지금까지의 절에서는 logistic regression을 사용해서 class 별 확률값을 추산했다.
이제 다른 방식을 사용해서 이 확률을 추정할 것이다.
X의 분포를 class 별로 각각 나누어지게 모델링하고, 이 분포를 확률값으로 바꾸기 위해 베이즈 정리를 사용할 것이다.
이미 logistic regression이 있는데, 왜 다른 모델을 사용해야 할까?

- 두 클래스 간의 상당한 분리가 있을 때(직역된 문장), logistic regression의 파라미터 추정치는 불안정하다.
- 분포가 정규분포를 따르고 데이터의 크기가 작을 경우, 더 정확할 수 있다.
- 다중클래스 문제로 변형없이 확장할 수 있다.

K개의 클래스 중 하나로 관측값을 분류한다고 가정했을 때, πk를 무작위로 선택된 관측값이 k번째 클래스에서 나올 전반적인 확률 또는 사전 확률이라고 하자.

fk(X)를 k번째 클래스에서 나온 관측값에 대한 X의 밀도 함수, 즉 Pr(X|Y = k)라고 하자.

k번째 클래스의 관측값이 X = x를 가질 확률이 높다면 fk(x)는 상대적으로 큰 값을 가지며, k번째 클래스의 관측값이 X = x를 가질 가능성이 매우 낮다면 fk(x)는 작은 값을 가진다.

pk(x) = Pr(Y = k|X = x)
이는 관측값 X = x가 k번째 클래스에 속할 사후 확률이다. 즉, 해당 관측값의 예측 변수값이 주어졌을 때, 그 관측값이 k번째 클래스에 속할 확률을 의미한다.

```
용어 정리

qualitative response variable : 정성적 반응 변수(Y)
prior probability : 사전확률
density function : 밀도 함수
posterior probability : 사후확률
```

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

이 식은 사후 확률 pk(x)를 직접 계산하는 대신, 단순히 πk와 fk(x)의 추정치를 (4.15)에 대입할 수 있음을 시사한다.
일반적으로, 모집단에서 무작위 표본을 가지고 있다면 πk를 추정하는 것은 쉽다(Sampling). 단순히 k번째 클래스에 속하는 훈련 관측치들의 비율을 계산하면 된다.
하지만, 밀도 함수 fk(x)를 추정하는 것은 훨씬 더 어려운 과제이다. 앞으로 보겠지만, fk(x)를 추정

하기 위해서는 일반적으로 몇 가지 단순화 가정을 해야 할 것이다.
--> 밀도 함수를 추정하는 것이 주요 과제!!

만약 모집단의 확률분포를 알아 계산할 수 있을 때, 이를 이용한 분류기를 베이즈 분류기라고 부른다.
베이즈 분류기:

1. 이론적으로 가장 낮은 오차율을 달성할 수 있는 이상적인 분류기입니다
2. 각 클래스의 실제 확률 분포를 알고 있다고 가정합니다
3. 새로운 데이터 포인트가 주어졌을 때, 각 클래스에 속할 사후 확률을 계산하고 가장 높은 확률을 가진 클래스로 할당합니다
4. 실제로는 진짜 확률 분포를 알 수 없기 때문에 구현이 불가능합니다
   모집단의 확률 분포에 따라 만든 베이즈 분류기에 근사하는 것이 목표!

4.4.1 Linear Discriminant Analysis for p =1

예측 변수가 하나라고 가정하고 서술하겠다. 또한 정규분포를 따른다.
우선 밀도 함수를 정의해야 한다.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

가우시안 분포를 따르게 만든다.

u = k 클래스의 평균 x값
시그마^2 = 분산
각 클래스마다 분산은 같다고 가정한다.
exp 안에 마이너스 항이라는 것을 기억하자!

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

클래스별 확률을 위와 같이 추산할 수 있다. 어떻게 보면 softmax 함수와 비슷해보이기도 하다.

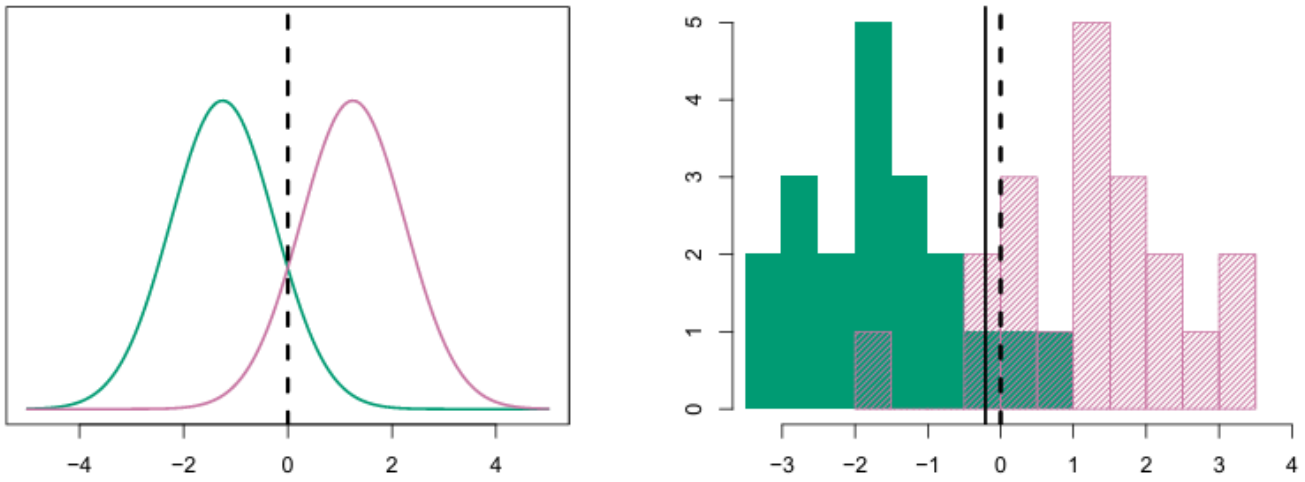


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

확률을 계산한 식에 log를 씌우면 다음과 같다:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

위 식이 클수록 어떤 입력값이 k클래스일 것이라는 추정의 정확도가 올라가야 한다.
또한 각 클래스에 대한 파이k 값이 같을 때 x 값도 계산할 수 있다.

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

linear discriminant analysis (LDA)
필요한 통계적 추정치에 대한 산식은 다음과 같다:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

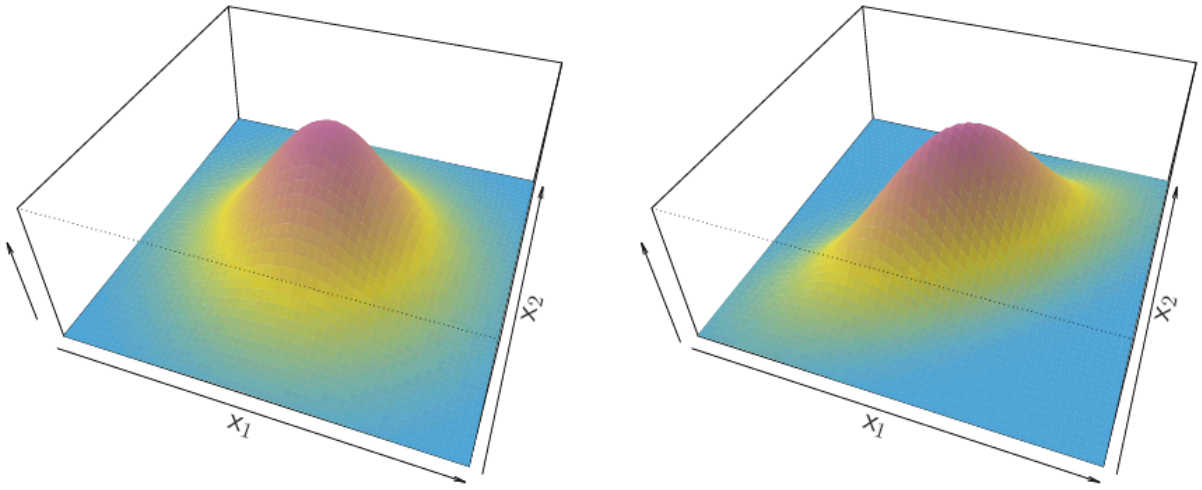
$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

πk는 무작위로 선택한 관측값이 특정 클래스일 확률이니, 다음과 같다:

$$\hat{\pi}_k = n_k/n.$$

4.4.2 Linear Discriminant Analysis for p>1

다변량 가우시안 분포는 변수들이 각자 정규분포를 따르고, 각각 상관관계를 가진다.
이 전제 하에서 p = 2일 경우 특정 클래스에 대한 가우시안 분포를 나타내면 다음과 같다:



예측 변수 x1, x2의 한 점에 대한 높이는 확률값이다.
왼쪽의 그래프는 두 변수의 분산이 같고, 상관관계가 0일 때이고, 오른쪽의 그래프는 상관관계가 0.7일 때의 그래프이다. 이 그래프처럼 두 변수간의 상관관계가 있다면 그래프의 bell 형태는 무너진다.

가우시안 밀도 함수를 정의하는 것은 단변량과 비슷하지만 다르다. 하지만 변수가 하나인 경우를 상정했을 때, 똑같은 식으로 귀결되는 것을 알 수 있다.

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

x : vector with p components(each data point)
시그마 : Covariance Matrix
u : vector with p components(mean of X) class 별로 있음
가장 다른 점은 기존의 분산 대신 공분산 행렬을 사용하는 것이라고 볼 수 있다.
-> 변수들간의 상관관계를 고려