Chapter 03 데이터의 통계적 분석

[학습목표]
이 장에서는 패턴인식의 첫 단계인 데이터의 분석 방법에 대하여 알아본다. 데이터의 통계적 분포 특성을 표현하기 위해서 사용되는 통계량과 확률분포에 대하여 간단히 알아보고, 데이터 집합으로부터 확률분포를 추정하기 위한 방법으로 모수적 방법과 비모수적 방법에 대하여 알아본다.

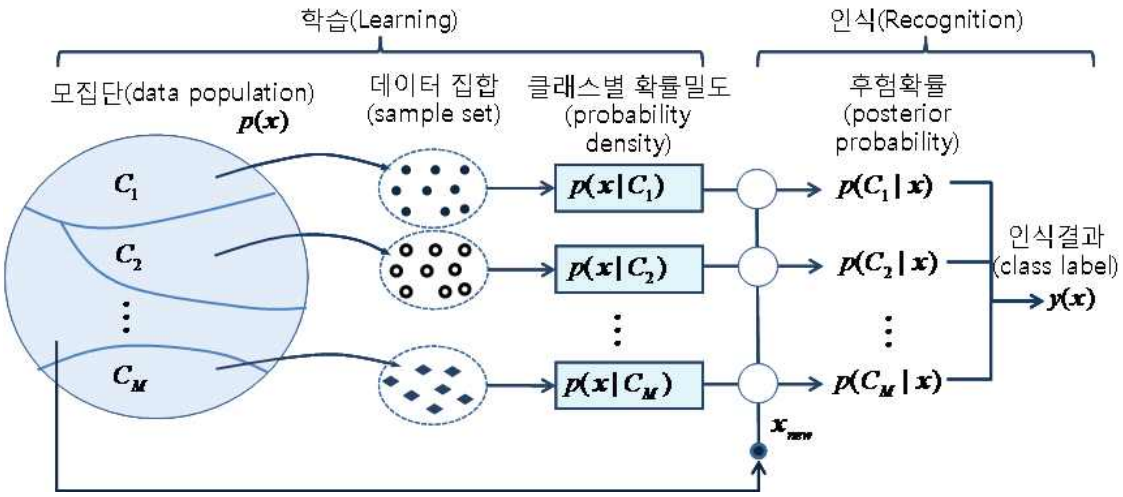
3. 데이터의 통계적 분석

3.1. 데이터의 확률분포와 패턴인식

패턴인식을 위한 통계적 접근법의 첫 단계는 데이터의 통계적 분포 특성을 분석하는 것이다. 이 장에서는 분류 문제에 초점을 맞추어, 각 클래스의 통계적 특성을 분석하고 확률분포를 추정하는 방법에 대하여 알아보겠다.

통계적 접근법에서는 우리가 실생활에서 관찰하는 데이터들에 대해, 그 기저에 존재하는 모집단이 있으며, 현재 관찰된 데이터는 그로부터 추출된 샘플이라고 본다. 따라서 현재 주어진 데이터에 대해 분석한다는 것은, 그 모집단에 대하여 확률분포 모델을 세우고 데이터를 이용하여 추정해 내는 과정이 된다. 예를 들어 얼굴영상 데이터의 경우, 모든 가능한 얼굴들과 그 변형을 포함하고 있는 모집단이 존재한다고 보고, 그 확률적 분포 특성을 설명할 수 있는 모델을 정의한다. 이를 수학적으로 표현하면, 하나의 얼굴 영상이 벡터 $\boldsymbol{x}$로 나타나고, 그 모집단의 확률분포 모델은 밀도함수 $p(\boldsymbol{x})$로 정의할 수 있다. $p(\boldsymbol{x})$의 구체적인 형태는 데이터의 성격에 맞추어 정의할 수 있는데, 예를 들면 얼굴영상 데이터의 경우, 모든 얼굴의 원형이 되는 평균얼굴($\boldsymbol{m}$)이 하나 존재하여, 그 평균을 중심으로 고르게 원형으로 분포되어 있다고 가정한다면, $p(\boldsymbol{x})$는 $\boldsymbol{m}$을 평균으로 가지는 가우시안 분포로 정의할 수 있을 것이다. 이 때 평균의 구체적인 값은 알지 못하므로, 관찰된 학습데이터를 통하여 추정하는 과정이 필요한데, 이것이 이 장에서 논의하는 데이터의 통계적 분석 과정이라고 볼 수 있다.

문제를 조금 더 발전시켜 분류 문제의 경우를 생각하면, 서로 다른 M개의 클래스가 주어지므로 각 클래스에 속한 데이터들이 각각 서로 다른 분포를 따른다고 가정하는 것이 타당할 것이다. 이를 표현하기 위하여 클래스 C_k에 대한 데이터 $\boldsymbol{x}$의 조건부 확률 $p(\boldsymbol{x}|C_k)$을 생각한다. 즉, 전체 데이터 집합에 대한 확률분포 $p(\boldsymbol{x})$을 생각하는 대신에 클래스별로 데이터 집합을 나누어 각각의 분포를 생각하는 것이다. 통계적 패턴인식에서는 우선 각 클래스에 대한 확률분포 모델을 정의하고, 이를 추정하게 되는데, 이것이 1장에서 배운 패턴인식의 처리과정에 있어서의 "학습" 단계에 해당된다. 앞서 2장에서는 각각의 클래스 집합에 대하여 정확한 확률분포 모델을 정의하는 대신 간단히 평균과 공분산이라는 통계량만을 추정하여 사용하였다. 이 장에서는 명확한 확률밀도함수를 정의하고 추정하는 방법에 대하여 논의하게 될 것이다. 학습 단계를 통해 각 클래스별 확률밀도함수 $p(\boldsymbol{x}|C_k)$가 추정되면, 이를 어떻게 분류에 활용할 것인지 생각해 보아야 할 것이다. 확률론을 바탕으로 분류를 수행한다는 것은, 결국 새로운 데이터 $\boldsymbol{x}_{new}$가 주어졌을 때, 이것이 각 클래스 $C_k(k=1,...,M)$로부터 관찰될 확률, 즉 $P(C_k|\boldsymbol{x}_{new})$을 계산하여 확률값이 가장 큰 클래스로 분류한다는 것을 의미한다. 이것이 패턴인식 처리과정에 있어서의 "인식/분류"단계에 해당한다. 이에 대한 전체적인 개념을 [그림 3-1]에 나타내었다. [그림 3-1]에서는 [그림 1-4]에 나타낸 전처리와 특징추출 과정은 생략하고, 학습과 인식 과정을 중심으로 나타내었다.



[그림 3-1] 통계적 패턴인식의 학습과 인식

이 절에서는 특히 학습단계, 즉 관찰된 학습데이터를 이용하여 클래스별 확률밀도함수 $p(\boldsymbol{x}|C_k)$을 추정하는 방법에 대하여 논의할 것이다. 그러나 그에 앞서 분류단계에 대해 조금 더 생각해 보자. 우리는 학습 단계에서 데이터를 이용하여 추정된 클래스별 확률밀도함수 $p(\boldsymbol{x}|C_k)$을 얻게 되므로, 이를 사용하여 분류의 기준이 되는 확률값 $P(C_k|\boldsymbol{x}_{new})$을 계산해 내는 과정이 필요하다. 이를 위해서 부록 B에서 소개된 베이즈 정리를 활용하면, 다음과 같이 간단히 표현할 수 있다.

$$P(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x})} \qquad \text{[식 3-1]}$$

여기서 $P(C_k), p(\boldsymbol{x})$ 등 구체적인 계산 방법에 대하여서는 4장의 베이즈 분류기에서 설명할 것이며, 이 장에서는 [식 3-1]의 핵심부분인 클래스별 분포 $p(\boldsymbol{x}|C_k)$을 추정하는 방법에 대하여 설명할 것이다.

한편, 이상에서 설명한 바와 같이 $p(\boldsymbol{x}|C_k)$을 먼저 추정하여 이로부터 $P(C_k|\boldsymbol{x}_{new})$을 계산하는 대신, 분류에 궁극적으로 필요한 값인 $P(C_k|\boldsymbol{x}_{new})$을 바로 추정하여 사용하는 방법도 생각해 볼 수 있을 것이다. 패턴인식 분야에서는 전자와 같이 $p(\boldsymbol{x}|C_k)$을 이용하는 방법을 <생성적 접근법 (generative approach)>이라고 하고, 후자와 같이 $P(C_k|\boldsymbol{x}_{new})$을 바로 이용하는 방법을 <식별적 접근법 (discriminative approach)>이라고 한다. 생성적 접근법의 경우, 관찰된 데이터들이 어떻게 생성되었는지를 생각하여 각 클래스별로 그러한 생성 시스템을 설명할 수 있는 확률분포를 먼저 추정하는 반면, 식별적 접근법에서는 이러한 고려 없이 바로 분류에 필요한 확률값 혹은 그와 유사한 판별함수의 값을 추정하여 사용한다. 후자의 경우는 간단히 판별에 기준이 되는 값만을 학습하므로 그 과정이 간단하다. 또한 클래스별 확률밀도가 매우 복잡한 경우 그를 추정할 필요가 없으므로 확률분포의 복잡도에 영향을 받지 않는다. 반대로 전자의 경우에는 클래스별 확률밀도가 매우 복잡한 경우에 그 추정 과정에서 발생하는 오차가 최종적인 분류 결과에 직접적인 영향을 주게 된다. 그러나 추정한 클래

스별 분포를 이용하면 단순히 분류를 위한 판단 기준 이외에도 데이터가 가지는 다양한 분포 특성을 설명할 수 있고, 분류 결과에 대한 보다 자세한 설명을 할 수 있을 뿐 아니라 이를 바탕으로 성능 향상을 위한 방법의 개선에도 활용할 수 있다.
이 장에서는 생성적 접근법을 위하여 각 클래스별 확률밀도함수 $p(\boldsymbol{x}|C_k)$을 추정하는 방법에 대하여 설명할 것이며, 이를 바탕으로 한 분류기를 4장과 5장에서 설명할 것이다. 6장에서는 식별적 접근법에 의한 분류기를 소개할 것이다.

3.2 모수적 확률밀도 추정

<모수적 확률밀도 추정법 (parametric density estimation)>은, 추정하고자 하는 확률분포가 어떤 분포 형태를 가지고 있는지를 먼저 가정하고, 그 형태를 정의하는데 사용하는 파라미터의 값을 데이터로부터 추정함으로서 구체적인 확률분포 함수를 얻는 방법이다.
예를 들어 하나의 클래스에 속하는 데이터 집합의 확률분포를 알고 싶은 경우, 먼저 이 데이터가 가우시안 분포 형태를 따른다고 가정해 볼 수 있다. 이렇게 가정하면 우리는 n차원 데이터 $\boldsymbol{x}$의 확률밀도함수 $p(\boldsymbol{x}|C_k)$의 함수 형태를 다음과 같이 나타낼 수 있다.

$$p(\boldsymbol{x}|C_k) = G(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{2\pi^n}\sqrt{|\boldsymbol{\Sigma}_k|}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right] \qquad \text{[식 3-2]}$$

그러나 이 함수식은 아직 명시적인 하나의 함수를 나타내는 것이 아니다. 왜냐하면 이 함수식에서는 명시적인 함수 형태를 결정하는 파라미터 $\boldsymbol{\mu}_k$와 $\boldsymbol{\Sigma}_k$가 들어 있어서 이 값이 변하면 그 형태도 변하게 된다. 따라서 우리는 데이터 집합을 이용하여 파라미터 값을 추정해 줄 필요가 있다. 이렇게 파라미터 추정을 통하여 최종적인 확률밀도함수를 찾는 방법을 파라미터 추정법이라고 한다. 다음 절에서는 데이터를 이용하여 파라미터를 추정하는 방법을 알아본다.

3.2.1 최우추정법

파라미터를 추정하기 위해서 우리는 주어진 데이터 집합을 사용한다. 이 데이터 집합이 관찰될 가능성, 즉 우도(likelihood)를 최대로 하는 파라미터를 찾아 추정치로 정하는 것을 <최우추정법 (Maximum likelihood estimation)>이라고 한다.
확률변수 $\boldsymbol{X}$가 파라미터 $\boldsymbol{\theta}$를 가진 확률밀도함수 $p(\boldsymbol{x};\boldsymbol{\theta})$를 따른다고 가정하고, 이를 추정하는데 사용할 데이터 집합이 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 으로 주어졌다고 하자. 우리의 목적은 이 데이터를 가장 잘 설명할 수 있는 파라미터 $\boldsymbol{\theta}$를 찾는 것이므로, 먼저 주어진 데이터들이 관찰될 가능성, 즉 우도를 생각해 볼 수 있다. 각각의 데이터가 서로 독립적으로 얻어진 것이라면 전체 데이터 집합에 대한 우도는 다음과 같이 주어진다.

$$p(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N;\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i;\boldsymbol{\theta}) \qquad \text{[식 3-3]}$$

이 우도값을 최대로 하는 파라미터 $\boldsymbol{\theta}$를 찾아 추정값으로 사용하는 방법이 바로 최우추정법이다. 그런데 우도는 0에서 1사이의 값을 가지는 확률값을 데이터의 개수만큼 곱해서 계산되므로, 데이터 수가 많은 경우에는 그 값이 지나치게 작아지는 문제가 발생한다. 따라서 우도값을 바로 사용하는 대신 반복된 곱셈에 의한 오차를 줄이기 위해 로그함수를 적용한 로그-우도 (log-likelihood)를 사용하게 되는데 이것을 식으로 표현하면 다음과 같다.

$$l(\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^{N} p(\boldsymbol{x}_i;\boldsymbol{\theta})\right) = \sum_{i=1}^{N} \ln\, p(\boldsymbol{x}_i;\boldsymbol{\theta}) \qquad \text{[식 3-4]}$$

로그함수는 단조증가함수 (monotonic increasing function)이므로 우리가 얻고자 하는 파라미터의 추정치 $\hat{\boldsymbol{\theta}}_{MLE}$는 로그-우도 함수 $l(\boldsymbol{\theta})$를 최대로 하는 값이 되어 다음과 같이 쓸 수 있다.

$$\hat{\boldsymbol{\theta}}_{MLE} = \mathrm{argmax}_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta}) \qquad \text{[식 3-5]}$$

이 추정치 $\hat{\boldsymbol{\theta}}_{MLE}$를 <최우추정량(maximum likelihood estimator, MLE)>이라고 한다. 구체적인 값은 $\partial l/\partial\boldsymbol{\theta} = \mathbf{0}$를 만족하는 파라미터 값을 찾음으로써 얻어질 수 있다. 로그-우도 함수 $l(\boldsymbol{\theta})$가 $\boldsymbol{\theta}$에 대한 이차식으로 주어진 경우 미분함으로써 얻어지는 연립일차방정식을 풀어서 $\boldsymbol{\theta}$의 추정치를 계산해 낼 수 있다. 그러나 일반적으로 $l(\boldsymbol{\theta})$는 복잡한 비선형 함수인 경우가 많고, 이때에는 별도의 최적화 알고리즘을 적용할 필요가 있다.

3.2.2 가우시안 확률밀도함수의 최우추정

각 클래스에 대한 확률밀도함수가 가우시안 분포를 따른다고 가정할 수 있는 경우, 최우추정치 계산을 위한 편미분방정식 $\partial l/\partial\boldsymbol{\theta} = \mathbf{0}$은 연립일차방정식이 되어 해석적인 해를 얻을 수 있다. 각 클래스 C_k에 대한 확률밀도함수가 [식 3-2]와 같이 가우시안 밀도함수로 주어진다고 가정하면 하나의 데이터 $\boldsymbol{x}$에 대한 로그-우도는 다음과 같이 얻어진다.

$$\begin{aligned} \ln p(\boldsymbol{x}|C_k) &= \ln\, G(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\ &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| \end{aligned} \qquad \text{[식 3-6]}$$

전체 데이터 집합 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 에 대한 로그-우도는 다음과 같다.

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln\, p(\boldsymbol{x}_i|C_k) = \sum_{i=1}^{N}\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right) - \frac{N}{2}\ln|\boldsymbol{\Sigma}_k| + const \qquad \text{[식 3-7]}$$

이 함수를 각각 파라미터 $\boldsymbol{\mu}_k$와 $\boldsymbol{\Sigma}_k$에 대해 미분하여 로그-우도를 최대로 하는 파라미터를 찾으면 최우추정량을 얻을 수 있다.
먼저 [식 3-7]을 평균 $\boldsymbol{\mu}_k$에 대하여 미분하면 다음 식을 얻을 수 있다.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = -\sum_{i=1}^{N}\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right) = 0 \qquad \text{[식 3-8]}$$

이를 $\boldsymbol{\mu}_k$에 대하여 풀면 다음과 같은 평균에 대한 최우추정량을 얻게 된다.

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \qquad \text{[식 3-9]}$$

이어서 공분산행렬에 대한 추정량을 계산하기 위해, 공분산 행렬의 역행렬을 $\boldsymbol{\Lambda}_k = \boldsymbol{\Sigma}_k^{-1}$로 정의하여 [식 3-7]을 다시 쓰면 다음과 같다.

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right) + \frac{N}{2}\ln|\boldsymbol{\Lambda}_k| + const \qquad \text{[식 3-10]}$$

[식 3-10]을 행렬 $\boldsymbol{\Lambda}_k$에 대해 미분하면 다음 관계식을 얻는다.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}_k} = -\sum_{i=1}^{N}\left(\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\right) + \frac{N}{2}(\boldsymbol{\Lambda}_k^{-1})^T = 0 \qquad \text{[식 3-11]}$$

여기서 행렬 미분에 대한 다음 관계식이 사용되었다.

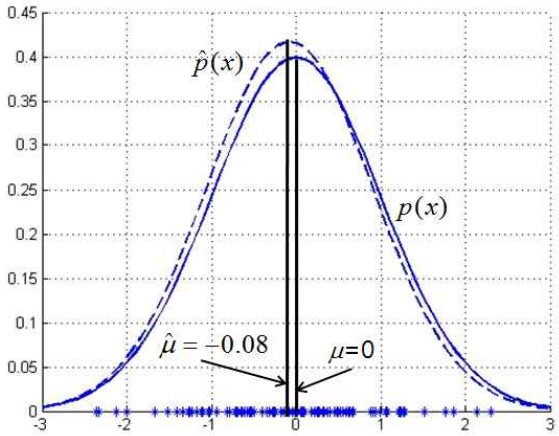
$$\frac{\partial}{\partial \boldsymbol{\Lambda}}\ln|\boldsymbol{\Lambda}| = (\boldsymbol{\Lambda}^{-1})^T \qquad \text{[식 3-12]}$$

[식 3-11]을 행렬 $\boldsymbol{\Sigma}_k$에 대해 정리하면 다음과 같이 파라미터의 최우추정량(maximum likelihood estimator) $\hat{\boldsymbol{\Sigma}}_k$를 얻을 수 있다.

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)^T \qquad \text{[식 3-13]}$$

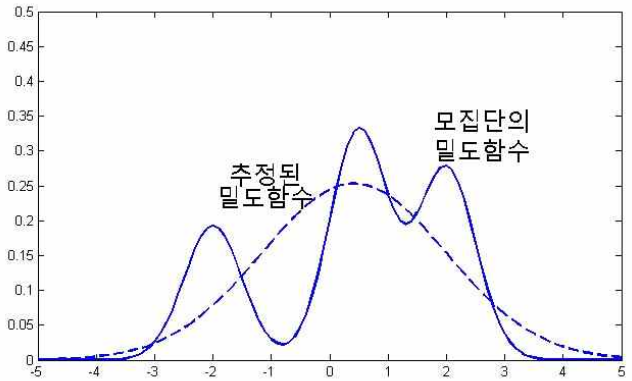
얻어진 파라미터의 값을 살펴보면 평균의 추정치는 데이터의 표본평균으로, 공분산의 추정치는 데이터의 표본공분산으로 쉽게 계산되는 값임을 알 수 있다.
간단한 예로 단변량 가우시안 분포함수를 따르는 데이터 집합을 이용하여 밀도함수를 추정해 보자. [그림 3-2]에 평균이 0이고 분산이 1인 가우시안 분포로부터 추출된 100개의 데이터를 직선상의 점으로 나타내고, 이 가우시안 확률밀도함수를 실선 그래프로 나타내었다. 이

어서 이 값들에 대하여 표본평균과 표본분산을 계산하면 그 추정값이 $\hat{\mu}=-0.08$, $\hat{\sigma}=0.96$로 계산되어, 이를 파라미터로 가지는 가우시안 밀도함수를 점선 그래프로 함께 표시하였다. 그림에서 확인할 수 있듯이 데이터를 생성하는데 사용된 실제 확률분포와 거의 유사한 확률분포를 찾아 낼 수 있음을 알 수 있다. 또한 그림에서 나타나는 오차는 파라미터 추정에 사용되는 데이터의 수를 늘림에 따라 점차 사라지고, 데이터의 수 N이 무한대에 가까워지면 추정 파라미터의 값도 실제 값에 수렴한다는 것이 알려져 있다.



[그림 3-2] 가우시안 확률분포함수에 대한 최우추정

그러나 최우추정법은 데이터의 분포에 대한 확률밀도함수를 미리 가정하고 시작하므로, 만약 데이터 분포에 대해 가정한 확률밀도함수 형태가 적절하지 못한 경우 추정한 파라미터도 의미를 갖지 못한다. [그림 3-2]에 그러한 경우의 예를 나타내었다. 원래 데이터 분포가 실선으로 나타낸 것과 같이 세 개의 언덕 모양으로 이루어진 경우에 이를 점선과 같이 가우시안 분포로 추정하게 되면, 그림에서 보는 바와 같이 평균과 분산만 같을 뿐 그 형태는 전혀 다른 분포를 가지게 된다. 이러한 문제점을 해결하기 위해서는 다음 절에서 설명한 비모수적 확률밀도 추정법을 사용하거나, 10장에서 소개하는 가우시안 혼합모델을 사용할 필요가 있다.



[그림 3-3] 실제 확률밀도함수와 가우시안 분포에 의해 최우추정된 확률밀도함수의 차이

3.3 비모수적 확률밀도 추정

앞서 언급한 바와 같이, 모수적 밀도 추정법은 확률 모델을 먼저 가정하고 시작해야한다는 단점이 있다. 만약 데이터 분포에 적합하지 않은 확률 모델을 가정하게 되면 아무리 좋은 파라미터를 추정하더라도 데이터 분포를 제대로 추정하기 어렵다. 이에 반하여, <비모수적 확률밀도 추정 (nonparametric density estimation)>에서는 특정한 확률 모델을 미리 가정하지 않는다. 따라서 추정해야할 파라미터가 따로 존재하지 않는다. 확률 모델을 미리 가정하지 않는다면, 어떻게 확률밀도함수를 표현할 수 있을지 생각해 보자. 우리가 밀도함수 추정에 사용할 수 있는 것은 오로지 관찰된 데이터 집합뿐이므로, 이 데이터 집합을 이용하여 밀도함수의 형태를 표현해야 한다. 이를 위한 가장 간단한 방법인 히스토그램법부터 시작하여 발전된 방법을 살펴보도록 하겠다.

3.3.1 히스토그램법

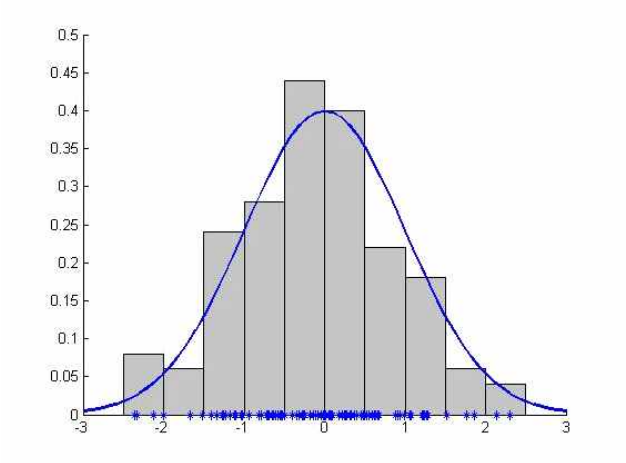
히스토그램 (histogram)이란 주어진 데이터가 가지는 값의 범위를 일정 간격의 구간들로 나누고, 각 구간에 존재하는 데이터 비율을 표현하는 막대그래프를 의미한다. [그림 3-1]에서 사용된 데이터에 대하여 히스토그램법을 적용하려면, 먼저 데이터가 존재하는 -2.5에서 2.5 사이의 구간을 0.5의 간격으로 10개 구간으로 나누고, 각 구간에 속하는 데이터가 몇 개인지를 세어 볼 수 있을 것이다. 주어진 데이터에서와 같이 0에서 0.5사이의 구간에 데이터가 20개 존재한다면 확률값 $P(0 \le x < 0.5) = 20/100 = 0.2$을 얻을 수 있다.
이를 일반화하면, 주어진 데이터 집합에 대한 히스토그램을 얻기 위해서는 먼저 히스토그램의 시작점 x_0과 각 구간의 간격, 즉 막대-폭(bin-width) h을 결정해 주어야 한다. 이것이 결정되면 i번째 구간 $(x_0+(i-1)h \le x < x_0+ih)$의 확률값은 다음과 같이 표현할 수 있다.

$$\Pr\left[x_0+(i-1)h \le x < x_0+ih)\right] = \frac{K(x \in [x_0+(i-1)h, x_0+ih))}{N} \qquad \text{[식 3-14]}$$

이 식에서 N은 전체 데이터의 수이고, $K(x \in [x_0+(i-1)h, x_0+ih))$은 해당 구간에 속하는 데이터의 수를 나타낸다. 그런데 우리는 개개의 확률값이 아닌 확률밀도함수를 얻고자 한다. 확률밀도함수를 해당 영역에서 적분한 값이 그 영역에 대한 확률값이 되어야 하므로, 막대 모양의 직사각영역을 사용하는 히스토그램법에서 각 구간별로 정의되는 불연속적 확률밀도함수는 다음과 같이 쓸 수 있다.

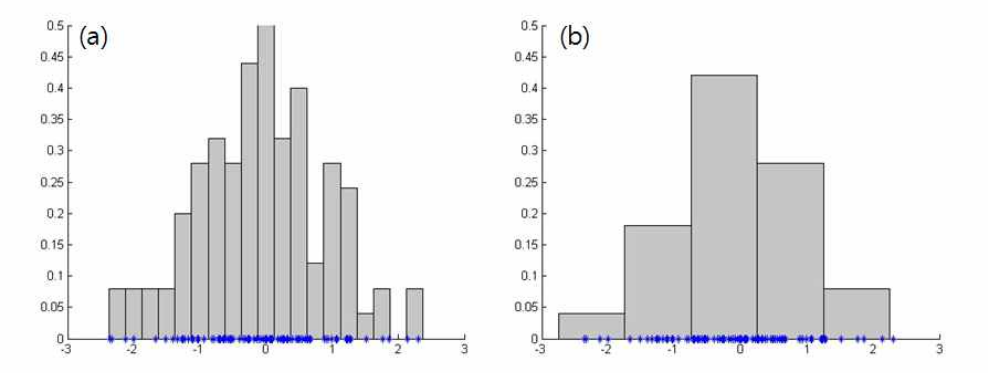
$$p_{(x_0+(i-1)h \le x < x_0+ih)} = \frac{1}{h}\frac{N(x \in [x_0+(i-1)h, x_0+ih))}{N} \qquad \text{[식 3-15]}$$

[그림 3-1]에서 사용된 데이터에 대하여 히스토그램법을 적용하여 찾아진 확률밀도함수를 [그림 3-4]에 나타내었다. 여기서 -2.5에서 2.5까지의 영역을 10개 구간으로 나누었으므로 막대-폭 h는 0.5가 된다. 각 구간에 속하는 데이터 수를 계산하여 [식 3-15]에 의해 전체 데이터 수 100과 막대-폭 0.5로 나누어 확률밀도함수 값을 얻었다. 그림에서 실선으로 나타난 실제 확률밀도함수를 대체로 잘 근사하고 있음을 볼 수 있다.



[그림 3-4] 히스토그램법에 의한 밀도함수 추정 예

[그림 3-5]에서는 막대-폭을 변화시켰을 때 얻어지는 추정 결과를 비교하였다. [그림 3-4a]의 경우와 같이 $h=0.25$로 줄이면 샘플수의 변화에 민감한 형태를 가지게 되고, 반대로 [그림 3-4b]의 경우와 같이 $h=1.0$로 늘어나면 지나치게 퍼져서 형태를 알아보기 힘들게 된다. 따라서 히스토그램법에 있어서는 적절한 h값을 찾는 것이 추정 성능에 큰 영향을 미친다. 이 밖에도 히스토그램법은 입력차원이 커지면 빈도수를 계산해야하는 구간 수가 기하급수적으로 늘어나는 문제와 함께 추정된 분포가 불연속적이고 매끄럽지 못하다는 단점도 가지고 있다.



[그림 3-5] 막대-폭의 변화에 따른 추정 결과의 변화

지금까지 소개한 1차원 데이터에 대한 히스토그램법을 n차원 데이터 $\boldsymbol{x}$로 확장하면 다음과 같은 확률밀도함수의 추정식을 얻을 수 있다.

$$p(\boldsymbol{x}) = \frac{1}{h^n}\frac{K}{N} \qquad \text{[식 3-16]}$$

여기서 h^n은 데이터 빈도수를 측정하는 구간의 부피가 되고, k는 그 구간에 속하는 데이터의 수가 된다.

3.3.2 히스토그램법의 일반화

히스토그램법이 가지는 문제를 해결하기 위하여 먼저 [식 3-16]을 좀 더 일반적인 형태로 바꾸면 n차원 랜덤벡터 $\boldsymbol{X}$에 대해 $\boldsymbol{X}=\boldsymbol{x}$일 때의 확률밀도함수는 다음과 같이 쓸 수 있다.

$$p(\boldsymbol{x}) = \frac{1}{V}\frac{K}{N} \qquad \text{[식 3-17]}$$

이 때 V는 $\boldsymbol{x}$값 주변에서 빈도수를 측정하기 위한 영역의 부피가 된다. 히스토그램법은 이것을 미리 정해진 크기의 격자 형태의 영역으로 규정하였다. 즉 1차원은 선분, 2차원은 정사각형, 3차원은 정육면체, 그리고 고차원의 경우 초입방체 형태의 영역으로 규정하였으나, 이 절에서는 이를 일반화한 보다 효율적인 방법을 소개한다.

히스토그램법에서 살펴본 바와 같이 정확한 확률밀도함수를 얻기 위해서는 데이터 빈도수를 측정하는 영역을 잘 선택해 주어야한다. 즉, V의 값을 어떻게 선택하느냐가 성공적인 밀도함수 추정의 열쇠가 된다. 우선 영역을 가능한 좁게 나누어 체적 V의 값을 좁게 하면 할수록 더 정확한 확률밀도 함수를 얻을 것으로 기대할 수 있다. 그러나 만약 V를 지나치게 적게 하면 각 영역에 속하는 데이터의 수 K가 0이 되는 영역들이 많아져서 바람직하지 못한 결과를 초래할 수 있다. 이와 같이 영역의 부피 V와 데이터의 수 K는 서로 의존적이므로 이를 적절히 조절하여 선택할 필요가 있다. 이 때 V 혹은 K 중 어떤 값을 고정하고 어떤 값을 계산하느냐에 따라 크게 두 가지 밀도 추정법이 존재한다.

첫 번째 방법은 확률밀도 추정의 일반식에서 모든 영역의 V를 하나의 적절한 값으로 고정하고, 각 영역에 속하는 데이터의 수 K를 측정하여 각 영역에서의 확률밀도를 추정하는 방법으로, 히스토그램법이 가장 간단한 예가 되고, 이를 변형한 <커널 밀도 추정법(Kernel Density Estimation, KDE)>이 가장 대표적인 예가 된다. 이 때 V를 정하는 기준으로는 전체 데이터의 수 N에 반비례하여 $V=1/\sqrt{N}$과 같은 값을 사용한다. 이렇게 하면 데이터 수가 많은 경우는 체적을 적게 하여 영역을 조밀하게 나누고, 데이터 수가 적은 경우는 체적을 크게 하여 영역을 넓게 나누어 하나의 영역 안에 일정 정도의 데이터 수가 포함될 수 있도록 조정하는 것이 가능하다.

두 번째 방법은 한 영역에 들어가는 데이터의 수 K를 고정하고, 각 영역에 K개의 데이터가 포함될 수 있도록 하는 영역을 찾아 그 체적값을 계산함으로써 확률밀도를 추정하는 방법으로, <K-NNR (K-근접이웃규칙, K-Nearest Neighbor Rule)>이 대표적인 방법이 된다. 이 때 K를 정하는 기준도 전체 데이터에 의존하는 것이 바람직한데, K-NNR 방법에서는 $K=\sqrt{N}$개의 데이터를 포함하는 체적 V를 찾는다. 이때 데이터의 수가 많아지면 대응되는

체적 V가 작아지게 되고, 데이터 수가 작아지면 대응되는 체적 V가 커지게 되어, 결국 커널밀도추정법과 유사한 효과를 가진다.
이 두 가지 방법은 모두 데이터 수 N이 무한대로 커지면 실제 확률밀도함수에 수렴하게 되는데, 이 절에서는 주로 사용되는 커널 밀도 추정법에 관하여 설명한다. K-근접이웃 규칙법은 확률밀도함수를 추정하는데 사용되기보다 직접적으로 분류에 적용되어 K-근접이웃분류기로 더 많이 알려져 있다. 이 때 이에 대한 내용은 5장에서 분류기를 중심으로 설명하겠다.

3.3.3 커널 밀도함수 추정법

커널 밀도함수 추정법에서는 영역의 부피가 V로 정해지고, 랜덤변수의 각 값 $\boldsymbol{x}$를 중심으로 하여 부피가 V인 영역 안에 속하는 데이터 수가 달라진다. 즉, 데이터 수 K는 $\boldsymbol{x}$에 의존하는 값이 되어, 이를 확률밀도함수로 명확히 표현하면 다음 식과 같다.

$$p(\boldsymbol{x}) = \frac{1}{V}\frac{K(\boldsymbol{x})}{N} \qquad \text{[식 3-18]}$$

결국 각 $\boldsymbol{x}$ 값에 대한 데이터 수 K값이 우리가 계산해 주어야 하는 값이 된다. 이 값을 계산하기 위하여 가장 간단하면서 널리 쓰이는 방법은 초입방체 형태의 창을 사용하는 <파젠창(Parzen Window) 방법>과 가우시안 함수를 사용하는 <가우시안 커널 방법>이 있다.
n차원 입력 $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^T$에 대한 파젠창 방법에서는 다음과 같은 초입방체 형태의 커널함수 $\varphi(\boldsymbol{x})$를 정의한다.

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & |x_i| < 1/2 \;\; (i=1,...,n) \\ 0 & otherwise \end{cases} \qquad \text{[식 3-19]}$$

이 커널함수는 부피가 1인 초입방체 형태의 창을 나타내는 것으로, 이를 이용하여 $\boldsymbol{x}$를 중심으로 하는 초입방체 안에 속하는 데이터의 수는 다음과 같이 계산할 수 있다.

$$K(\boldsymbol{x}) = \sum_{i=1}^{N}\varphi(\boldsymbol{x}-\boldsymbol{x}_i) \qquad \text{[식 3-20]}$$

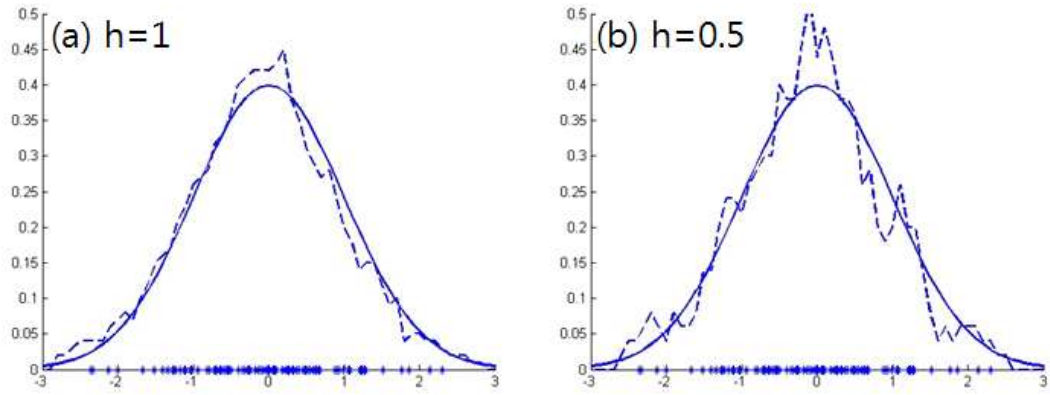
이 초입방체는 부피 V가 1인 경우이므로, 이제 이 부피를 조정하는 방법을 고려해야 한다. 초입방체의 부피를 조정하는 것은 각 면의 너비를 조정하는 것이므로, 너비가 h인 초입방체는 [식 3-19]의 밀도함수를 이용하여 $\varphi(\boldsymbol{x}/h)$로 나타낼 수 있다. 이를 이용하여 부피가 $V=h^n$인 초입방체 안에 속하는 데이터 수를 계산하는 함수는 다음과 같이 정의된다.

$$K(\boldsymbol{x}) = \sum_{i=1}^{N}\varphi\left(\frac{\boldsymbol{x}-\boldsymbol{x_i}}{h}\right) \qquad \text{[식 3-21]}$$

이를 이용하면 [식 3-18]로부터 다음과 같은 확률밀도함수의 추정식을 얻을 수 있다.

$$p(\boldsymbol{x}) = \frac{1}{Nh^n}\sum_{i=1}^{N}\varphi\left(\frac{\boldsymbol{x}-\boldsymbol{x_i}}{h}\right) \qquad \text{[식 3-22]}$$

[그림 3-6]에는 앞에서 사용한 1차원 데이터에 대해 파젠창을 사용한 커널 밀도함수 추정법으로 밀도함수를 추정한 결과를 나타내었다. [그림 3-6a]는 $h=1$인 경우이고, [그림 3-6b]는 $h=0.5$인 경우이다. 그림에서 h값이 적으면 그 불연속성 두드러짐을 볼 수 있다.



[그림 3-6] 파젠창에 의한 밀도함수 추정 예

[식 3-22]를 히스토그램법에 대한 [식 3-16]과 비교하면, 그 차이는 커널함수에 있음을 알 수 있다. 그런데 [식 3-19]에서 정의된 커널함수는 히스토그램법과 같은 초입방체이므로 동일한 결과를 가져올 것으로 예상할 수 있다. 그러나 커널법을 사용하는 경우에는 임의의 $\boldsymbol{x}$에 대한 확률값을 계산하는데 필요한 계산량은 데이터의 수 N에 의존하여 결정되는 반면, 히스토그램법의 경우에는 확률값을 계산해야 하는 영역의 수가 입력의 차원 수에 의존하여 기하급수적으로 늘어나므로 계산량의 측면에서 커널법이 우월함을 알 수 있다. 또한 파젠창에서 사용하는 입방체 커널 대신 연속함수로 정의되는 일반적인 커널함수로 확장하면, 확률밀도함수의 불연속적인 문제도 해결할 수 있다. 이 때 새롭게 정의되는 커널함수는 다음 두 조건을 만족해야 한다.

$$\varphi(\boldsymbol{x}) \geq 0, \int \varphi(\boldsymbol{x})d\boldsymbol{x} = 1 \qquad \text{[식 3-23]}$$

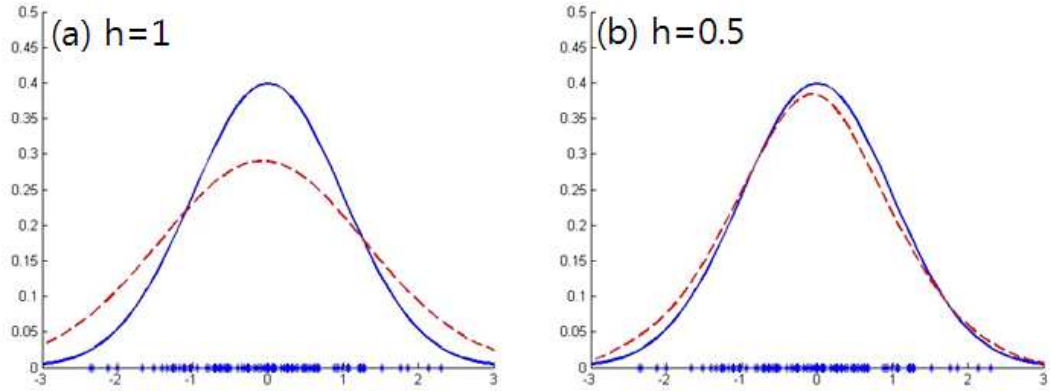
가우시안밀도함수는 이러한 특성을 만족하는 훌륭한 커널이 될 수 있다. 이를 정의하면 다음과 같다.

$$\varphi(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n}}\exp\left\{-\frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{x}\right\} \qquad \text{[식 3-24]}$$

이 커널함수를 이용하여 확률밀도함수식을 다시 쓰면 다음과 같다.

$$p(\boldsymbol{x}) = \frac{1}{Nh^n}\sum_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^n}}exp\left(-\frac{1}{2}\frac{(\boldsymbol{x}-\boldsymbol{x_i})^T(\boldsymbol{x}-\boldsymbol{x_i})}{h^2}\right) \qquad \text{[식 3-25]}$$

[그림 3-7]에서는 앞에서 사용한 것과 같은 데이터에 대하여 가우시안 커널에 의한 밀도함수 추정결과를 보여주고 있다. 가우시안 커널을 사용한 경우에는 파젠창 커널을 사용한 경우에 비해 부드러운 곡선형태의 추정함수를 가짐을 알 수 있다. 그러나 h값이 1인 경우에는 오히려 실제 분포보다 지나치게 완만한 형태를 찾는 문제가 발생했다. 이와 같이 비모수 밀도추정에 있어서는 적절한 구간의 크기 h를 정하는 것이 중요하다.



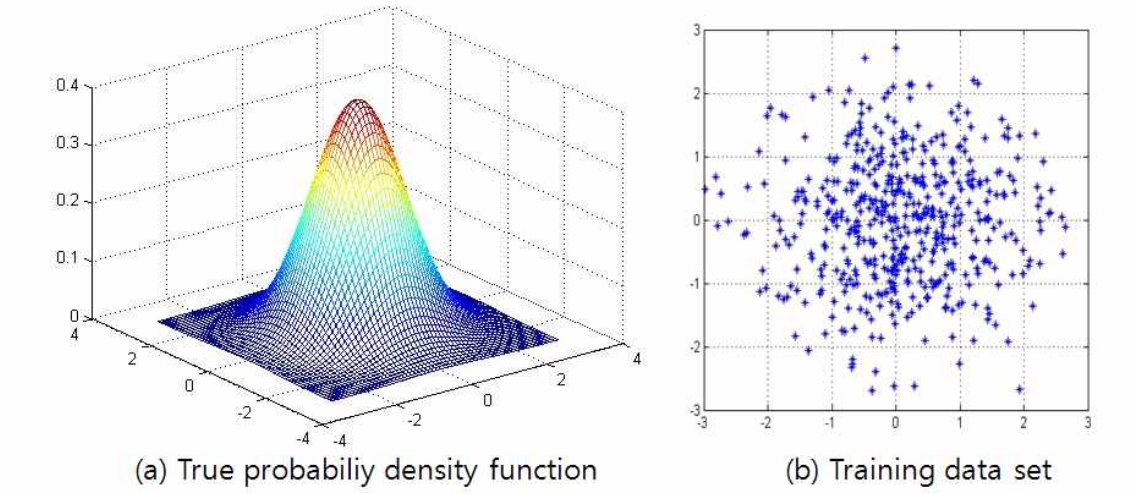
[그림 3-7] 가우시안 커널에 의한 확률밀도 추정

3.4 매트랩을 이용한 밀도함수 추정 실험

간단한 2차원 데이터에 대하여 지금까지 살펴본 밀도함수 추정법을 매트랩으로 구현해 보겠다. 사용한 데이터는 평균이 0이고 공분산이 단위행렬인 가우시안 분포를 따르는 모집단으로부터 추출된 500개의 샘플로 이루어 졌으며, [그림 3-8]에 모집단의 확률밀도 함수와 추출된 데이터 집합을 그림으로 나타내었다. 데이터를 생성하고, 밀도함수와 산점도를 그리는 프로그램을 [프로그램 3-1]에 나타내었다. 데이터 생성 방법과 산점도를 그리는 방법은 2장에서 수행한 것과 동일하다. 모집단의 확률밀도함수를 그리기 위하여서는 입력공간의 범위를 적절하게 정한 다음, 그 영역을 일정간격으로 세밀하게 분할하여, 각 점에서의 밀도값을 계산한 후, 그것을 3차원 공간상에 나타내는 방법을 취하였다. 이를 위하여 매트랩에서 제공하는 함수 meshgrid와 mesh가 사용되었다.

주어진 데이터에 대하여 모수적 밀도함수 추정을 하는 경우, 확률모델을 가우시안으로 세운다면 2장에서 수행한 것과 마찬가지로 매트랩의 함수 mean과 cov를 이용하여 표본평균과 표본공분산 행렬을 계산할 수 있을 것이다. 1.2절에서 살펴본 바와 같이 표본평균과 표본공분산이 최우추정법에서 얻어지는 최우추정치가 되므로, 이 과정을 통하여 간단히 확률밀도함수를 얻어낼 수 있다. 따라서 모수적 방법에 의한 추정은 이 절에서는 생략하고, 비모수적

방법에 대하여 구현해 보겠다.



[그림 3-8] 모집단의 확률밀도함수와 학습용 샘플데이터 집합

프로그램 3-1 Data Generation and Drawing true pdf

가우시안 분포를 따르는 데이터를 생성하고, 2차원 평면에 데이터의 산점도를 나타낸 후 모집단의 확률밀도함수를 그래프로 나타냄

```
001 N=500;                                          %데이터 수 설정
002 X = randn(N,2);                                 %데이터 생성
003 figure(1); plot(X(:,1), X(:,2), '*');           %데이터를 2차원 공간에 나타냄
004 axis([-3 3 -3 3]); grid on
005 % 모집단의 확률밀도함수를 그림
006 [x,y]=meshgrid([-3:0.1:3],[-3:0.1:3]);          % 입력공간 설정
007 XY=[x(:), y(:)];
008 m=[0 0]; s=eye(1);                             % 평균과 공분산 설정
009 for i=1:size(XY,1)                             % 확률밀도함수의 값 계산
010 pxy(i,1)=1/sqrt(2*pi*det(s))*exp(-1/2*(XY(i,:)-m)*inv(s)*(XY(i,:)-m)');
011 end
012 pxy=reshape(pxy,size(x));
013 figure(2); mesh(x, y,pxy); hold on              % 확률밀도 그래프 그리기
014 save data3_1 X;                                % 데이터 저장
```

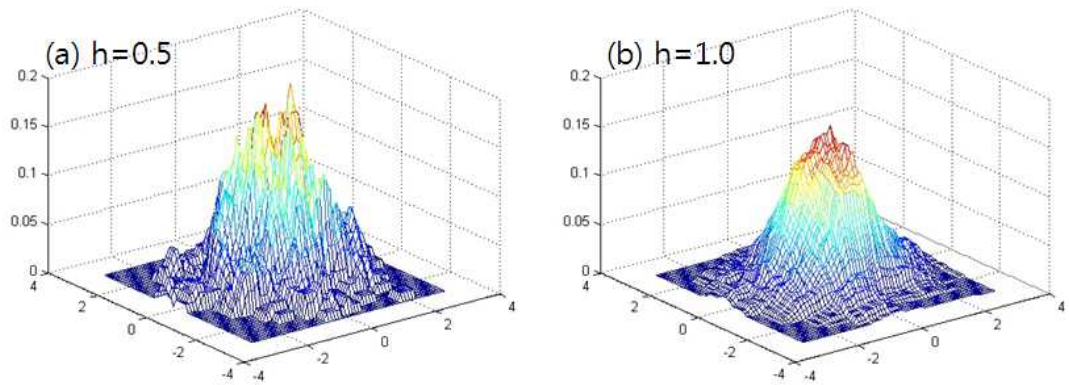
[프로그램 3-2]에는 파젠창 방법에 의한 밀도함수 추정을 수행하는 프로그램을 나타내었다. 먼저 [프로그램 3-1]에 의해 생성하여 저장해 둔 데이터 집합을 불러들인 후, 밀도함수의 값이 추정될 입력공간을 세분된 영역으로 나눈다. 또한 h의 값도 설정해 둔다. 세분된 입력공간의 각 점에 대해, 확률밀도함수의 값을 계산해야 하므로, 저장된 데이터들을 하나씩 불러와 커널함수의 값을 계산한 후 모두 더해주는 과정을 수행한다. 마지막으로 커널함수의 합

을 데이터 수 N와 h^d(이 경우 $d=2$)로 나누어 밀도함수의 값을 얻는다. 얻어진 밀도함수의는 3차원 공간상에 그래프로 나타낸다.

프로그램 3-2 Density estimation using Parzen window
파젠창에 의한 밀도함수 추정

```
001 load data3_1                                    %데이터 불러들임
002 h=1.0; N=size(X,1);                             %h값과 데이터 수 설정
003 [x,y]=meshgrid([-3:0.1:3],[-3:0.1:3]);  %입력공간 설정
004 % 입력공간의 모든 입력에 대해 확률밀도값 계산
005 for i=1:size(x,1)
006     for j=1:size(x,2)
007         k(i,j)=0;
008         %커널함수 값 계산
009         for n=1:N
010             if ((abs(X(n,1)-x(i,j))<(h/2))&(abs(X(n,2)-y(i,j))<(h/2)))
011                 k(i,j)= k(i,j)+1;
012             end
013         end
014     end
015 end
016 pxyp = k/(h^2*N);                              %밀도함수 값 계산
017 figure(3); mesh(x,y,pxyp); hold on             %그래프로 나타냄.
```

[그림 3-9]에 서로 다른 h값을 사용하여 추정한 결과를 나타내었다. 전체적으로 [그림 3-8]에 나타난 모집단의 분포보다 옆으로 퍼져있으며, 파젠창의 특성인 함수의 불연속성이 나타남을 알 수 있다. 특히 h값이 작은 경우에 불연속성은 더 뚜렷하게 나타났다.

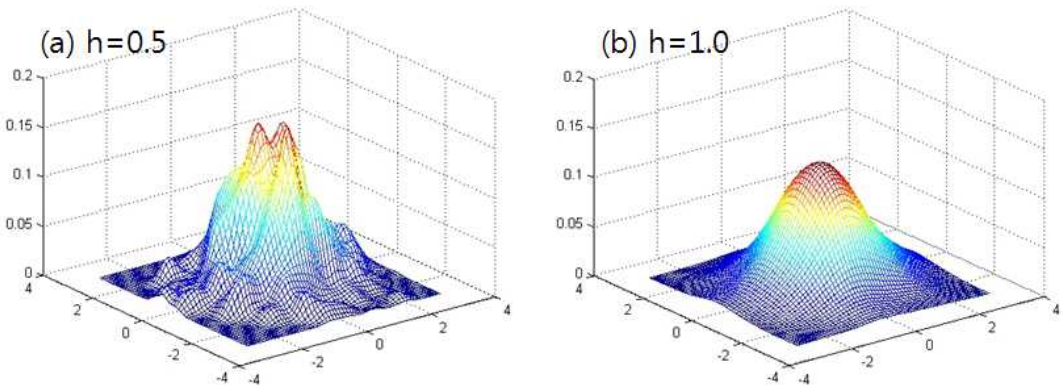


[그림 3-9] 파젠창에 의한 밀도함수 추정 결과

이어서 가우시안 커널을 사용하여 밀도함수를 추정하는 프로그램을 [프로그램 3-3]에 나타내었다. 전체적인 과정은 파젠창을 이용한 방법과 동일하며, 커널함수의 값을 계산하는 방법에 차이가 존재한다. 이변량 가우시안 분포에 대한 커널 값을 계산하는 과정이 9번째 줄에서 12번째 줄까지 나타나 있다. 최종적으로 커널함수에 h^2과 N을 나누어서 밀도함수의 값을 얻는 것은 파젠창 방법과 동일하다. 얻어진 밀도함수를 그래프로 나타낸 것이 [그림 3-10]에 보이고 있다. 파젠창 방법에서 보이는 불연속성은 없어졌으나, h값이 작은 경우 밀도함수의 형태가 복잡해짐을 알 수 있다.

프로그램 3-3 Density estimation using Gaussian kernel
가우시안 커널에 의한 밀도함수 추정

```
001 load data3_1                                    %데이터 불러들임
002 h=1.0; N=size(X,1);                             %h값과 데이터 수 설정
003 [x,y]=meshgrid([-3:0.1:3],[-3:0.1:3]);          %입력공간 설정
004 % 입력공간의 모든 입력에 대해 확률밀도값 계산
005 for i=1:size(x,1)
006     for j=1:size(x,2)
007         txy=[x(i,j), y(i,j)];
008         %커널함수 값 계산
009         for n=1:N
010             ksum(n)=(exp((-(X(n,:)-txy)*(X(n,:)-txy)')/(2*h^2)))/(2*pi);
011         end
012         k(i,j)=sum(ksum);
013     end
014 end
015 pxyg = k/(h^2*N);                               %밀도함수 값 계산
016 figure(4); mesh(x,y,pxyg); hold on              %그래프로 나타냄.
```



[그림 3-10] 가우시안 커널에 의한 밀도함수 추정 결과

연습 문제

1. 두 클래스 집합 C_1, C_2는 각각 다음과 같은 평균과 공분산을 가지는 가우시안 분포를 따른다. 매트랩을 이용하여 각 클래스별로 데이터를 100개씩 생성하시오. 생성된 데이터를 2차원 평면상의 한 점으로 표시한 산점도를 그리시오.

$\boldsymbol{\mu}_1 = [0,0]^T,\ \boldsymbol{\mu}_2 = [4,4]^T$

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

2. 각 클래스의 데이터 분포가 가우시안 함수를 따른다는 가정 하에, 1번에서 생성한 데이터 집합을 이용하여 각 클래스의 확률밀도함수 $p(\boldsymbol{x}|C_i)$의 파라미터 $\boldsymbol{\mu}_i$와 $\boldsymbol{\Sigma}_i$를 추정하시오. (힌트: 매트랩을 이용하여 데이터집합의 표본평균과 표본공분산을 계산한다.)

3. 2번 문제에서 추정된 파라미터를 이용하여 [식 3-1]에서 소개된 판별함수 $G(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 의 식을 찾으시오.

4. 1번에서 사용한 데이터에 대해, 커널밀도함수 추정법에 의해 밀도함수를 추정하고자 한다. 다음을 각각 수행하시오.

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

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

(3) (1), (2)에서 얻어진 결과와 3번에서 얻은 결과들을 비교하여 보시오.

5. 1차원 데이터 N개로 이루어진 집합 $X = \{x_1, x_2, ..., x_N\}$가 주어졌을 때, 그 분포가 가우시안을 따른다고 가정하고 그 파라미터를 추정하면 표본평균과 표본분산이 된다. 이 유도과정을 보이시오.

6. 다음과 같이 1차원 데이터가 5개 주어졌다.

$$X = \{-1.3,\ -0.7,\ 0.4,\ 1.1,\ 1.5\}$$

파젠창 방법에 의해 추정되는 확률밀도 함수식을 쓰고, $x = 1.3$일 때의 확률밀도함수의 값을 계산하시오. (단, $h = 0.5$로 둠)

7. 파젠창에 의한 밀도함수 추정법과 히스토그램법의 공통점과 차이점에 대하여 생각해 보시오.

8. 이 책에서 소개한 파젠창과 가우시안 커널 이외에 다양한 커널함수에 대하여 생각해 보

고 이를 이용하여 밀도함수를 추정해 보시오.

참고 자료

이 책에서는 주로 가우시안 함수를 이용한 밀도추정을 수행하였으나, 매우 다양한 확률밀도 함수들이 존재한다. 다양한 확률밀도 함수와 그 특성, 그리고 최우추정을 비롯한 모수적 확률밀도 추정법에 대한 이론적인 내용들은 수리통계학의 교과서에서 찾아볼 수 있다. 대표적인 교과서로 [Hogg, McKean & Craig 05]가 있다. 커널밀도함수 추정법에 대한 내용은 통계적 추론에 관한 교과서 [Wasserman 05] 등에서 찾아볼 수 있으며 1장에서 소개한 [Duda, Hart & Stork 01]에도 비교적 자세히 소개되어 있다.

[Hogg, McKean & Craig 05] R. V. Hogg, J. W. McKean, A. T. Craig. Introduction to Mathematical Statistics (6th ed.). Prentice Hall, 2005.
[Wasserman 05] L. Wasserman. All of Statistics: A Concise Course in Statistical Inference. Springer, 2005.
[Duda, Hart & Stork 01] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2ed.). Wiley, 2001