Classification

Chapter 4

2013122044 응용통계학과 최연수

Yonsei University, Department Of Applied Statistics

Contents

1 An Overview of Classification

Multiple Linear Regression of Y on $X_1, \ldots, X_{p-1}$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where $\beta_0, \beta_1, \ldots, \beta_{p-1}$: parameters
$X_1, \ldots, X_{p-1}$: known constant predictors
ϵ_i's are independent $N(0, \sigma^2)$, $\quad i = 1, \ldots, n$

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}$$

▶ Response function represents a *hyperplane*.

3장에서는 다음과 같이 독립변수인 X와 종속변수인 Y간의 인과관계를 선형 모형으로 정의하여 모형을 추정하고자 , 더 나아가 알려지지 않은 종속변수 Y에 대해 예측하고자 하는 선형 회귀 모형을 공부했다. 이러한 선형 회귀 모형의 종속 변수는 Quantative, 즉 정량적인 연속형 자료라고 가정한다. 하지만 많은 경우에 이러한 종속 변수가 정량적인 자료가 아닌 질적인 범주형 자료일 때가 많다. 우리는 이렇게 종속변수 **Y**가 범주형 자료일 때의 이 값을 예측하는, 예측 관점의 문제를 **4**장에서 다루고, 이를 **Classification.** 즉 분류 모형으로 정의한다.

중요한 것은, 종종 분류 예측 문제에 대한 방법론들이 우선적으로 질적 자료의 범주형 변수들의 '확률'을 예측하기 때문에 그들은 앞서 배운 '회귀 분석'의 방법론을 공유한다.

우리는 다양한 분류 모형들 중에서 대표적으로 '*Logistic Regression*, *Linear Discriminant Analysis*, $K-$*Nearest Neighbors*' 다음과 같은 세 가지 모형에 대한 기초를 공부했다.

2 Why Not Linear Regression?

분류 예측 문제에 대한 방법론들을 소개하기 전에, '왜' 3장에서 공부한 선형 회귀 모형이 종속 변수가 범주형 자료일 때 적절하지 않은 지에 대한 설명이 필요하다.

예를 들면,

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

우리는 다음과 같이 환자의 증상들을 통해 3가지로 나눌 수 있는 환자의 의학적 상태를 예측하는 모형을 만들고 싶다. 하지만, 이렇게 종속 변수 값을 연속형 자료로 정의하여 회귀 모형을 설명하게 되면 결과값들의 '순서'가 존재하게 되는 것을 알 수 있다.

예를 들어 만일 종속 변수 Y의 값을 '1=약간의, 2=보통의, 3=심각한' 으로 순서를 설명할 수 있는 값으로 정의하거나, '0=stroke(뇌졸중), 1=drug overdose(약물 과다 복용)' 으로 오직 두 개의 level로 설명 가능한, 즉 *dummy variable* 방법을 사용할 수 있는 경우에는 문제가 없지만, 위의 그림과 같이 1의 값으로 나타나는 stroke(뇌졸중)과 2의 값으로 drug overdose(약물 과다 복용) 의 정도의 차이가 2의 값으로 나타나는 drug overdose(약물 과다 복용) 과 3의 값으로 나타나는 epileptic seizure(간질발작) 의 차이와 동일하다고 설명하는 꼴이 된다. 따라서 이러한 연속형 자료로 종속 변수를 정의하는 것은 우리가 다루고자 하는 '질적 자료'의 '질'을 바르게 설명할 수 없게 됨을 의미한다. 실제로 많은 경우에 있어서 2개 이상의 범주로 정의되는 질적 자료의 질을 연속형 자료로 설명하는 것에는 한계가 있다.

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

다음과 같이 우리는 $dummify$된 종속 변수를 선형 회귀 모형으로 예측하는 경우를 생각할 수 있지만, 우리의 예측치가 0~1사이의 값을 벗어날 수도 있는 경우를 생각해야 하며 위와 같이 두 단계의 변수가 아닌 그 이상이라면 위에서 설명한 예와 같이 순서형, 연속형 자료로 우리의 자료를 설명하기 어려워진다.

3 Logistic Regression

3.1 The Logistic Model

이러한 문제를 해결하기 위한 첫번째 방법으로, *Logistic Regression*, 로지스틱 회귀 모형을 소개한다.

일반적인 이항형 로지스틱 회귀에서는 2개의 범주로 존재하는 종속 변수의 결과를 직접적으로 추정, 예측하는 것이 아닌 종속 변수 Y가 특정 범주에 속할 '확률'로 설명하며, 종속 변수 Y_i는 일반적으로 베르누이 분포의 자료로서 표현된다.

```
install.packages('ISLR',repos = "http://cran.us.r-project.org")

## package 'ISLR' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\tti20\AppData\Local\Temp\RtmpSeg1tA\downloaded_packages

library(ISLR)
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

위의 ISL 교재 자료는 '개인의 수입 등 신상 정보를 비롯한 신용카드 사용 정보에 따른 부도 여부' 를 나타낸다. 이러한 자료에서 로지스틱 회귀는 개인의 부도 여부를 추정하고, 예측하기 위해 '부도 확률'을 정의한다.

예를 들면, 개인의 신용카드 연체 금액에 따른 부도 확률은 다음과 같다.

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

위의 조건부 확률값은 확률이기 때문에 0에서 1사이의 값으로 나타날 것이며, 어떠한 독립변수 값에도 이러한 범위 내에서 예측이 이루어 질 것이다. 누군가는 이러한 부도 확률이 0.5 보다 크다면 부도 여부를 $'Yes'$ 라고 정의할 수있고, 부도 위험에 대한 보수적인 시각에서는 부도 확률이 0.1보다 크면 부도 여부를 $'Yes'$ 라고 정의할 수도 있을 것이다.

그렇다면 우리는 어떻게 이러한 독립변수와 독립변수 값에 따른 종속변수 값의 확률간의 관계를 설명하는 모형을 만들 수 있을까?

만일 우리가 앞서 설명했듯 이러한 관계를 다음과 같이 선형 회귀 모형으로 설명하려 한다면,

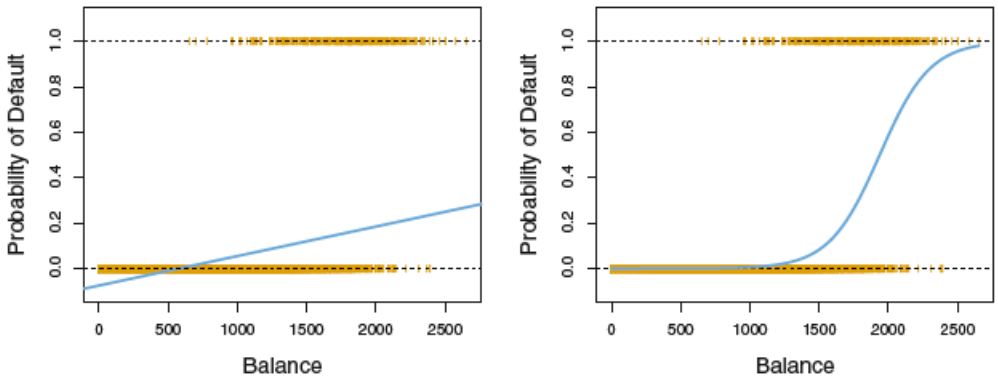
$$p(X) = \beta_0 + \beta_1 X.$$

신용카드 연체 금액이 0에 가깝게 작다면 우리는 음의 값을 가지는 부도 확률을 예측할 수 있고, 반대의 경우로 신용카드 연체 금액이 매우 크다면, 부도 확률은 1을 초과하게 될 것이며 이는 확률의 정의에 위배된다.

이러한 문제를 피하기 위해, 우리는 확률값의 정의에 위배되지 않는 함수를 정의해야하는데 로지스틱 회귀에서는 다음의 $'Logistic\ Function'$ 으로써 이를 설명한다.

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

위의 모델을 추정하기 위해서는 $'Maximum\ Likelihood\ Method'$ (최대우도추정법) 을 사용하는데, 이에 대해서는 다음 섹션에서 다루기로 하자.



좌측은 본 자료의 부도 확률을 선형 회귀 모형으로 설명한 모형의 추정 결과이고, 우측은 로지스틱 함수로 정의되는 로지스틱 회귀로 설명한 모형의 추정 결과이다. 앞서 말했듯이 좌측의 선형 회귀 모형에서는 음의 확률값이 추정되는 경우가 발생하지만, 로지스틱 함수로 정의되는 로지스틱 회귀 모형에서는 0에서 1사이의 값을 취하는 '$S-shaped$' 곡선으로 추정값이 나타난다.

로지스틱 함수를 변형하면 다음과 같이 우리는 0에서 ∞사이의 값을 취하는, '$odds$' 를 정의할 수 있고, 이는 쉽게 설명하면 성공 확률이 실패 확률에 비해 몇 배 더 높은가를 나타낸다고 할 수 있다. 이는 간단히 말하면 0에 가까운 $odds$값은 매우 낮은 종속변수의 확률 값을, 무한대에 가까운 $odds$값은 매우 높은 종속변수의 확률 값을 나타낸다고 할 수 있다.

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

이러한 $odds$값에 로그 변환을 취하면 다음과 같은 형태를 띄게 되는데,

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

이를 우리는 $log-odds$ 혹은 $logit$ 변환이라고 정의하고, 위의 식에서 우리의 로지스틱 회귀 모형이 독립변수 X에 선형인 $logit$을 갖게 됨을 확인할 수 있다.

하지만 우리는 $logistic\,function$ 하에서 정의되는 우리의 로지스틱 회귀 모형이 위 그림에서와 같이 독립변수 X와 $p(X)$ 와의 관계가 직선의 형태가 아니라는 것, 그리고 그렇기 때문에 독립 변수 X가 한 단위 변화할때의 $p(X)$ 값의 변화율이 X값에 따라 다르다는 것을 알 수 있다.

3.2 Estimating the Regression Coefficients

로지스틱 회귀 모형 또한 회귀계수 β_0,β_1이 우리의 training data를 통해서 추정되어야 하는 모수임을 알 수 있는데, *Least Squares Method* (최소제곱법)을 사용했던 선형 회귀와는 다르게 보다 일반적이고 비선형 모형들을 추정하는데에 있어서 통계적으로 더 좋은 성질을 가지는 *likelihood function*을 통해서 구할 수 있는 *maximum likelihood*를 통해 추정한다. 사실 최소제곱법 또한 설명할 최대우도법의 특별한 케이스이다. *maximum likelihood method* 는 쉽게 말해서, *logistic function*을 통해서 예측한 부도 확률값 $\hat{p}(x_i)$가 샘플의 부도 여부와 가능한 가깝게 하는 β_0과 β_1을 추정하는 것을 말한다.

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

즉, 다음의 *likelihood function* 의 값을 최대화 하는 추정 계수값을 추정하겠다는 것이다.

***우도(가능도, **likelihood**) : 보통 모수와 모집단이 이미 알려져있고 여기서 어떠한 사건이 발생할 가능성을 확률이라고 하는데, 우도는 반대의 개념이다. 관측치가 고정되고, 그러한 관측치가 나오게 하는 가장 그럴 듯한 모수값을 추정하는 것이다. 이 때 '이 관측치가 관찰될 가능성'을 '우도'라고 하고, 함수로 표현하며, 우도가 가장 높아지게 하여 모수를 추정하는 방법이 최대우도법이고, 이는 사실상 어떤 현상에 대해서 우리가 어떤 모집단을 관찰할 수 없거나 모수를 알 수 없기 때문에 나온 방법론이라고 생각할 수 있다.. ***

```
attach(Default)
logreg<-glm(formula = default ~ balance, family = binomial, data = Default)
summary(logreg)
```

```
##
## Call:
## glm(formula = default ~ balance, family = binomial, data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
```

```
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

로지스틱 회귀 모형에서는 선형 회귀 모형과 비슷하게 추정 계수의 정확도를 측정하기 위해 계수의 표준오차를 계산하고, 선형 회귀 모형에서의 $t-statistic$과 같은 역할을 하는 $z-statistic$을 사용해 모형의 계수가 0이라는 귀무 가설에 대한 가설 검정을 시행하여 유의수준 0.05하에서 귀무가설의 기각 여부를 판단할 수 있다. 모델의 진단 과정을 제외하고 단순히 교재 예제 자료를 다음과 같이 모델 추정 결과를 출력했고, 이를 통해 두 변수 간의 관계가 존재한다고 판단할 수 있다.

3.3 Making Predictions

회귀 계수들이 추정되었으면 추정한 모형을 통해 종속 변수의 확률을 계산하는 것은 어렵지 않다.

3.4 Multiple Logistic Regression

독립변수가 2개 이상일 때의 다중 로지스틱 회귀 모형의 경우에는,

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

이와 같이 p개의 독립변수들로 일반화시킬 수 있고, 이는 곧

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

을 의미한다.

```r
install.packages('dummies',repos = "http://cran.us.r-project.org")
library(dummies)
dummy<-dummy(student)
Default2<-cbind(Default,dummy)
attach(Default2)
multilogreg<-glm(default ~ balance + income + studentYes, family = binomial, data = Default2
summary(multilogreg)
```

```
Call:
glm(formula = default ~ balance + income + studentYes, famil
y = binomial,
    data = Default2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

학생의 여부를 나타내는 student 변수를 더미 변수화 하여 실행한 이러한 다중 로지스틱 회귀 모형 추정의 결과를 통해서 계수의 유의성을 확인하여 해석할 수 있다. 예를 들면, 더미화된 학생 변수 studentYes 의 계수값이 음수라는 것은, 다른 balance와 income 변수들의 종속변수 default에 대한 선형성을 제거했을 때, 학생은 학생이 아닌 사람보다 부도가 날 확률이 적다고 해석할 수 있다.

3.5 Logistic Regression for >2 Response Classes

그렇다면 종속변수가 2개 이상의 범주로 나타날 때는 어떻게 해야할까? 2개 이상의 범주로 나타나는 종속변수를 다루는 다항형 로지스틱 회귀 모형이 있긴 하지만, 2개 이상의 범주를 다룰 때에는 *Discriminant Analysis*(판별 분석) 이 잘 알려져있다.

4 Linear Discriminant Analysis (LDA)

4.1 Using Bayes' Theorem for Classification

로지스틱 회귀 모형이 로지스틱 함수를 이용해 직접적인 조건부확률 $Pr(Y = k \mid X = x)$ 를 추정했다면,

Linear Discriminant Analysis 에서는 보다 간접적인 방법으로 이러한 확률을 추정한다.

선형 판별 분석에서는 종속변수의 각각의 범주에 해당하는 독립변수의 분포를 정의하고 이를 $Bayes' theorem$ 을 이용하여 $Pr(Y = k \mid X = x)$ 의 추정값으로 설명한다. 이러한 분포들이 정규분포를 따른다면, 우리의 모델은 로지스틱 회귀 모형의 형태와 매우 비슷하게 되는 것을 확인한다.

베이즈 정리를 이용하는 과정은 다음과 같다.

우리의 관측치들을 K 개의 그룹으로 분류하고 싶을 때, 우리는 무작위로 추출된 관측치가 k 번째 그룹에서 나왔을 확률을 말하는 사전 확률 π_k 를 정의하고 이는 주어진 관측치가 종속변수의 k 번째 범주와 관련이 있을 확률을 말한다.

$f_k(x) = Pr(X = x \mid Y = k)$ 를 k 번째 범주에서 나온 관측치 X 의 확률밀도함수라 한다면, 베이즈 정리를 이용하면

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{4.10}$$

다음과 같이 나타낼 수 있고, $p_k(X) = Pr(Y = k \mid X)$ 를 사후 확률이라고 한다면 우리는 베이즈 정리를 통해 이를 직접 계산하지 않고 사전 확률 π_k 와 $f_k(x)$ 의 추정치를 계산하여 구할 수 있고, 더욱이 π_k 를 추정하는 것은 어렵지 않고 X 의 확률밀도함수는 조금 어려울 수 있지만 간단한 형태들을 가정한다고 한다.

2장에서 언급되었던 $Bayes\ Classifier$ 는 사후 확률 $p_k(X)$ 가 최대이고 error rate가 최소인 관측치들을 범주에 분류한다고 했다. 따라서 우리가 $f_k(x)$ 를 추정할 수 있다면, 베이즈 분류기에 근사하는 분류기를 만들 수 있다.

4.2 Linear Discriminant Analysis for p = 1

우리는 X 의 확률 밀도 함수를 추정하고 이를 통해 사후 확률을 추정해야한다. 그리고 범주에 속할 가장 큰 사후 확률을 기준으로 관측치를 분류한다.

이러한 $f_k(x)$ 를 추정하기 위해 몇가지 가정이 앞서야하는데, 대표적인 것이 바로 정규분포 가정이다.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right), \tag{4.11}$$

정규분포의 확률밀도함수는 위와 같고, 모든 각각의 범주들이 분산을 공유한다는 추가적인 가정 하에 우리는 사후 확률 $p_k(x)$를 다음과 같이 정의할 수 있다.

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}. \tag{4.12}$$

위의 식에 로그를 취하면,

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{4.13}$$

위의 식과 같고, 관측치가 범주에 속할 가장 큰 사후 확률에 의해 분류하게 된다. 예를 들면, $K = 2$ 이고 $\pi_1 = \pi_2$이면 베이즈 분류기는 만약 $2x(\mu_! - \mu_2) > \mu_1^2 - \mu_2^2$ 이면 범주 1에, 아니면 2에 지정하게 된다. 이러한 경우에 *Bayes decision boundary*

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}. \tag{4.14}$$

는 다음과 같은 지점과 일치한다.

LDA 방법은 사전확률 π_k과 μ_k 그리고 σ^2의 추정치를 이용하여 베이즈 분류기에 근사시킬 수 있고, 다음과 같이 정의할 수 있다.

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \end{aligned} \tag{4.15}$$

$n = \#\, of\ training\ observations$

$n_k = \#\, of\ training\ observations\ in\ the\ kth\ class$

$\mu_k = average\ of\ all\ the\ training\ observations\ from\ the\ kth\ class$

$\hat{\sigma^2} = weighted\ average\ of\ the\ sample\ variances\ for\ each\ of\ the\ K\ classes$

우리가 사전 확률을 알 때도 있겠지만 모를 때에, LDA는 사전 확률 π_k를 k번째 범주에 속하는 training 관측치들의 비율을 이용하여 다음과 같이 추정한다.

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

LDA 분류기는 위의 두 가지 식에서 구한 추정치들을 이용해 관측치 $X = x$ 를 밑의 식의 결과가 최대일 때를 만족하는 범주에 지정해준다.

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

정리하면, LDA 분류기는 각각의 범주에 해당하는 관측치들의 분포가 정규분포를 따른다는 가정이 앞서야하며, 이를 만족하면 추정치들을 이용해 베이즈 분류기에 근사시킬 수 있다는 것이다.

4.3 Linear Discriminant Analysis for p > 1

독립변수의 갯수가 여러개일 때에는, 독립변수 X 가 특정 범주의 평균 (벡터)과 공분산 (행렬)을 따르는 다변량 정규분포에서 뽑힌 관측치라고 가정한다.

p차원의 *random variable* 이 다변량 정규 분포를 따르는 것을 우리는 다음과 같이 나타내고,

$$X \sim N(\mu, \sum)$$

확률밀도함수는 다음과 같이 정의된다.

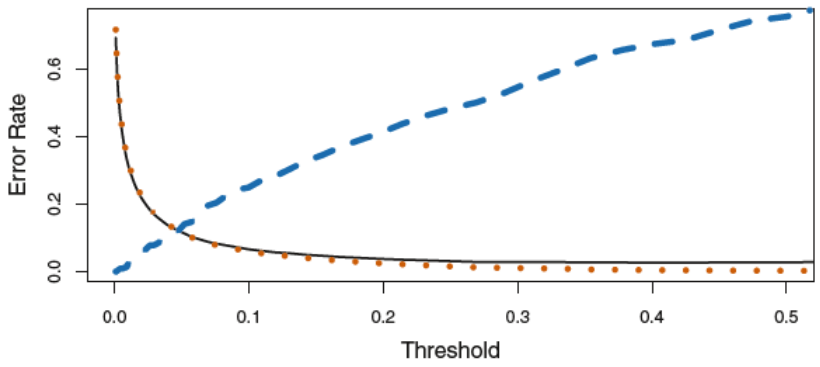
$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

우리는 또한 위에서와 같이 k번째 범주의 확률밀도함수와 사전 확률의 추정치를 결합하는 과정을 선형대수학적으로 풀면

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

다음과 같이 식 4.13의 벡터/행렬 버전으로 나타나게 되며, 사후 확률로써 베이즈 분류기에 근사시킬 수 있다.

우리는 일차원이든 p차원이든 사후 확률에 로그를 씌운 $\delta_k(x)$ 가 x에 의한 선형 함수 형태로 나타난다는 것을 확인할 수 있다. 이러한 LDA decision rule은 오직 독립변수 x의 원소들의 선형 결합으로 나타나며, 왜 *Linear Discriminant Analysis* 가 '*Linear*' 인지 알 수 있다.

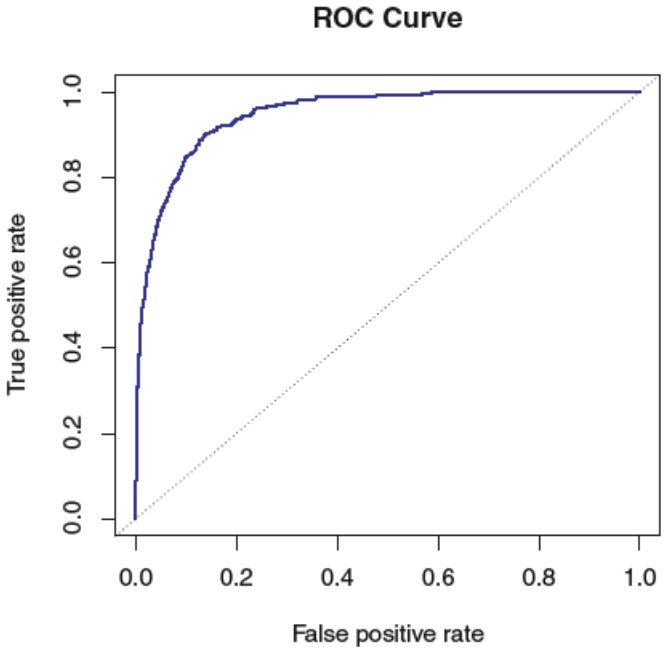


한편 LDA가 *total* error rate 가 가장 낮은 분류를 설명하는 베이즈 분류기에 근사시키는 방법이기 때문에,

잘못 분류된 관측치들의 총 갯수는 작을 수 있으나 어떤 범주에서 error 가 발생하는지에 대해서는 무시한다. 부도 문제에 예에서 부도인 사람을 부도라고 맞추는 것을 *sensitivity* , 부도가 아닌 사람을 부도가 아닌 사람이라고 맞추는 것을 *specificity* 라고 하는데, 부도인 사람을 맞추는 *sensitivity* 측면에서 떨어지는 성능을 보여준다.

위의 그림은 다음과 같이 $\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5.$ 에서의 0.5, 즉 기준 사후 확률(범주에 속할 확률)을 설정하는 threshold의 값에 따른 error rate 와의 trade-off 관계를 나타내고 있다. 검은색 선은 전체적인 error rate, 주황색 점선은 non-defaulting customers들에 대한 error이고 파란색 선은 defaulting customers들에 대한 error를 말한다.

우리는 어떠한 threshold value가 최적일지 어떻게 결정해야할까? 이러한 경우 *domain knowledge* 가 기초해야 한다고 책에서는 설명하고 있다.



ROC Curve로 가능한 threshold value들에 대해 발생하는 두 가지 종류의 error를 동시에 보여줄 수 있다.

$true\,positive\,rate = defaulters\,correctly\,identified$

$false\,positive\,rate = 1 - specificity = non - defaulters$들을 $defaulter$로 잘못 분류한 오류

4.4 Quadratic Discriminant Analysis (QDA)

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

LDA 처럼 각 범주에 해당하는 관측치들이 정규분포에서 추출되어야 하고 추정치들을 베이즈 정리에 plugging해서 예측을 실행하는 것은 같지만, 각각의 범주들이 각각의 고유 공분산 행렬을 가진다고 가정한다.

즉, k번째 범주에서 관측된 자료가 $X \sim N(\mu_k, \sum_k)$ 의 형태를 취하는 것을 말한다.

그렇다면 언제 QDA, LDA?

$$'Bias - Variance\,Trade\,off'$$

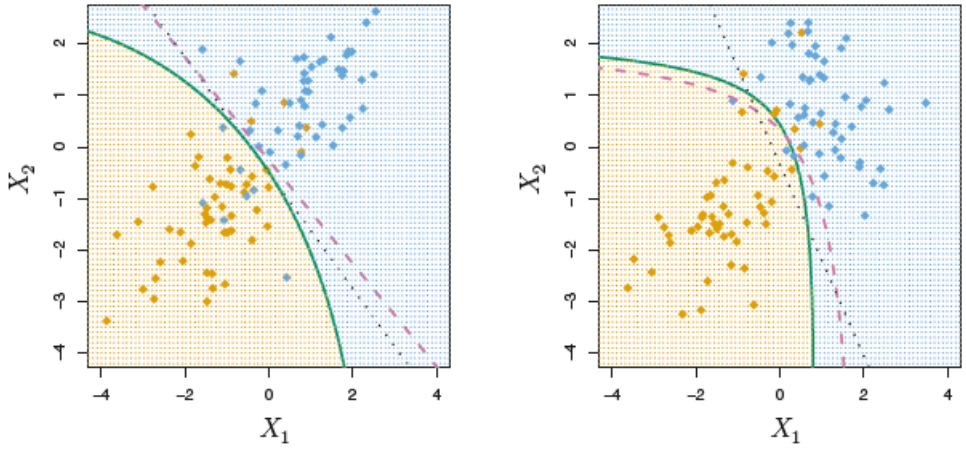
우리가 p개의 독립 변수가 있을 때, 공분산 행렬을 추정하기 위해서는 $\frac{p(p+1)}{2}$ 개의 모수를 추정해야한다. QDA는 각각의 다른 공분산 행렬을 추정해야 하기 때문에 총 $K\frac{p(p+1)}{2}$ 개의 모수를 추정해야 한다. 이는 즉 QDA는 much more flexible, bias는 적겠지만 variance는 높다. 반면 LDA는 선형 모델이기 때문에 Kp 개의 선형 계수들을 추정하면 되기 때문에 QDA만큼 유연하지는 못해 bias는 높을 수 있어도 variance는 낮다.

어떨 때 bias가 높을까? LDA는 K 개의 범주들이 같은 공분산 행렬을 갖는다고 가정하고 이러한 가정이 맞지 않을 때 높은 bias가 발생한다.

즉, LDA같은 경우에는 적은 수의 training observations (reducing variance is crucial) 가 있는 상황일 때,

QDA는 sample size가 크고 K 개의 범주들이 같은 공분산 행렬을 갖는다고 가정할 수 없을 때 더 좋다고 할 수 있겠다.

ex) two-class problem



좌측(linear) : Bayes (purple dashed), LDA(black dotted), QDA(green solid) decision boundaries

우측 (non-linear): Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.

5 A Comparison of Classification Methods

모든 상황에서 특정 방법론이 더 탁월한 성능을 보인다거나 우수하다라는 말은 할 수 없다.

실제 decision boundary 들이 선형이라면, LDA 나 로지스틱 회귀 모형이 좋은 성능을 보일 것이고, 선형이 아닐 경우에는 QDA 가 더 좋은 결과를 보일 수 있을 것이다. 마지막으로 과적합 문제를 잘 조절한다면, 상대적으로 훨씬 복잡한 decision boundaries 들의 경우에는 KNN 같은 비모수적 방법론이 더 우수할 것이다.