4강. 베이지안 분류기

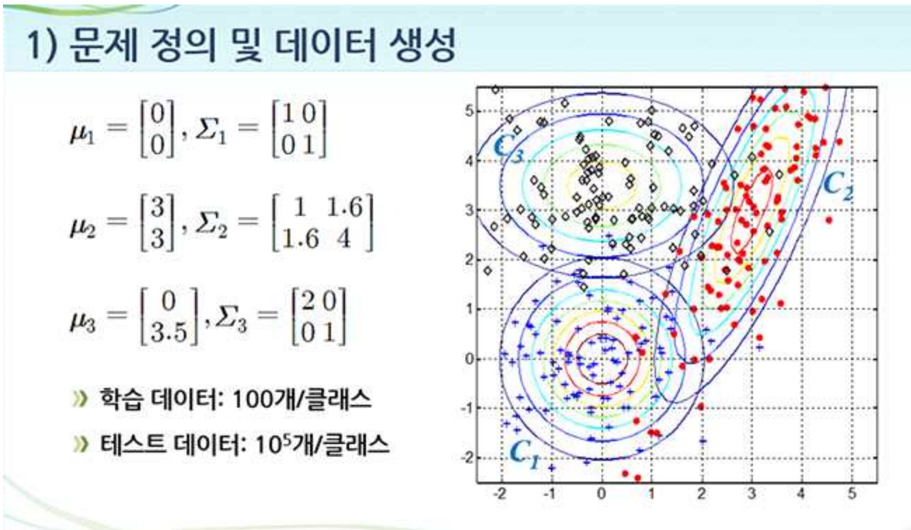
※ 점검하기

Q1. 다음에 주어진 단계에 따라 데이터를 생성하고 베이지안 분류기를 이용하여 분류를 수행하시오.

(1) 매트랩을 이용하여 다음과 같은 평균과 공분산을 가지는 가우시안 분포를 따르는 2차원 데이터를 각각 100개씩 가지는 두 클래스 집합를 생성하시오. 생성된 데이터를 2차원 평면상의 점으로 표시한 그래프를 그리시오.

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

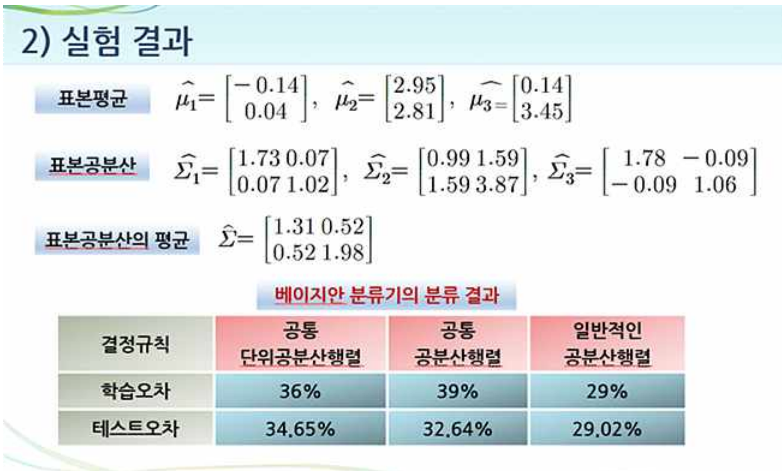
<관련학습보기>



교재 2장 [프로그램 2-1, 데이터 생성]에서 데이터 개수(N),
평균(m1,m2)와 공분산(s1,s2)의 값을 조정하면 된다.
[참조] 4. 매트랩을 이용한 베이지안 분류기 실험의「1) 문제 정의 및 데이터 생성」

(2) 각 클래스의 데이터 분포가 가우시안 함수를 따른다는 가정 하에, (1)에서 생성한 데이터 집합을 이용하여 각 클래스의 확률밀도함수 $p(x|c_k)$의 파라미터(평균과 공분산)를 추정하시오.

<관련학습보기>



2) 실험 결과

표본평균 $\hat{\mu}_1=\begin{bmatrix}-0.14\\0.04\end{bmatrix}$, $\hat{\mu}_2=\begin{bmatrix}2.95\\2.81\end{bmatrix}$, $\hat{\mu}_3=\begin{bmatrix}0.14\\3.45\end{bmatrix}$

표본공분산 $\hat{\Sigma}_1=\begin{bmatrix}1.73 & 0.07\\0.07 & 1.02\end{bmatrix}$, $\hat{\Sigma}_2=\begin{bmatrix}0.99 & 1.59\\1.59 & 3.87\end{bmatrix}$, $\hat{\Sigma}_3=\begin{bmatrix}1.78 & -0.09\\-0.09 & 1.06\end{bmatrix}$

표본공분산의 평균 $\hat{\Sigma}=\begin{bmatrix}1.31 & 0.52\\0.52 & 1.98\end{bmatrix}$

베이지안 분류기의 분류 결과

결정규칙	공통 단위공분산행렬	공통 공분산행렬	일반적인 공분산행렬
학습오차	36%	39%	29%
테스트오차	34.65%	32.64%	29.02%

교재 2장 [프로그램 2-2, 데이터 분석]를 이용하여 집합의 표본평균과 표본공분산을 계산한다. [참조] 4. 매트랩을 이용한 베이지안 분류기 실험의 「2) 실험 결과」

(3) (2)에서 추정된 파라미터를 이용하여 교재 4.3절에서 소개된 판별함수를 찾으시오.

<관련학습보기>

2) 실험 결과

표본평균 $\hat{\mu}_1=\begin{bmatrix}-0.14\\0.04\end{bmatrix}$, $\hat{\mu}_2=\begin{bmatrix}2.95\\2.81\end{bmatrix}$, $\hat{\mu}_3=\begin{bmatrix}0.14\\3.45\end{bmatrix}$

표본공분산 $\hat{\Sigma}_1=\begin{bmatrix}1.73 & 0.07\\0.07 & 1.02\end{bmatrix}$, $\hat{\Sigma}_2=\begin{bmatrix}0.99 & 1.59\\1.59 & 3.87\end{bmatrix}$, $\hat{\Sigma}_3=\begin{bmatrix}1.78 & -0.09\\-0.09 & 1.06\end{bmatrix}$

표본공분산의 평균 $\hat{\Sigma}=\begin{bmatrix}1.31 & 0.52\\0.52 & 1.98\end{bmatrix}$

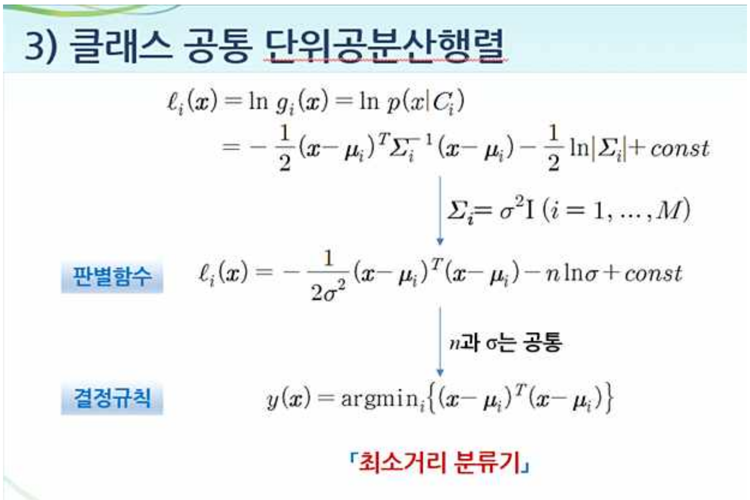
베이지안 분류기의 분류 결과

결정규칙	공통 단위공분산행렬	공통 공분산행렬	일반적인 공분산행렬
학습오차	36%	39%	29%
테스트오차	34.65%	32.64%	29.02%

추정된 파라미터를 클래스 공통 단위공분산행렬의 결정규칙 ([식 4-26]), 클래스 공통 공분산행렬의 결정규칙([식 4-28]), 일반적인 공분산행렬의 결정규칙([식 4-29])에 대입하여 판별함수를 유도한다. [참조] 4. 매트랩을 이용한 베이지안 분류기 실험의 「2) 실험 결과」

(4) 새롭게 주어진 데이터 $x = [2, 1]^T$에 대해, (3)에서 얻어진 판별함수를 이용하여 어떤 클래스에 속하는지를 매트랩을 이용하여 판단하시오.

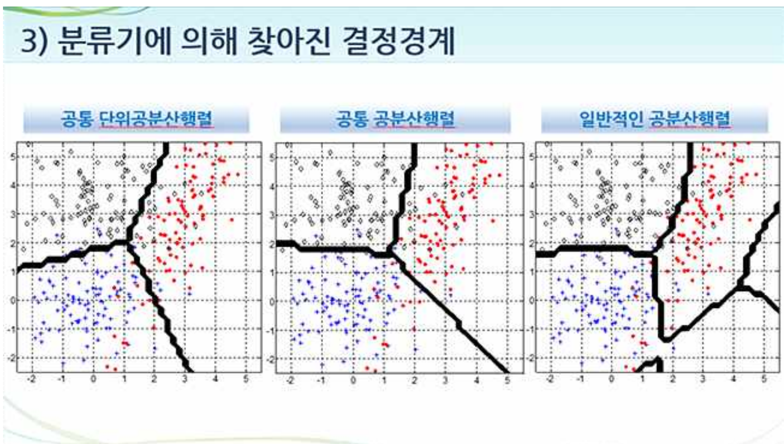
<관련학습보기>



교재 4장 [프로그램 4-1, Bayes Classifier]를 이용한다.
[참조] 3. 가우시안 확률분포와 베이지안 분류기의 「3) 클래스 공통 단위공분산행렬」

(5) (3)에서 계산된 판별함수로부터 얻어지는 결정경계를 찾아 매트랩을 이용하여 (1)에서 그린 그래프 위에 표시하시오.

<관련학습보기>



교재 4장 [프로그램 4-2, Drawing Decision Boundary]를 이용한다.
[참조] 4. 매트랩을 이용한 베이지안 분류기 실험의 「3) 분류기에 의해 찾아진 결정경계」

※ 정리하기

1. 베이지안 분류기
   1) 우도비 분류는 각 클래스에서 데이터 x가 관찰된 확률밀도의 비율(우도비, $p(x|C_1)/p(x|C_2)$)을 이용한 분류이며, 특히 후험확률에 대한 베이즈 정리로부터 유도된 판별함수 $g_i(x)=p(x|C_i)p(C_i)$를 이용하여 분류하는 방식을 베이지안 분류기라고 함
   2) 베이지안 분류기에서는 새로운 데이터가 주어지면 각 클래스에 대해서 $g_i(x)$의 값을 계산한 후, 그 값이 가장 큰 클래스로 데이터를 할당함

2. 베이지안 분류기의 결정경계와 오차
   1) 이진 분류에서의 베이지안 분류기의 오류확률은 $p(C_1)\times$($p(x|C_1)$함수의 영역 R1에 속하는 면적)과 $p(C_2)\times$($p(x|C_2)$함수의 영역 R2에 속하는 면적)의 합으로 계산됨
   2) 한편 다중 클래스인 경우에는 오류확률을 생각하기보다 바르게 결정할 확률을 계산하는 것이 더 용이함
   3) 단순한 오류확률이 아닌 각 오분류에 대한 판단 결과가 초래할 비용을 함께 고려한 오류확률을 베이즈 위험이라고 함

3. 가우시안 확률분포와 베이지안 분류기
   1) 확률밀도함수가 가우시안 분포를 따르는 경우에는 공분산 행렬의 형태에 따라 다른 판별 함수가 결정됨
      - 모든 클래스의 공분산이 동일하게 단위행렬의 상수 배인 행렬을 가지는 경우의 결정경계는 각 클래스의 평균을 잇는 직선에 수직이면서 그 중점을 지나는 직선이 되며, 이때의 분류기는 최소거리 분류기가 됨
      - 모든 클래스에 대해 동일한 공분산을 가지지만 그 형태는 일반적인 행렬이 되는 경우의 결정경계는 여전히 직선 형태가 됨
        → 하지만 결정규칙이 공분산도 함께 고려하여 거리를 계산하는 마할라노비스 거리가 되어 보다 좋은 성능을 기대할 수 있음
      - 각 클래스의 공분산이 서로 다른 일반적인 형태를 가지는 경우에는 자연스러운 곡선 형태의 결정경계를 가지기 때문에 보다 다양한 데이터 분포에 대해 적절한 결정규칙을 제공할 수 있을 것임
        → 하지만 클래스에 대해서 추정해야 할 파라미터가 많아져서 추정 오차에 의해서 최종 결정규칙이 나쁜 영향을 받을 가능성도 높아지는 문제를 가짐