

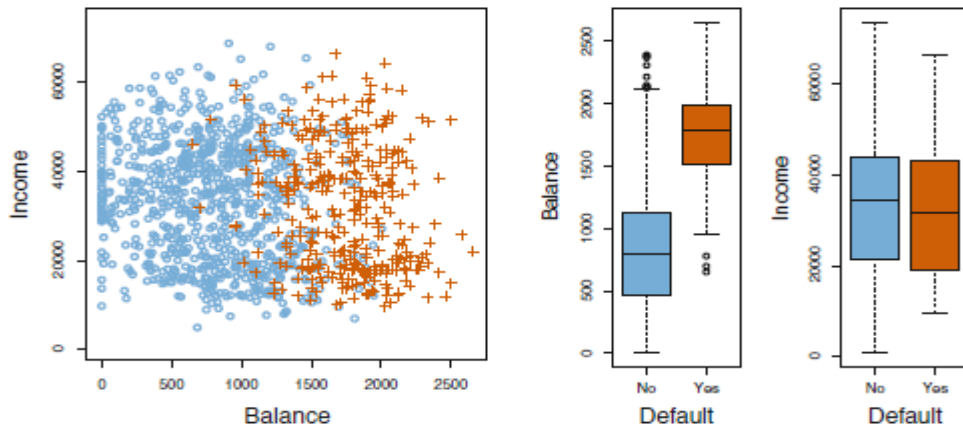
iv. Statistical Learning

4.1 분류의 개요

🚦 분류 : 관측치에 대한 질적 반응 변수를 예측하는 것

🚦 간단한 예시

- ✓ 온라인 뱅킹 서비스는 사용자의 IP 주소, 과거 거래이력 등을 바탕으로 현지에서 진행되고 있는 거래가 사기성인지를 결정할 수 있어야 한다.
- ✓ 응급실에 오는 환자는 3가지 의료상태중 하나일 때 이 환자는 3가지 중 어느 상태에 해당되는가?



- ✓ 오렌지색 : 주어진 달에 연체했던 사람들
- ✓ 파란색 : 주어진 달에 연체하지 않았던 사람들
- ✓ Balance : 월간 신용카드 대금, Income : 연간 수익
- ✓ 주어진 balance X1과 income X2에 대해 default Y를 예측하는 모델 배움

4.2 왜 선형회귀를 사용하지 않는가?

🚦 이유

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

- ✓ 이러한 코딩은 결과에 순서가 있다는 것을 의미한다. 각각의 변수사이의 차이가 동일 한 것이다.

- ✓ 이진 질적 반응변수의 경우에는 가변수 방법을 사용하면 반응변수를 다음과 같이 코딩한다.

$$Y = \begin{cases} 0 & \text{if stroke;} \\ 1 & \text{if drug overdose.} \end{cases}$$

- ✓ $\hat{Y} > 0.5$ 면 drug overdose, 그렇지 않으면 stroke로 예측할 수 있다.
- ✓ 만약 선형회귀를 사용하면 추정치 중 일부는 [0,1] 범위 밖에 놓일 수 있어 확률로 해석하기 어렵다.
- ✓ 가변수 방식은 질적 변수가 3-레벨 이상인 경우 확장이 쉽지 않다.

4.3 로지스틱 회귀(Logistic Regression)

- 📌 로지스틱 회귀는 반응변수 Y를 직접 모델링하지 않고 Y가 특정 범주에 속하는 확률을 모델링 한다.

$$\Pr(\text{default} = \text{Yes} | \text{balance}).$$

4.3.1 로지스틱 모델

- 📌 $P(X) = \Pr(Y=1 | X)$ 사이의 관계를 어떻게 모델링 할까?

- ✓ $P(X) = \beta_0 + \beta_1 X \rightarrow$ 문제점: 일부 X 값에 대해서 $P(X) < 0$ 이고 일부 다른 경우는 $P(X) > 1$ 이 될수 있다.
- ✓ 이러한 문제를 해결하기 위해 모든 X값에 대해 0과 1 사이의 값을 제공하는 로지스틱 함수 사용

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

- ✓ 모델의 적합을 위해 최대가능도 사용(다음 절에서 다룸)

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}.$$

- ✓ $p(X)/[1-p(X)]$ 은 공산(odds)이라 하며 항상 0과 무한대사이의 값을 가진다. 공산이 0에 가까울수록 연체 확률이 매우 낮고 무한대일수록 연체확률이 아주 높다는 것이다. (배팅 전략에 사용 ex) 경마)

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X.$$

- ✓ 로지스틱 회귀모델은 X에 선형적인 로짓을 가진다.
- ✓ P(X)와 X사이에 직선 상관관계가 없고 X의 유닛 변화당 p(X) 변화율이 X의 현재 값에 따라 다르다.

4.3.2 회귀계수의 추정

🌈 최대가능도(Maximum Likelihood)

$$\iota(\beta_0, \beta_1) = \prod_{i: y_i=1} p(x_i) \prod_{i': y_{i'}=0} (1 - p(x_{i'}))$$

- ✓ 추정치 β_0, β_1 은 이 가능도 함수를 최대화하도록 선택된다.

4.3.3 예측하기

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1,000}}{1 + e^{-10.6513 + 0.0055 \times 1,000}} = 0.00576,$$

$$\widehat{\Pr}(\text{default=Yes} | \text{student=Yes}) = \frac{e^{-3.5041 + 0.4049 \times 1}}{1 + e^{-3.5041 + 0.4049 \times 1}} = 0.0431,$$

$$\widehat{\Pr}(\text{default=Yes} | \text{student=No}) = \frac{e^{-3.5041 + 0.4049 \times 0}}{1 + e^{-3.5041 + 0.4049 \times 0}} = 0.0292.$$

- ✓ 가변수도 조건부확률로 구할 수 있다.

4.3.4 다중로지스틱

🌈 다수의 설명변수들을 사용하여 이진 반응변수 값을 예측

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}.$$

- ✓ 추정치 β_0, β_1 은 이 가능도 함수를 최대화하도록 선택된다.

🌈 교란(confounding)

- ✓ 선형회귀에서와 같이 하나의 설명변수를 사용하여 얻은 결과는 다수의 설명변수를 사용하여 얻은 결과와 상당히 다를 수 있다.(특히 설명변수들 사이에 상관성이 있는 경우)

4.3.5 반응변수의 클래스가 2개보다 많은 로지스틱 회귀

🌈 반응 변수가 세개라면 예시로 뇌졸중, 약물 과다복용, 간질발작

- ✓ $\Pr(Y = \text{뇌졸중} \mid X), \Pr(Y = \text{약물과다복용} \mid X),$
- ✓ $1 - \Pr(Y = \text{뇌졸중} \mid X) - \Pr(Y = \text{약물과다복용} \mid X)$
- ✓ 하지만 실제로는 판별분석(discriminant analysis) 방법이 다중클래스 분류에 일반적으로 사용되어서 잘 안 씀

4.4 선형판별분석(Linear Discriminant Analysis)

🌈 기존의 로지스틱 함수 : $\Pr(Y = k \mid X = x)$ 를 직접 모델링

- ✓ 주어진 설명변수 X 에 대해 반응변수 Y 의 조건부 분포를 모델링

🌈 선형판별분석

- ✓ 반응변수 Y 의 각 클래스에서 설명변수 X 의 분포를 모델링하고, 베이지 정리를 사용하여 $\Pr(Y = k \mid X = x)$ 의 추정치를 구한다.
- ✓ 이때 X 의 분포를 정규분포로 가정하면 로지스틱 모델과 유사

🌈 선형판별분석의 필요성

- ✓ 클래스들이 잘 분리될 때 로지스틱 모델에 대한 모수 추정치는 불안정
- ✓ n 이 작고 각 클래스에서 X 의 분포가 근사적으로 정규분포이면 선형판별모델이 더 안정
- ✓ 반응변수의 클래스가 2보다 클 때 일반적으로 선형판별분석 이용

4.4.1 분류를 위한 베이즈 정리의 사용

- ✓ π_k : 무작위로 선택된 관측치가 k번째 클래스에서 나올 전체 확률(사전 확률)
- ✓ $f_k(X) = \Pr(X = x | Y = k)$: k번째 클래스에 속하는 관측치에 대한 X의 밀도함수

📌 베이즈 정리

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

- ✓ $p_k(X) = \Pr(Y = k | X)$: π_k 와 $f_k(X)$ 의 추정치를 대입하겠다는 의미
- ✓ $f_k(X)$ 추정은 밀도에 대한 어떤 단순한 형태를 가정
- ✓ $p_k(X) = \Pr(Y = k | X)$ 는 관측치 $X=x$ 가 k번째 클래스에 속하는 사후 확률

4-4-2. 선형판별 분석 ($p = 1$)

$f_k(X)$ 가 정규분포 또는 가우스분포라고 가정

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

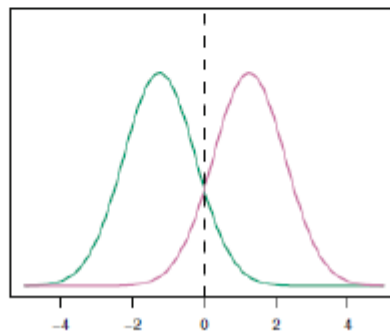
- μ_k, σ_k^2 : k번째 클래스에 대한 평균과 분산

- $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$: 모든 K 개 클래스에 대한 공통의 분산 존재 $= \sigma^2$

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

베이지 분류기는 $p_k(x)$ 가 최대가 되는 클래스에 관측치 $X = x$ 를 할당한다. 이를 로그를 취하고 정리하면

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



- 초록색: $f_1(x)$ $\mu_1 = -1.25$ $\sigma_1^2 = 1$ $\pi_1 = 0.5$
- 보라색: $f_2(x)$ $\mu_2 = 1.25$ $\sigma_2^2 = 1$ $\pi_2 = 0.5$
- $x < 0$ 이면 베이지 분류기는 클래스 1(초록색)에 할당함
- 가우스 분포 및 관련 파라미터를 모두 알고 있기 때문에 베이지 분류기 사용 가능
- 하지만 실제 환경에서는 불가능 (모두 추정해야 됨)

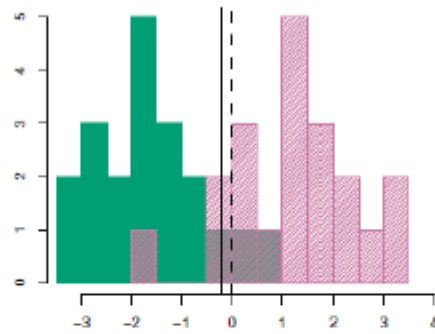
선형판별분석 (LDA) 방법

$$\begin{aligned}\hat{\mu}_k &= \frac{1}{n_k} \sum_{i: y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^K \sum_{i: y_i=k} (x_i - \hat{\mu}_k)^2 \\ \hat{\pi}_k &= n_k/n.\end{aligned}$$

위와 같이 필요한 파라미터들을 추정한 후 대입해서 사용함

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

위 식이 최대가 되는 클래스에 관측치 $X = x$ 를 할당한다.



각 클래스에서 랜덤 추출한 20개의 관측치의 히스토그램

$n_1 = n_2 = 20$ 이므로 $\hat{n}_1 = \hat{n}_2 \rightarrow$ 결정경계 $\frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$

- 베이즈 분류기의 결정경계보다 미세하게 왼쪽
- LDA vs. 베이즈 오차율의 차이는 단지 0.5%

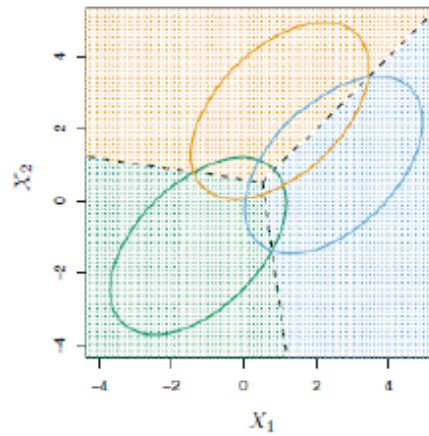
4-4-3. 선형판별 분석 ($p > 1$)

- $X = (X_1, X_2, \dots, X_p)$ 는 클래스 특정 평균벡터와 공통의 공분산행렬을 가지는 다변량가우스분포(혹은 다변량정규분포)를 따른다고 가정하자.
- $X \sim N(\mu, \Sigma)$: p 차원 랜덤변수 X 는 다변량가우스분포
 - $E(X) = \mu$: X 의 평균 (p 개 원소를 가진 벡터)
 - $Cov(X) = \Sigma$: X 의 $p \times p$ 공분산행렬

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

k 번째 클래스에 대한 밀도함수 $f_k(X = x)$ 를 정리하면 베이즈 분류기는 관측치 $X = x$ 를 다음 식이 최대가 되는 클래스에 할당한다.

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$



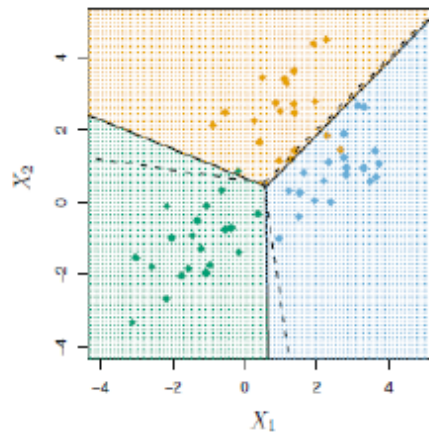
• 파선 : 베이즈 결정경계

◦ $\delta_k(x) = \delta_l(x) \quad k \neq l$

$$x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k = x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l$$

이 상황에서도 마찬가지로 파라미터 $\mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K$, 그리고 Σ 를 추정해야 한다.

LDA는 추정된 파라미터를 위 식에 대입하여 $\delta_k(x)$ 가 최대가 되는 클래스로 분류한다.



오차율 : 2.75%

- 주의점1 : 표본 n 대비 파라미터 p 의 비율이 높을수록 overfitting에 취약하다
- 주의점2 : 연체자는 고작 3.33%로 imbalanced data에서 오차율이 2.75%라는 것은 훌륭한 분류가 아니다.

Confusion Matrix

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9,644	252	9,896
	Yes	23	81	104
Total		9,667	333	10,000

민감도 (Sensitivity)가 너무 낮다! (24.3%)

LDA는 모든 분류기 중에서 accuracy가 가장 낮은 베이스 분류기에 접근하고자 한다. (즉 scoring=accuracy)

- FN : 실제로 Negative (비연체자)인데 Positive(연체자)로 예측한 경우
- FP : 실제로 Positive(연체자)인데 Negative(비연체자)로 예측한 경우
- 카드사 입장에선 FP를 최소화하는 것이 중요
- FN의 증가를 감수하고서라도 FP를 줄여야 한다
- 임계값을 내림으로써 (연체의 사후확률이 20%만 넘더라도 연체자로 예측해버리기)

4-4-4. 이차선형판별분석 (Quadratic Discriminant Analysis)

QDA

- 각 클래스의 관측치들이 가우스분포를 따른다고 가정
- 각 클래스가 자체 공분산행렬을 갖는다 Σ_k
- $X \sim N(\mu_k, \Sigma_k)$

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log \pi_k \\ &= -\frac{1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \log |\Sigma_k| + \log \pi_k\end{aligned}$$

- Σ_k, μ_k, π_k 에 대한 추정치를 위 식에 대입 후, 식이 최대가 되는 클래스에 관측치 $X = x$ 를 할당한다.

왜 K 개의 클래스들이 공통 공분산행렬을 갖는 가정이 중요한가?

Bias - Varaince Trade Off

- 설명변수가 p 개일 때 하나의 공분산행렬을 추정하는 데는 $p(p+1)/2$ 개의 파라미터에 대한 추정 필요
- QDA는 총 $Kp(p+1)/2$ 개의 파라미터에 대한 추정이 필요함
- 즉, LDA의 Kp 개의 선형계수보다 훨씬 많은 수의 선형계수가 생성되고
- 이는 유연성은 높지만 너무 높은 분산을 지니기 때문에 예측 성능이 떨어질 위험이 있다.
- 훈련 관측치의 수가 비교적 작아 분산을 줄이는 것이 중요하다면 LDA가 유용
- 반대로 훈련셋이 아주 커 분산이 주요 우려사항이 아니거나 공통 공분산행렬을 갖는다는 가정이 명백히 맞지 않다면 QDA가 유용

4-5 분류방법의 비교

1. LDA

- $p_1(x), p_2(x) = 1 - p_1(x)$
- $\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x$
- $c_0, c_1 : \mu_1, \mu_2, \sigma^2$ 에 대한 함수
- 관측치들이 각 클래스에 공통인 공분산행렬을 갖는 가우스분포를 따른다고 가정

2. 로지스틱 회귀

- $\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$

LDA & 로지스틱 회귀

- 공통점 : 선형의 결정경계
- 차이점 : 정규분포로부터 추정된 평균과 분산을 사용 vs 최대가능도를 사용

3. K-NN

- 관측치 x 에 가장 가까운 K 개의 훈련 관측치가 식별되고 이 관측치들이 속하는 클래스에 x 가 할당됨
- 비모수적 방법
- 결정경계의 형태에 대한 가정 X
- 결정경계가 상당히 비선형적일 때 우세

4. QDA

- LDA 및 로지스틱 & K-NN의 절충안
- 이차 결정경계를 가정

Scenario 1