방법임

2) 최우추정법
: 주어진 데이터 집합이 관찰된 가능성, 즉 우도를 최대로 하는 파라미터를 찾아 추정치로 정하는 방법이며, 가우시안 확률밀도함수의 최우추정량은 데이터의 표본평균과 표분공분산으로 쉽게 계산됨

3) 비모수 밀도 추정 방법
: 부피 V를 고정하고 밀도함수를 계산한 커널 밀도함수 추정법과 한 영역에 들어가는 데이터의 수 K를 고정하고 밀도함수를 계산하는 K-NNR 방법이 있음

4) 대표적인 커널 밀도 추정 방법
: 초입방체 형태의 창을 사용하는 파젠창 방법은 찾아진 확률밀도함수의 불연속적인 문제가 있으며, 이러한 문제를 해결하기 위한 다른 방법으로 입방체 커널 대신 연속함수로 정의되는 가우시안 함수를 사용하는 가우시안 커널 방법이 있음