Discriminant Analysis

(판별분석)

H. Park

HUFS

Discriminant Analysis

- 목적: 주어진 과거의 자료를 근거로 현재 어떤 개체(subject)가 어느 group에 속하게 되는지 판별하는 통계적 기법

  (예): 홍적세 층에서 발견된 두개골(skull)화석의 크기를 측정($x_1, x_2, \ldots x_q$)하여 어느 그룹에 속한 화석인지 판별

  ★ 네안데르탈인

  ★ 크로마뇽인

  ★ 원숭이

(예) 은행의 대출정보를 가지고 새롭게 대출을 신청하는 고객의 대출가능/불가 판정

★ 고객정보 (x_1, x_2, ... x_q)

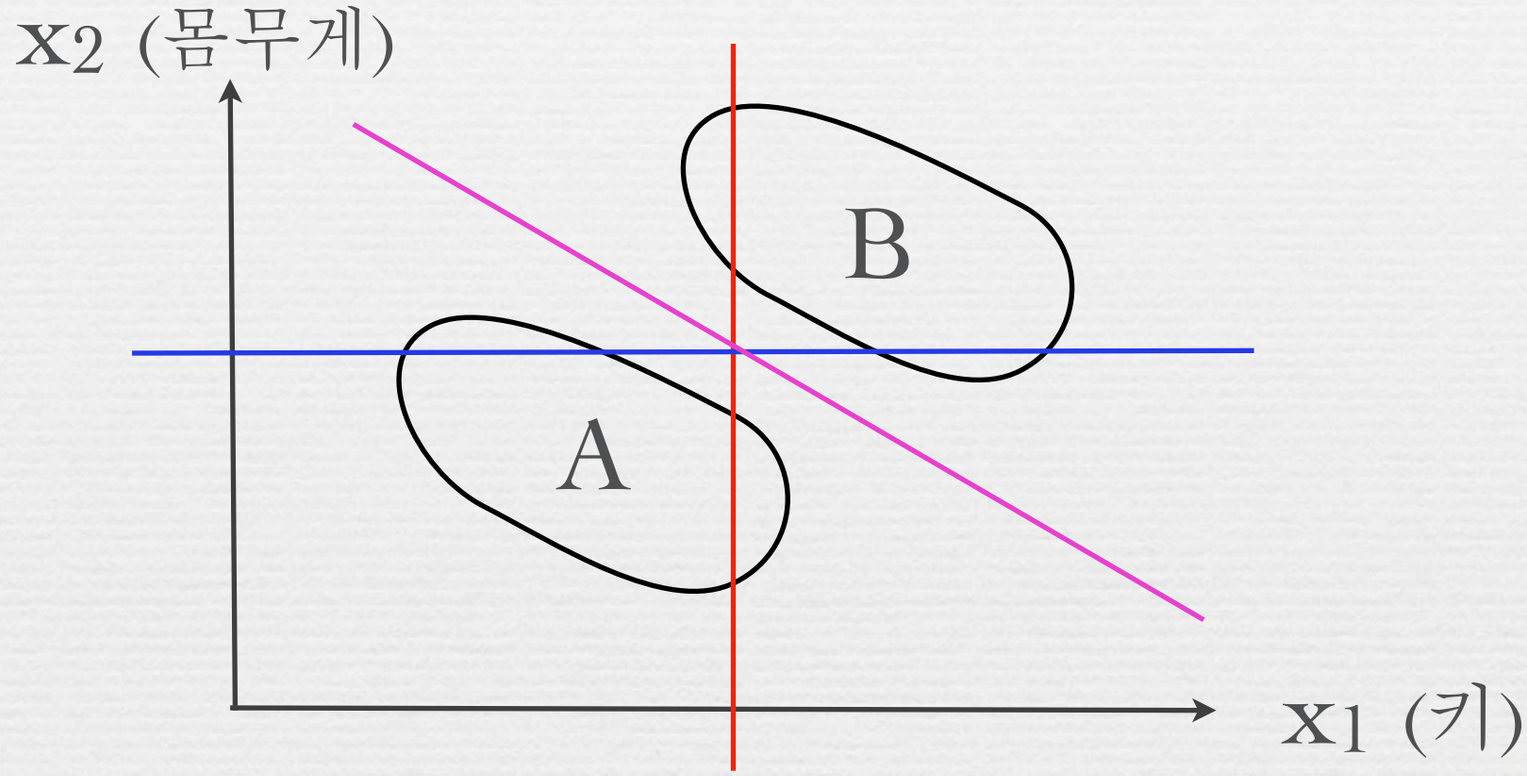
★ 3개월 연체(default) >> (y=1 or 0)

- 박사과정 학생들의 과거자료를 가지고 어떤 신입생이 5년이내에 학위를 마칠 수 있는지 판별
  - 신입생 정보(학부GPA, GRE, SAT)
  - 5년 이내 학위 >> (y=1 or 0)

?

판별함수(Discriminant Function)

- 어떤 개체가 두 개 이상의 그룹 중에 어디에 속하는지를 규명해 주는 함수



- x_1 (키) < c 이면 , x_1 (키) > c 이면
- x_2 (몸무게) < c 이면 , x_2 (몸무게)> c 이면
- $f(x_1, x_2) < c$ 이면 A 그룹, $f(x_1, x_2) > c$ 이면 B 그룹

Fisher의 판별함수

	Group 1 (Y=1)			
sub 1	x_{11}	x_{12}	...	x_{1p}
sub 2	x_{21}	x_{22}	...	x_{2p}
...	...	... x_{G1}	...	...
sub n_1	x_{n11}	x_{n12}	...	x_{n1p}

	Group 2 (Y=0)			
sub 1	x_{11}	x_{12}	...	x_{1p}
sub 2	x_{21}	x_{22}	...	x_{2p}
...	...	x_{G2}	...	...
sub n_2	x_{n21}	x_{n22}	...	x_{n2p}

Assume

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\{(n_1 - 1)S_1 + (n_2 - 1)S_2\}$$

$S_1 =$

$S_2 =$

Fisher의 판별함수

선형판별함수(linear discriminant function)

$$\boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{b} =$$

- Group 1 의 판별함수 평균값 ⟶ $E(\boldsymbol{x}_{G1}^T \boldsymbol{b}) =$
- Group 2 의 판별함수 평균값 ⟶ $E(\boldsymbol{x}_{G2}^T \boldsymbol{b}) =$
- 판별함수의 분산 ⟶ $V(\boldsymbol{y}) = V(\boldsymbol{x}^T \boldsymbol{b}) = V(\boldsymbol{b}^T \boldsymbol{x}) =$

- Ideal discriminant function ?

(1) 판별함수 평균 차이를 시키면서

(2) 판별함수 분산은 시키는 함수

Fisher의 판별함수

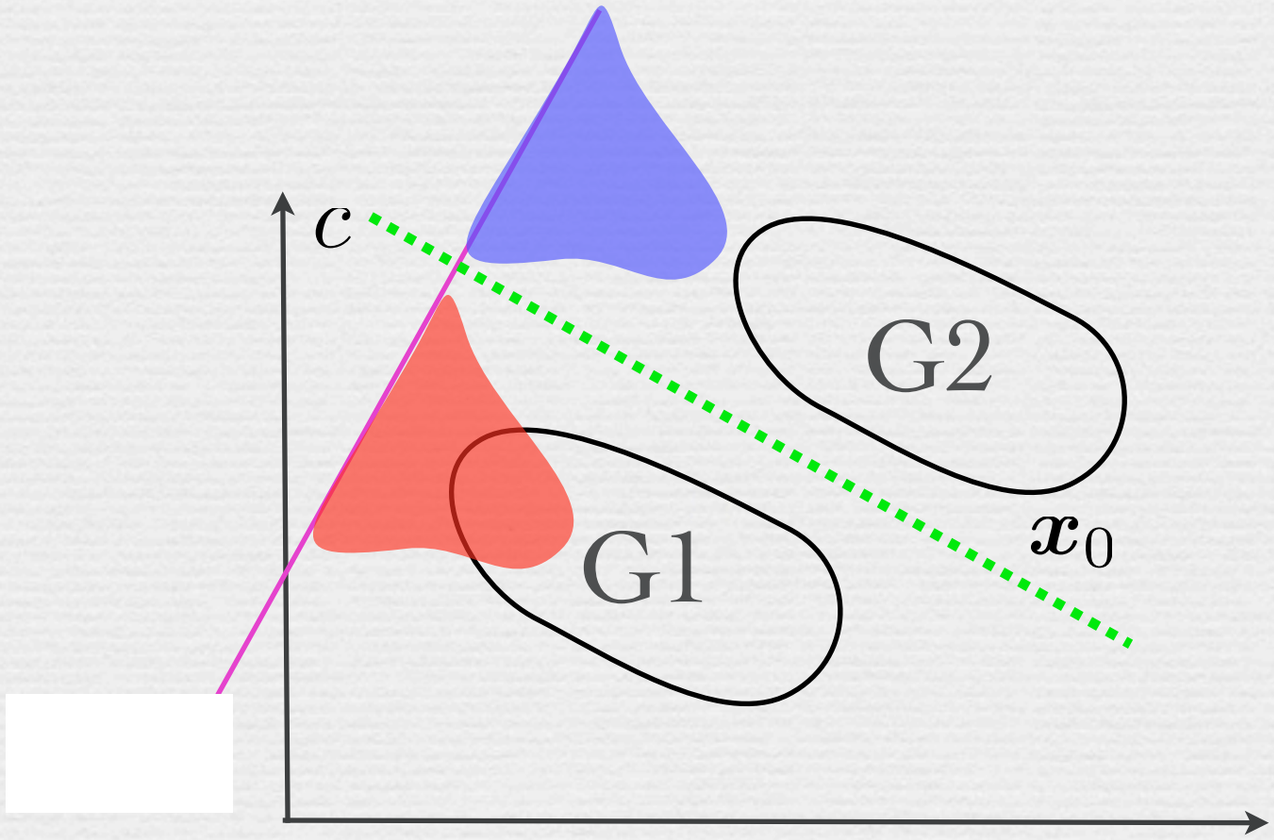
$$\max_{b}$$

R.A. Fisher

$$\hat{\boldsymbol{b}} =$$

Fisher 의 판별함수 :

$$y = \boldsymbol{x}^T \hat{\boldsymbol{b}} =$$



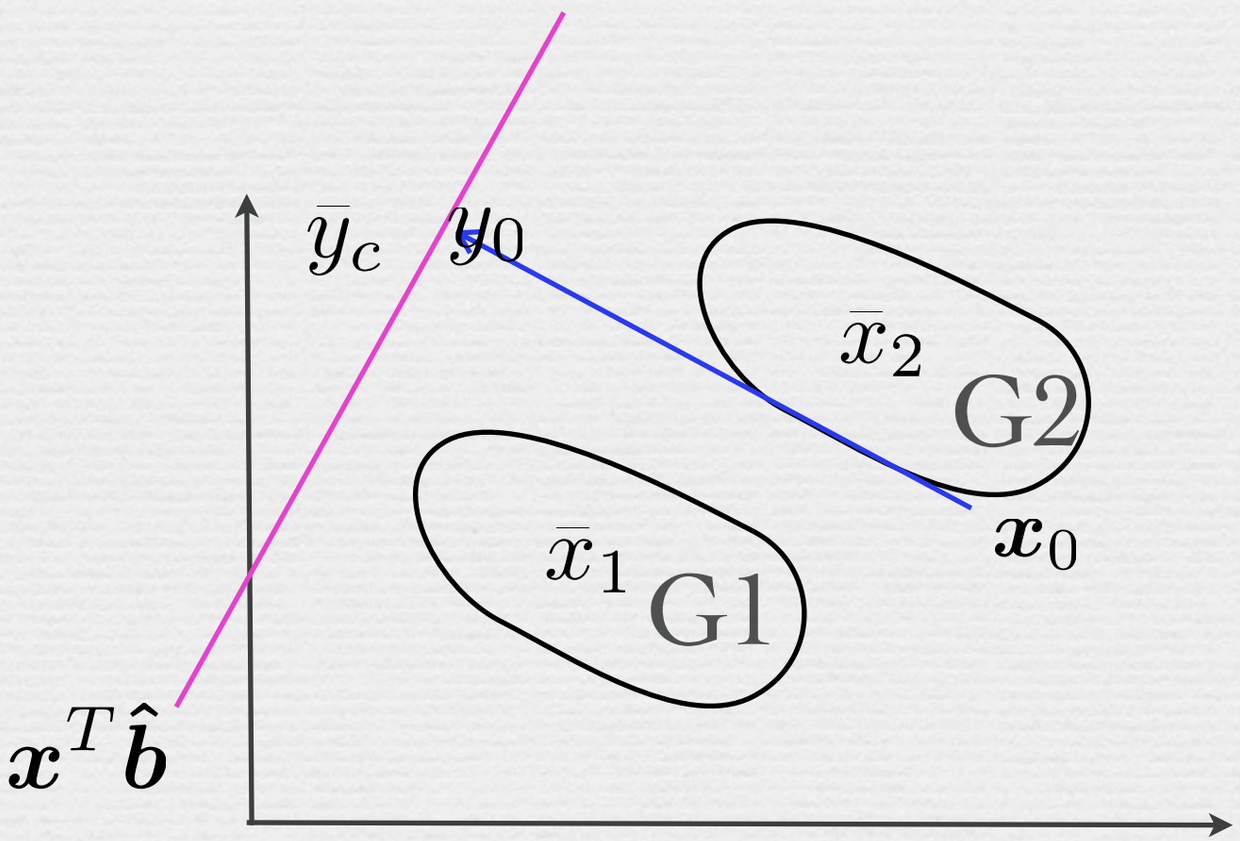
$$If \quad \boldsymbol{x}_0^T \hat{\boldsymbol{b}} > c$$

Cut-Off Value 의 결정

- Group 1 의 판별함수 평균값 $\bar{y}_1 = \bar{\boldsymbol{x}}_1^T \hat{\boldsymbol{b}} =$
- Group 2 의 판별함수 평균값 $\bar{y}_2 = \bar{\boldsymbol{x}}_2^T \hat{\boldsymbol{b}} =$
- Group 1 이 n_1 개, Group 2 가 n_2 개로 이루어진 경우

$$\bar{y}_c = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}$$

(가중평균: weighted average)



If $y_1 < y_2$,

$y_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{b}} > \bar{y}_c$ 라면

$\boldsymbol{x}_0 \in$

$y_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{b}} < \bar{y}_c$ 이면

$\boldsymbol{x}_0 \in$

PROC DISCRIM

- designed to classify data into
  (contrast to "cluster analysis")
- is required
- each obs. has a prob. belonging to a group

quad. discrim. ft.

linear discrim. ft.

```
proc discrim outstat=<..>  crosvalidate pool=<Yes/No> ;
        class <..>;
        var  <var1 var2 ... >;
        priors <equal/proportional>;
```

group variable

example: flour data

- RVA(Rapid Visco Analyzer) 의 6개 관측변수
- peak_visc, trough_visc, final_visc, breakdown, total_setback, timepeak_visc
- group cultivar = Ida Rose / Russet Burbank

Ida Rose

Russet Burbank

```
PROC DISCRIM CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=YES;
      CLASS CULTIVAR;
      VAR PEAK_VISC TROUGH_VISC FINAL_VISC BREAKDOWN TOTAL_SETBACK
            TIMEPEAK_VISC;
      PRIORS EQUAL;
```

The DISCRIM Procedure

Observations	450	DF Total	449
Variables	6	DF Within Classes	448
Classes	2	DF Between Classes	1

Class Level Information

Cultivar	Variable Name	Frequency	Weight	Proportion	Prior Probability
IR	IR	225	225.0000	0.500000	0.500000
RB	RB	225	225.0000	0.500000	0.500000

Linear Discriminant Function for Cultivar

Variable	Label	IR	RB
Constant		-17.80777	-10.84606
Peak_Visc	Peak_Visc	-1.54603	-10.51420
Trough_Visc	Trough_Visc	0.90295	0.57117
Final_Visc	Final_Visc	0.66918	9.96298
Breakdown	Breakdown	1.53603	10.52007
Total_Setback	Total_Setback	-0.60204	-9.93633
TimePeak_Visc	TimePeak_Visc	3.00518	2.26983

$\hat{\Sigma}^{-1}\bar{\boldsymbol{x}}_{IR}$ $\hat{\Sigma}^{-1}\bar{\boldsymbol{x}}_{RB}$

$$\boldsymbol{x}^T\hat{\boldsymbol{b}} = \boldsymbol{x}^T\hat{\Sigma}^{-1}(\bar{\boldsymbol{x}}_{RB} - \bar{\boldsymbol{x}}_{IR})$$

$$= (-10.51 + 1.54)PeakVisc + (0.57 - 0.9)TroughVisc + (9.96 - 0.66)FinalVisc$$
$$+(10.52 - 1.53)Breakdown + (-9.93 + 0.60)TotalSetback + (2.26 - 3.00)TimePeakVisc$$

$$c = -17.80 - (-10.84) = -6.96$$

$\boldsymbol{x}_0^T\hat{\boldsymbol{b}} > c$ →

$\boldsymbol{x}_0^T\hat{\boldsymbol{b}} < c$ →

misclassification rate

Number of Observations and Percent Classified into Cultivar

From Cultivar	IR	RB	Total
IR	190	35	225
	84.44	15.56	100.00
RB	15	210	225
	6.67	93.33	100.00
Total	205	245	450
	45.56	54.44	100.00
Priors	0.5	0.5	

Error Count Estimates for Cultivar

	IR	RB	Total
Rate	0.1556	0.0667	0.1111
Priors	0.5000	0.5000	

CROSSVALIDATION

- classification accuracy for each observation
- discrim. ft. is obtained by taking the obs. out of data

Number of Observations and Percent Classified into Cultivar

From Cultivar	IR	RB	Total
IR	188	37	225
	83.56	16.44	100.00
RB	16	209	225
	7.11	92.89	100.00
Total	204	246	450
	45.33	54.67	100.00
Priors	0.5	0.5	

Error Count Estimates for Cultivar

	IR	RB	Total
Rate	0.1644	0.0711	0.1178
Priors	0.5000	0.5000	

for new unknown data..

- use DIS_FUNC data set to build the discrim. ft.
- to classify the NEW data set
- TESTLIST provides each obs. with it classified value

```
PROC DISCRIM DATA=DIS_FUNC TESTDATA=NEW TESTLIST;
        CLASS CULTIVAR;
```