

분류

서울대학교
통계학과

2018년 8월

“분류” 수업에서 다룰 내용

1. 분류에 대한 소개
2. 베이즈 분류기
3. 최근접이웃방법
4. 로지스틱 모형
5. 선형판별분석
6. 혼동행렬, 오분류율, 민감도와 특이도, ROC 커브

분류(classification)

1. **분류란** 반응변수가 범주형 변수(categorical variable)인 회귀분석
2. **분류의 예들로는**
 - 2.1 응급실에 도착한 환자의 상태 : 심각, 양호
 - 2.2 온라인 결제 은행은 IP 주소, 과거 결제기록 등에 기반해 사기인가, 아닌가 판단
 - 2.3 DNA를 기반해 병의 발생을 예측
3. **분류기(classifier)**는 분류의 방법을 말한다.

4. 분류기(classifier)의 종류로는

- ▶ 로지스틱 회귀분석(logistic regression),
- ▶ 선형판별분석(linear discriminant analysis, LDA),
- ▶ 최근접이웃방법(K-nearest neighbors, KNN),
- ▶ 나무모형(tree models),
- ▶ 랜덤숲(random forest),
- ▶ 부스팅(boosting),
- ▶ 서포트벡터머신(support vector machine)

등이 있다.

5. 본 수업에서는 로지스틱 회귀분석, 선형판별분석, 최근접이웃방법을 다룬다.

반응변수가 범주형 자료일 때 선형회귀가 적합하지 않은 이유

1. y 의 순서가 의미가 없을 때, y 의 순서가 다르면 다른 모형을 나타낸다. 예를 들면, 다음과 같다.

$$y = \begin{cases} 1, & \text{stroke} \\ 2, & \text{drug overdose} \\ 3, & \text{epileptic seizure} \end{cases}$$

2. y 의 순서가 의미가 있다고 할지라도 (예, mild, moderate, severe) 값들의 차이가 1,2,3 으로 혹은 1,2, 10으로 해야하는지 확실하지 않다.
3. y 가 이항변수일 때는 좀 낫다, 그러나 $\mathbb{E}y = X\beta$ 의 범위가 $[0, 1]$ 을 벗어나 해석이 어렵다. 그러나 흥미롭게도 선형회귀를 통한 분류가 LDA와 같다는 것이 알려졌다.

오류 혹은 오차의 정의

훈련오류율(training error rate)

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

훈련 오차율에서 \hat{y}_i 는 자료 y_1, y_2, \dots, y_n 을 이용하여 예측한 값이다.

시험오류율(test error rate)

$$P(y_0 \neq \hat{y}_0)$$

여기서 y_0 는 미래의 관측치를 의미한다. 즉, 시험오류율은 미래의 관측값에 대해 몇 %를 맞추는가에 대한 확률이다.

베이즈 분류기(Bayes classifier)

베이즈 분류기

X 의 값이 x_0 일 때 Y (class)의 예측을 조건부 확률

$$\hat{y}_0 = \operatorname{argmax}_j \mathbb{P}(Y = j | X = x_0)$$

을 이용하고

$$\hat{y}_0 = \operatorname{argmax}_j \mathbb{P}(Y = j | X = x_0)$$

와 같이 정의된 것을 베이즈 분류기라 한다.

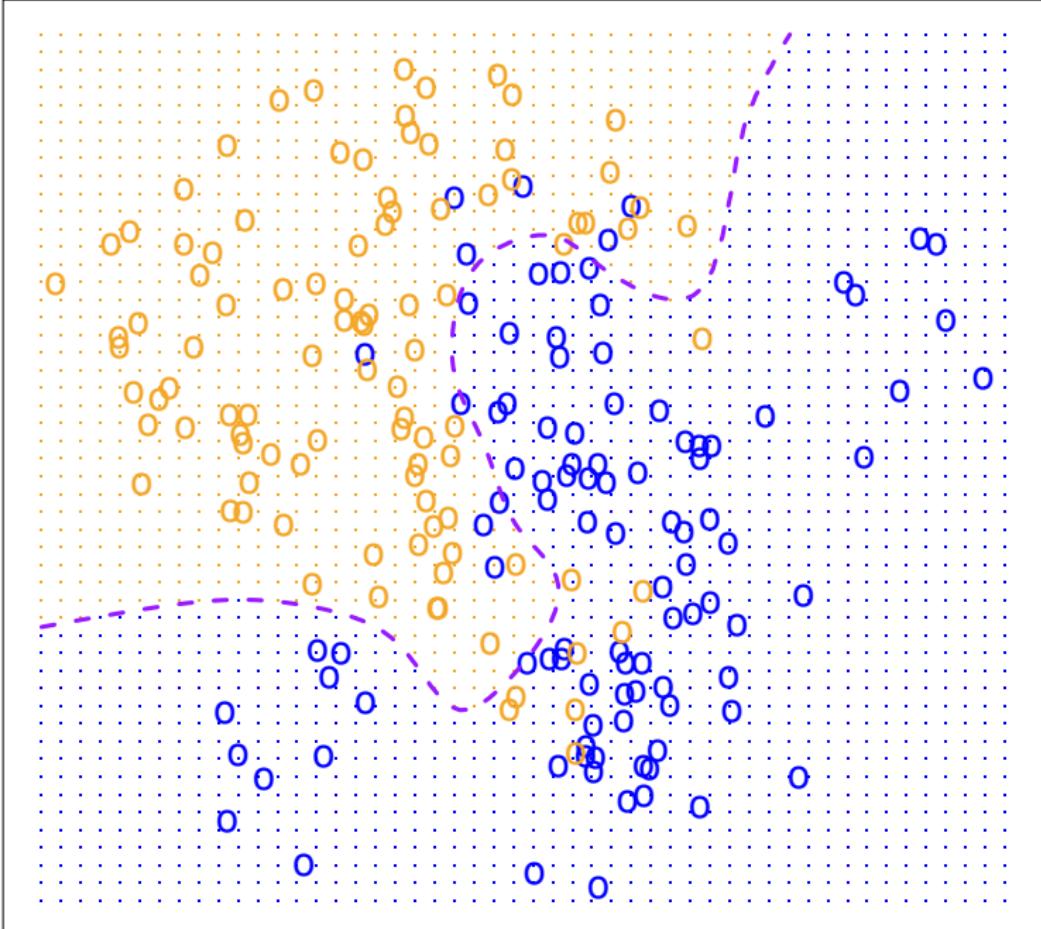
1. 베이즈분류기는 시험오류율을 최소화한다.
2. y 가 2 개의 값(0 또는 1)을 갖으면,

$$\hat{y}_0 = \begin{cases} 1, & \text{if } \mathbb{P}(y_0 = 1|x_0) > 0.5 \\ 0, & \text{if } \mathbb{P}(y_0 = 1|x_0) \leq 0.5 \end{cases}$$

와 같이 정의된다.

3. 베이즈 결정경계(Bayes decision boundary) :

$\mathbb{P}(Y = 1|X) = 0.5$ 인 x 를 나타낸다. 다음 쪽의 그림은
추정된 베이즈 결정경계의 한 예이다.

X_2 

4. 베이지 분류기에서 $\mathbb{P}(Y = j | X = x_0)$ 는 알려져 있지 않고 훈련자료(training data set)로 부터 학습되어야 한다.

5. 베이즈 오류율 $X = x_0$ 에서 오류율은

$$1 - \max_j \mathbb{P}(Y = j | X = x_0)$$

이므로, 전체 x 에서 베이즈 오류율은

$$1 - \mathbb{E}\left(1 - \max_j \mathbb{P}(Y = j | X = x_0)\right)$$

로 정의된다.

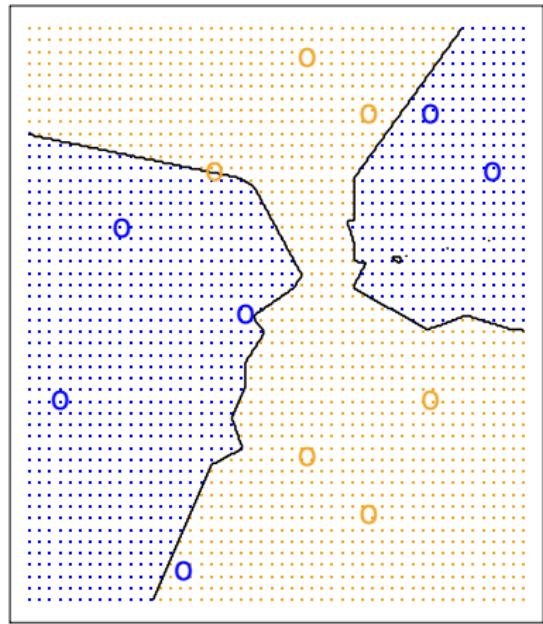
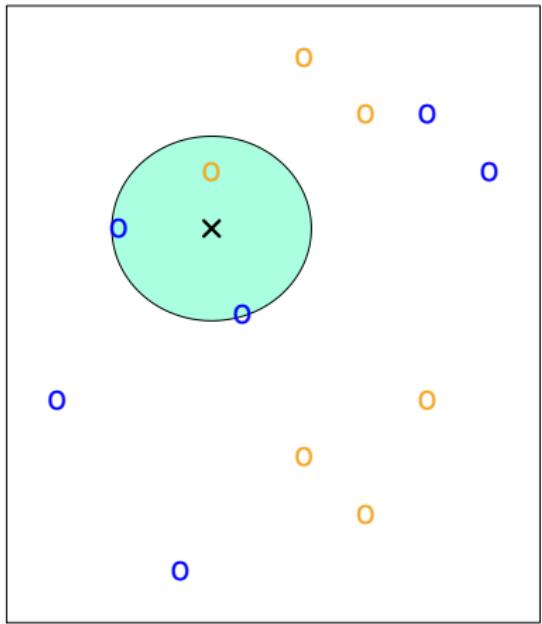
- ▶ 회귀분석에서 irreducible error와 동일한 개념이다.
- ▶ 조건부확률을 모르므로 실제로 구할 수는 없다.
- ▶ 구현할 수 없는 gold standard이다.

최근접이웃방법

최근접이웃방법(K-nearest neighbors, KNN)

주어진 x_0 에서 가장 가까운 K 개의 관측치들로 x_0 에서 각 범주의 확률을 추정하고 가장 많이 나온 범주로 추정한다.

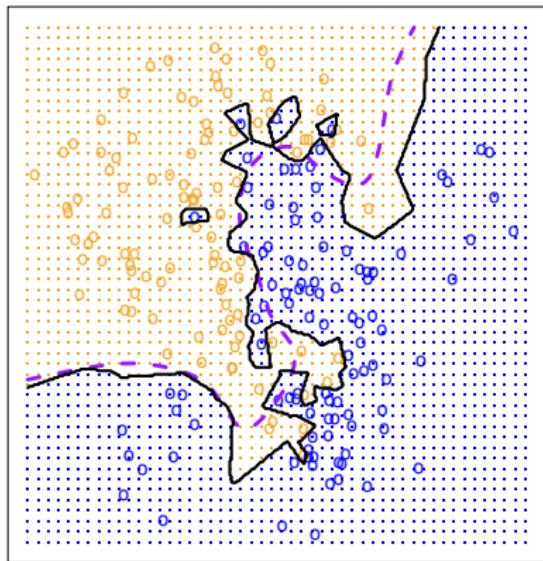
아래의 그림은 $K = 3$ 일 때의 예를 보여준다.



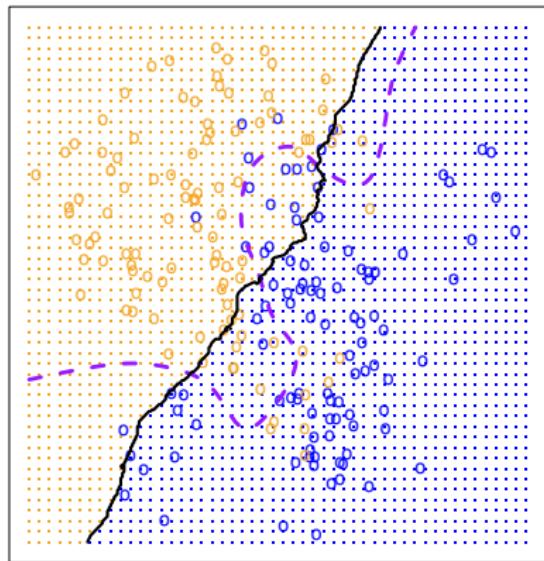
특징들

K가 커질수록 유연성을 떨어진다.

KNN: K=1



KNN: K=100



S & P 500 자료 I

S&P 500 지수

1. S&P 500 혹은 Standard and Poors 500은 뉴욕증시와 나즈닥에 상장되어 있는 500개의 큰 회사들의 주식 가격을 바탕으로 만든 지수(index)이다. 미국 경기를 가장 잘 나타낸다고 알려져 있다.
2. 이 자료는 S&P 500 지수를 2001년부터 2005년까지 1250일 자료 바탕으로 작성된 것이다. 7개의 변수가 있다.

Today가 현재의 1일 수익(return) 퍼센트이고,
Volume은 거래량(단위 : 10억 주),
lag 1은 하루 전 수익, lag5는 5일전 수익,
Year는 년도,
Direction은 수익의 방향을 나타낸다. Down은 1 Up은
2로 코딩이 되었다.

S & P 500 자료 I

```
library(ISLR)
str(Smarket)

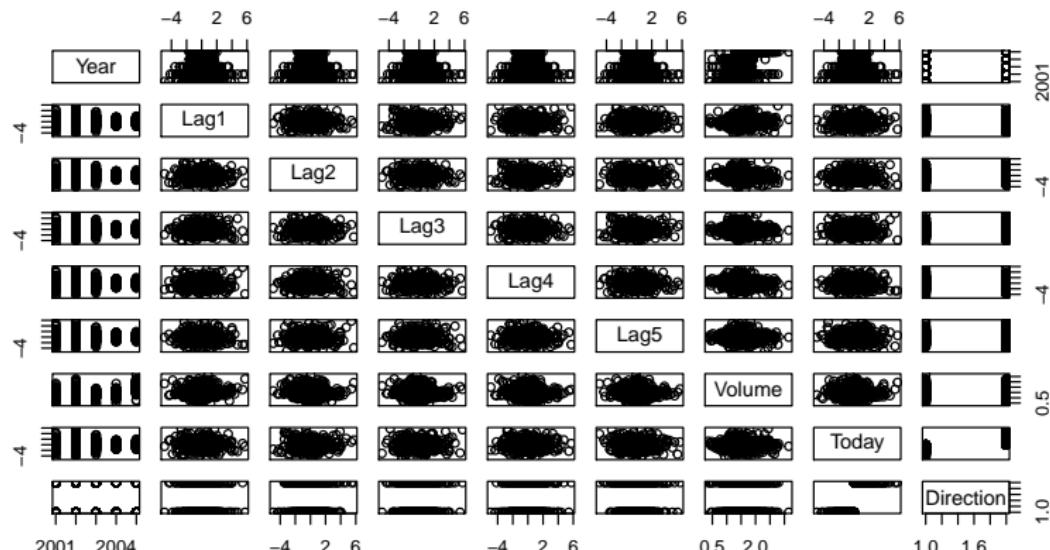
## 'data.frame': 1250 obs. of 9 variables:
## $ Year      : num  2001 2001 2001 2001 2001 ...
## $ Lag1       : num  0.381 0.959 1.032 -0.623 0.614 ...
## $ Lag2       : num  -0.192 0.381 0.959 1.032 -0.623 ...
## $ Lag3       : num  -2.624 -0.192 0.381 0.959 1.032 ...
## $ Lag4       : num  -1.055 -2.624 -0.192 0.381 0.959 ...
## $ Lag5       : num  5.01 -1.055 -2.624 -0.192 0.381 ...
## $ Volume     : num  1.19 1.3 1.41 1.28 1.21 ...
## $ Today      : num  0.959 1.032 -0.623 0.614 0.213 ...
## $ Direction: Factor w/ 2 levels "Down","Up": 2 2 1 2 2 2 1 2 2 2 ...

summary(Smarket[,c(1,7,8,9)])
```

	Year	Volume	Today	Direction
## Min.	:2001	:0.3561	Min. :-4.922000	Down:602
## 1st Qu.	:2002	1st Qu.:1.2574	1st Qu.:-0.639500	Up :648
## Median	:2003	Median :1.4229	Median : 0.038500	
## Mean	:2003	Mean :1.4783	Mean : 0.003138	
## 3rd Qu.	:2004	3rd Qu.:1.6417	3rd Qu.: 0.596750	
## Max.	:2005	Max. :3.1525	Max. : 5.733000	

S & P 500 자료 II

pairs(Smarket)



S & P 500 자료 III

```
cor(Smarket[,-9])
```

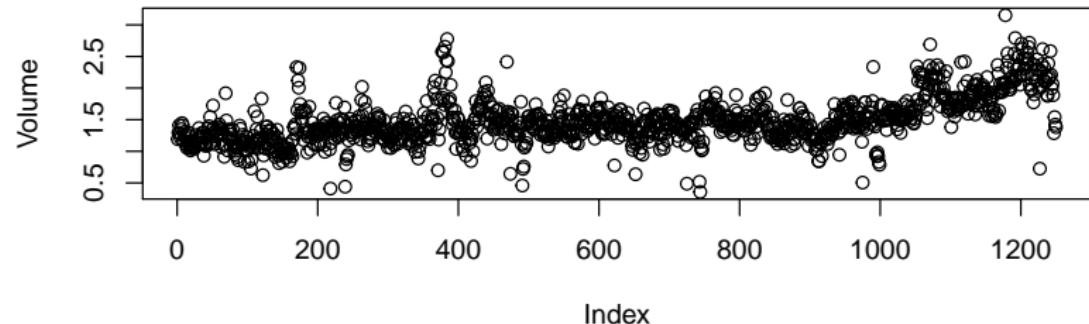
```
##          Year      Lag1      Lag2      Lag3      Lag4
## Year  1.00000000  0.029699649  0.030596422  0.033194581  0.035688718
## Lag1  0.02969965  1.000000000 -0.026294328 -0.010803402 -0.002985911
## Lag2  0.03059642 -0.026294328  1.000000000 -0.025896670 -0.010853533
## Lag3  0.03319458 -0.010803402 -0.025896670  1.000000000 -0.024051036
## Lag4  0.03568872 -0.002985911 -0.010853533 -0.024051036  1.000000000
## Lag5  0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641
## Volume 0.53900647  0.040909908 -0.043383215 -0.041823686 -0.048414246
## Today  0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527
##          Lag5      Volume     Today
## Year   0.029787995  0.53900647  0.030095229
## Lag1  -0.005674606  0.04090991 -0.026155045
## Lag2  -0.003557949 -0.04338321 -0.010250033
## Lag3  -0.018808338 -0.04182369 -0.002447647
## Lag4  -0.027083641 -0.04841425 -0.006899527
## Lag5   1.000000000 -0.02200231 -0.034860083
## Volume -0.022002315  1.00000000  0.014591823
## Today  -0.034860083  0.01459182  1.000000000
```

S & P 500 자료 IV

Direction은 팩터 변수이므로 cor 함수에 인수로 들어가면
에러가 난다. 에러를 피하기 위해서 팩터 변수를 제거하고
cor 함수를 돌렸다.

```
attach(Smarket)  
plot(Volume)
```

S & P 500 자료 V



S & P 500 자료 VI

훈련자료와 시험자료의 분리

```
train=(Year<2005)
Smarket.2005=Smarket[!train,]
dim(Smarket.2005)

## [1] 252    9

Direction.2005=Direction[!train]
```

2001년부터 2004년까지를 훈련자료로, 2005년 자료를 시험자료로 구성하였다.

최근접이웃방법 R 코드 |

```
library(class)
train.X=cbind(Lag1,Lag2)[train,]
test.X=cbind(Lag1,Lag2)[!train,]
train.Direction=Direction[train]
set.seed(1)
knn.pred=knn(train=train.X,test=test.X,cl=train.Direction, k=1, prob=TRUE)
table(knn.pred,Direction.2005)

##          Direction.2005
## knn.pred Down Up
##      Down   43 58
##      Up     68 83

head(attributes(knn.pred)$prob)

## [1] 1 1 1 1 1 1
```

최근접이웃방법 R 코드 II

- 함수 knn은 class 패키지 안에 있다.
- 다른 통계함수들은 모형을 적합한 후에, 적합한 결과를 이용해서 예측을 2단계로 따로 하는데, knn은 한번에 예측을 한다.
- 인수 : train은 훈련자료의 설명변수를 행렬이나 데이터프레임으로, test는 예측을 할 시험자료의 설명변수값들을 행렬이나 데이터프레임으로 ci은 훈련자료의 반응변수로 팩터이다. k는 예측에 이용하는 이웃의 개수이다.
- 결과는 팩터이다. 확률은 attributes를 이용해 얻을 수 있다.
- 위의 코드는 K = 1인 경우 아래는 K = 3인 경우이다.
- set.seed를 쓴 이유 : knn은 tie가 있는 경우 랜덤하게 tie를 깨는데, 이를 고정시키기 위해 썼다.

최근접이웃방법 R 코드 III

```
knn.pred=knn(train.X,test.X,train.Direction,k=3)
table(knn.pred,Direction.2005)

##          Direction.2005
## knn.pred Down Up
##      Down   48 54
##      Up     63 87

mean(knn.pred==Direction.2005)

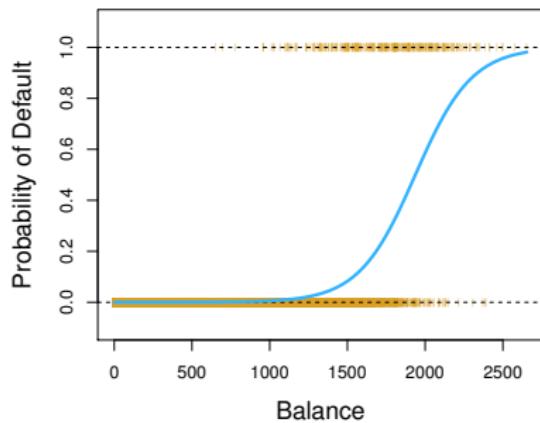
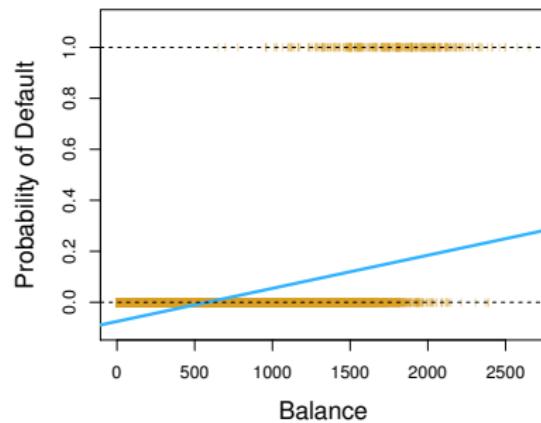
## [1] 0.5357143
```

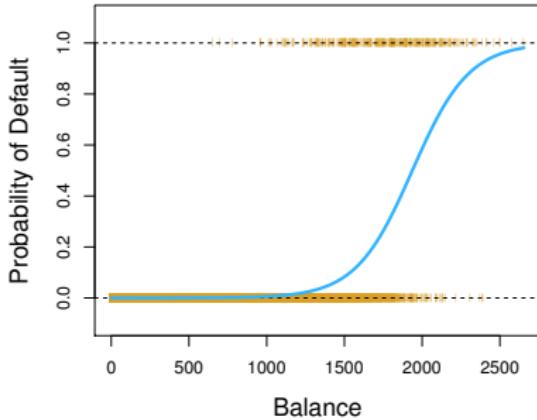
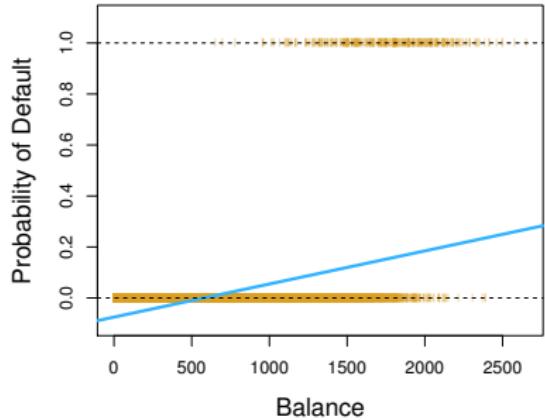
- ▶ 여러 K값에 대하여 시도하여 본다.
- ▶ 임의로 정의한 거리함수에 대하여 시도하여 본다.

로지스틱 회귀모형

$Y = 1$ 인 확률

$$p(x) = \text{logistic}(\beta_0 + \beta_1 x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$
$$\text{logistic}(x) = \frac{e^x}{1 + e^x}$$





1. 왼쪽은 선형회귀를 이용하여 확률을 추정한 것이다.
어떤 값들은 음수가 된다.
2. 오른쪽은 로지스틱 회귀모형을 이용하여 확률을 추정한 것이다. 함수의 모양이 S자 커브이다.

오즈(odds)

반응변수의 값이 x 일 때 오즈는

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

와 같이 나타내진다. 오즈가 $1/4$ 라는 것은 $p(x) = 0.2$ 이고
5명 중 1명이 디폴트가 된다는 뜻이다. 또는 디폴트와
디폴트가 아닌 사람들의 비율이 $1 : 4$ 가 된다는 뜻이다.

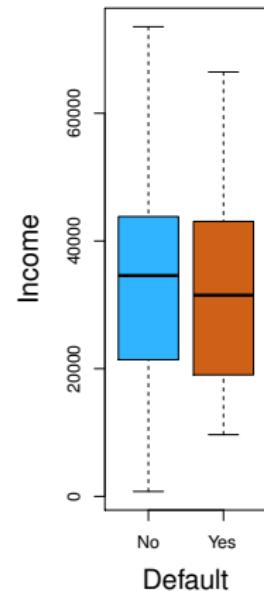
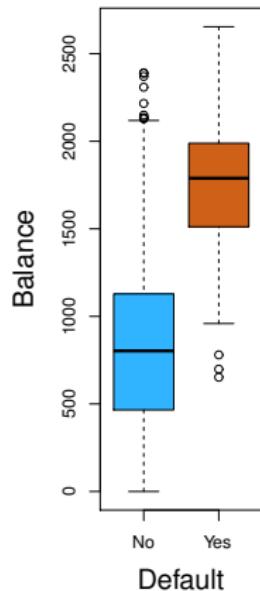
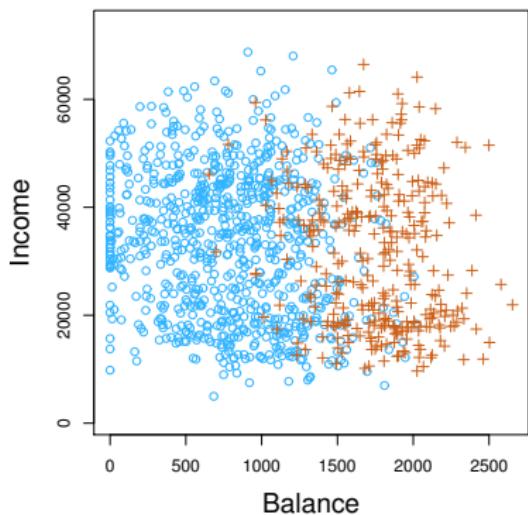
로짓(logit) 함수

로지스틱 함수의 역함수이다.

$$\text{logit}(p(x)) = \log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x$$

로그오즈는 $p(x)$ 의 로짓함수가 된다. β_1 은 x 가 한 단위 증가할 때 증가하는 로그오즈의 크기이다.

디폴트 자료



1. 모의 자료이다. 로지스틱모형을 이 자료로 설명한다.
2. 변수가 y (default), x_1 : annual income, x_2 : monthly balance (잔고)가 있다.

오렌지 색은 신용카드 채무 디폴트가 된 고객, 푸른색은 디폴트가 아닌 고객을 나타낸다.

최대가능도 추정법 I

가능도함수

$X \sim f(x; \theta)$ 를 따르고 $X = x$ 를 관측했을 때,

$$\mathcal{L}(\theta) = f(x; \theta)$$

를 가능도 함수라 한다. 밀도함수를 θ 의 함수로 본 것이다.
 $\ell(\theta) = \log \mathcal{L}(\theta)$ 는 로그가능도 함수(log-likelihood ftn)라 한다.

최대가능도 추정량

$$\hat{\theta}^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta)$$

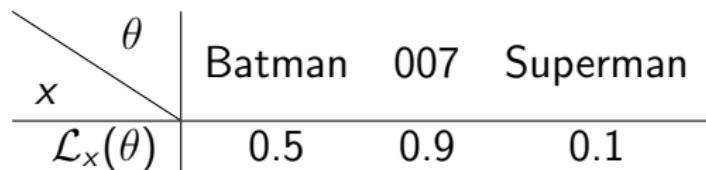
최대가능도 추정법 II

최대가능도 추정량의 근거

젊은 아가씨가 보낸 사람의 이름이 적혀 있지 않은 꽃다발을 받았다. 이 아가씨에게는 현재 세 명의 구혼자가 있다. 세 명의 구혼자들의 행적을 볼 때 이들에게 구애의 대상이 생겼을 때 꽃다발을 보낼 확률은 다음과 같다.

θ	Batman (average)	007 (womanizer)	Superman (shy)
x	0.5	0.9	0.1
꽃다발 보냄	0.5	0.1	0.9
꽃다발 보내지 않음	0.5	0.1	0.9

이 때 가능도함수를 보면



최대가능도 추정법 III

$$\hat{\theta}^{MLE} = \operatorname{argmax}_{\theta} \mathcal{L}_x(\theta) = 007.$$

MLE는 상식적이다. MLE를 배우지 않은 사람도 MLE를 사용한다.

회귀계수의 추정 : 최대가능도 방법

최대가능도 추정법

가능도 함수

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

를 최대로 하는 β_0 와 β_1 값을 구한다.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

확률의 추정

$p(x)$ 의 식에 $\hat{\beta}_0, \hat{\beta}_1$ 의 값을 대입하여 추정량을 구한다.

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

1. $balance = 1000$ 일 때, $\hat{p}(x) = 0.00576$
2. $balance = 2000$ 일 때, $\hat{p}(x) = 0.586$

회귀계수의 설명

디폴트 자료의 로지스틱 회귀모형의 회귀추정량 값들이다.

1. $\hat{\beta}_1 = 0.0055$ 라는 것은 balance가 1불 증가할 때 디폴트 로그오즈가 0.0055씩 증가한다는 뜻이다.
2. $\hat{\beta}_0 = -10.6513$ 라는 것은 balance가 0일 때, 디폴트의 로그오즈가 -10.6513 혹은 오즈가 $e^{-10.6513} = 2.367005 \times 10^{-5}$ 이라는 뜻이다.
3. 이 테이블의 표준오차를 이용하여 β_0 와 β_1 에 대한 가설검정과 신뢰구간을 구할 수 있다.

다중 로지스틱 회귀모형(multiple logistic regression)

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

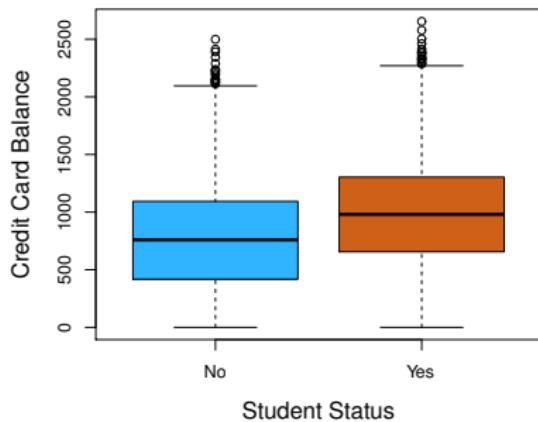
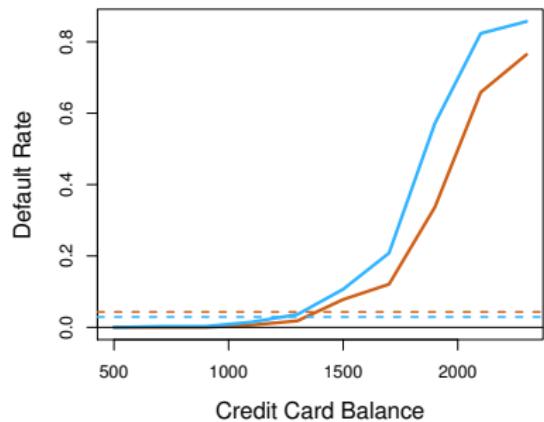
문제

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	<0.0001
student [Yes]	0.4049	0.1150	3.52	0.0004

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

학생만을 예측변수로 썼을 때
회귀계수는 0.4이고
유의했는데, 다른 변수를
포함시켰을 때는 학생의
회귀계수는 음수이고
유의하지 않다. 이 현상을
어떻게 설명해야 할까?

설명



노트.

위의 그림에서 학생들은 오렌지, 일반인은 파란색이다. 오른쪽 박스그림을 학생들이 디폴트할 확률이 크다는 것을 보인다. 하지만 왼쪽 그림은 다른 반응변수들의 값들이 고정되었을 경우, 즉 학생들이 동일한 반응변수의 값을 갖는 일반인들 보다는 디폴트할 확률이 작다. 학생들은 보통 다른 사람들 보다 많은 빛(balance)을 지고 있어서 이런 현상이 나타난다. 신용카드 회사의 입장에서는 학생에 대한 다른 정보가 없다면 학생들이 디폴트할 확률이 크기 때문에 다른 사람들에게 신용카드를 허가하는 것보다 신중해야 한다. 그러나 만약 잔고와 같은 학생의 다른 정보가 있다면 다른 변수들의 값이 동일한 일반인들 보다 학생들이 더 안전하다는 뜻이다.

로지스틱 모형의 R 코드: 모형의 적합 |

```
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Smarket,family=binomial)
summary(glm.fit)

##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Smarket)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.446  -1.203   1.065   1.145   1.326
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.126000  0.240736 -0.523   0.601
## Lag1        -0.073074  0.050167 -1.457   0.145
## Lag2        -0.042301  0.050086 -0.845   0.398
## Lag3         0.011085  0.049939  0.222   0.824
## Lag4         0.009359  0.049974  0.187   0.851
## Lag5         0.010313  0.049511  0.208   0.835
## Volume       0.135441  0.158360  0.855   0.392
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1731.2 on 1249 degrees of freedom
## Residual deviance: 1727.6 on 1243 degrees of freedom
## AIC: 1741.6
##
## Number of Fisher Scoring iterations: 3
```

로지스틱 모형의 R 코드: 모형의 적합 II

노트

회귀계수에 다음과 같은 함수들을 적용할 수 있다.

로지스틱 모형의 R 코드: 모형의 적합 III

```
coef(glm.fit)

## (Intercept)          Lag1          Lag2          Lag3          Lag4
## -0.126000257 -0.073073746 -0.042301344  0.011085108  0.009358938
##           Lag5          Volume
##  0.010313068  0.135440659

summary(glm.fit)$coef

##                               Estimate Std. Error     z value Pr(>|z|)
## (Intercept) -0.126000257  0.24073574 -0.5233966 0.6006983
## Lag1        -0.073073746  0.05016739 -1.4565986 0.1452272
## Lag2        -0.042301344  0.05008605 -0.8445733 0.3983491
## Lag3         0.011085108  0.04993854  0.2219750 0.8243333
## Lag4         0.009358938  0.04997413  0.1872757 0.8514445
## Lag5         0.010313068  0.04951146  0.2082966 0.8349974
## Volume      0.135440659  0.15835970  0.8552723 0.3924004
```

```
summary(glm.fit)$coef[,4]

## (Intercept)      Lag1      Lag2      Lag3      Lag4      Lag5
## 0.6006983  0.1452272  0.3983491  0.8243333  0.8514445  0.8349974
##       Volume
## 0.3924004
```

```
glm.probs=predict(glm.fit,type="response")
glm.probs[1:10]

##      1      2      3      4      5      6      7
## 0.5070841 0.4814679 0.4811388 0.5152224 0.5107812 0.5069565 0.4926509
##      8      9     10
## 0.5092292 0.5176135 0.4888378
```

1. predict 함수에 newdata 인수는 디폴트는 NULL이다. 이때는 적합한 자료에서의 예측값을 준다.
2. type의 디폴트는 "link"이다. 이때는 예측값의 단위가 로그 오즈 즉 $\log \frac{p(x)}{1 - p(x)}$ 이다. type의 값을 "response"로 하면 예측값의 단위가 $p(x)$ 가 된다.

```
glm.pred=rep("Down",1250)
glm.pred[glm.probs>.5]="Up"
table(glm.pred,Direction)
```

```
##          Direction
## glm.pred Down  Up
##      Down   145 141
##      Up     457 507
```

```
mean(glm.pred==Direction)
```

```
## [1] 0.5216
```

```
train=(Year<2005)
Smarket.2005=Smarket[!train,]
dim(Smarket.2005)
```

```
## [1] 252    9
```

```
Direction.2005=Direction[!train]
```

시험자료와 훈련자료로 나누었다.

```
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Smarket,  
family=binomial,subset=train)
```

훈련자료를 이용해 모형을 적합했다.

```
glm.probs=predict(glm.fit,Smarket.2005,type="response")  
glm.pred=rep("Down",252)  
glm.pred[glm.probs>.5]="Up"  
table(glm.pred,Direction.2005)  
  
## Direction.2005  
## glm.pred Down Up  
##     Down    77 97  
##     Up      34 44  
  
mean(glm.pred==Direction.2005)  
  
## [1] 0.4801587  
  
mean(glm.pred!=Direction.2005)  
  
## [1] 0.5198413
```

시험자료에서 예측을 해서 실제값과 비교한다.

```
glm.fit=glm(Direction~Lag1+Lag2,data=Smarket,
family=binomial,subset=train)
glm.probs=predict(glm.fit,Smarket.2005,type="response")
glm.pred=rep("Down",252)
glm.pred[glm.probs>.5]="Up"
table(glm.pred,Direction.2005)

##          Direction.2005
## glm.pred Down  Up
##      Down    35   35
##      Up     76 106

mean(glm.pred==Direction.2005)

## [1] 0.5595238

106/(106+76)

## [1] 0.5824176
```

```
predict(glm.fit,newdata=data.frame(Lag1=c(1.2,1.5),Lag2=c(1.1,-0.8)),  
       type="response")  
  
##           1           2  
## 0.4791462 0.4960939
```

Lag1과 Lag2 두 개의 설명변수만 이용해서 다시 적합해보았다.

선형판별분석(linear discriminant analysis)

목적

1. Y 는 K 개의 범주를 갖는 반응변수.
2. X 는 예측변수.
3. $\mathbb{P}(Y = k|X = x) = p(x)$ 를 추정하고자 한다.

선형판별분석을 배우는 이유

1. 반응변수의 값들이 분리가 잘되어 있는 경우 로지스틱 회귀모형의 변수 추정값은 안정적이지 못하다.
2. X 의 분포가 정규분포에 가까울 때 선형판별분석의 추정량들이 안정적이다.
3. 반응 변수의 범주가 3 개 이상일 때 많이 쓰인다.

아이디어

- 로지스틱 모형과 같이 $p_k(X)$ 를 직접 모형화하지 않고

$$f_k(x) = \mathbb{P}(X = x | Y = k)$$

를 **다면량 정규분포**로 모형화한다.

- 베이즈 정리를 이용하여 $\mathbb{P}(Y = k | X = x)$ 를 추정한다.

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

여기서 $\pi_k = \mathbb{P}(Y = k)$ 이며, 이의 추정은 표본에서 $Y = k$ 의 비율로 추정한다.

- $\hat{y} = \operatorname{argmax}_k \mathbb{P}(Y = k | X = x) = \operatorname{argmax}_k \pi_k f_k(x)$

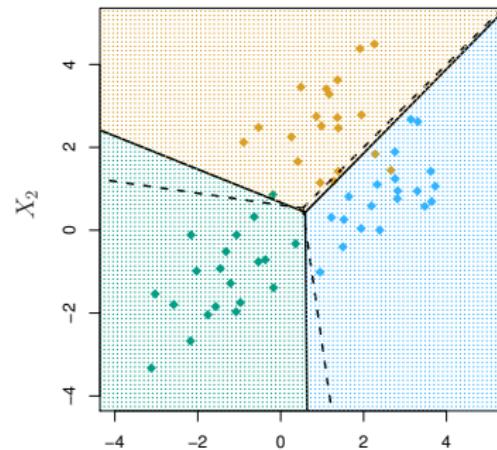
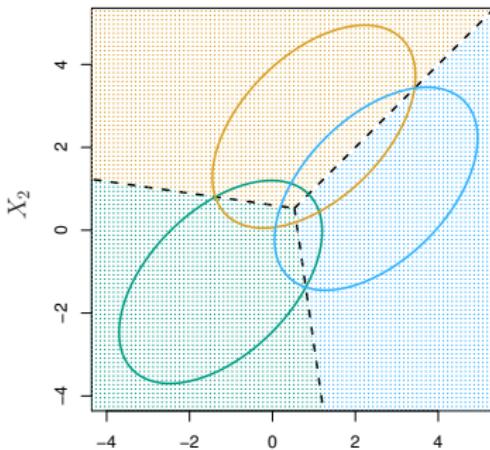
결정경계

결정경계

$\pi_k = \pi_l$ 일 때, $\delta_k(x) = \delta_l(x)$ 가 같은 조건은

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

이다.



선형판별분석 R 코드 I

```
library(MASS)
lda.fit=lda(Direction~Lag1+Lag2,data=Smarket,subset=train)
lda.fit

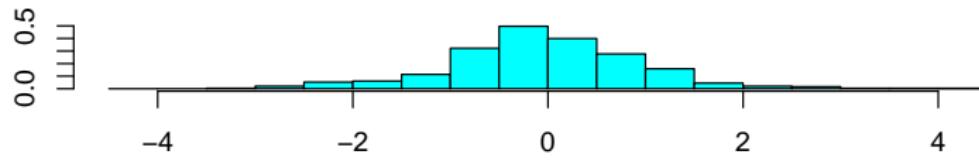
## Call:
## lda(Direction ~ Lag1 + Lag2, data = Smarket, subset = train)
##
## Prior probabilities of groups:
##       Down      Up
## 0.491984 0.508016
##
## Group means:
##           Lag1      Lag2
## Down  0.04279022  0.03389409
## Up   -0.03954635 -0.03132544
##
## Coefficients of linear discriminants:
##           LD1
## Lag1 -0.6420190
## Lag2 -0.5135293
```

선형판별분석 R 코드 II

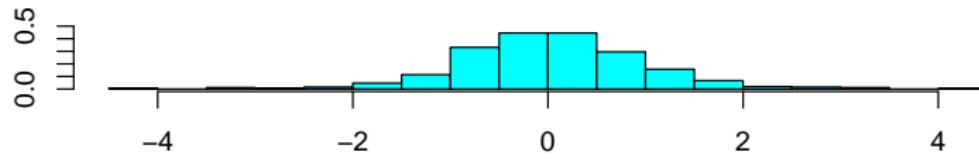
선형판별분석을 적합하는 함수 lda는 MASS 패키지 안에 있다. 반응변수와 설명변수를 모형식을 이용해 대입한다.

```
plot(lda.fit)
```

선형판별분석 R 코드 III



group Down



group Up

선형판별함수의 값을 각 그룹에 대해 히스토그램을 그렸다.

선형판별분석 R 코드 IV

```
lda.pred=predict(lda.fit, Smarket.2005)
names(lda.pred)

## [1] "class"      "posterior"   "x"

lda.class=lda.pred$class
```

lda.pred의 결과는 세 개의 구성요소를 갖는다.

선형판별분석 R 코드 V

```
table(lda.class,Direction.2005)

##          Direction.2005
## lda.class Down   Up
##      Down    35   35
##      Up     76 106

mean(lda.class==Direction.2005)

## [1] 0.5595238

sum(lda.pred$posterior[,1]>=.5)

## [1] 70

sum(lda.pred$posterior[,1]<.5)

## [1] 182

# lda.pred$posterior[1:20,1]
# lda.class[1:20]
sum(lda.pred$posterior[,1]>.9)

## [1] 0
```

혼동행렬(confusion matrix)과 오분류율

아래는 디폴트 자료의 혼동행렬을 나타낸다. 대각원소는 옳은 분류를, 비대각 원소는 오류를 의미한다.

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

1. 전체오차율

$$= \frac{\text{오분류된 자료의 개수}}{\text{전체 자료의 개수}} = \frac{23 + 252}{10,000} = 2.75\%$$

$$2. \text{ 디폴트하지 않은 고객 중 오분류율} = \frac{23}{9667} = 0.2\%$$

$$3. \text{ 디폴트한 고객 중 오분류율} = \frac{252}{333} = 75.7\%$$

민감도(sensitivity)와 특이도(specificity)

분류기의 성능을 나타내는 지표로 쓰인다. 두 개의 값 모두 1에 가까워야 좋은 것이다.

1. 민감도(sensitivity) : $y = 1$ 인 관측치 중 정확히 분류된 비율. $\frac{81}{333} = 24.3\%$
2. 특이도(specificity) : $y = 0$ 인 관측치 중 정확히 분류된 비율. $\frac{9644}{9667} = 99.8\%$

디폴스자료: 질문

전체 오분류은 2.75%로 매우 낮아서 분류기의 성능이 좋아 보이지만, 디폴트를 한 고객들에 대한 오분류율은 75.7%로 매우 높다. 신용카드회사의 입장에서는 디폴트할 고객들을 찾아내는 것이 중요하기 때문에 이 분류기의 성능은 오히려 나쁘다고 할 만하다. 왜 이런 일이 생기는가?

노트. 답변

베이즈 분류기는

$$\mathbb{P}(\text{default} = \text{yes} | X = x) > 0.5$$

인 경우 디폴트라고 분류한다. 디폴트한 고객인가 하지 않은 고객인가에 대한 구별은 하지 않고 전체 오분류율을 최소화 한다. 대부분의 고객들이 디폴트하지 않은 고객이기 때문에 디폴트하지 않은 고객에 대해 오분류율을 줄여주면 전체 오분류율이 줄어들기 때문에 디폴트하지 않은 고객들에 대해서는 정교하게 분류하고 디폴트한 고객에 대해서는 정교하게 분류하지 않아서 생기는 문제이다.

해결책 : 한계점(threshold, 분계점)의 변화

한계점 축소

디폴트가 난 사람들을 좀 더 정교하게 골라내고 싶으면 한계값을 0.5에서 다음과 같이 좀 줄인다. 즉

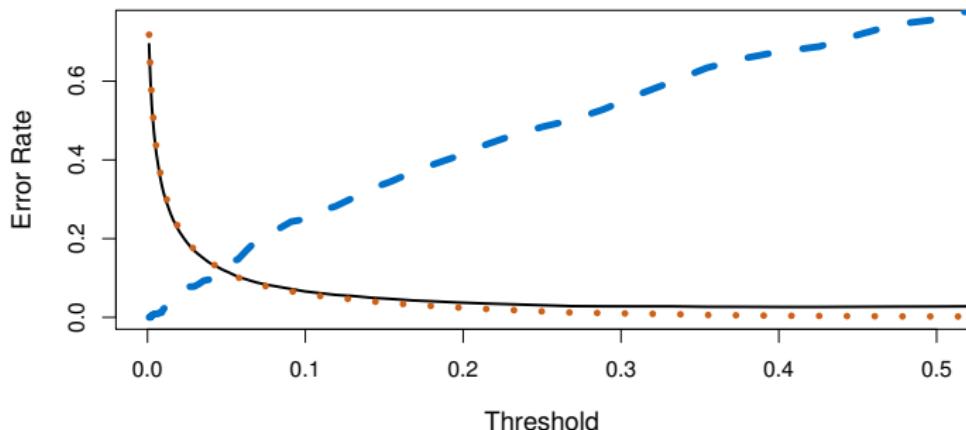
$$\mathbb{P}(\text{default} = \text{yes} | X = x) > 0.2$$

이면 디폴트로 분류한다. 이 때의 혼동행렬은 다음과 같다.

		<i>True default status</i>		Total
		No	Yes	
<i>Predicted default status</i>	No	9,432	138	9,570
	Yes	235	195	430
Total	9,667	333	10,000	

한계점과 오차율

전체오차율(검은색 실선), 디폴트한 고객 중 오차율(파란색 대시), 디폴트 안한 고객 중 오차율(오렌지)이 어떻게 변하는지 보여준다. 전체오차율은 디폴트 안한 사람들 중 오차율에 지배를 받는다. 한계점이 낮아질 수록 디폴트한 사람 중 오차율은 작아진다.

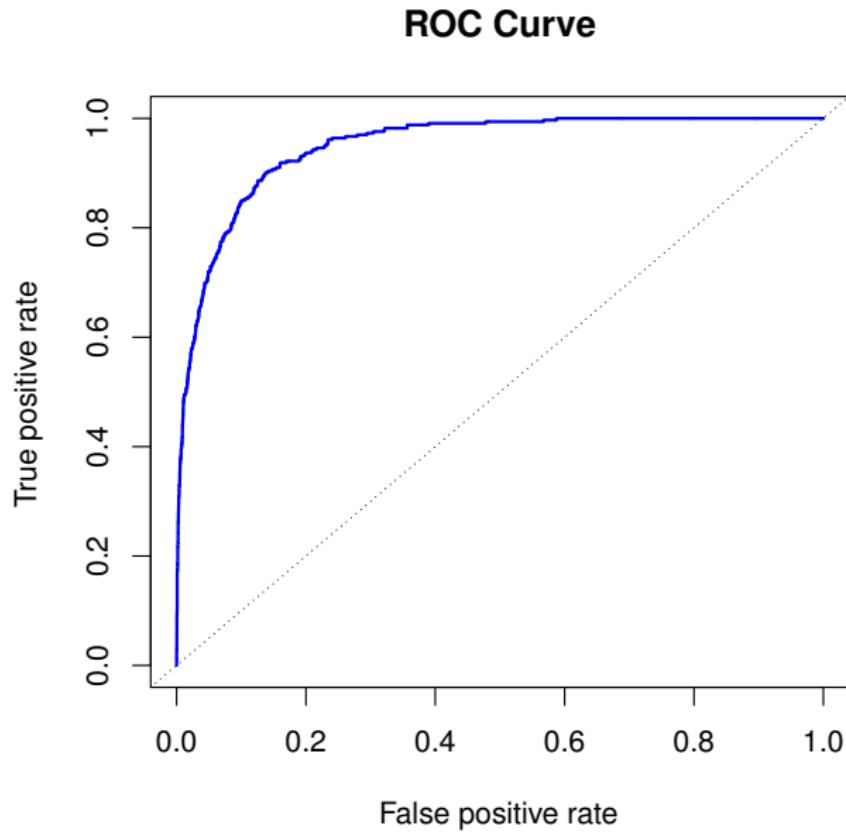


성능측도의 종류

1. 진양성율(true positive rate) : $\frac{TP}{P}$
민감도와 같다. + 중 +로 예측한 비율. 1에 가까울수록 좋다. H_1 중 H_1 이라 예측한 비율. $\mathbb{P}(H_1\text{를 선택}|H_1\text{가 사실})$. 검정력과 같다.
2. 위양성율(false positive rate) : FP/N
- 중 +로 예측한 비율. H_0 중 H_1 이라 예측한 비율.
 $\mathbb{P}(H_1\text{를 선택}|H_0\text{가 사실})$. 제1종의 오류 확률과 같다. 0에 가까울수록 좋다.
3. positive predicted rate : +로 예측한 것 중 +인 것. TP/P^*
4. negative predicted rate : -로 예측한 것 중 -인 것. TN/N^*

		Predicted class		
		- or Null	+ or Non-null	Total
True class	- or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
Total		N*	P*	

ROC 커브



노트.

1. ROC는 receiver operating characteristic의 약자이다. 이 개념은 세계 제2차 대전 때 레이더 기술자들이 적군의 군함을 감지하는 문제에 처음 적용하면서 만들어진 개념이다. 영어의 실제 의미는 현재는 큰 의미가 없다.
2. 위양성을 대 진양성을 그림을 그린 것이다.
3. 한 점이 한계값이 고정된 한 개의 분류기의 오차를 나타낸다.
4. 한계값을 크게하면 +로 예측한 관측치의 개수가 늘어난다. 따라서 진양성을, 위양성을 모두 커진다.
5. 자료와 전혀 상관 없이 분류기를 만들면 예를 들어 동전을 던져 앞면이 나오면 디폴트라고 하고 뒷면이 나오면 디폴트가 아니라고 하면, 진양성을 =위양성을 이게 된다. 관측치가 +인지 -인지 것과 상관없이 +, -를 분류하기 때문이다.(분류기의 예측값과 +, - 값이 서로 독립이게 된다.) 이 때 동전의 앞면이 나올 확률이 0에서 1로 움직이면서 ROC커브에서 $y = x$ 직선을 나타낸다. 보통 ROC 커브는 $y = x$ 직선의 위쪽에 곡선이 있고 왼쪽 상단에 붙어있는 곡선일 수록 좋은 성능을 나타낸다.
6. ROC 커브 전체를 평가하는 기준으로 AUC(area under curve)가 있다. 보통 0.5와 1사이의 값이고 1에 가까울 수록 좋은 것이다.

이차판별분석(quadratic discriminant analysis)

1. 아이디어

$X|Y = k$ 의 분포를 $N(\mu_k, \Sigma_k)$ 로 모형화하여 베이즈 정리를 적용한다.

2. 예측식

$$\hat{y} = \operatorname{argmax}_k \delta_k(x)$$

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

3. δ_k 가 x 에 관한 2차식이어서 2차판별분석이라한다.
 k 마다 Σ_k 를 추정해야 하므로 추정해야 할 파리미터의 개수가 선형판별분석보다 $(k - 1)\frac{p(p + 1)}{2}$ 개 많다.

이차판별분석 R 코드 I

```
qda.fit=qda(Direction~Lag1+Lag2,data=Smarket,subset=train)
qda.fit

## Call:
## qda(Direction ~ Lag1 + Lag2, data = Smarket, subset = train)
##
## Prior probabilities of groups:
##      Down       Up
## 0.491984 0.508016
##
## Group means:
##           Lag1       Lag2
## Down  0.04279022  0.03389409
## Up   -0.03954635 -0.03132544
```

이차판별분석 R 코드 II

```
qda.class=predict(qda.fit,Smarket.2005)$class
table(qda.class,Direction.2005)

##          Direction.2005
## qda.class Down  Up
##       Down    30   20
##       Up      81 121

mean(qda.class==Direction.2005)

## [1] 0.5992063
```

참고문헌

아래의 책에서 제공되는 그림을 써서 슬라이드를 만들었다.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning. Springer, 2013.