Chapter 04 베이지안 분류기

[학습목표]
3장에서는 데이터를 이용하여 각 클래스별 확률밀도함수를 추정하는 방법을 살펴보았고, 이것을 이용하면 새로운 데이터가 각 클래스에 속할 확률을 계산할 수 있다. 이와 같이 확률값을 이용하여 패턴을 분류하는 대표적 분류기로 베이지안 분류기가 있다. 이 장에서는 베이지안 분류기의 기본 개념에 대하여 알아보고, 간단한 2차원 데이터를 베이지안 분류기로 분류해 보면서 그 성질을 알아보겠다.

4. 베이지안 분류기

4.1. 베이지안 분류기

3장에서 각 클래스별 확률밀도함수 $p(\boldsymbol{x}|C_k)$를 추정하는 방법에 대하여 알아보았다. 이 장에서는 추정된 확률밀도함수를 이용하여 패턴을 분류하는 방법에 대하여 알아볼 것이다. 새로운 데이터 $\boldsymbol{x}_{new}$가 어떤 클래스에 속하는 지 판단하는 기준으로, 그 데이터가 각 클래스로부터 생성되었을 확률을 계산하고, 이것들 중에서 보다 높은 확률을 가지는 클래스로 분류를 수행한다. 먼저 클래스가 두 개인 간단한 이분류 문제에 대하여 생각해 보고, 이어서 다중 클래스 문제로 확장하여 알아보겠다.

4.1.1 우도비에 의한 패턴분류

두 개의 서로 다른 패턴을 구별하는 이분류(binary classification) 문제를 생각한다. 하나의 특징 벡터 $\boldsymbol{x}$가 입력으로 주어졌을 때, 이 데이터가 두 개의 서로 다른 클래스 C_1과 C_2 중 어느 클래스에 속하는 지를 결정하는 문제이다. 즉, 데이터 $\boldsymbol{x}$에서 클래스 라벨값 $y(\boldsymbol{x})$로 매핑하는 결정규칙을 찾아야 한다. 확률을 바탕으로 이를 결정하기 위해서는 먼저 주어진 데이터가 각 클래스에 속할 확률 $P(C_1|\boldsymbol{x})$와 $P(C_2|\boldsymbol{x})$를 각각 계산하여야 한다. 이 확률값은 데이터가 관찰된 후에 그로부터 추정할 수 있는 각 클래스에 대한 확률을 의미하여 후험확률(posterior probability)이라고 부른다. 이에 반해 데이터가 관찰되기 이전의 각 클래스에 대한 확률 $p(C_1)$, $p(C_2)$를 선험확률 (prior probability)이라고 부른다. 일단 두 클래스에 대해 후험확률이 계산되면, 그 값이 큰 클래스에 현재 주어진 데이터 $\boldsymbol{x}$가 속한다고 결론을 내릴 수 있다. 즉, 다음과 같은 판별함수를 정의할 수 있을 것이다.

$$g(\boldsymbol{x}) = p(C_1|\boldsymbol{x}) - p(C_2|\boldsymbol{x}) = 0 \qquad \text{[식 4-1]}$$

만약 $g(\boldsymbol{x})$값이 0보다 크면 클래스 라벨 $y(\boldsymbol{x}) = 1$이 되어 $\boldsymbol{x}$는 C_1에 할당되고, 0보다 작으면 $y(\boldsymbol{x}) = -1$이 되어 C_2에 할당된다.

그런데 이 후험확률 값은 일반적으로 바로 계산될 수 없는 값이다. 따라서 베이즈 규칙을 이용하여 후험확률 값을 다시 쓰면 다음과 같이 나타낼 수 있다.

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})} \qquad \text{[식 4-2]}$$

[식 4-2]를 이용하여 판별함수 [식 4-1]을 다시 표현해 보면 다음과 같다.

$$g(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x})} - \frac{p(\boldsymbol{x}|C_2)p(C_2)}{p(\boldsymbol{x})} = 0 \qquad \text{[식 4-3]}$$

여기서 양변에 공통으로 들어있는 $p(\boldsymbol{x})$는 결정에 영향을 미치지 않으므로 제외하고 각 항에 $p(\boldsymbol{x}|C_2)p(C_1)$을 나누어주면 다음과 같은 새로운 결정경계를 얻을 수 있다.

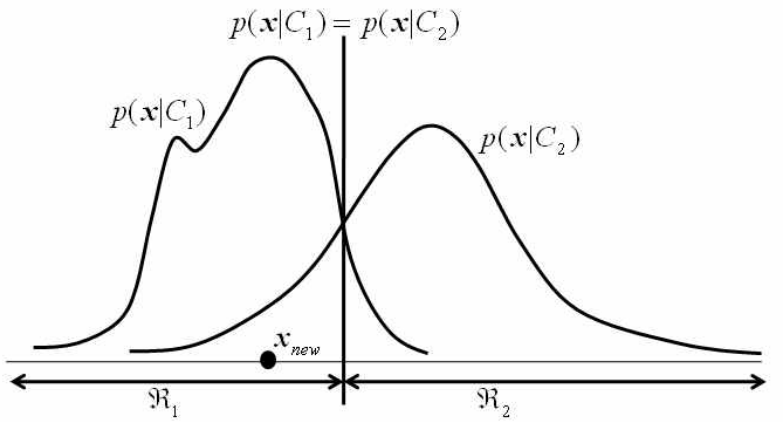
$$g_{LRT}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} - \frac{p(C_2)}{p(C_1)} = 0 \qquad \text{[식 4-4]}$$

[식 4-4]의 우변에서 첫 번째 항 $p(\boldsymbol{x}|C_1)/p(\boldsymbol{x}|C_2)$은 각 클래스에서 데이터 $\boldsymbol{x}$가 관찰될 확률밀도의 비율로서, 이를 우도비(Likelihood ratio)라고 한다. 이 값은 3장에서 소개한 확률밀도함수 추정법을 이용하여 클래스별 확률밀도함수를 추정함으로써 얻을 수 있다. 이 판별함수를 이용하면 $g_{LRT}(\boldsymbol{x})$가 0보다 크면 클래스 C_1으로, 0보다 작으면 C_2로 분류하게 된다. 이를 결정규칙으로 나타내면 다음 식으로 표현가능하다.

$$y(\boldsymbol{x}) = \begin{cases} 1 & \text{if } g_{LRT}(\boldsymbol{x}) > 0 \\ -1 & \text{otherwise} \end{cases} \qquad \text{[식 4-5]}$$

이와 같은 우도비에 의한 분류를 <우도비 분류(Likelihood ratio classifier)>라고 한다. 또한 [식 4-4]는 후험확률에 대한 베이즈 정리로부터 유도된 것으로, 이 결정경계를 이용하여 분류하는 것을 이분류 문제를 위한 <베이즈 분류기(Bayes classifier)>라고 한다.
[식 4-4]의 두 번째 항에 있는 $p(C_1)$과 $p(C_2)$는 각각 전체 데이터 집합에서 각 클래스가 차지하는 비율을 말하는 것으로, 데이터가 관찰되기 전의 각 클래스의 확률이므로 선험확률(prior probability)이라고 부른다. 만약 두 클래스가 같은 비율로 전체 데이터 집합을 구성하고 있다면 이 항은 1의 값을 가지게 되어 결정규칙은 다음과 같이 더욱 간단한 형태로 나타낼 수 있다.

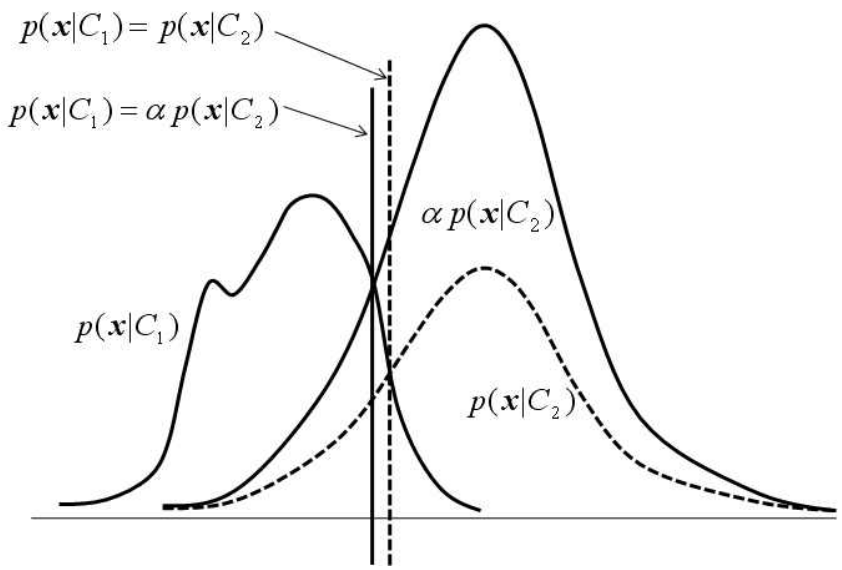
$$y(\boldsymbol{x}) = \begin{cases} 1 & \text{if } p(\boldsymbol{x}|C_1) > p(\boldsymbol{x}|C_2) \\ -1 & \text{otherwise} \end{cases} \qquad \text{[식 4-6]}$$



[그림 4-1] 1차원 데이터에 대한 베이즈 분류

[그림 4-1]에 간단한 1차원 데이터에 대한 분류 예를 나타내었다. 왼쪽 곡선이 클래스 C_1에 속하는 데이터의 확률밀도함수, 오른쪽 곡선이 클래스 C_2에 속하는 데이터의 확률밀도함수를 나타낸다. 예를 들어 $p(C_1)$과 $p(C_2)$가 각각 전체 데이터의 50%씩을 차지한다고 했을 때, [식 4-4]로부터 결정경계는 $p(\boldsymbol{x}|C_1)=p(\boldsymbol{x}|C_2)$가 되는 지점([그림 4-1]에서 수직선 부분)이라는 것을 알 수 있다. 이렇게 결정경계가 정해지면, 그림에서와 같이 새로운 데이터 $\boldsymbol{x}_{new}$가 주어졌을 때 결정경계의 왼쪽에 있는 경우는 C_1, 오른쪽에 있는 경우는 C_2로 분류를 수행한다. 이때, 전체 입력 공간 중에서 C_1으로 할당되는 부분을 클래스 C_1의 결정영역(decision region) $\Re_1$으로 나타내고, C_2로 할당되는 부분을 클래스 C_2의 결정영역 $\Re_2$로 나타낸다.

그런데 만약 $p(C_1)$과 $p(C_2)$의 값이 서로 같지 않다면 어떻게 될 것인가? 예를 들어 병을 진단하는 분류 문제의 경우에는 병에 걸린 환자(클래스 C_1)에 비해 정상인의 수가 월등이 높은 비율을 차지하는 경우가 일반적이다. 이와 같이 $p(C_1)$과 $p(C_2)$의 값이 다른 경우로, 그 비율이 $p(C_2)=\alpha p(C_1)$라고 하면 결정경계는 $p(\boldsymbol{x}|C_1)=\alpha p(\boldsymbol{x}|C_2)$를 만족하는 지점이 된다. 따라서 만약 α가 1보다 크다면(즉, $p(C_2)$가 $p(C_1)$에 비해 크다면) 앞서 알아본 결정경계에 비해 C_1쪽으로 치우치게 될 것이다 [그림 4-2]에 그 예를 나타내었다. 그림에서는 α가 1보다 큰 경우로, 점선으로 나타난 확률밀도 함수 $p(\boldsymbol{x}|C_2)$에 비해 α배 증가한 곡선 $\alpha p(\boldsymbol{x}|C_2)$과 곡선 $p(\boldsymbol{x}|C_1)$의 값이 같아지는 부분이 새로운 결정경계가 되어, 점선으로 나타난 원래 결정경계보다 클래스 C_1쪽에 더 가까워지게 됨을 볼 수 있다. 바꾸어 말하면, 클래스 C_2의 결정영역이 넓어진 것이 되는데, 이는 α값이 1보다 커서 관찰 가능한 전체 데이터에서 클래스 C_2가 차지하는 비중이 높아진 것을 반영한 결과로 해석할 수 있다.



[그림 4-2] 선험확률에 따른 결정경계의 변화

4.1.2 다중 클래스 분류기

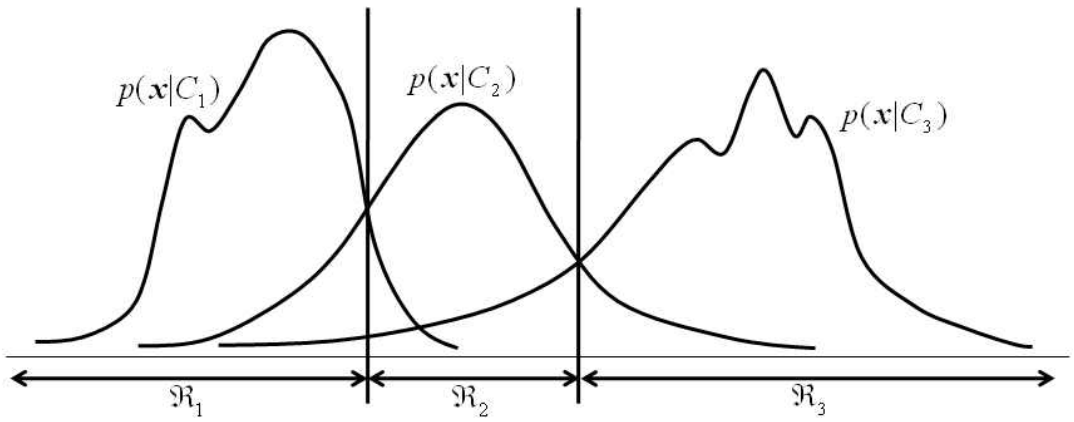
지금까지는 분류 대상이 되는 클래스의 수가 두 개인 경우에 대해서 살펴보았다. 클래스의 수가 3개 이상이 되는 다중 클래스 분류 문제에서는, 먼저 각 클래스 C_i에 대한 판별함수를 다음과 같이 정의한다.

$$g_i(\boldsymbol{x}) = p(x|C_i)p(C_i) \qquad \text{[식 4-7]}$$

이 함수는 이진분류기의 경우와 마찬가지로 데이터 $\boldsymbol{x}$가 주어졌을 때 그것이 클래스 C_i로부터 얻어졌을 확률, 즉 C_i에 대한 후험확률 $P(C_i|\boldsymbol{x})$에 베이즈 정리를 적용하고, 클래스에 상관없이 일정한 값을 가지는 분모 부분인 $p(\boldsymbol{x})$를 제외함으로써 얻어진 것이다. 따라서 각 클래스에 대하여 이 함수의 값을 계산한 후, 그 값이 가장 큰 클래스로 데이터를 할당하면 된다. 결국, 클래스 라벨 $y(\boldsymbol{x})$를 결정하는 결정규칙은 다음과 같이 나타낼 수 있다.

$$y(\boldsymbol{x}) = \mathrm{argmax}_i\{g_i(\boldsymbol{x})\} \qquad \text{[식 4-8]}$$

[그림 4-3]에서는 1차원 입력 데이터에 대하여 3개의 클래스가 존재하여, 각 클래스 C_k에 대한 확률밀도함수와 이를 바탕으로 [식 4-6]에 의해 결정되는 결정영역 $\Re_i (i = 1,2,3)$을 나타내었다. 여기서는 각 클래스의 선험확률 $p(C_i)$는 모두 동일하다는 가정 하에 결정영역을 찾았으나, 만약 이 값이 달라진다면 [그림 4-2]와 유사하게 결정영역에 변동이 발생할 것이다.



[그림 4-3] 다중 클래스 분류에서의 결정경계

4.2 베이지안 분류기의 결정경계와 오차

앞에서 사용한 결정규칙은 어디까지나 확률값에 의존하는 것으로 절대적으로 올바른 결과를

낸다고 할 수 없으며, 항상 판단에는 오류가 있을 수 있다. 이 절에서는 주어진 결정규칙에 의한 분류기가 잘못된 분류결과를 낼 오류확률을 정의하고, 베이지안 결정규칙이 오류확률의 측면에서 어느 정도의 성능을 보장하는지 알아보겠다.

4.2.1 분류기의 오류확률

패턴 분류에 있어서 오류란, 주어진 입력 $\boldsymbol{x}$가 어떤 클래스에 속할지 판단할 때 그 판단이 잘못되어 실제와는 다른 클래스로 할당하는 경우를 말한다. 앞 절에서와 같이 일단 판별함수에 의한 결정규칙이 정해지면 전체 입력공간은 각 클래스에 대응되는 부분 영역으로 나뉘게 되는데, 이 때 클래스 C_i에 대응되는 영역을 결정영역 R_i라고 한다. 이 때 서로 다른 클래스에 해당하는 결정영역들 사이의 경계가 결정경계를 이룬다. 또한 이 결정경계를 이용하여 얻어진 결정규칙이 오류를 일으킬 확률을 계산해 볼 수 있는데, 이를 <오류확률(Probability of error)>이라고 한다.

논의를 간단히 하기 위해 먼저 이진분류기에 대하여 생각해보자. 두 개의 클래스를 분류하는 경우 오류에도 두 가지 경우가 있을 수 있다. 첫 번째 경우는 클래스 C_1에 속하는 데이터 $\boldsymbol{x}$를 C_2에 속한다고 판단한 경우이다. 다시 말하면 C_1에 속하는 데이터 $\boldsymbol{x}$가 우리가 만든 결정규칙에 의해 정해진 결정영역 중 $\Re_2$에 속하게 되는 경우로, 이러한 일이 일어날 확률은 결합확률 $\mathrm{Prob}(x \in \Re_2, \boldsymbol{x} \in C_1)$로 나타낼 수 있다. 두 번째 경우는 반대의 경우로서, 클래스 C_2에 속하는 데이터를 C_1에 속한다고 판단한 경우이며, 이도 역시 결합확률 $\mathrm{Prob}(\boldsymbol{x} \in \Re_1, \boldsymbol{x} \in C_2)$로 표현할 수 있다. 따라서 이 결정규칙에 의해 발생될 수 있는 전체 오류는 이러한 두 가지의 오류를 합하여 다음과 같이 얻어진다.

$$\mathrm{P}_{\mathrm{err}} = \mathrm{Prob}(\boldsymbol{x} \in \Re_2, \boldsymbol{x} \in C_1) + \mathrm{Prob}(\boldsymbol{x} \in \Re_1, \boldsymbol{x} \in C_2) \qquad \text{[식 4-9]}$$

이 때 $\mathrm{Prob}(x \in \Re_2, \boldsymbol{x} \in C_1)$이 가지는 의미를 생각해 보면, 클래스 C_1의 확률분포를 따르는 데이터가 영역 R_2에 속할 확률이 되므로, 결합확률밀도함수 $p(\boldsymbol{x}, C_1)$을 영역 R_2상에서 적분한 값이 된다. 또한 베이즈 정리를 이용하면 결합확률밀도함수 $p(\boldsymbol{x}, C_1)$는 다음과 같이 나타낼 수 있다.

$$p(\boldsymbol{x}, C_1) = p(\boldsymbol{x}|C_1)p(C_1) \qquad \text{[식 4-10]}$$

따라서 $\mathrm{Prob}(x \in \Re_2, \boldsymbol{x} \in C_1)$는 [식 4-10]을 이용하여 다음과 같이 나타낼 수 있다.

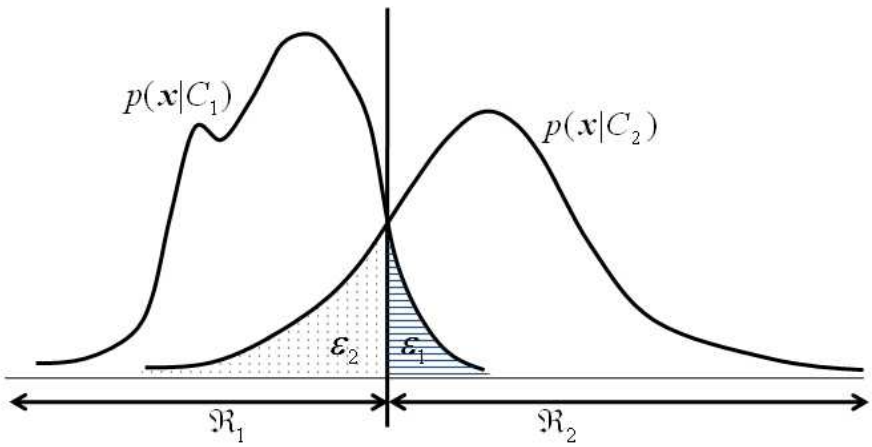
$$\mathrm{Prob}(x \in \Re_2, \boldsymbol{x} \in C_1) = \int_{\Re_2} p(\boldsymbol{x}|C_1)p(C_1)d\boldsymbol{x} = p(C_1)\int_{\Re_2} p(\boldsymbol{x}|C_1)d\boldsymbol{x} \qquad \text{[식 4-11]}$$

여기서 적분변수 $\boldsymbol{x}$와 상관없는 $p(C_1)$ 부분을 적분 밖으로 뺀 나머지 부분 $\int_{R_2} p(\boldsymbol{x}|C_1)d\boldsymbol{x}$의 의미를 생각해 보면, 이는 $p(\boldsymbol{x}|C_1)$ 함수의 영역 R_2에 속하는 부분의 면적이므로, [그림 4-4]에서 $\mathcal{E}_1$에 해당하는 부분이다. 마찬가지로 $\mathrm{Prob}(\boldsymbol{x} \in \Re_1, \boldsymbol{x} \in C_2)$에 대해서도 생각해 보면

$\int_{\Re_1} p(\boldsymbol{x}|C_2)d\boldsymbol{x}$가 같은 그림에서 ε_2가 되어, 전체 오류 확률은 결국 다음과 같이 쓸 수 있다.

$$\mathrm{P}_{\mathrm{err}} = p(C_1)\int_{\Re_2} p(\boldsymbol{x}|C_1)d\boldsymbol{x} + p(C_2)\int_{\Re_1} p(\boldsymbol{x}|C_2)d\boldsymbol{x} = p(C_1)\varepsilon_1 + p(C_2)\varepsilon_2 \qquad [식\ 4\text{-}12]$$

여기서 선험확률 $p(C_1)$과 $p(C_2)$가 동일하게 0.5라고 하면 오류는 $(\varepsilon_1 + \varepsilon_2)/2$가 된다.



[그림 4-4] 우도비 검증에 의한 분류의 오류확률

한편, 클래스가 두 개 이상인 경우에는 오류확률을 생각하기보다 바르게 결정할 확률을 생각해 보는 것이 더 용이하다. 클래스 C_i에 속하는 데이터를 바르게 판단할 확률은 다음 식과 같다.

$$\mathrm{Prob}(x \in \Re_i, \boldsymbol{x} \in C_i) = \int_{\Re_i} p(\boldsymbol{x}|C_i)p(C_i)d\boldsymbol{x} = p(C_i)\int_{\Re_i} p(\boldsymbol{x}|C_i)d\boldsymbol{x} \qquad [식\ 4\text{-}13]$$

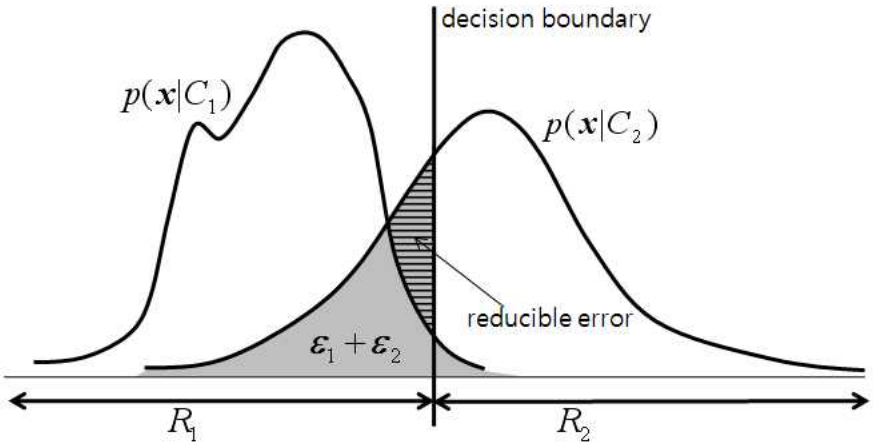
따라서 모두 M개의 클래스가 있는 경우, 각각의 클래스에 속하는 데이터를 바르게 판단할 확률을 모두 합하여 정분류 확률 $\mathrm{P}_{\mathrm{correct}}$를 계산할 수 있다.

$$\mathrm{P}_{\mathrm{correct}} = \sum_{i=1}^{M} p(C_i)\int_{\Re_i} p(\boldsymbol{x}|C_i)d\boldsymbol{x} \qquad [식\ 4\text{-}14]$$

4.2.2 최소 오류확률 결정경계

오류확률이 계산되면, 이를 이용하여 오류확률을 최소로 하는 결정경계를 찾을 수 있다. 먼저 이분류 문제에서 선험확률 $p(C_1)$과 $p(C_2)$가 동일하게 0.5라는 가정 하에서 오류확률이

최소가 되는 결정경계를 찾아보자. [그림 4-5]에서 결정경계가 수직선과 같이 주어졌을 때, 오류확률을 결정하는 영역 $\mathcal{E}_1 + \mathcal{E}_2$가 음영으로 표시되었다. 이 결정경계와 [그림 4-4]에서 찾아진 결정경계를 비교해 보면, [그림 4-5]에서 가로선으로 표시된 영역이 추가적으로 오류를 발생시키는 영역임을 알 수 있다. 이와 같이 결정경계를 적절히 움직여서 오류를 최소화하는 지점을 찾아보면, 두 확률밀도함수의 곡선 $p(\boldsymbol{x}|C_1)$과 $p(\boldsymbol{x}|C_2)$가 만나는 곳, 즉 [그림 4-4]에 표시된 영역이 최소 오류확률을 갖는 경계가 됨을 알 수 있다. 이는 앞 절에서 소개한 베이즈 분류기에 의해서 얻어지는 결정경계가 되어, 이 오류확률을 <베이즈 오류율(Bayes Error Rate)>이라고 한다.



[그림 4-5] 결정경계에 따른 오류확률의 변화

최소 오류확률을 주는 결정규칙과 베이즈 분류기와의 관계에 대해 수식을 통해 좀 더 자세히 알아보자. 앞에서 주어진 오류율에 대한 [식 4-12]에서, $\Re_1 \cup \Re_2$가 전체집합임을 이용하면 다음과 같은 관계식을 얻을 수 있다.

$$\begin{aligned} \mathrm{P}_{\mathrm{err}} &= p(C_1)\int_{\Re_2} p(\boldsymbol{x}|C_1)d\boldsymbol{x} + p(C_2)\int_{\Re_1} p(\boldsymbol{x}|C_2)d\boldsymbol{x} \\ &= p(C_1)\left(1 - \int_{\Re_1} p(\boldsymbol{x}|C_1)d\boldsymbol{x}\right) + p(C_2)\int_{\Re_1} p(\boldsymbol{x}|C_2)d\boldsymbol{x} \\ &= p(C_1) + \int_{\Re_1} (p(\boldsymbol{x}|C_2)p(C_2) - p(\boldsymbol{x}|C_1)p(C_1))d\boldsymbol{x} \end{aligned} \qquad \text{[식 4-15]}$$

이 오류값이 최소가 되는 결정경계(즉, 결정영역 $\Re_1$)를 찾기 위해서는 [식 4-15]의 마지막 적분항의 값을 최소로 만들면 된다. 이를 위해 먼저 적분 대상이 되는 함수가 0이 되는 점, 즉 $p(\boldsymbol{x}|C_2)p(C_2) = p(\boldsymbol{x}|C_1)p(C_1)$를 결정경계로 찾는다. 이 결정경계를 중심으로 적분대상 함수의 값이 음수가 되는 부분, 즉 $p(\boldsymbol{x}|C_2)p(C_2) < p(\boldsymbol{x}|C_1)p(C_1)$인 부분만을 결정영역 R_1으로 두어 적분의 대상이 되도록 함으로써 최소화가 가능하다. 따라서 오류확률을 최소로 하는 결정경계는 $p(\boldsymbol{x}|C_2)p(C_2) = p(\boldsymbol{x}|C_1)p(C_1)$를 만족하는 $\boldsymbol{x}$값이 되고, 이것을 정리하면 앞 절에서 정의된 우도비 분류에 의한 결정경계와 같은 [식 4-4]를 얻음을 알 수 있다. 결론적으로, 최소 오류확률을 가지는 결정경계는 [식 4-4]의 베이즈 결정경계와 동일하다.

이어서 클래스가 두 개 이상인 다중 클래스 분류에서 바르게 분류할 확률을 최대로 하는 결정규칙을 생각해 보자. [식 4-14]를 $p(\boldsymbol{x}|C_i)p(C_i) = p(\boldsymbol{x}, C_i) = p(C_i|\boldsymbol{x})p(\boldsymbol{x})$인 관계를 이용하여 다시 쓰면 다음과 같다.

$$\mathrm{P}_{\mathrm{correct}} = \sum_{i=1}^{M} \int_{\Re_i} p(\boldsymbol{x}|C_i)p(C_i)d\boldsymbol{x} = \sum_{i=1}^{M} \int_{\Re_i} p(C_i|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \qquad \text{[식 4-16]}$$

이 식으로부터 각 클래스별 결정영역에 대한 적분값들의 합을 최대로 하기 위해서는 각 클래스에 속할 확률 $p(C_i|\boldsymbol{x})$들을 계산하여 이 값을 최대로 하는 클래스를 선택하면 된다. 결국 이로부터 [식 4-8]에서 정의된 베이즈 결정규칙이 최대 분류율(혹은 최소 오류확률)을 주는 결정경계를 만들게 됨을 알 수 있다.

4.2.3 오류확률의 확장 - 베이즈 위험

결정규칙을 정할 때, 기본적으로는 앞에서 설명한 오류확률을 최소로 하는 결정경계를 찾는 것이 최선이라고 볼 수 있다. 그러나 때에 따라서는 단순한 오류확률이 아닌 문제의 특성에 맞는 기준을 사용해야할 필요가 있다.
예를 들어, 로켓 개발실에서 개발된 로켓의 각종 자료를 바탕으로 결함이 있는지 없는지를 판별하는 경우를 생각해 보자. 결함이 있는 경우를 없는 것으로 오분류하는 경우와 그 반대로 결함이 없는 것을 있는 것으로 오분류하는 경우는 그 판단 결과가 초래할 피해 정도는 매우 다르다. 따라서 현재 만들어진 패턴 분류기의 성능을 평가하기 위해서는, 단순한 오류확률이 아니라 두 가지 서로 다른 오분류에 대해 각각의 판단 결과가 초래하는 비용을 고려한 새로운 평가 기준의 설정이 필요하다. 이를 <베이즈 위험(Bayes risk)>이라고 한다.
이를 위해 먼저 비용계수를 정해야 한다. 즉, 클래스 C_1을 C_2로 오인식했을 때 발생하는 비용을 ρ_{12}라 하고, 반대의 경우를 ρ_{21}이라고 하여, 이 값을 각각 앞 절에서 계산한 오류확률에 곱하여 새로운 평가 기준인 베이즈 위험을 정의한다. [식 4-12]를 이용하여 이를 식으로 나타내면 다음과 같다.

$$\mathrm{P}_{\mathrm{risk}} = \rho_{12}p(C_1)\int_{\Re_2} p(\boldsymbol{x}|C_1)d\boldsymbol{x} + \rho_{21}p(C_2)\int_{\Re_1} p(\boldsymbol{x}|C_2)d\boldsymbol{x} \qquad \text{[식 4-17]}$$

앞 절에서 오류확률을 최소로 하는 결정경계를 찾을 때와 마찬가지로 베이즈 위험을 최소화하는 결정경계를 유도해 보면 다음과 같은 결정경계를 얻을 수 있다.

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\,C_1)}{p(\boldsymbol{x}|\,C_2)} - \frac{\rho_{21}p(C_2)}{\rho_{12}p(C_1)} = 0 \qquad \text{[식 4-18]}$$

베이즈 위험을 다중 클래스의 경우로 확장해 보자. 클래스 C_i에 속하는 데이터를 잘못하여 클래스 C_j에 속한다고 판단하였을 때 발생하는 비용을 ρ_{ij}라고 하면, 이 비용을 고려하여 C_i에 속하는 데이터를 다른 클래스로 잘못 분류하였을 때의 베이즈 위험은 다음과 같이 쓸

수 있다.

$$\mathrm{P}_i(\boldsymbol{x}) = \sum_{j \neq i} \rho_{ij} p(C_j|\boldsymbol{x}) = \sum_{j \neq i} \rho_{ij} p(C_i) \int_{\Re_j} p(\boldsymbol{x}|C_i) d\boldsymbol{x} \qquad \text{[식 4-19]}$$

이 베이즈 위험을 최소로 하는 결정규칙은 다음과 같이 정의할 수 있다.

$$y(\boldsymbol{x}) = \mathrm{argmin}_i \{\mathrm{P}_i(\boldsymbol{x})\} \qquad \text{[식 4-20]}$$

4.3 가우시안 확률분포와 베이지안 분류기

지금까지는 확률밀도함수의 구체적인 형태를 가정하지 않고, 단지 클래스별 데이터 분포 $p(\boldsymbol{x}|C_k)$가 주어졌다는 가정 하에서 논의하였다. 이 절에서는 각 클래스별 데이터 분포가 가우시안 분포를 따르는 경우에 대하여 구체적으로 결정경계와 판별함수가 어떻게 정해지는지 알아보겠다.

각 클래스가 가우시안 분포를 따르므로, 클래스 C_i에 대한 가우시안 분포의 파라미터인 평균과 공분산을 각각 $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$로 둔다. 이 파라미터를 가지는 가우시안 분포의 확률밀도함수는 다음과 같이 주어진다.

$$p(\boldsymbol{x}|C_i) = G(\boldsymbol{x};\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{2\pi^n}\sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \qquad \text{[식 4-21]}$$

계산을 간단히 하기 위하여 각 클래스의 선험확률은 동일하다고 가정한다. 따라서 이진분류기의 경우 판별함수는 [식 4-4]로부터 다음과 같이 정의된다.

$$g(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_1)}{P(\boldsymbol{x}|C_2)} - 1 = 0 \qquad \text{[식 4-22]}$$

가우시안 분포의 경우, [식 4-22]를 그대로 사용하는 대신 자연로그를 취하면 보다 간단한 형태의 판별함수를 얻을 수 있다. 즉, 로그를 취하여 얻어진 새로운 판별함수는 다음과 같이 정의된다.

$$\begin{aligned} \ell(\boldsymbol{x}) &= \ln g(\boldsymbol{x}) = \ln P(\boldsymbol{x}|C_1) - \ln P(\boldsymbol{x}|C_2) \\ &= -\frac{1}{2}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} - \frac{1}{2}\left\{\ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right\} = 0 \end{aligned} \qquad \text{[식 4-23]}$$

이를 확장하여 다중 클래스 분류를 위한 판별함수에 대해서도 [식 4-9]에 로그를 취한 형태를 생각한다. 이에 더하여 선험확률 $P(C_i)$는 모두 동일하다는 가정 하에 각 클래스에 대한 판별함수는 다음과 같이 주어진다.

$$\ell_i(\boldsymbol{x}) = \ln g_i(\boldsymbol{x}) = \ln p(x|C_i) \qquad \text{[식 4-24]}$$
$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + const$$

지금부터는 가우시안 분포에 있어서 공분산 행렬의 여러 가지 형태에 따른 구체적인 판별함수에 대해 살펴보겠다.

4.3.1 클래스 공통 단위공분산 행렬

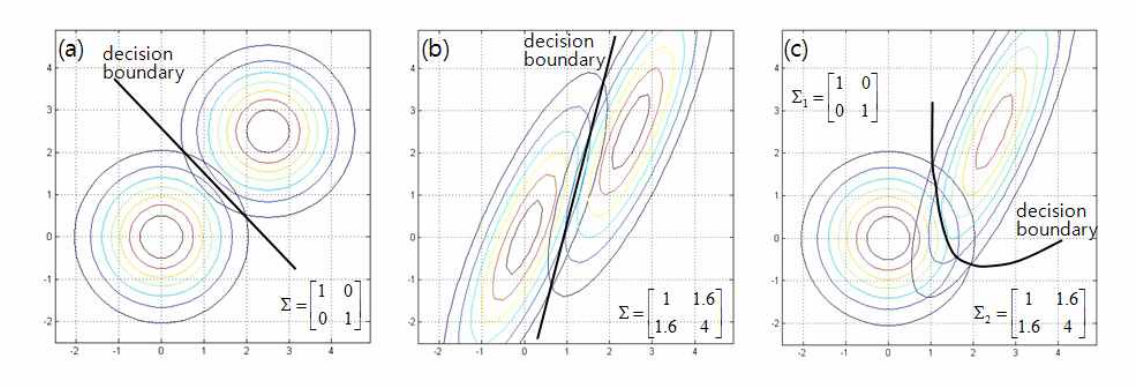
먼저 가장 간단한 경우로, 모든 클래스의 공분산이 동일하게 단위행렬의 상수 배인 행렬을 가지는 경우, 즉 $\boldsymbol{\Sigma_i} = \sigma^2\mathrm{I}\ (i = 1, \ldots, M)$를 만족하는 경우를 생각한다. 다중 클래스 분류를 위한 판별함수는 [식 4-24]로부터 다음과 같이 간단한 형태로 얻어진다.

$$\ell_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}-\boldsymbol{\mu}_i) - n\ln\sigma + const \qquad \text{[식 4-25]}$$

여기서 입력 데이터의 차원을 나타내는 n값과 분산 σ값은 모든 클래스에 공통되는 값이므로 분류에 영향을 주지 않는다. 따라서 이를 생략하여 보다 간단히 정리하면 다음과 같은 결정규칙을 얻을 수 있다.

$$y(\boldsymbol{x}) = \mathrm{argmin}_i\left\{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \qquad \text{[식 4-26]}$$

이 식이 가지는 의미는 입력 데이터에서 각 클래스의 평균까지의 거리를 계산하여 그 거리가 가장 작은 클래스로 분류하는 것으로, 이러한 분류기를 <최소거리 분류기 (Minimum Distance Classifier)>라고 한다. 이 분류기는 앞서 2장에서 살펴본 간단한 통계량을 이용한 분류기와 동일함을 알 수 있다. 이로부터, 처음으로 패턴인식기를 만들고자 할 때 누구나 쉽게 생각할 수 있는 평균과의 거리에 기반을 둔 분류기는, 이론적으로는 각 클래스가 모두 동일한 공분산을 가지고 그 공분산이 정방행렬의 상수배로 주어진다는 가정 하에서 최소 오류확률을 보장하는 분류기가 됨을 알 수 있다. 물론 이와 함께 각 클래스에 대한 선험확률도 모두 동일하다는 가정도 전제로 한다.



[그림 4-6] 클래스의 확률분포 형태에 따른 결정경계의 변화 (a) 두 클래스가 동일한 단위 공분산 행렬을 가지는 경우 (b) 두 클래스가 동일한 공분산 행렬을 가지는 경우 (c) 두 클래스가 서로 다른 공분산 행렬을 가지는 경우

이와 함께 [식 4-24]에 의해 정해지는 결정경계의 형태에 대해서도 생각해 보자. 앞서 2장에서 이진분류기에 대한 예에서 살펴본 바와 같이, 결정경계는 각 클래스의 평균을 잇는 직선에 수직이면서 그 중점을 지나는 직선이 된다. [그림 4-6]에 2차원 입력 데이터에 대해 각 클래스별 분포를 다양한 형태로 가정하여 각 분포의 등고선과 함께 결정경계를 나타내었다. [그림 4-6a]에서 두 클래스의 분포가 동일하고 원형으로 퍼지는 형태의 공분산을 가지는 경우에 결정경계가 직선으로 얻어짐을 확인할 수 있고, 이는 2장의 결과와 일치한다.
이와 같은 통계적인 고찰을 통해서 경험적으로 개발된 방법론에 대한 설명을 제공할 수 있고, 나아가 그 한계를 극복할 수 있는 방법을 제시할 수 있다. 여기서 살펴본 바에 따르면, 2장에서 제시한 간단한 분류기는 특정한 경우에 최소 오류확률을 보장하기는 하나, 그 가정(두 클래스가 모두 단위 공분산 행렬을 가지는 정규분포를 따름) 이 매우 제한적이기 때문에 실제 데이터가 이를 따를 것이라고는 생각하기 어렵다. 따라서 이 방법에 의한 분류기가 실제 응용에서 좋은 성능을 낼 것이라고 기대하기는 힘들다. 이러한 문제를 해결하기 위해서는 [식 4-26]을 유도하는 과정에서 주어진 가정을 완화하여 보다 일반적인 데이터 분포에 대해서도 적용 가능한 분류기를 만들어야 할 것이다.

4.3.2 클래스 공통 공분산 행렬

이제 앞 절에서의 가정을 조금 완화하여, 모든 클래스에 대해 동일한 공분산을 가지지만 그 형태는 일반적인 행렬이 되는 경우, 즉 $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$인 경우를 생각한다. 이 공분산에 따른 데이터 분포는 타원형 형태가 된다([그림 4-6b] 참조). 앞 절에서와 마찬가지로 선험확률 $P(C_i)$는 모두 동일하다는 가정 하에 생략하고, [식 4-23]로부터 분류에 영향을 미치지 않는 상수항을 제거하면 다음과 같은 간단한 판별함수와 결정규칙을 얻는다.

$$\ell_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) \qquad \text{[식 4-27]}$$

$$y(\boldsymbol{x}) = \text{argmin}_i\left\{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \qquad \text{[식 4-28]}$$

이 결정규칙은 [식 4-26]의 경우와 달리 클래스의 공분산 $\boldsymbol{\Sigma}$를 고려하고 있다. 즉, 데이터 $\boldsymbol{x}$에서 평균까지의 거리를 계산할 때 공분산도 함께 고려하여 거리를 계산하는 것으로, 이와 같은 거리를 <마할라노비스 거리(Mahalanobis distance)>라고 한다. 만약 공분산 $\boldsymbol{\Sigma}$가 대각행렬이 되어 데이터의 각 요소들 사이의 상관관계는 존재하지 않고 각 요소들의 분산만 고려가 되는 경우에는, 먼저 각 요소별로 그 표준편차 값을 나누어 준 후에 일반적인 유클리디안 거리를 계산하는 방식이 되어 이러한 거리를 <정규화된 유클리디안 거리 (Normalized Euclidean Distance)>라고 한다.

[그림 4-6b]에 [식 4-28]에 의한 결정경계를 예로 나타내었다. 그림에서 알 수 있듯이 결정경계는 여전히 직선 형태를 가지게 된다. 그러나 [식 4-26]대신 [식 4-28]을 사용하면, 입력벡터의 각 요소들 사이의 상관관계를 고려하게 되어 보다 좋은 성능을 기대할 수 있다. 예를 들어, 각 사람의 신체 정보를 받아서 남, 여 성별을 분류하는 문제를 생각해 보자. 이 때 신체 정보로서 신장과 발 크기에 대한 데이터를 이용하여 2차원 벡터를 만들어 입력으로 사용한다고 하자. 같은 길이 단위를 사용하는 경우 발 크기 데이터의 분산에 비해 신장 데이터의 분산이 큰 값을 가지므로 이를 그대로 사용하면 [식 4-26]의 거리 계산식에서는 신장 데이터가 더 큰 영향을 미치게 된다. 다시 말해 각 요소의 중요도와는 상관없이 분산값에 의해 각 요소의 반영 정도가 달라진다. 이러한 문제를 해결하기 위해서는 각 요소의 분산을 1로 정규화 하는 과정이 필요하다. 이를 공분산에도 확장하여 두 요소들 간의 상관관계가 없는 형태로 데이터 변형을 수행한 후 유클리디안 거리를 계산함으로써 입력 데이터 자체가 가지는 표현 범위나 분산 등에 의해 거리가 왜곡되는 것을 방지할 수 있다. [식 4-28]은 공분산이 단위행렬이 되도록 데이터를 정규화 한 후 거리를 계산하는 것과 같은 효과를 가진다.

4.3.3 일반적인 공분산 행렬

이제 가장 일반적인 경우로, 각 클래스의 공분산이 서로 다른 일반적인 형태를 가지는 경우($\boldsymbol{\Sigma_i} \neq \boldsymbol{\Sigma_j}$)를 생각한다. 이는 각 클래스의 분포 형태가 서로 다른 타원형을 가지는 경우이다. 각 클래스의 선험확률이 같은 경우, 공분산 행렬이 각 클래스마다 모두 다르므로 판별함수는 [식 4-24]에서와 같이 주어진다. 이 판별함수를 간단히 정리하여 결정규칙을 만들면 다음과 같은 식을 얻을 수 있다.

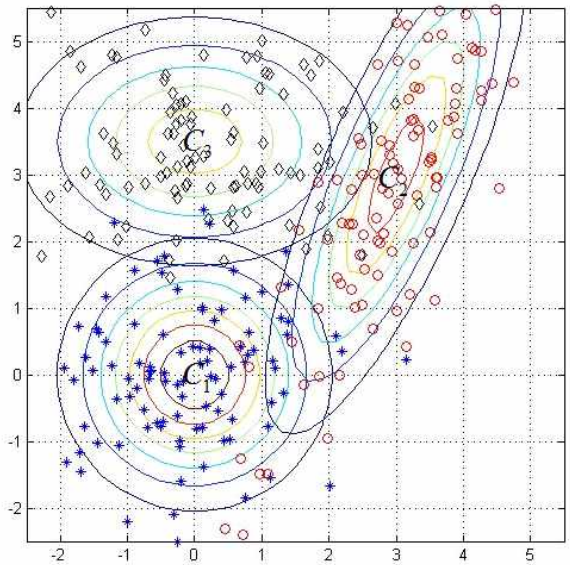
$$y(\boldsymbol{x}) = \mathrm{argmin}_i \left\{ (\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln |\boldsymbol{\Sigma}_i| \right\} \qquad \text{[식 4-29]}$$

이 결정규칙을 이용하면, 하나의 데이터 $\boldsymbol{x}$가 주어진 경우 각 클래스와의 거리를 계산할 때 각 클래스에 대해 정해지는 공분산 행렬 $\boldsymbol{\Sigma_i}$도 함께 고려한 거리를 계산해 주어야 한다. [그림 4-6c]에 일반적인 공분산을 가지는 데이터 집합에서 각 데이터의 분포함수에 대한 등고선과 결정경계함수를 나타내었다. 그림에서 알 수 있듯이, 앞의 두 경우에서처럼 각 클래스가 같은 공분산을 가지는 경우는 직선 결정경계를 가지는 데 반해, 서로 다른 공분산을 가지는 경우는 자연스러운 곡선 형태의 결정경계를 가지기 때문에 보다 다양한 데이터 분포에 대해 알맞은 결정규칙을 제공할 수 있을 것이다.

그러나 데이터 집합으로부터 결정규칙 [식 4-29]를 얻기 위해서는 각 클래스별로 파라미터 μ_i와 Σ_i를 추정해 주어야 하므로 앞의 두 경우에 비해 추정 파라미터의 수가 많아져서 파라미터 추정 시에 발생하는 추정 오차가 최종 결정규칙에 나쁜 영향을 줄 가능성이 높아지는 문제를 가지고 있다.

4.4 매트랩을 이용한 베이지안 분류기 실험

지금까지 살펴본 내용을 바탕으로 간단한 2차원 데이터의 분류를 위한 베이지안 분류기를 만들고 성능을 평가해 보겠다. 서로 다른 공분산을 가지는 3개의 클래스로부터 각각 100개씩의 데이터를 생성하여 학습에 활용하고, 성능 평가를 위해서는 같은 분포로부터 각 클래스당 10^5개의 데이터를 생성하여 테스트 오차를 계산하였다. 테스트를 위해 충분히 많은 양의 데이터를 사용하는 이유는 가능한 일반화 오차에 가까운 오차값으로 성능을 평가하기 위함이다. [그림 4-7] 에는 학습에 사용된 데이터의 분포의 보여주고 있다.



[그림 4-7] 베이즈 분류기를 이용한 데이터 분류 문제

각 클래스별 데이터 생성을 위하여 사용된 평균과 공분산은 다음과 같다.

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 1.6 \\ 1.6 & 4 \end{bmatrix}, \mu_3 = \begin{bmatrix} 0 \\ 3.5 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

학습 데이터를 생성할 때 가정한 분포는 각 클래스들이 모두 서로 다른 공분산을 가지는 경우이므로 4.3.3절의 결정규칙 [식 4-29]를 사용하는 것이 적합할 것이다. 그러나 여기서는 간

단한 두 가지 방법도 함께 적용하여 그 성능을 비교한다.
각 방법에 대한 결정규칙, 즉 [식 4-26], [식 4-28], [식 4-29]를 찾기 위해서는 먼저 학습 데이터를 이용하여 필요한 파라미터를 추정해야 한다. 우선 각 클래스의 평균은 세 가지 규칙에서 모두 사용되므로 각 클래스의 표본평균을 계산하여 사용한다. 주어진 학습 데이터로부터 추정된 표본평균의 값은 다음과 같다.

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

이 표본평균만으로 가장 간단한 [식 4-26]의 결정규칙을 얻을 수 있다. 이어서 [식 4-28]과 [식 4-29]의 결정규칙을 얻기 위하여 표본공분산을 계산하면 각 클래스에 대한 표본 공분산은 다음과 같이 주어진다.

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

이를 이용하여 [식 4-29]의 결정규칙을 얻을 수 있다. 그런데 [식 4-28]의 경우는 각 클래스의 공분산이 모두 동일하다는 가정 하에 유도된 것이므로, 이 문제에서는 각 클래스의 공분산들의 평균을 계산하여 대신 사용한다. 추정된 공분산은 다음과 같다.

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

이렇게 얻어진 결정규칙을 이용하여 학습 데이터에 대한 분류를 수행하는 프로그램을 [프로그램 4-1]에 나타내었다. 먼저 학습 데이터를 불러들여 표본평균과 표본공분산을 계산하고, 판별식 [식 4-28]을 위한 표본공분산들의 평균을 계산해 둔다. 이를 이용하여 각 학습 데이터에 대해 세 클래스에 대한 판별함수의 값, 즉 [식 4-26], [식 4-28], [식 4-29]에서 각각 정의된 거리함수의 값을 계산하여, 이 값이 가장 작은 클래스로 데이터를 할당한다. 만약 할당된 클래스가 원래 데이터를 생산한 클래스와 다르면 오분류된 데이터의 수를 증가시킨다. 모든 데이터에 대한 처리가 끝나면 오분류율을 계산한다. 학습 데이터에 대하여 [프로그램 4-1]을 수행한 결과와 테스트 데이터에 대해서 동일한 처리 과정을 수행한 결과를 [표 4-1]에 나타내었다. 표에서 적절한 가정에 의해 유도된 결정규칙 [식 4-29]를 사용한 결과가 가장 좋은 성능을 보임을 확인할 수 있다. 학습 데이터에 대해 [식 4-26]에 비해 [식 4-28]의 결과가 더 좋지 못한 것은 부적절한 가정으로 인해 잘못 추정된 공분산행렬이 오히려 더 좋지 못한 결정경계를 제공했기 때문으로 추측할 수 있다.

[표 4-1] 베이즈 분류기의 분류결과

결정규칙	[식 4-26]	[식 4-28]	[식 4-29]
오분류율(학습오차)	12%	13%	9.67%
오분류율(테스트오차)	11.55%	10.88%	9.67%

프로그램 4-1 Bayes Classifier

베이즈 분류기를 이용하여 데이터를 분류하고 학습 오차를 출력

```
001 load dataCh4_7                                  %데이터 불러오기
002 K=3;                                            %클래스의 수
003 M=[mean(X1);mean(X2);mean(X3)]                  %클래스별 표본평균 계산
004 S(:,:,1)=cov(X1);                               %클래스별 표본공분산 계산
005 S(:,:,2)=cov(X2);
006 S(:,:,3)=cov(X3);
007 smean=(cov(X1)+cov(X2)+cov(X3))/3               %클래스별 표본공분산의 평균
008 Dtrain=[X1;X2;X3];                              %학습 데이터 구성
009
010 Etrain=zeros(3,1);                              %오분류 데이터의 수를 셈
011 N = size(X1,1);                                 %각 클래스별 데이터의 수
012 for k=1:K                                       %각 클래스별로 분류 시작
013     X=Dtrain((k-1)*100+1:k*100,:);
014     for i=1:N                                   %각 데이터에 대해 분류 시작
015         for j=1:K                               %세 개의 판별함수의 값 계산
016             %단위 공분산행렬을 가정한 경우의 판별함수
017             d1(j,1)=(X(i,:)-M(j,:))*(X(i,:)-M(j,:))';
018             %모든 클래스에 동일한 공분산행렬을 가정한 경우의 판별함수
019             d2(j,1)=(X(i,:)-M(j,:))*inv(smean)*(X(i,:)-M(j,:))';
020             %일반적인 공분산행렬을 가정한 경우의 판별함수
021             d3(j,1)=(X(i,:)-M(j,:))*inv(reshape(S(:,:,j),2,2))*(X(i,:)-M(j,:))';
022         end
023         [min1v, min1i]=min(d1);                 %각 판별함수 값에 따라 분류
024         if (min1i~=k) Etrain(1,1) = Etrain(1,1)+1; end %분류결과가
025         [min2v, min2i]=min(d2);                 %원래 클래스와 다르면 오류증가
026         if (min2i~=k) Etrain(2,1) = Etrain(2,1)+1; end
027         [min3v, min3i]=min(d3);
028         if (min3i~=k) Etrain(3,1) = Etrain(3,1)+1; end
029     end
030 end
031 Error_rate = Etrain/(N*K)                       %오분류율 (학습오차) 출력
```

[프로그램 4-2]에는 베이지안 분류기에 의해 찾아지는 결정경계를 그려보는 매트랩 코드를 나타내었다. 데이터가 생성되는 2차원 입력 공간 영역을 등간격의 세밀한 격자로 나누어 입력행렬 (프로그램 4-2에서 정의된 행렬 XY)을 만들고, 각 입력값에 대하여 베이지안 분류를 수행하여 그 결과에 따라 결정경계를 나누었다. 이때 [프로그램 4-1]에서와 같이 공분산에 대한 서로 다른 세 가지 가정에 대해 각각 베이지안 분류기를 만들어 분류를 수행하였으며,

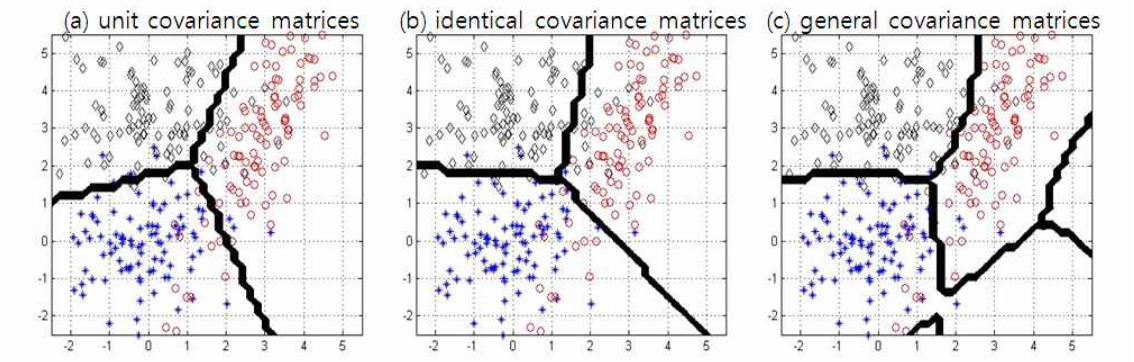
그 결과가 [그림 4-8]에 나타나 있다.

프로그램 4-2 Drawing Decision Boundary

베이즈 분류기에 의해 얻어지는 결정경계를 2차원 평면상에 나타냄

```
001 load dataCh4_7                                  %데이터 로드
002 K=3;                                            %클래스의 수
003 M=[mean(X1);mean(X2);mean(X3)]                  %클래스별 표본평균 계산
004 S(:,:,1)=cov(X1);                               %클래스별 표본공분산 계산
005 S(:,:,2)=cov(X2);
006 S(:,:,3)=cov(X3);
007 smean=(cov(X1)+cov(X2)+cov(X3))/3               %클래스별 표본공분산들의 전체 평균
008 % 2차원 입력공간을 격자 형태로 나누어 입력 행렬을 만듦
009 [x,y]=meshgrid([-2.5:0.2:5.5],[-2.5:0.2:5.5]);
010 XY=[x(:), y(:)];
011 % 입력 행렬의 각 점에 대해 세 가지 베이지안 분류기의 분류결과 계산
012 for i=1:size(XY,1)
013     for j=1:K
014         d1(j,1)=(XY(i,:)-M(j,:))*(XY(i,:)-M(j,:))';
015         d2(j,1)=(XY(i,:)-M(j,:))*inv(smean)*(XY(i,:)-M(j,:))';
016         d3(j,1)=(XY(i,:)-M(j,:))*inv(reshape(S(:,:,j),2,2))*(XY(i,:)-M(j,:))';
017     end
018     [min1v, min1i]=min(d1); res_classify(1,i)=min1i;
019     [min1v, min2i]=min(d2); res_classify(2,i)=min2i;
020     [min1v, min3i]=min(d3); res_classify(3,i)=min3i;
021 end
022 for m=1:3                          % 분류 결과에 따라 데이터와 결정경계를 나타냄
023     figure(m); hold on
024     axis([-2.5 5.5 -2.5 5.5]); grid on
025     plot(X1(:,1), X1(:,2), '*');        % 클래스 C1 데이터 표시
026     plot(X2(:,1), X2(:,2), 'ro');       % 클래스 C2 데이터 표시
027     plot(X3(:,1), X3(:,2), 'kd');       % 클래스 C3 데이터 표시
028     res=reshape(res_classify(m,:),size(x));
029     contour(x,y,res);                   % 결정경계 그리기
030 end
```

[그림 4-8]로부터, [그림 4-6]에서 설명한 바와 같이 각 클래스의 공분산이 동일하다고 가정한 경우에는 직선 형태의 결정경계가 찾아지고 그렇지 않은 일반적인 가정 하에서는 곡선형태의 결정경계가 찾아짐을 확인할 수 있다. 이때 결정경계가 매끄러운 직선이나 곡선으로 나타나지 않은 이유는, 결정경계를 그리기 위해 입력공간을 격자로 나눌 때 그 간격이 세밀하지 못한 때문으로, 실험에서는 이를 0.2로 설정하였으나 이 값을 줄이면 보다 매끄러운 직선(혹은 곡선)을 얻을 수 있다.



[그림 4-8] 베이즈 분류기에 의해 찾아진 결정경계
(a) 단위 공분산 행렬을 가정한 경우 (b) 모든 클래스에 동일한 공분산 행렬을 가정한 경우 (c) 각 클래스에 서로 다른 공분산 행렬을 가정한 경우

이상에서 살펴본 바와 같이 베이지안 분류기를 사용할 경우에는 데이터의 분포에 대해 어느 정도 사전 지식을 가지고 그에 적절한 확률모델을 설정하여 결정규칙을 유도해 내는 것이 필요하다. 특히 가우시안 밀도함수를 가정한 베이지안 분류기는 그 구현이 간단하고 직관적으로도 그 의미가 확실하여 널리 사용되고 있으나, 데이터 분포가 가우시안 함수와는 거리가 먼 복잡한 형태를 가지고 있는 경우에는 좋은 성능을 기대하기 힘들다. 이러한 문제를 해결하기 위하여 3장의 비모수적 밀도 추정에 의한 방법이나 9장의 가우시안 혼합모델을 이용한 밀도추정 방법을 적용할 수 있다.

연습 문제

1. 두 클래스의 확률밀도함수가 각각 가우시안 확률분포를 따르고 그 평균과 분산이 다음과 같이 주어진 경우, 우도비 검증에 의한 결정규칙을 찾으시오. 단, 선험확률은 $P(C_1) = P(C_2)$라고 가정한다.

   $\mu_1 = 1,\ \sigma_1 = 1, \qquad \mu_2 = 5, \sigma_2 = 1$

2. 1번 문제에 대하여, 오분류에 대한 베이즈 위험이 다음과 같이 주어졌다고 할 때 결정경계는 어떻게 되는지 찾으시오.

   $\rho_{12} = 1,\ \rho_{21} = 2$

3. 다음에 주어진 단계에 따라 데이터를 생성하고 베이즈 분류기를 이용하여 분류를 수행하시오.

(1) 매트랩을 이용하여 다음과 같은 평균과 공분산을 가지는 가우시안 분포를 따르는 2차원 데이터를 각각 100개씩 가지는 두 클래스 집합 C_1, C_2를 생성하시오. 생성된 데이터를 2차원 평면상의 점으로 표시한 그래프를 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, (1)에서 생성한 데이터 집합을 이용하여 각 클래스의 확률밀도함수 $p(\boldsymbol{x}|C_k)$의 파라미터(평균과 공분산)를 추정하시오. (힌트: 매트랩을 이용하여 데이터 집합의 표본평균과 표본공분산을 계산한다.)

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

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

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

4. [식 4-5]에 의해 분류를 수행한 결과와 [식 4-6]에 의해 분류를 수행한 결과가 현저하게 달라질 수 있는 응용 사례를 생각해 보시오.

5. 베이즈 위험에 의한 결정경계를 선택하는 것이 필요한 실제 응용 사례를 생각해 보시오.

참고 자료

이 장에서는 확률밀도 함수를 추정하고 이를 바탕으로 분류를 수행하는 베이지안 분류기에 대하여 알아보았다. 베이지안 분류기를 설계하기 위해서는 확률밀도함수를 추정하는 것이 선행되어야 하는데, 이와 관련된 내용은 3장에서 살펴보았으며, 심화된 내용에 대해서는 2장에서 소개한 참고문헌을 활용할 수 있을 것이다. 또한, 이 장에서는 주로 모수적 확률밀도함수 추정법에 기반하여, 각 클래스가 가우시안 분포를 따른다고 가정하여 구체적인 판별함수를 유도해 보았다. 그러나 주어진 데이터가 이산형태의 데이터인 경우에는 가우시안 분포를 가정하는 것이 적절치 못할 수 있으며, 또한 데이터 특성을 고려하여 다양한 확률모델을 활용할 수 있을 것이다. 다양한 확률 모델에 대해서는 3장에서 소개한 수리통계학 교과서 [Hogg, McKean & Craig 05]를 참조하기 바란다. 마지막으로, 이 장에서 소개한 베이즈 결정이론을 시작으로 하여, 베이즈 정리에 기반한 매우 다양한 기계학습 이론들이 연구되어 왔다. 이에 대해서는 1장에서 소개한 [Bishop 06]에 자세히 소개되어있다.

[Hogg, McKean & Craig 05] R. V. Hogg, J. W. McKean, A. T. Craig. Introduction to Mathematical Statistics (6th ed.). Prentice Hall, 2005.
[Bishop 06] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006