R로 배우는 데이터사이언스 - Lecture 8: 분류

Classification and Discrimination

서울대학교 통계연구소

이권상

이번 강좌에서 다룰 내용

- 분류 소개
- 분류 방법 (Algorithms, models, classifiers)
  - k-Nearest neighbors
  - Fisher의 선형분류판별분석 (Linear discriminant analysis)
  - Quadratic discriminant analysis
  - 로지스틱회귀모형 (Logistic regression model)
- 분류 모형의 평가
  - Confusion matrix, sensitivity, specificity, ROC curve
  - Training and testing data, generalization error, cross validation
- 실습 (R package caret)

혼돈행렬, 민감도 / 특이도

시작하기 전에…

- 실습시간에 사용할 R script와 자료를 통계연구소 강좌자료실에서 download 받고, working directory 설정하기.
- R 패키지 `caret, pROC, GGally, ggplot2, dplyr`를 설치하자.

```
install.packages("caret") # for classification
install.packages("pROC")  #
install.packages("dplyr") # data manipulation
install.packages("ggplot2")
install.packages("GGally") # Graphing supplement

library(MASS)
library(caret)
library(dplyr)
library(ggplot2)
library(GGally)
library(pROC)
```

Windows에서 한글 깨짐 현상 해결방법

RStudio에서

1. Tools > Global Options > Code > Saving > Default text encoding을 UTF-8으로 변경하고 Apply
2. Practice.R 파일을 RStudio에 불러온 후에 File > Save with Encoding > UTF-8을 선택하고 아래 Set as default encoding for source files를 체크
3. RStudio에서 Practice.R를 닫고 Practice.R 를 다시 불러오기

소개: 분류, 분류기준, 분류기준의 평가

분류

- 분류기준에 따라 분류하기

source: EBS 수학이 야호 https://clipbank.ebs.co.kr/clip/view?clipId=VOD_20171006_00444

질문

- 색이 같은 세 그룹으로 분류할 때, 빨간 세모는 어디로 분류할까?
- 모양이 같은 세 그룹으로 분류할 때, 빨간 세모는 어디로 분류할까?

분류

- 분류기준에 따라 분류하기
- 분류기준이 주어지지 않은 경우, **(데이터로부터) 분류기준을 추론**하여 분류법을 만들기
- 예: 타이타닉 생존여부 분류

예측
각각의 사람들에 대한 정보 (성별, 좌석 등) -> 변수
생존은 제외한 변수들로 생존 여부를 예측할 수 있나?

예) 1등석의 여자는 살아남는다.

(image source: github/grantmlong)

변수	뜻	
Survival	생존 여부	0 = No, 1 = Yes
Pclass	좌석등급	1 = 1등석, 2 = 2등석, 3 = 3등석
Sex	성별	

질문

- 3등석 여자 손님은 살아남을까?
- 이 대답을 하기 위해서 자료를 참조하여 분류기준을 도출해야 한다

타이타닉 자료

- 여기에 소개하는 자료는 호화여객선 타이타닉에 탑승한 891명들에 대한 기록으로 개인별로 14개의 변수가 있다.
- 개인별 변수는 pclass(좌석등급), survived (1: 생존, 0:사망), name (이름), sex(성별), age (나이), sibsp (동승한 형제자매/배우자 수), parch (동승한 부모/자녀 숫자), ticket (티켓번호), fare(티켓가격), cabin (방번호), embarked (승선항), boat(구명보트 번호), body (사망자 인식번호), home.dest (목적지) 로 구성되어 있다.
- 다음과 같이 자료를 살펴본다.

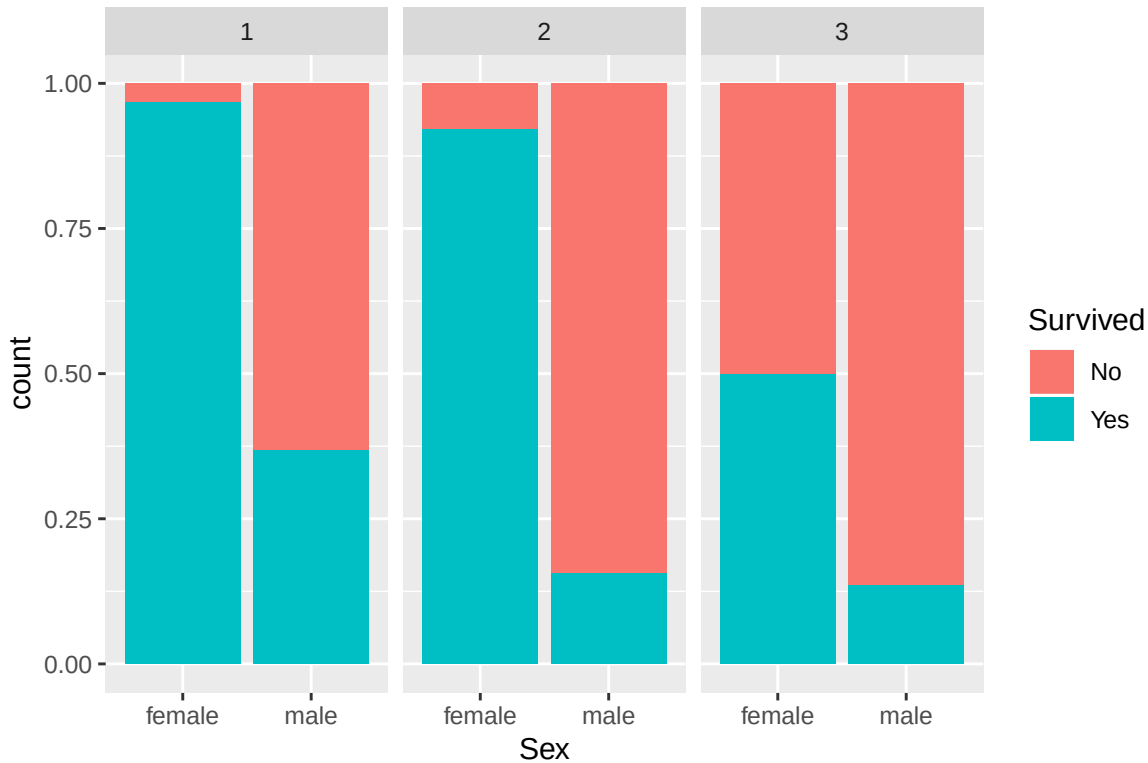
```
titan <-read.csv("train.csv")
View(titan)
```

좌석등급과 성별을 중심으로..

타이타닉 자료: 성별과 좌석등급으로 생존여부 예측

- 성별, 좌석 등급별로 Survived의 분포를 보고 분류기준 도출
- 두 개 (이상)의 범주형 변수로 다른 범주형 변수의 값을 예측

```
library(tidyverse)
titan <- titan %>% dplyr::select(Sex, Pclass, Survived) %>%
  mutate(Survived = as.factor(ifelse(Survived == 1,"Yes","No")))
titan %>% ggplot(aes(x = Sex)) +
  geom_bar(aes(fill = Survived), position = "fill") + facet_wrap(~Pclass)
```



타이타닉 자료: 성별과 좌석등급으로 생존여부 예측

- 아래의 표에 의하면
- 여자는 생존, 남자는 비생존으로 분류하는것이 최선

좌석 등급은 중요하지 않음.
성별에 따라 여성 생존, 남성 사망으로 분류

최선인 이유? 베이지 분류

Sex	Pclass	Surv.Rate
female	1	0.9680851
female	2	0.9210526
female	3	0.5000000
male	1	0.3688525
male	2	0.1574074
male	3	0.1354467

- 더 많은 변수 (예: 나이) 가 있으면 더 좋은 분류 가능
- **이 분류 기준이 왜 최선인가?**

대학원 입학 자료

- 미국 대학원 입학 자료: 400명의 TOEFL.score (영어 점수)와 CGPA (대학교 학점)으로 대학원 입학 허가/불가 (Admit/No) 예측

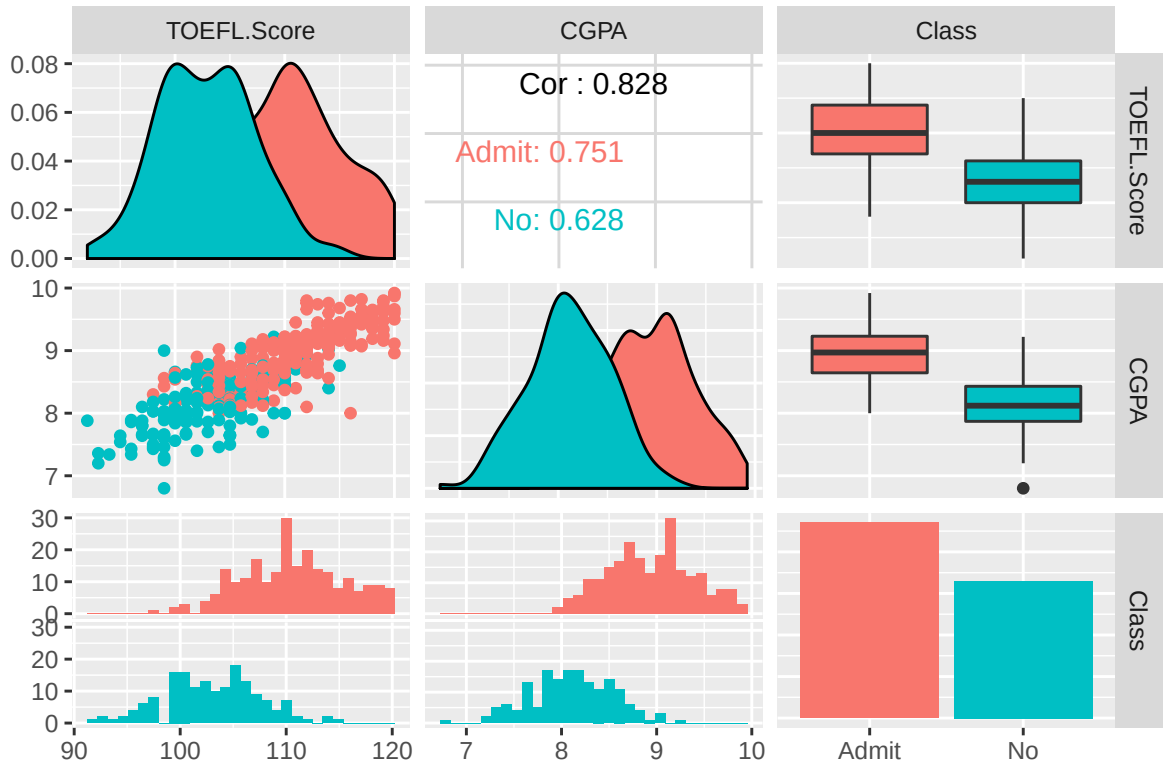
TOEFL.Score	CGPA	Class
118	9.65	Admit
107	8.87	Admit
104	8.00	Admit
110	8.67	Admit
103	8.21	No
115	9.34	Admit

변수가 연속형일 때 어떻게 볼 수 있니?
반응변수 Admit / No
설명변수 토플, 학점 (연속형)

- TOEFL.score, CGPA, 그리고 Class의 분포를 보고 분류기준 도출
- 두 개 (이상)의 연속형 변수로 범주형 변수의 값 예측

```
library(GGally) # for pairwise scatterplot
ggpairs(data, mapping = aes(color = Class))
```

타이타닉은 설명변수의 조합이 가능함 (범주형이기 때문에)
밀도함수, 박스플롯 등으로 데이터 탐색



분류 분석

- 목적: p개의 입력변수 $(X_1, \ldots, X_p) = (x_1, \ldots, x_p)$의 값으로부터 J개의 값을 가지는 범주형 변수 Y의 값을 예측
- 자료: n개의 입력 - 출력 변수 값 쌍

X_1	X_2	$\cdots$	X_p	Y
x_{11}	x_{12}	...	x_{1p}	y_1
$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
x_{n1}	x_{n2}	...	x_{np}	y_n

- 분류 함수: 입력변수 $x = (x_1, \ldots, x_p)$를 받아 Y의 범주 중 하나를 출력하는 함수 $y = f(x)$.

분류 함수의 예

- 타이타닉 자료 (X_1 = 성별, X_2 = 좌석등급)

$$f(X_1, X_2) = \begin{cases} \text{생존}, & x_1 = \text{여자일때 ;} \\ \text{사망}, & x_1 = \text{남자일때.} \end{cases}$$

- 대학원 자료 (X_1 = 토플 점수, X_2 = 대학 학점)

$$f(X_1, X_2) = \begin{cases} \text{합격}, & \text{만약 } x_1 > 100, x_2 > 8; \\ \text{불합격}, & \text{만약 } x_1 \leq 100 \text{ 또는 } x_2 \leq 8. \end{cases}$$

용어 정리

- 분류: 반응변수 Y가 범주형인 회귀분석 <- > 비지도학습 (군집화), 반응변수 X
- 분류법은 **지도학습**이다 - 추론하여 도출된 분류기준의 품질평가 가능 (주어진 반응변수 (또는 출력변수)의 값을 이용해 평가).
- Classification (분류)와 Discrimination (판별)은 같은 뜻
- Y의 범주가 2개인 경우: 이항분류 (Binary classification)
  - 때로는 $Y = 0, 1$로 코딩함
  - 예: Survived = Yes (1), No (0); Class = Admit (1), No (0)
- Y의 범주가 2개 이상인 경우: 다항분류 (Multi-category classification)
  - 범주의 순서가 있는 경우: Ordinal (multicategory) classification
  - 범주의 순서가 없는 경우: Multicategory classification
- 분류기 (Classifier): 자료로부터 분류기준 "$f(x)$" 을 만드는 방법 또는 만들어진 분류함수 $f(x)$
  - 예: LDA, kNN, Random Forest 등은 다른 자료에 적용하면 다른 **분류기준을 만드는 방법**이면서,
  - 한 자료로 국한시키면, 서로 다르게 **만들어진 분류함수**들을 지칭.

최선의 분류기준?

- 분류기준 또는 분류기로 분류 후 분류기 $\hat{y} = f(x)$의 값과 실제 y값을 비교
- 분류기 $f(x)$의 오분류율 $Pr(\hat{y} \neq y) = Pr(f(x) \neq y)$
- 주어진 자료에서의 **오분류율을 가장 작게 만드는 분류**기가 가장 좋은 분류기
- 만약 입력변수 $x = (x_1, \ldots, x_p)$에서의 $y = 1$ (또는 $y = 0$)의 발생확률을 알 수 있다면, 다음의 분류기가 가장 좋음이 알려져 있다.

$$f(x) = \begin{cases} 1, & \text{만약 } P(Y = 1 \mid X = x) > P(Y = 0 \mid X = x); \\ 0, & \text{만약 } P(Y = 1 \mid X = x) \leq P(Y = 0 \mid X = x). \end{cases}$$

- 베이즈 정리를 이용하여 증명할 수 있으므로, 때로는 **베이즈분류기**로 불리운다.

텍스트

타이타닉 자료의 베이즈 분류기

- 타이타닉 자료는 x (Sex, Pclass)의 값이 6개 중 하나 (성별과 좌석등급이 모두 범주형)
- 각각의 값 x에 대해 $P(Y = \text{생존} \mid X = x)$ 과 $P(Y = \text{사망} \mid X = x)$를 비교 가능

Sex	Pclass	P_survive	P_demise
female	1	0.9680851	0.0319149
female	2	0.9210526	0.0789474
female	3	0.5000000	0.5000000
male	1	0.3688525	0.6311475
male	2	0.1574074	0.8425926
male	3	0.1354467	0.8645533

위 3줄 = 전원 생존
아래 3줄 = 전원 사망

$P(Y = \text{생존} \mid X = x) = P(Y = \text{사망} \mid X = x)$일때는 동전던지기(50/50) 또는 전부 생존 또는 전부 사망으로 분류

참조:

- 대부분의 경우 $P(Y = \text{생존} \mid X = x)$ 를 알 수 없고, 자료로부터 추정!
- 베이즈분류기는 실제 도출 불가능 (자료를 무한히 볼 수 있을 때, 또는 생성과정을 정확히 알 때만 가능)

다양한 분류기 (Classifiers)

오늘 다룰 분류기

- 최근접 이웃 방법 (k-nearest neighbors; k-NN)
- 선형판별분석 (Linear Discriminant Analysis; LDA)
- 이차판별분석 (Quadratic Discriminant Analysis; QDA)
- 로지스틱회귀분석 (Logistic regression)

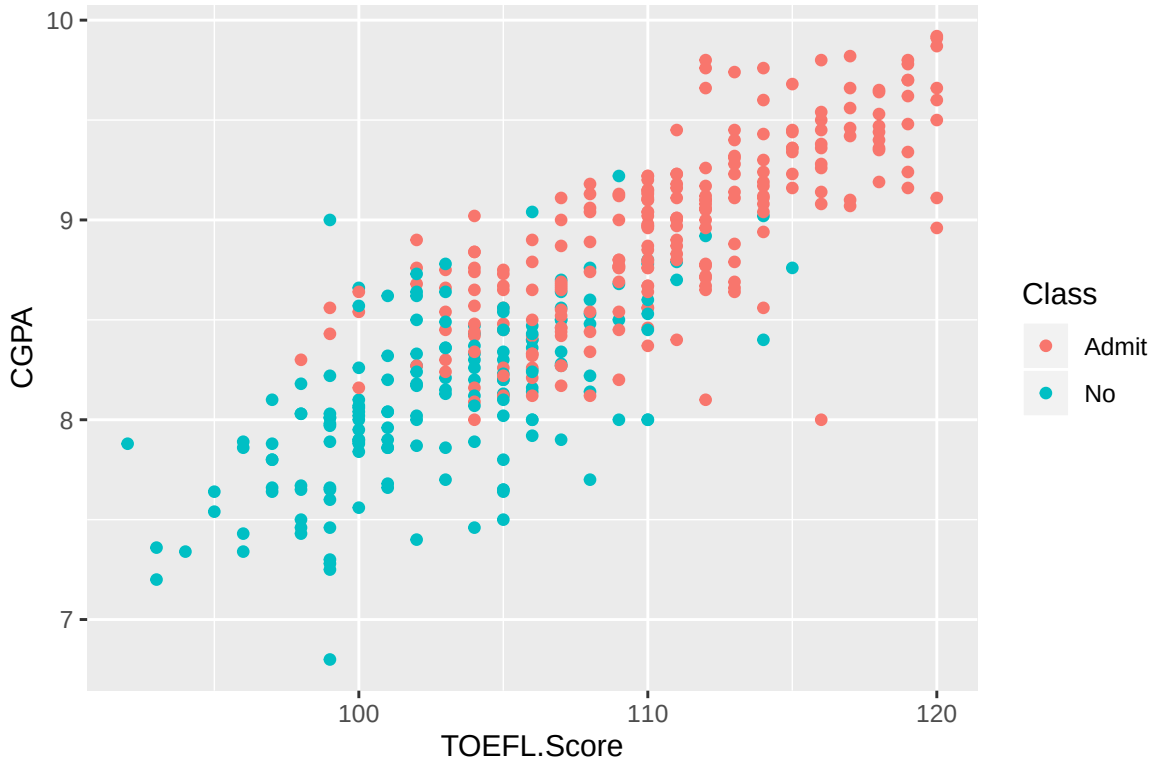
오늘 다루지 않을 분류기

- 분류 나무 (Classification Tree)
- Support Vector Machine
- Naive Bayes Classifier
- Boosting 기법
- Bagging (Bootstrap aggregating) 기법

최근접 이웃 방법 (k-nearest neighbors; k-NN)

- 베이즈분류기를 추정하는 한 방법
- 입력변수들 모두 연속형임을 가정 (예: 대학원 입시 자료)

어떤 점이랑 가장 가까운 k개의 점을 조사하여 과반수가 나타내는 반응변수를 채택



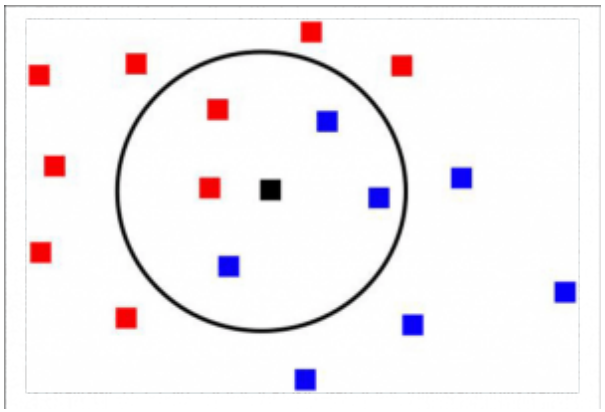
- 연속형 변수이기 때문에 주어진 x에서의 관측값이 많아야 1개.
  - 예를 들어 $x = (110, 8.5)$에서의 관측값은 없음
  - $\Pr(\text{Admit} \,|x = (110, 8.5))$를 추정하기 위해 x주변의 **이웃**을 조사

최근접 이웃 방법 (k-nearest neighbors; k-NN)

주어진 x에서 가장 가까운 k개의 관측치들로 x에서 각 범주의 확률을 추정하고 가장 많이 나온 범주를 k-NN classifier $f(x)$의 값으로 정한다.

- 아래 그림에서 k = 5인 경우, 검은 점 (x_0)에서의 파란 범주 확률은 3/5 (60%), 빨간 범주 확률은 2/5 (40%)이므로

$$\text{5-NN classifier } f(x_0) = \text{파랑}.$$

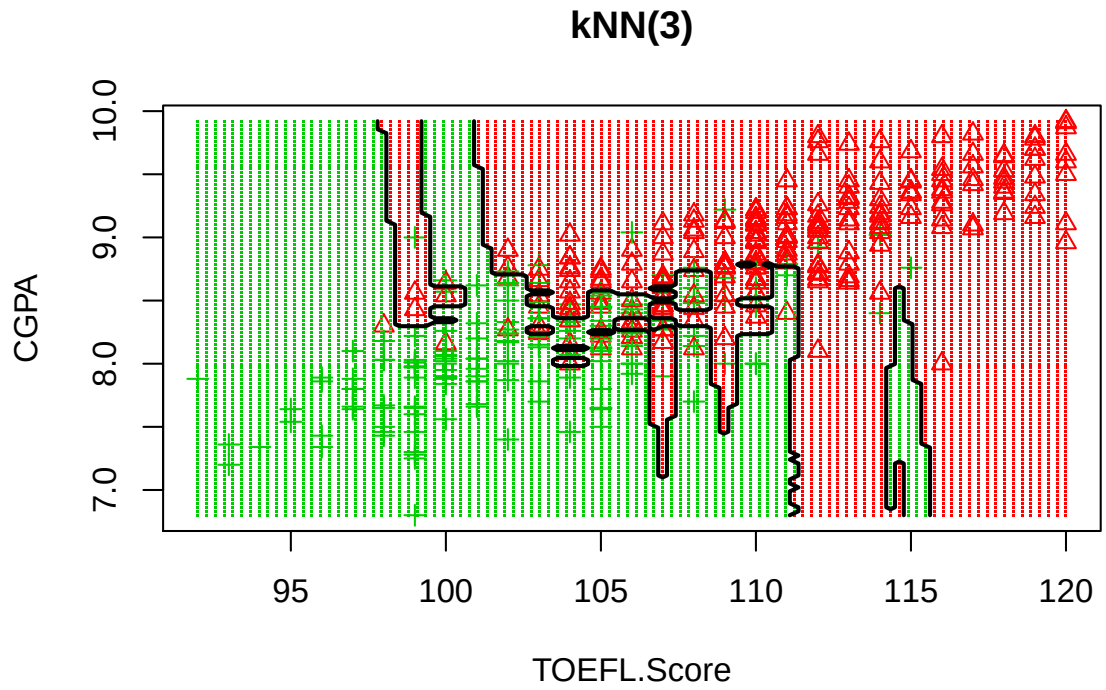


(image source: https://www.ednology.co.uk/

k-NN 예제 (대학원 입시 자료)

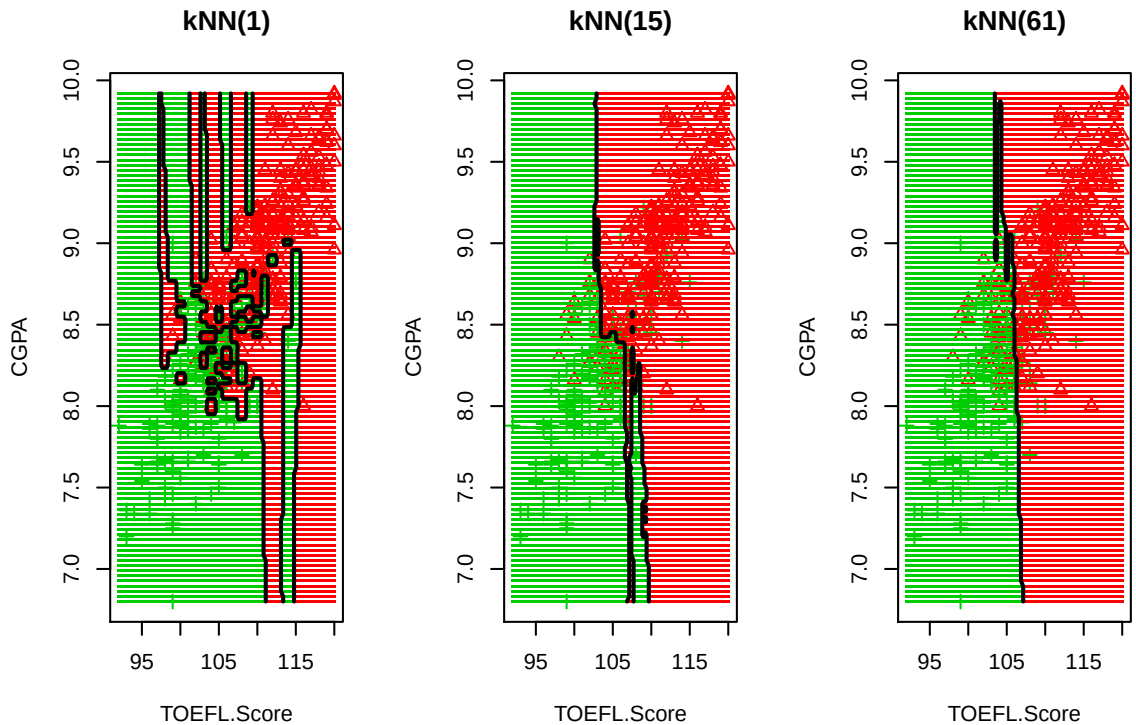
- $k = 3$인 k-NN classifier

k가 작으면 결정경계가 지저분해진다.



- 불합격과 합격을 나누는 선을 결정 경계 (Decision Boundary)라고 한다

- 일반적으로, k는 홀수로 정한다 (동점 방지).
- 분류 결과가 k의 값에 민감하다.
- k의 값이 클 수록 유연성이 떨어진다.

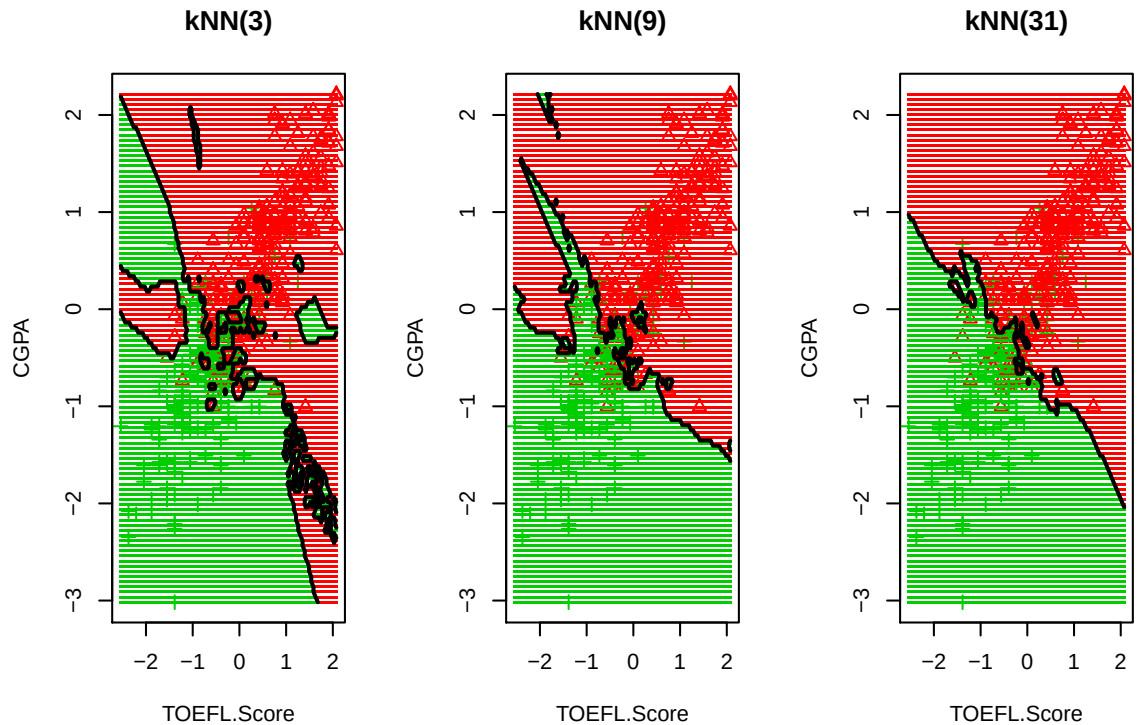


경계가 단순해진다. ->
분류 기준이 단순해진다.
and 오류가 많아진다. (국지적 부분 무시, 전체적 경향만)

오분류율이 가장 적은 k를 선택하자..!

- 이웃의 정의(definition)에 따라 분류 결과가 달라진다
  - 윗 장의 결과에서, TOEFL 점수 10점 차이와 CGPA 10 차이는 상대적으로 매우 다르지만, 모두 같은 거리의 이웃이다.
  - 일반적으로 변수의 scale을 조정하여, 모든 변수의 표준편차가 1이 되도록 변환해 준다. Z 정규화

```
data2 <- data %>% mutate(TOEFL.Score = c(scale(TOEFL.Score)),
                         CGPA = c(scale(CGPA)))
```



최근접 이웃 방법 (k-nearest neighbors; k-NN): 간단한 정리

- 베이즈분류기를 추정하는 매우 기초적인 방법
- 이웃의 갯수 k를 정해주어야 한다 (Model validation을 이용; 곧 다룸)
- 누가 이웃인가(이웃을 정하는 거리의 기준)에 따라 결과가 다르다 (Model Validation을 이용)

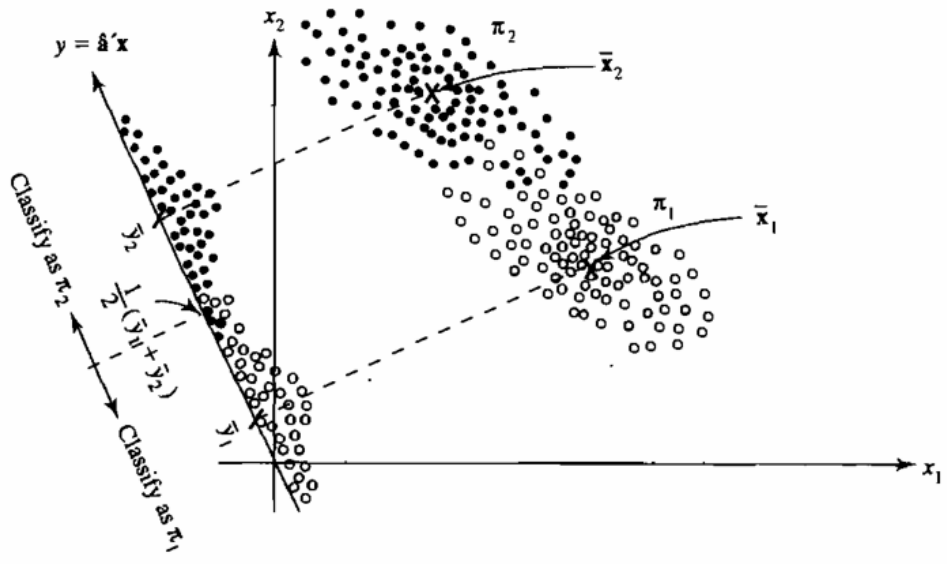
k=5 일 때 y= 1이 1개, 2가 2개, 3이 3개 있다면? -> y= 2, 3 중 랜덤으로 선택

- 출력변수 (또는 반응변수)의 범주의 갯수가 몇 개이든지 사용 가능
- k-NN 분류기는 모든 자료를 다 저장하고 있어야 함 (비효율적)

모든 점 끼리의 거리 계산하는 과정...

선형판별분석 (Linear Discriminant Analysis; LDA)

- 연속형 변수들이 입력변수
- 이항분류, 다항분류 문제에 모두 적용가능
- 두 범주를 선 (3차원 이상이면 면)으로 나누게 됨
- 자료를 자료공간의 선 위에 정사영 (projection)시킨 뒤 한 점을 기준으로 나눔



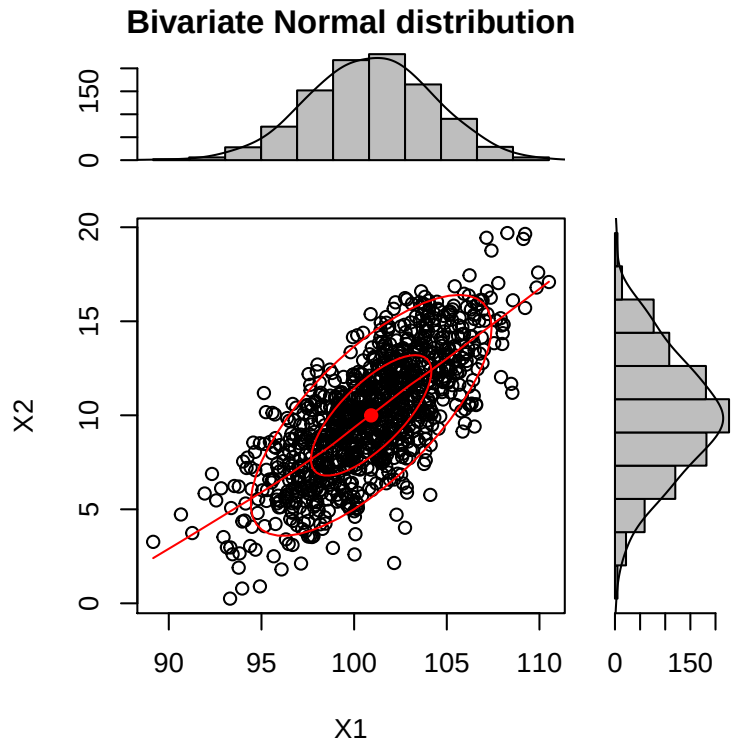
(image source: Johnson and Wichern, Applied Multivariate Statistical Analysis, Pearson.)

다변량 정규분포

- 다변량 정규분포: 입력변수 (X_1, X_2)의 분포를 평균이 (μ_1, μ_2), 분산이 σ_1^2, σ_2^2, 상관계수가 ρ인 정규분포로 모형화

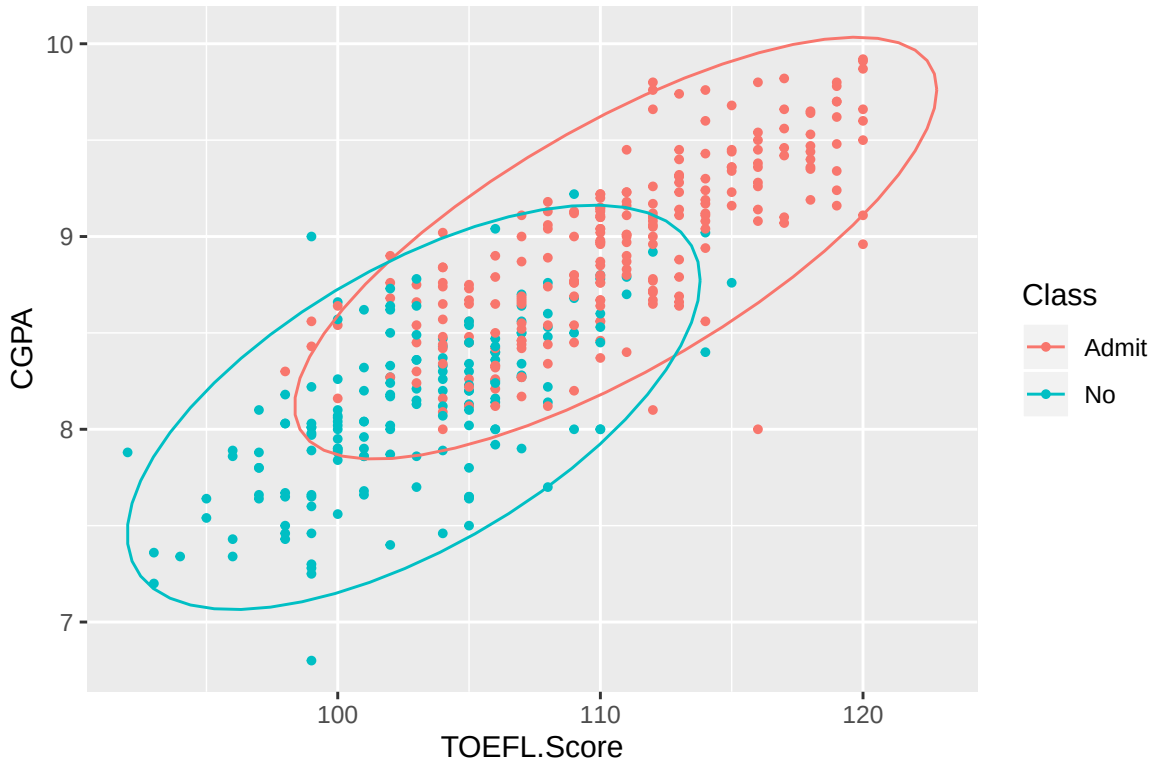
아래에서 평균 (101,10), 분산-공분산은 타원의 모양과 크기로 결정

입력변수 x1, x2가 각각 정규분포한다.



선형판별분석 (Linear Discriminant Analysis; LDA)의 아이디어

- 그룹 1의 분포 (X | Y = Admit)와 그룹 2의 분포 (X | Y = No) 를 **정규분포로 모형**화
- 두 그룹의 **평균은 다르고**, **분산-공분산은 같다**고 가정



선형판별분석 (Linear Discriminant Analysis; LDA)의 아이디어

베이즈 정리를 이용하여, 베이즈분류기에서 쓰이는 $Pr(Y = j|X = x)$를 추정:

$$P(Y = j|X = x) = \frac{\pi_j \phi_j(x)}{\sum_{i=1}^{J} \pi_j \phi_j(x)} \approx \frac{P(X = x|Y = j)P(Y = j)}{P(X = x)},$$

여기서

- Y의 범주는 $\{1, \ldots, J\}$,
- $\pi_j = P(Y = j)$ (자료에서 $Y = j$인 관측값의 비율로 추정),
- $\phi_j(x)$은 추정된 다변량정규분포의 x에서의 밀도함수값
- 모형이 근사적으로 맞을 경우, 베이즈분류기의 $P(Y = y|X = x)$를 추정하는 효율적인 방법.

선형판별분석

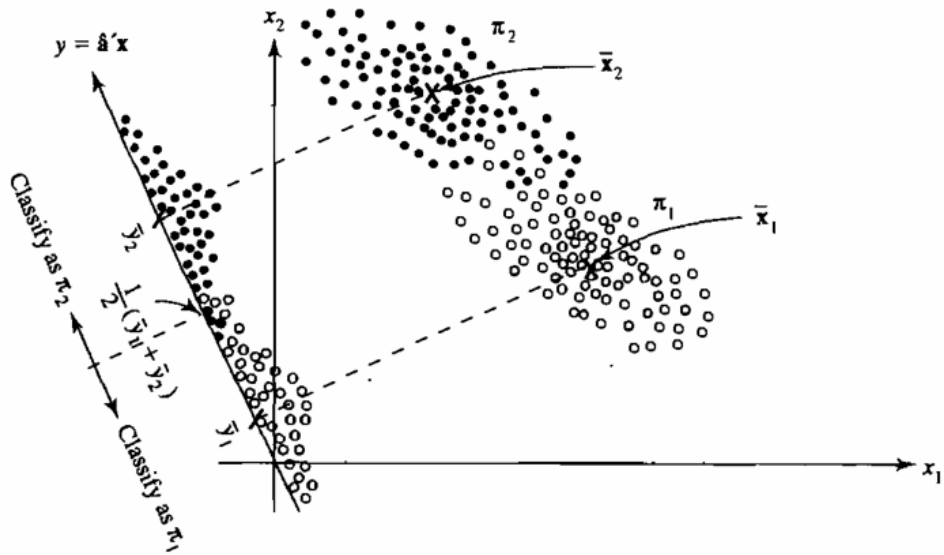
- 이항 분류 문제의 경우,
- $X|Y=1$의 평균이 $\bar{x}_1$, 공분산 행렬이 Σ이며,
- $X|Y=2$의 평균이 $\bar{x}_2$, 공분산 행렬 또한 Σ라면:
- LDA의 방향:

$$\mathbf{a}_{LDA} = \Sigma^{-1}(\bar{x}_2 - \bar{x}_1)$$

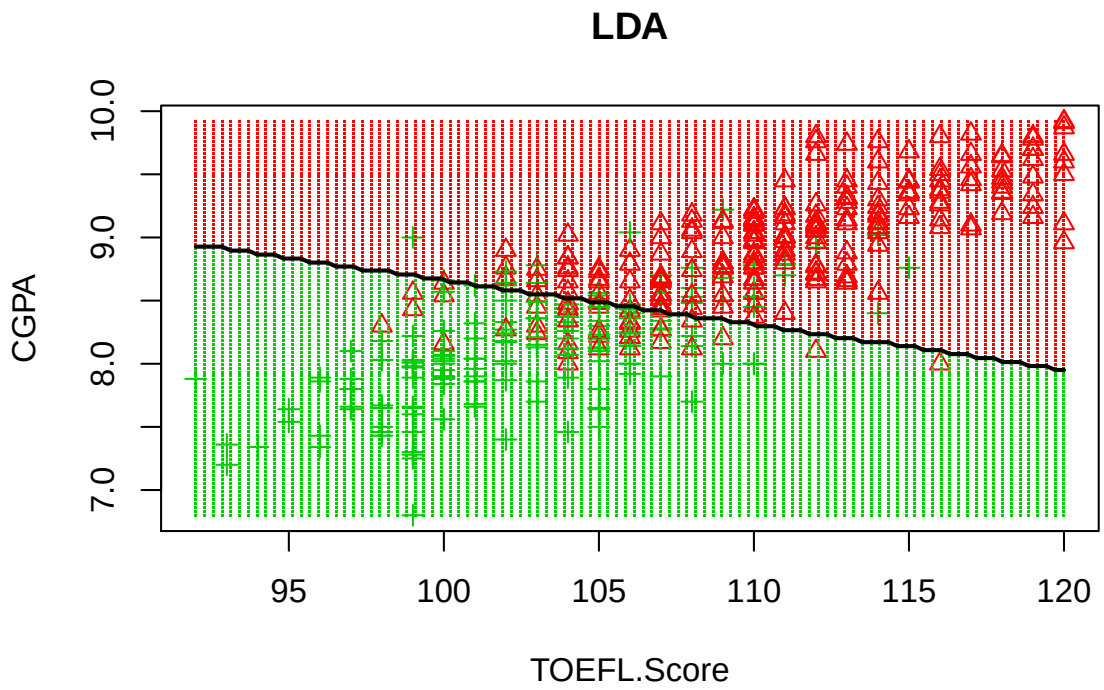
(공분산으로 표준화한 뒤) $\bar{x}_1$에서 $\bar{x}_2$를 가리키는 벡터의 방향

- LDA 분류기 $f_{LDA}(x)$의 값은 다음을 만족할 때 1:

$$\mathbf{a}'_{LDA}(x - \frac{\bar{x}_1 + \bar{x}_2}{2}) < \log(\pi_1/\pi_2)$$



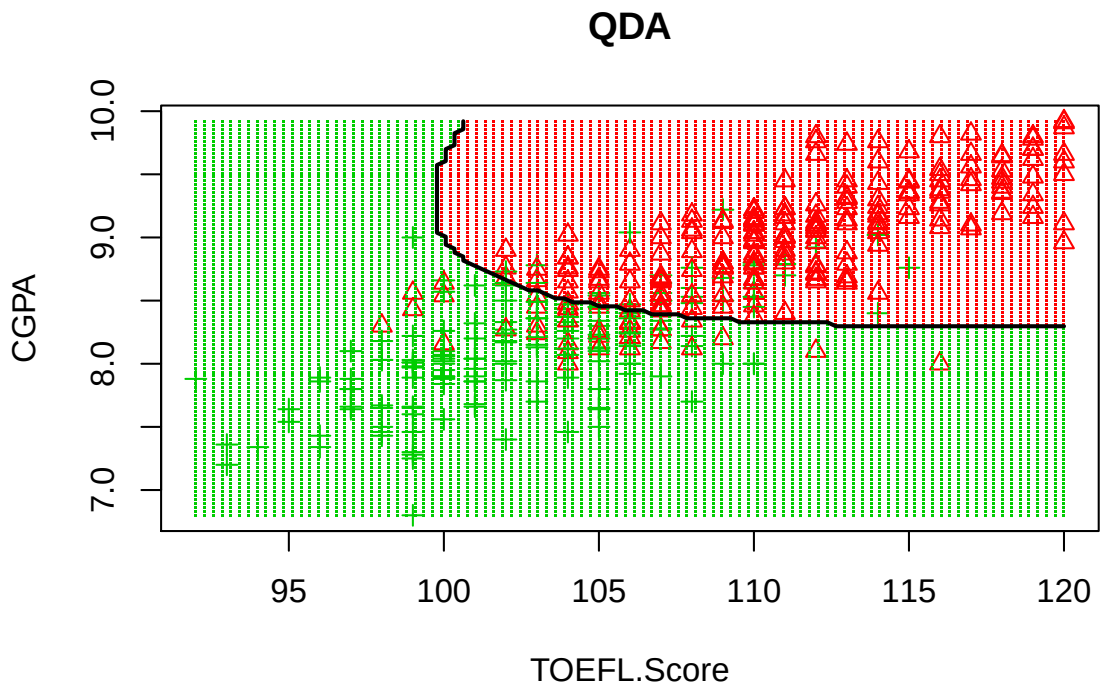
대학원 입시 자료 예



이차판별분석 (Quadratic Discriminant Analysis; QDA)

- 연속형 변수들이 입력변수
- 이항분류, 다항분류 문제에 모두 적용가능
- 선형판별분석과 마찬가지로, 그룹 1의 분포 (X | Y = Admit)와 그룹 2의 분포 (X | Y = No) 를 **정규분포로 모형**화
- 두 그룹의 **평균도 다르고**, **분산-공분산도 다르다**고 가정
- 이항 분류 문제의 경우,
- $X|Y = 1$의 평균이 μ_1, 공분산 행렬이 Σ_1이며,
- $X|Y = 2$의 평균이 μ_2, 공분산 행렬이 Σ_2이며, $\Sigma_1 \neq \Sigma_2$임을 가정.

대학원 입시 자료 예



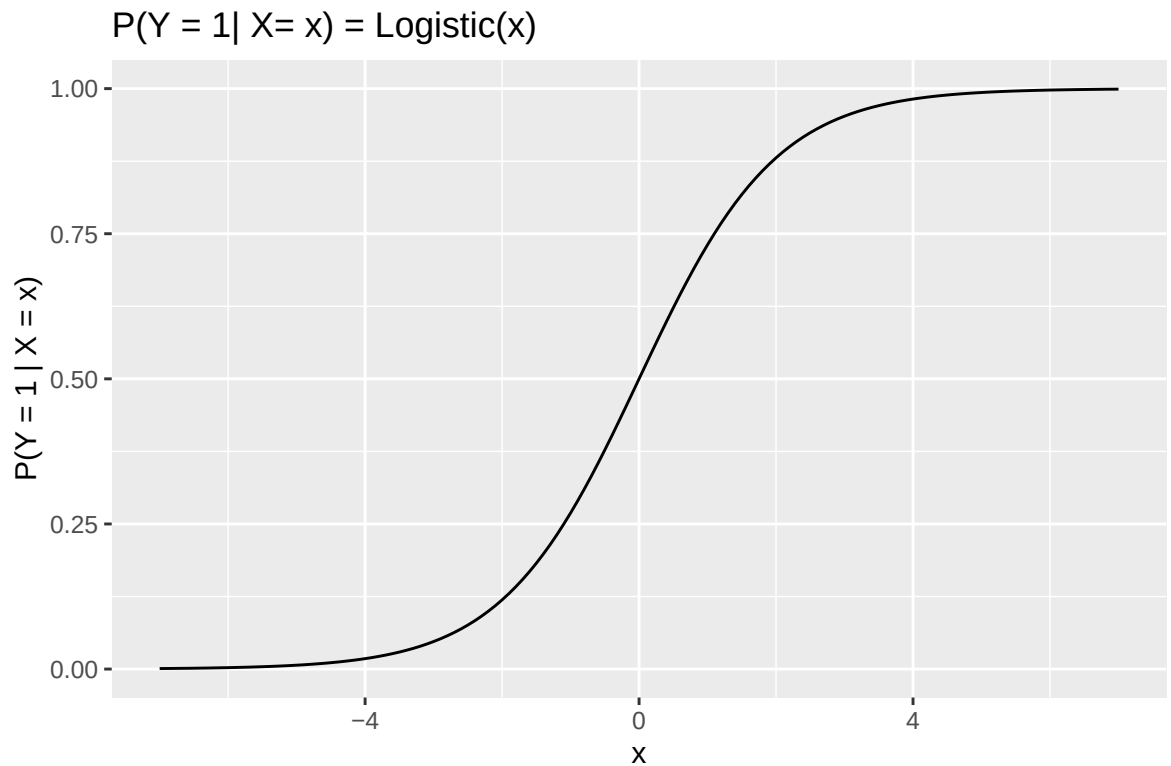
LDA와 QDA 간단한 정리

- 범주별 자료가 다변량 정규분포를 따른다는 가정이 맞다면, 베이즈분류기를 추정하는 매우 강력한 방법
- 변수 변환 (예를 들어 log변환) 등을 이용해 변수의 정규화 뒤 시도 가능
- 입력변수가 모두 연속형인 경우에만 적용 가능
- 분류기가 일차함수 (방향벡터, cutoff 값) 또는 이차함수 (행렬, 벡터, cutoff값)의 형태이므로, 계산이 빠르고 분류가 효율적

knn과 다르게 모든 점에 대한 정보 알 필요 없음

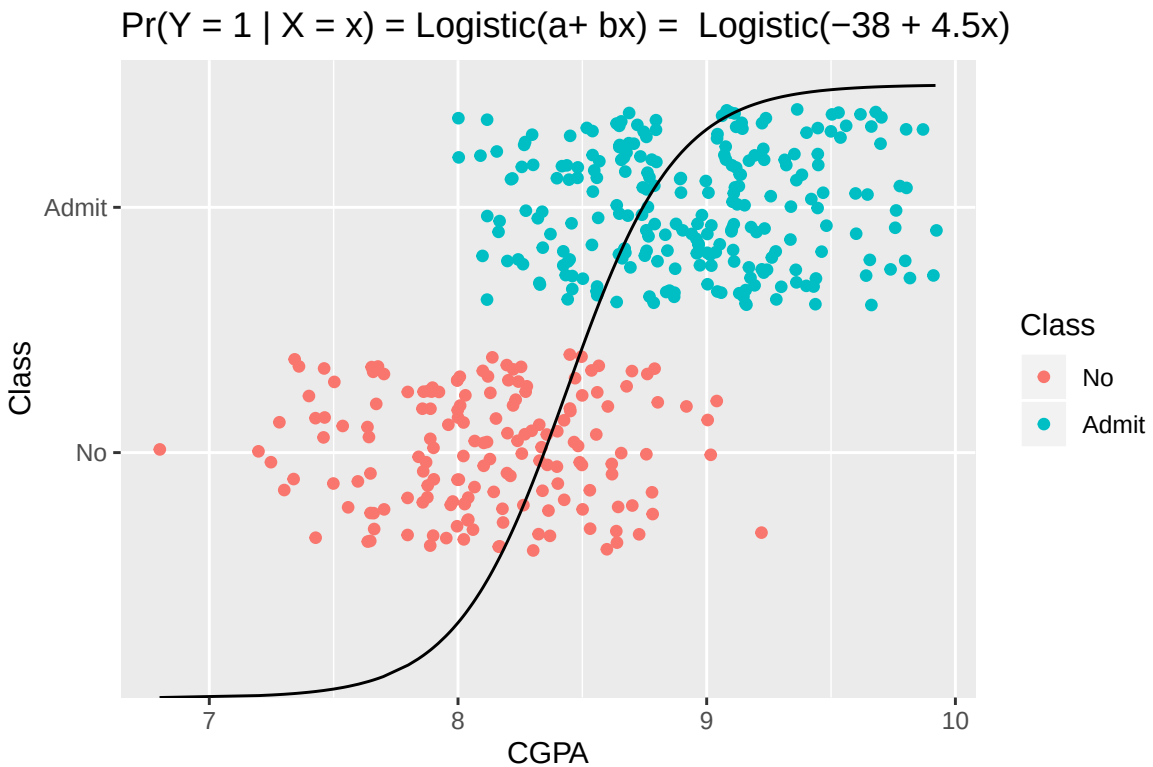
로지스틱회귀분석 (Logistic regression)

- 이항분류법.
- 반응변수 Y의 범주가 단 두 개 ($Y = 0$ 또는 $Y = 1$)일 때의 (일반화선형)회귀모형.
- $P(Y = 1|X = x)$를 로지스틱 함수로 모형화



로지스틱회귀분석

- 하나의 입력변수 x만 있을 때, P(Y = 1 | X = x) = Logistic$(a + bx)$의 a와 b를 추정



로지스틱회귀분석

- 두 개 이상의 입력변수 $x_1, \ldots, x_p$가 있을 때,

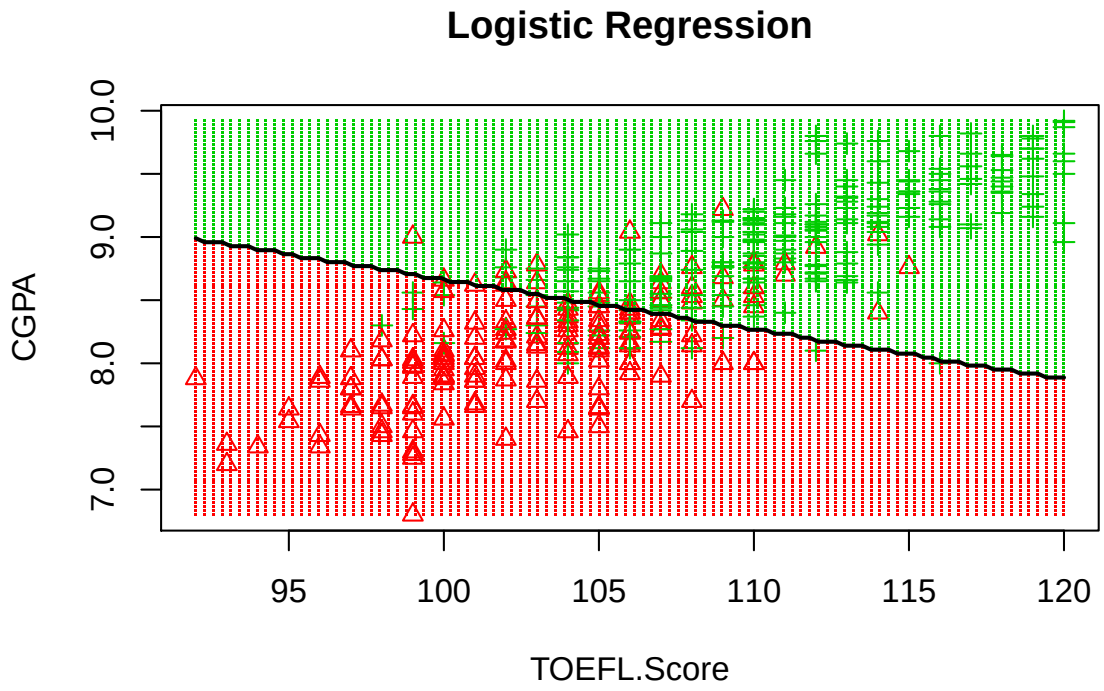
$$\Pr(Y = 1 \mid X = x) = \text{Logistic}(a + \mathbf{b}'\mathbf{x})$$

로 모형, $\mathbf{b}'\mathbf{x} = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$. $(a, \mathbf{b})$ 추정.

- 이때, 로지스틱회귀분석의 분류법은

$$f(x) = \begin{cases} 1, & \text{만약 } P(Y = 1 \mid X = x) > P(Y = 0 \mid X = x), \\ & \text{또는 } P(Y = 1 \mid X = x) > 0.5, \\ & \text{또는 } a + \mathbf{b}'\mathbf{x} > 0; \\ 0, & \text{반대의 경우.} \end{cases}$$

즉, $a + \mathbf{b}'\mathbf{x}$의 값 만으로 분류기가 작동


Logistic Regression
CGPA
10.0
9.0
8.0
7.0
95
100
105
110
115
120
TOEFL.Score

로지스틱회귀분석은 (일반화된) 회귀분석

- 회귀 계수 $\mathbf{b}$에 대한 가설 검정 가능
- 각 변수의 유의성 검정 가능
- 변수 선택 시도 가능 (Deviance, AIC, 가설검정 등을 이용)
  - LDA, QDA, knn도 가능!
- 연속형과 범주형 변수 동시에 설명변수가 될 수 있음
- 해석 상의 편리 (회귀계수와 오즈비)
- 이차항을 만들면 비선형 분류 가능

로지스틱회귀분석 회귀계수의 해석

- 입력변수가 두 개인 경우

$$p(x) \equiv \text{P(Y = 1 | X = x)} = \text{Logistic}(a + b_1 x_1 + b_2 x_2)$$

- Logistic(z) = $\frac{\exp(z)}{1+\exp(z)}$이므로
  $\text{Odds}(x) \equiv \frac{p(x)}{1-p(x)} = \exp(a + b_1 x_1 + b_2 x_2).$
- 예를 들어, x_1 = TOEFL.Score, x_2 = CGPA일 때, CGPA가 같을 때 TOEFL.Score가 1점이 오를 때의 오즈비가 $\exp(b_1)$;

$$\log \text{OddsRatio}(x_1) = b_1.$$

  - TOEFL.Score가 1점이 오르면 합격 로그오즈가 b_1만큼 증가 (또는 합격 오즈가 e^{b_1}배만큼 증가)

잠깐 쉬어가기

최고의 분류법

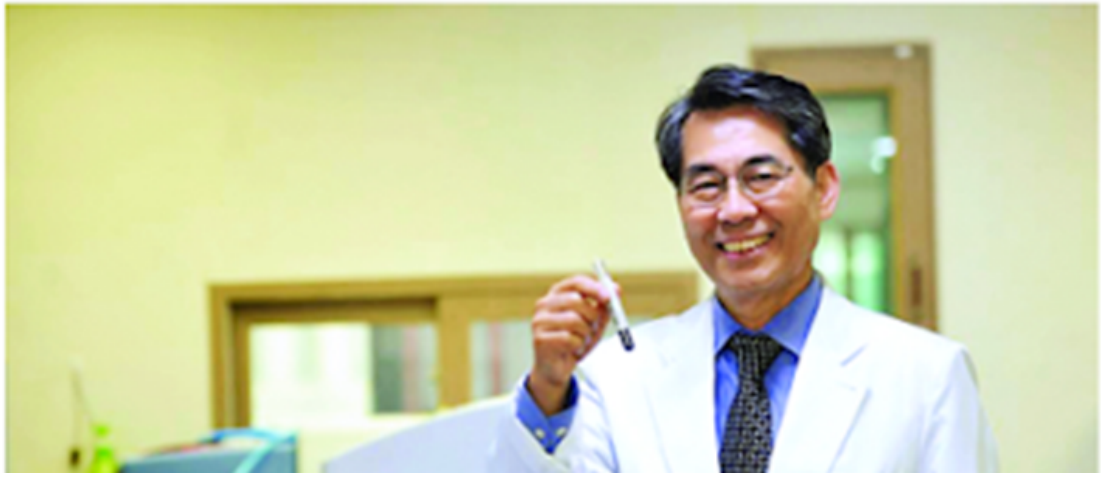
오분류가 아닐 확률

암검사 정확도 90퍼센트!!

국민일보
www.kmib.co.kr

바이오인프라생명과학 김철우 대표 "스마트 암검사 정확도 90%... 8대암 한번에 검진"

입력 2018-08-12 20:32



성능평가 시에는 여러 지표를 봐야 한다.

비교의 원칙!

- 비교를 위해, 2011년 한겨레 기사의 암 종류별 양성예측도와 비교

http://www.hani.co.kr/arti/society/health/477238.html

90% vs 0.64% ???

정확도 vs 양성예측도

국가 암 검진사업의 암 종류별 양성

	위암	간암	대장암	유방암
평균	3.28	5.65	1.69	0.64
30~39살	–	–	–	–
40~49살	1.33	2.90	–	0.5
50~59살	2.68	5.23	1.14	0.7
60~69살	4.87	7.4	1.97	0.94
70살 이상	7.13	6.51	2.72	0.95

※양성예측도: 암 검진에서 암이 의심된다고
람 가운데 최종 검사에서도 암으로 판정되는
칸은 연령별 검진사업 대상이 아닌 항목임.

자료: 〈국가 암 검진사업의 비용과 효과〉(박은철 연세대 의
학교실 교수)

P(Y=1 | Y^=1)

정확도, 양성예측도, 민감도, 특이도.

- 위 국민일보의 기사에는

  "유방암의 경우에서도 83%의 민감도와 90%의 특이도를 나타낸다"

P(Y^=1 | Y=1)

- 민감도(sensitivity): 실제 질병을 가진 사람 중 검사결과가 양성인 비율.
- 특이도(specificity): 실제 질병을 가지지 않은 사람 중 검사결과가 음성인 비율.

P(Y^= 0 | Y= 0)

정확도, 양성예측도, 민감도, 특이도.

- 국가지표체계의 암 발생 및 사망 현황 자료 (http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2770)를 보면 2011년의 유방암 발생율은 100,000명 중 32.3명, 2016년은 42.7명.
- 10만명 당 평균 검사 결과.

	검사결과 양성	검사결과 음성	
유방암 있음	42.7 x 83% = 35.4	7.3	42.7
유방암 없음	9995.7	99957.3 x 90% = 89961.6	99957.3
	10031.1	89968.9	100,000

정확도, 양성예측도, 민감도, 특이도.

정확도 = (35.4 + 89961.6) / 100,000 ~ 90%

국민일보

바이오인프라생명과학 김철우 대표 "스마트 암검사 정확도 90%... 8대암 한번에 검진"

	검사결과 양성	검사결과 음성	
유방암 있음	42.7 x 83% = 35.4	7.3	42.7
유방암 없음	9995.7	99957.3 x 90% = 89961.6	99957.3
	10031.1	89968.9	100,000

매우 스마트한 검사법을 개발해 보자

질문: 정확도가 더 높은 검사방법을 제시하라 (30초)

	검사결과 양성	검사결과 음성	
유방암 있음	42.7 x 83% = 35.4	7.3	42.7
유방암 없음	9995.7	99957.3 x 90% = 89961.6	99957.3
	10031.1	89968.9	100,000

매우 스마트한 검사법을 개발해 보자

답: 검사 결과를 모두 음성으로 만들자!

민감도 0%. 특이도 100%. **정확도 = 99.9573%**

	검사결과 양성	검사결과 음성	
유방암 있음	42.7 x 0% = 0	42.7	42.7
유방암 없음	0	99957.3 x 100% = 99957.3	99957.3
	0	100,000	100,000

~~정확도~~, 양성예측도, 민감도, 특이도.

1:1 정도의 흔한 질병에는 정확도가 큰 의미가 있음

- 희귀한 질병의 예측에는 분류의 정확도를 쓸 수 없다 (매우 잘 알려진 사실)

	검사결과 양성	검사결과 음성	
유방암 있음	42.7 x 83% = 35.4	7.3	42.7
유방암 없음	9995.7	99957.3 x 90% = 89961.6	99957.3
	10031.1	89968.9	100,000

$$\text{양성예측도} = \frac{\text{암판정수}}{\text{검사결과양성의수}} = \frac{35.4}{10031.1} = 0.0035 = 0.35\%$$

공정한 비교의 원칙

- 스마트 암검사 유방암의 양성예측도 = 0.35%
- "정확도 90%", 무엇과 비교해서?
- 암검사 성능의 척도로 부적절한 "정확도"

(Disclaimer) 위암, 간암, 대장암 등 한국인에게 더 많이 나타나는 암의 경우, 스마트 암검사의 양성예측도, 민감도, 특이도가 기존의 방법보다 더 좋을 수가 있다.

국가 암 검진사업의 암 종류별 양성

	위암	간암	대장암	유방암
평균	3.28	5.65	1.69	0.64
30~39살	–	–	–	–
40~49살	1.33	2.90	–	0.5
50~59살	2.68	5.23	1.14	0.7
60~69살	4.87	7.4	1.97	0.94
70살 이상	7.13	6.51	2.72	0.95

※양성예측도: 암 검진에서 암이 의심된다고
람 가운데 최종 검사에서도 암으로 판정되는
칸은 연령별 검진사업 대상이 아닌 항목임.

자료: 〈국가 암 검진사업의 비용과 효과〉(박은철 연세대 의
학교실 교수)

분류 모형의 평가

혼동행렬과 오분류율

- 혼동행렬 (Confusion Matrix): 분류기에 의해 분류를 한 뒤, 각 관측값의 실제 Y값과 예측된 $\hat{Y}$값으로 만든 분할표
- 대학원 입학 자료를 LDA에 의해 분류한 결과

TOEFL.Score	CGPA	Class	predicted.Class
118	9.65	Admit	Admit
107	8.87	Admit	Admit
104	8.00	Admit	No
110	8.67	Admit	Admit
103	8.21	No	No
115	9.34	Admit	Admit

- 대학원 입학 자료를 LDA에 의해 분류한 뒤의 혼동행렬

```
##            Reference
## Prediction  No Admit
##       No    132    38
##       Admit  33   197
```

- 혼동행렬 (Confusion Matrix): 분류기에 의해 분류를 한 뒤, 각 관측값의 실제 Y값과 예측된 $\hat{Y}$값으로 만든 분할표

Y의 두 범주가 Positive, Negative 임을 가정. (병/무병, 합격/불합격 등)

		실제 Positive	실제 Negative	
예측	Positive	TP	FP	P^{pred}
	Negative	FN	TN	N^{pred}
		P	N	Total

- 오분류율 (Missclassification error rate)

$$= \frac{\text{오분류된 자료의 갯수}}{\text{전체 자료의 갯수}} = \frac{FP + FN}{\text{Total}}$$

- 정확도 (Accuracy) = 1 - 오분류율 $= \frac{TP+TN}{\text{Total}}$

양성예측도, 음성예측도

전체적인 오분류율, 정확도도 중요하지만, 각 범주별 오분류율도 중요하다. 암 검사 예제와 같은 **불균형자료**인 경우 더욱 중요하다.

- 양성예측도(Postivie Predictive Value): Positive로 예측분류된 개체가 정말로 Positive인 비율$= \frac{TP}{TP+FP}$ (Precision)
- 음성예측도(Negative Predictive Value): Negative로 예측분류된 개체가 정말로 Negative인 비율 $= \frac{TN}{FN+TN}$

		실제: Positive	실제: Negative	
예측	Positive	TP	FP	→ PPV
	Negative	FN	TN	→ NPV
		P ↘ Sensitivity	N ↘ Specificity	

민감도, 특이도

- 민감도(sensitivity): 실제 Positive인 개체가 정확히 분류된 비율
  $= \frac{TP}{TP+FN} = \frac{TP}{P}$ = 진양성율 (True Positive Rate)
- 특이도(Specificity): 실제 Negative인 개체가 정확히 분류된 비율
  $= \frac{TN}{TN+FP} = \frac{TN}{N}$
  - 위양성율(False Positive Rate) $= \frac{FP}{N} = 1 - \text{Specificity}$

		실제: Positive	실제: Negative	
예측	Positive	TP	FP	→ PPV
	Negative	FN	TN	→ NPV
		P ↘ Sensitivity	N ↘ Specificity	

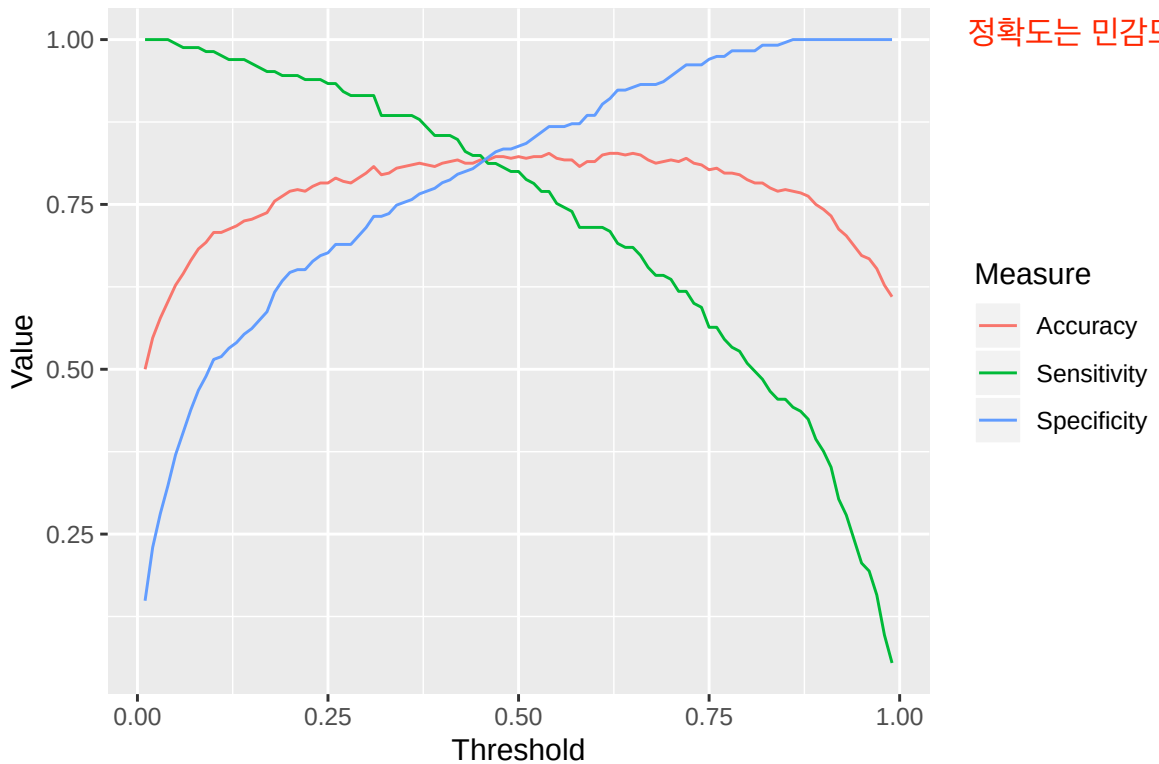
한계점의 변화

- kNN, LDA, QDA, Logistic regression 모두 $\hat{P}(\text{Y = Positive} \mid \text{X = x}) > 0.5$ 일때 $\hat{Y} = \text{Positive}$로 정하는 분류기

이때, 한계점 (cutoff 또는 threshold) 0.5를 줄이면, 더 많은 개체를 Positive로 예측하게 된다. 이때, 진양성(True Positive)율과 민감도는 증가, 특이도와 진음성(True Negative)율은 감소.

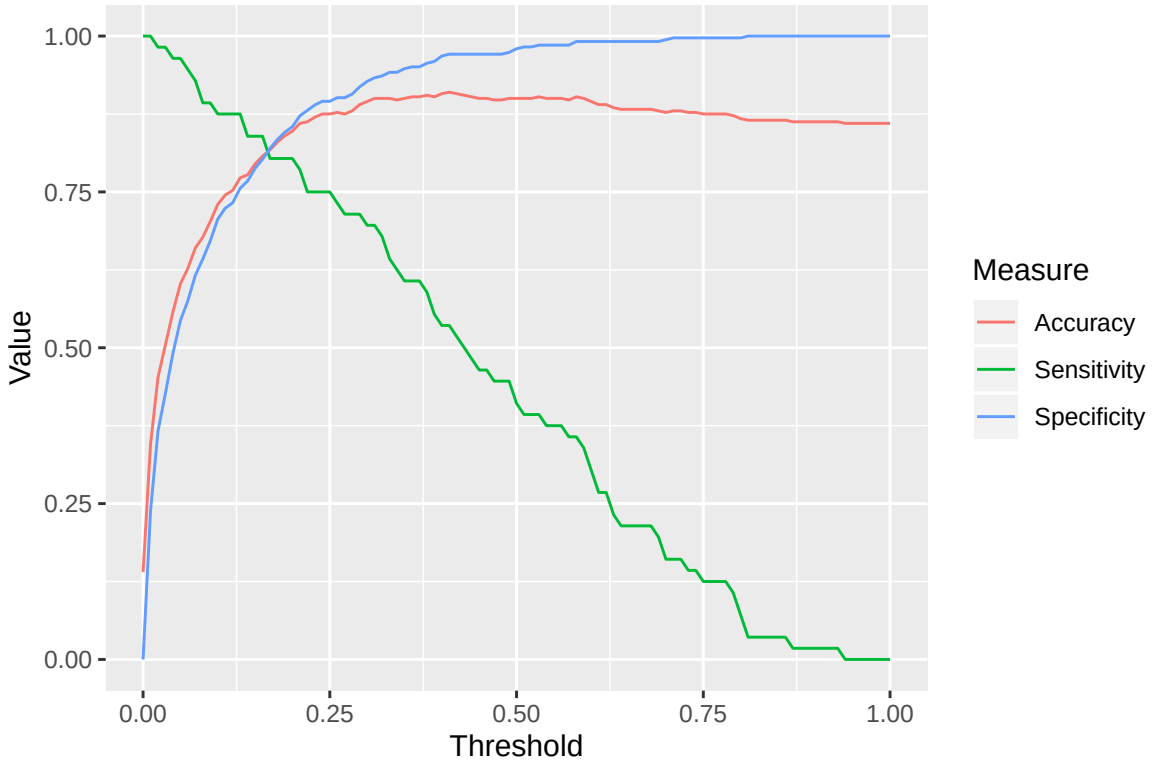
쓰레쉬 증가 -> 민감도 감소, 특이도 증가

정확도는 민감도와 특이도 사이에 존재함



불균형자료에서 한계점의 변화

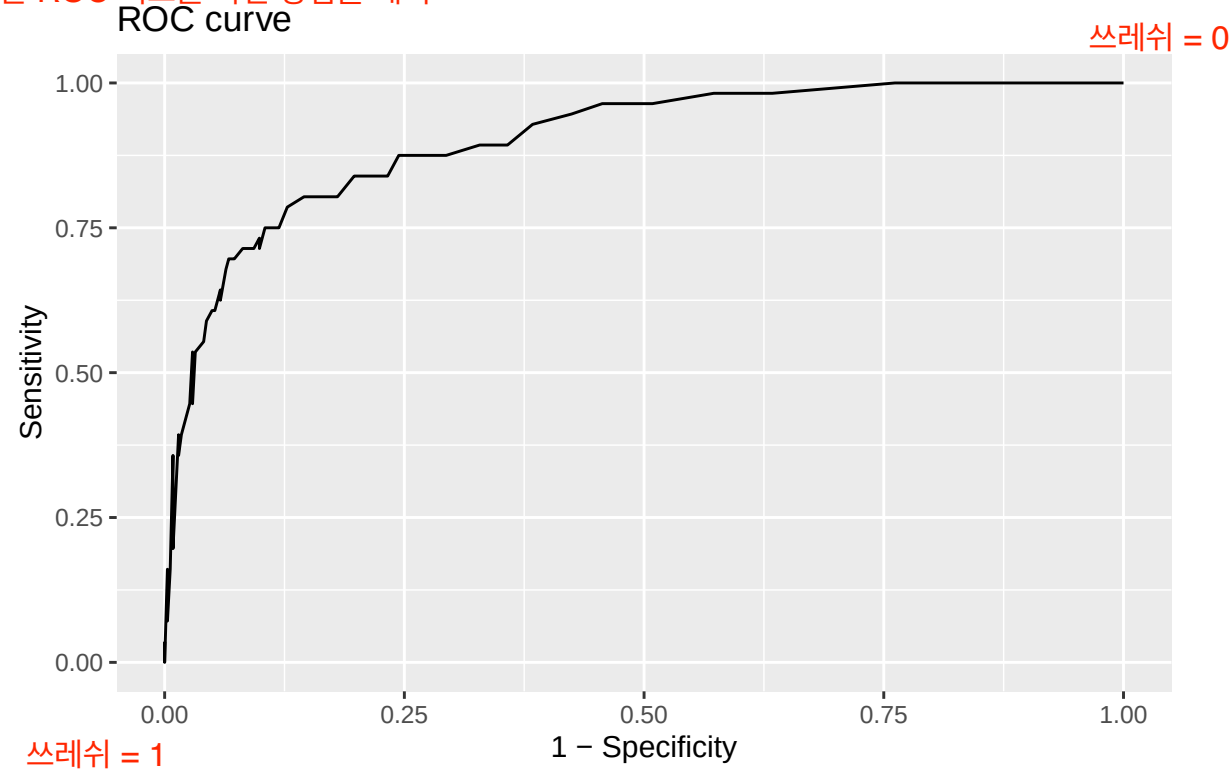
- 예를 들어, “Positive”의 비율이 실제로는 10%였다면, 특이도를 조금 희생하여 민감도를 크게 늘리는게 가능. 쓰레쉬를 0.5에서 0.3 정도로 내림



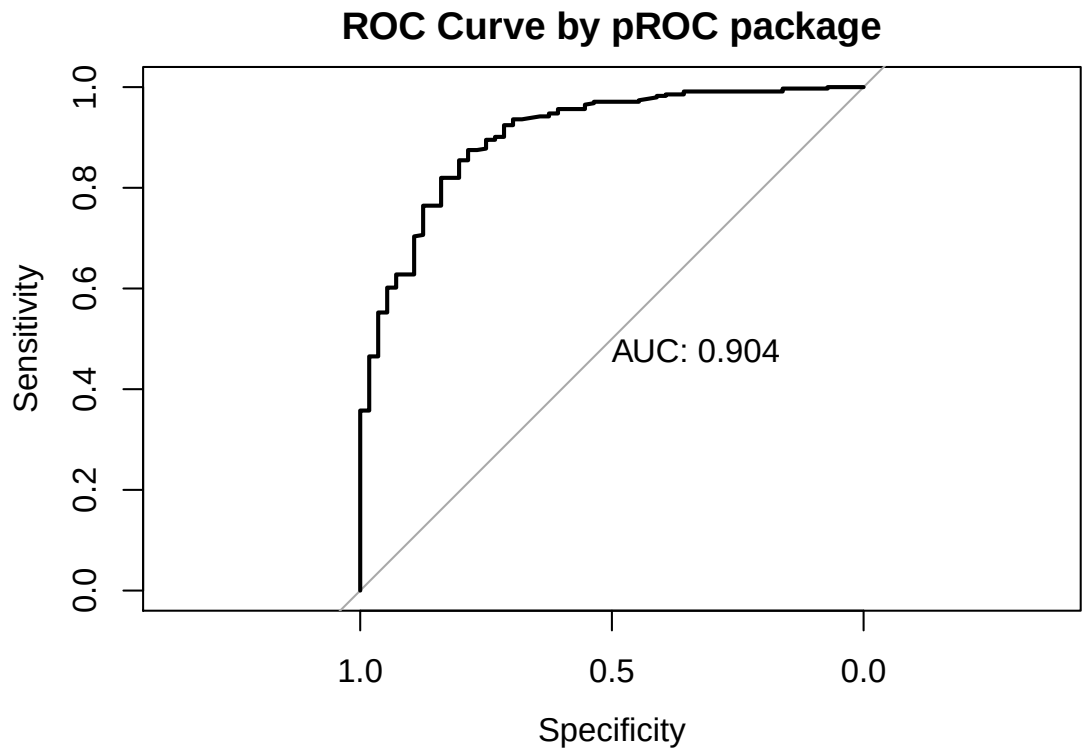
ROC Curve

- (특히 불균형자료인 경우) 한계점의 변화에 따른 민감도와 특이도의 변화를 동시에 보는 것이 중요하다
- 하나의 Threshold에 따라 (민감도, 특이도)의 값이 주어지므로 이를 그래프로 표현 (x축은 1-특이도 = False Positive Rate, y축은 민감도 = True Positive Rate)

(0, 1)에 바짝 붙어있는 ROC 커브를 가진 방법을 채택



- ROC Curve 아래의 면적을 AUC (Area Under the ROC curve)라고 하며 AUC = 0.5 이면 예측력 0. AUC = 1이면 완벽한 예측 가능 (오분류율 0).



분류의 성능 평가

텍스트

- 정확도: 전체 자료 중 분류기가 제대로 예측한 개체의 비율
- 민감도, 특이도: 범주별 정확도
- 양성예측도: Positive로 예측한 개체 중 진짜 Positive인 비율
- AUC: ROC Curve를 종합적으로 평가한 값. 1에 가까울수록 좋은 분류법.

균형자료의 경우, 일반적으로 정확도 (또는 오분류율)을 기준으로 성능 평가

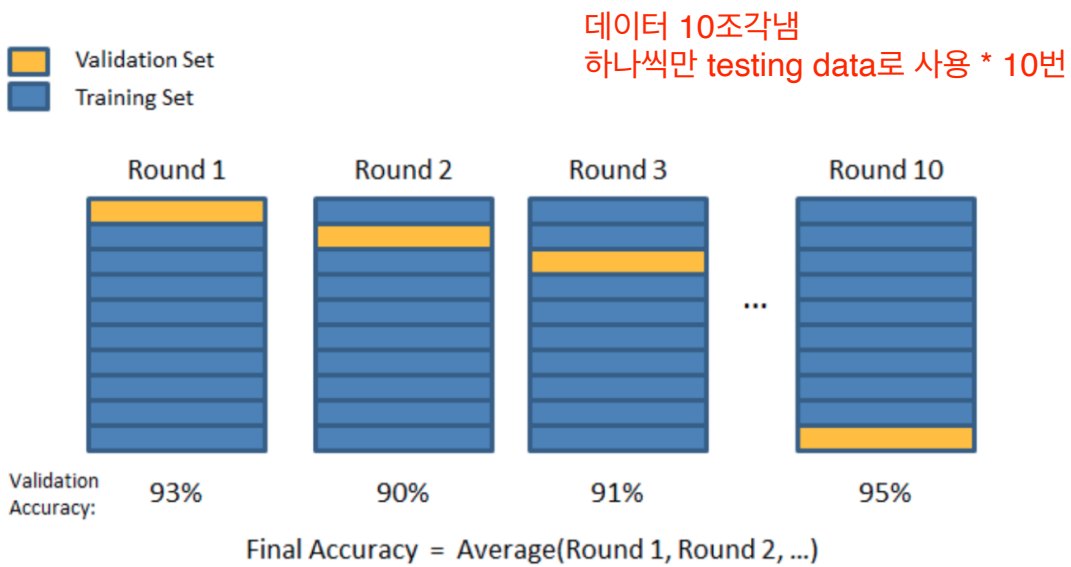
불균형자료인 경우, AUC를 이용 (또는 자료를 강제로 균형있게 만들어줌)

Training vs Testing

- 자료로부터 분류기 $f(x)$를 추정하고, 같은 자료로 정확도 등을 평가하면, 정확도가 과대평가된다.
- 분류기 $f(x)$의 오분류율 $P(\hat{y} \neq y) = P(f(x) \neq y)$을 작게 만드는 분류기가 좋은 분류기
- 이때, $f(x)$를 추정할 때 쓰는 자료 (Training data)와 $P(f(x) \neq y)$를 계산할 때 쓰는 자료 (Testing data)를 분리해야 한다
- 일반적으로, Training data와 Testing data는 같은 모집단에서 샘플된 서로 독립인 자료
- 분류기 비교 평가의 과정:
  - Training data로부터 분류기 (kNN, LDA, QDA, Logistic, etc) $f(x)$를 추정
  - Testing data로부터 $f(x)$의 오분류율, 정확도, AUC 등을 계산

Cross-validation

- 일반적인 자료에는 Training / Testing split이 되어 있지 않다.
- Testing data를 랜덤하게 선택하게 되므로 f(x)와 AUC 등의 계산에 임의성이 포함됨.
- Cross-validation: 반복을 이용하여 임의성을 줄여주는 방법



source: https://medium.com/@josephofiowa

실습

실습 Overview

주로 사용하는 R package

1. `caret` (**C**lassification **A**nd **RE**gression **T**raining): 여러 방법을 이용하여 예측모형을 만들 때 쉽게 이용
2. `pROC`: ROC Curve와 AUC 계산
3. `dplyr`, `GGally`, etc

이용할 자료

- Lower Back Pain Symptoms Dataset
- 310명 개인으로부터 실측된 12개의 척추관련 변수를 이용하여 Y = Abnormal 또는 Normal 예측

분석과정

1. 탐색적자료분석
2. Training/Testing Split
3. 분류기 추정 (kNN, LDA, QDA, Logistic regression)
4. 분류기 성능 계산 및 비교