13강

분류분석 (1)

서강대학교 경영학부 이윤동교수

목 차

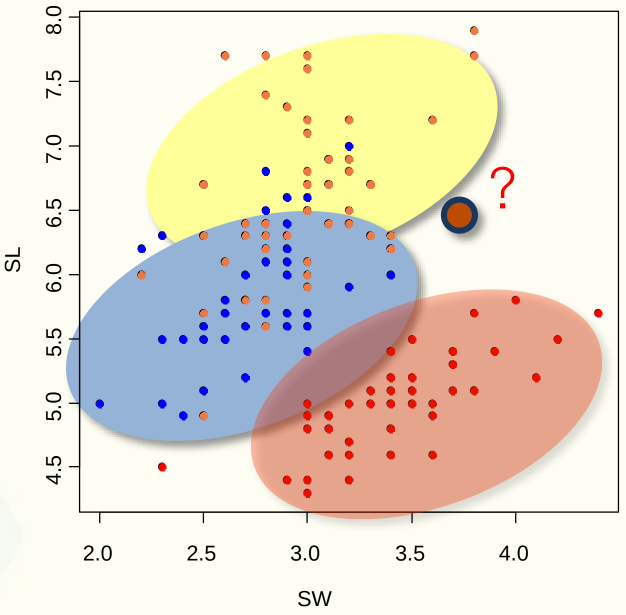
1 판별분석

붓꽃

setosa versicolor virginica

?

판별분석



판별분석

판별분석 (**D**iscriminant **A**nalysis) :

예 : $D= 0.83*SL+1.53*SW-2.2*PL-2.8*PW$

만약 $D> 0$ 이면 “세토사”

선형판별분석(**L**inear **DA**) : 판별식 D가 독립변수들에 대하여 선형

이차판별분석(**Q**uadratic **DA**) : 판별식 D가 독립변수들에 대하여 2차식

판별분석의 기본 틀

종속변수 : $y \in \{y_{(0)}, \ldots, y_{(K)}\}$ 예 : 붖꽃의 종류

독립변수 : $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ 예 : (꽃잎, 꽃받침)x(폭,길이)

$$\mathbf{x}|y_{(k)} \quad \sim \quad f(\mathbf{x}|y_{(k)}) \qquad k = 0, 1, \ldots, K$$

$$\hat{k} = \arg\max_k f(y_{(k)}|\mathbf{x}) = \arg\max_k \log f(y_{(k)}|\mathbf{x})$$

$$\log f(y_{(k)}|\mathbf{x}) \propto \log f(\mathbf{x}|y_{(k)}) + \log f(y_{(k)})$$ 베이즈 정리

$$\hat{y} = y_{(\hat{k})}$$

베이즈 정리

$$x|y \sim f(x|y) \quad \longrightarrow \quad y|x \sim f(y|x)$$

$$f(x,y) = f(x)\,f(y|x) = f(x|y)\,f(y)$$

$$f(x) = \sum_y f(x|y)\,f(y)$$

$$f(y|x) = \frac{f(x|y)\,f(y)}{f(x)}$$

$$f(y|x) \propto f(x|y)\,f(y) \qquad \log f(y|x) \propto \log f(x|y) + \log f(y)$$

정규분포 가정

$$\mathbf{x}|y_{(k)} \quad \sim \quad N_p\left(\mu_k, \Sigma_k\right) \qquad k = 0, 1, \ldots, K$$

$$\mu_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{kp})' \qquad \Sigma_k = \begin{pmatrix} \sigma_{k11} & \cdots & \sigma_{k1p} \\ \cdots & \cdots & \cdots \\ \sigma_{kp1} & \cdots & \sigma_{kpp} \end{pmatrix}$$

$$\begin{aligned} Q_k(\mathbf{x}) &= -2\ \log f(y_{(k)}|\mathbf{x}) \\ &= -2\ \log f(\mathbf{x}|y_{(k)}) - 2\log f(y_{(k)}) + constant \\ &= (\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) + \log|\Sigma_k| - 2\ \log f(y_{(k)}) \end{aligned}$$

LDA & QDA : 정규분포

QDA : $Q_k(\mathbf{x}) = (\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) + \log|\Sigma_k| - 2\,\log f(y_{(k)})$ ⬇

LDA : $\Sigma_k = \Sigma \qquad k = 0, 1, \ldots, K$

$$L_k = \mathbf{x}'\Sigma^{-1}\mu_k - (1/2)\mu_k'\Sigma^{-1}\mu_k + \log f(y_{(k)})$$ ⬆

LDA : 2 그룹

LDA : $\Sigma_k = \Sigma \qquad k = 0, 1, \ldots, K$

$$L_k = \mathbf{x}'\Sigma^{-1}\mu_k - (1/2)\mu_k'\Sigma^{-1}\mu_k + \log f(y_{(k)})$$

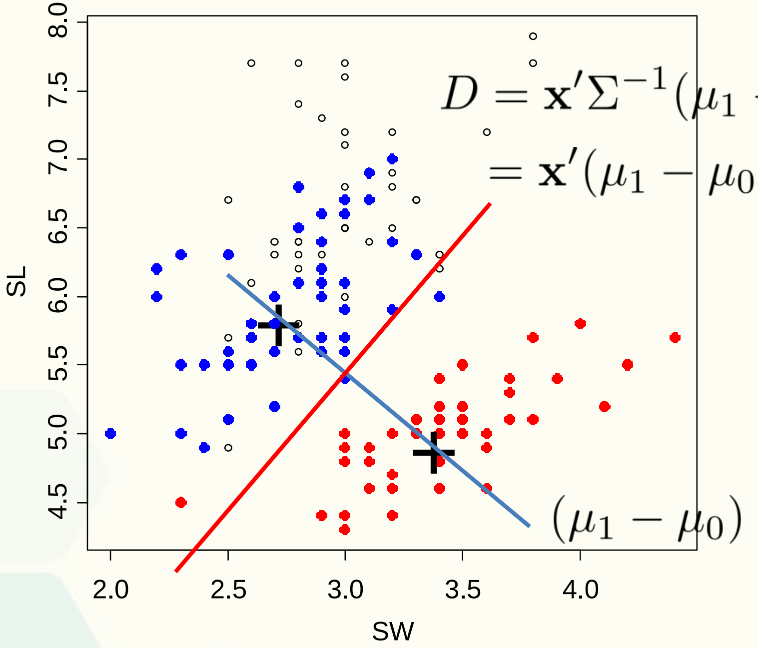
두 그룹의 경우 ($K = 1$) 선형판별식 :

$$D = L_1 - L_0 \boxed{= \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_0)} + constant$$

- $D > c_0 (= 0) \;\Rightarrow\; \hat{k} = 1$
- $D < c_0 (= 0) \;\Rightarrow\; \hat{k} = 0$

LDA : 2 그룹

단순화하여 $\Sigma = I$ 라고 가정하고,
setosa 와 versicolor 경우.



$$D = \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_0) + const$$
$$= \mathbf{x}'(\mu_1 - \mu_0) + const = 0$$

LDA : 다그룹

그룹내 분산 : $\Sigma = E\left[Var(\mathbf{x}|y)\right]$

그룹간 분산 : $\Sigma_b = Var\left[E(\mathbf{x}|y)\right]$

$D = w'\mathbf{x}$ 라 하면,

$$S = \frac{Var\left[E(w'\mathbf{x}|y)\right]}{E\left[Var(w'\mathbf{x}|y)\right]} = \frac{w'\Sigma_b w}{w'\Sigma w}$$

$$D_k = w_k'\mathbf{x} \qquad k = 1, \ldots, K$$

w_k : $\Sigma^{-1}\Sigma_b$ 의 고유벡터들(eigen vectors) .

LDA와 LRA

$$p_k = f(y_{(k)}|\mathbf{x}) \qquad k = 0, 1, \ldots, K$$

LRA : 로지스틱 회귀분석(**L**ogistic **R**egression **A**nalysis)

$$(y_{(0)}, y_{(1)}, \ldots, y_{(K)}) \sim Multinom(n, p_0, \ldots, p_K)$$

$$\log(p_k/p_0) = \mathbf{x}'\beta_k + \alpha_k$$

LDA : $\mathbf{x}|y_{(k)} \sim N_p(\mu_k, \Sigma)$ $\quad$ $\mathbf{x}$: 양적변수

$$\log(p_k/p_0) = \log\ f(\mathbf{x}|y_{(k)}) - \log\ f(\mathbf{x}|y_{(0)}) + const$$

$$= \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_0) + const$$

$$= \mathbf{x}'\beta_k + \alpha_k \qquad \beta_k = (\beta_{k1}, \ldots, \beta_{kp})'$$

2 판별분석의 적용 예

R의 lda 와 qda 함수

```
> library(MASS)
```

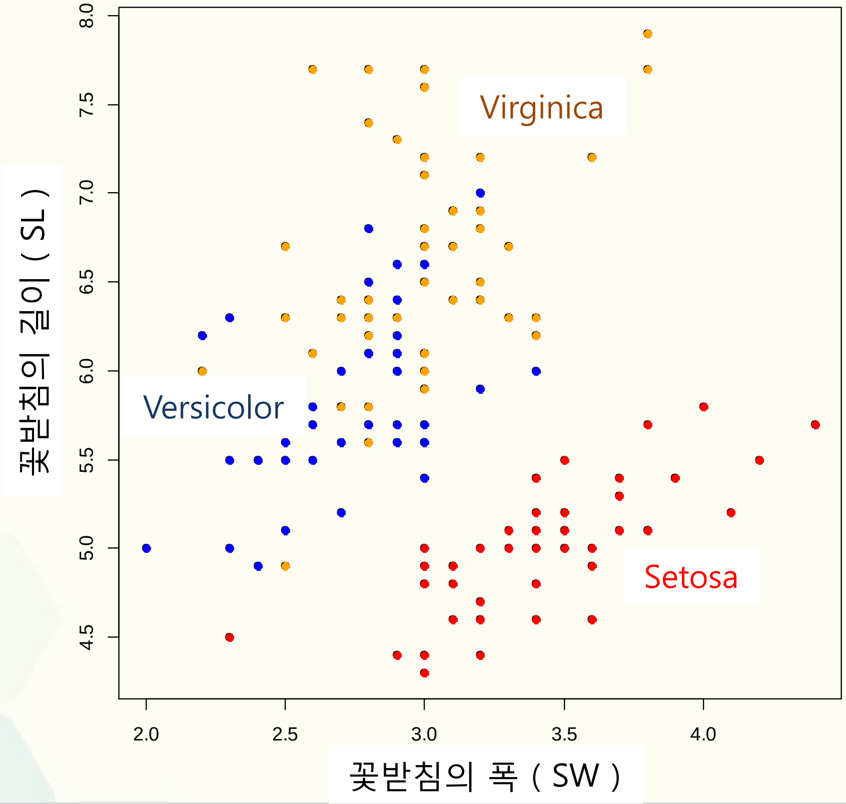
lda(formula, data, ...)

lda(x, grouping, …)

기타 주요 전달인자 :

- prior : 사전확률, 기정값은 “propotional” (표본수 비례)
- Method : “moment”, “mle”, “mve”, “t” 기정값은 “moment”
- CV : leave-one-out 상호검증에 의한 “사후확률”제공
  기정값은 FALSE

붓꽃 꽃받침의 폭과 길이


Virginica
Versicolor
Setosa
꽃받침의 길이 (SL)
꽃받침의 폭 (SW)
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
2.0
2.5
3.0
3.5
4.0

Prior

QDA : $Q_k(\mathbf{x}) = (\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) + \log|\Sigma_k| - 2\log f(y_{(k)})$

LDA : $\Sigma_k = \Sigma \qquad k = 0, 1, \ldots, K$

$$L_k = \mathbf{x}'\Sigma^{-1}\mu_k - (1/2)\mu_k'\Sigma^{-1}\mu_k + \log f(y_{(k)})$$

- $f(y_{(k)}) = 1/(K+1) \qquad k = 0, 1, \ldots, K$
- $f(y_{(k)}) = n_k/n \qquad k = 0, 1, \ldots, K$

2 그룹 판별의 경우

```
> newSP<-iris$SP
> newSP[101:150]<-iris$SP[51:100]
> newSP<- newSP[, drop=TRUE]
```

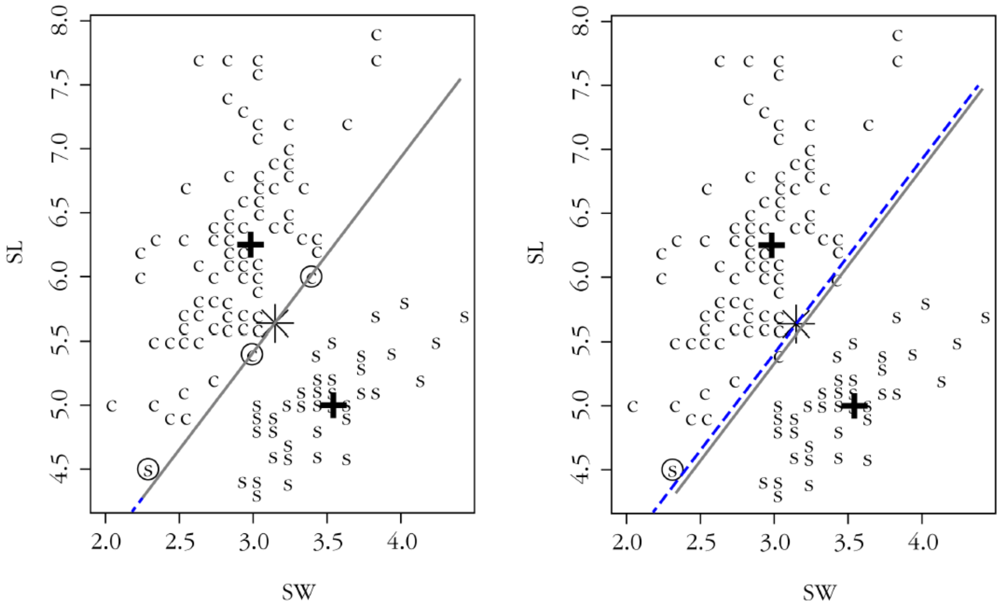
Versicolor , Virginica → “c”

Setosa → “s”

```
> library(MASS)
> (  iris2a<- lda( newSP~SW+SL, prior=c(1,1)/2, data=iris ) )
> (  iris2b<-lda( newSP~SW+SL, data=iris )  )

> par(mfrow=c(1,2))  # for black and white
> plot.iris.lda(iris2a, c("s","c"), newSP, line=1, border=1)
> plot.iris.lda(iris2b, c("s","c"), newSP, line=1, border=1)
```

plot.iris.lda() 함수는 부록에 수록되어 있음

Prior의 비교



[그림 7.1] 두 그룹 LDA 적용 예. (좌측은 사전확률 (0.5, 0.5), 우측은 사전확률 (1/3, 2/3). 잘못 분류된 결과들에 동그라미(O) 표시가 되어 있다.)

P. 275

LDA, QDA & LRA

```
> ( iris3a<- lda( SP~SW+SL,data=iris)  )          # LDA
>  iris3q<- qda( SP~SW+SL, data=iris)              # QDA
> ( iris3m<- multinom( SP~SW+SL,data=iris)  )   # LRA

>  par(mfrow=c(1,3))

>  plot.iris.lda(iris3a, border=1)
>  plot.iris.lda(iris3q,  border=1)
>  plot.iris.mnom(iris3m,  border=1)
```

plot.iris.mnom() 함수는 부록에 수록되어 있음

결과 : LDA & QDA

```
> lda(SP ~ SW + SL, data = iris)

Prior probabilities of groups:
       st        vc        vg
0.3333333 0.3333333 0.3333333

Group means:
      SW    SL
st 3.428 5.006
vc 2.770 5.936
vg 2.974 6.588

Coefficients of linear discriminants:
         LD1        LD2
SW  2.768109 -2.0960764
SL -2.141178 -0.8152721

Proportion of trace:
   LD1    LD2
0.9628 0.0372
```

```
> qda(SP ~ SW + SL, data = iris)

Prior probabilities of groups:
       st        vc        vg
0.3333333 0.3333333 0.3333333

Group means:
      SW    SL
st 3.428 5.006
vc 2.770 5.936
vg 2.974 6.588
```

결과 : LRA

```
> library(nnet)
> summary( multinom( SP ~ SW + SL, data = iris) )
```

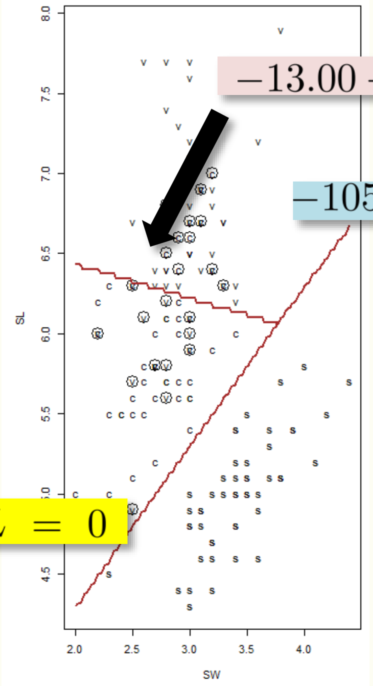
Coefficients:

	(Intercept)	SW	SL
vc	-92.09925	-40.58756	40.40327
vg	-105.10097	-40.18800	42.30095

Std. Errors:

	(Intercept)	SW	SL
vc	26.27830	27.77771	9.142715
vg	26.37025	27.78874	9.131118

Residual Deviance: 110.425
AIC: 122.425

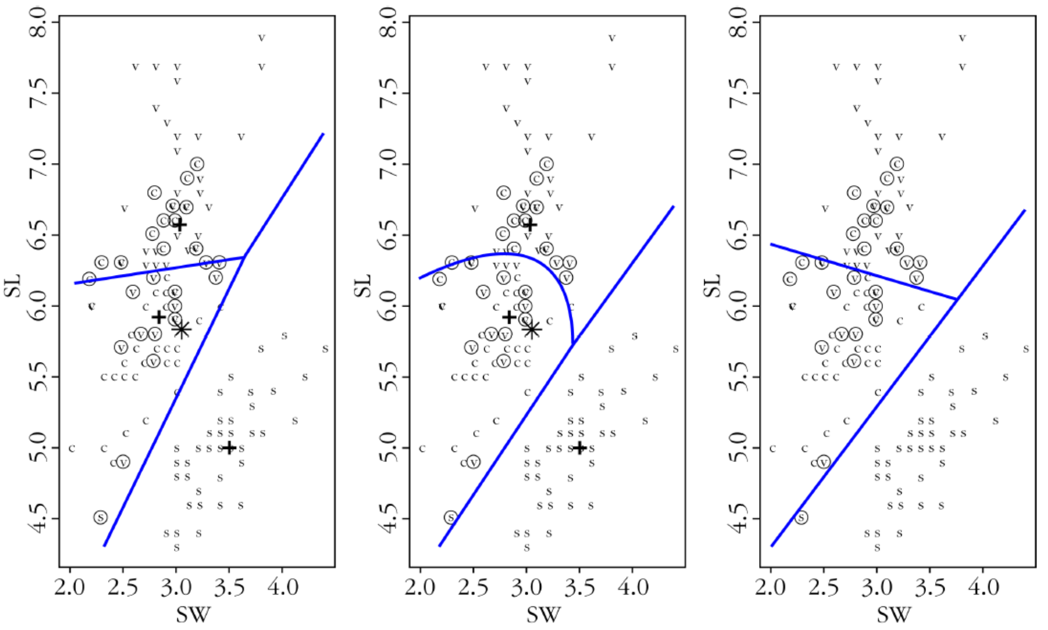


$$-13.00 + 0.40\,SW + 1.90\,SL = 0$$

$$-105.10 - 40.19\,SW + 42.30\,SL = 0$$

$$-92.10 - 40.59\,SW + 40.40\,SL = 0$$

LDA, QDA & LRA



[그림 7.2] LDA, QDA, LRA 를 이용한 붓꽃자료 세 그룹 분류 결과
(좌측은 LDA, 가운데는 QDA, 우측은 LRA 잘못 분류된
결과들에 동그라미(O) 표시가 되어 있다.)

LDA, QDA & LRA : 2변수

```
> table(iris$SP, predict(iris3a)$class )
> table(iris$SP, predict(iris3q)$class )
> table(iris$SP, predict(iris3m) )
```

〈표 7.1〉 LDA, QDA, LRA 의 분류 결과 비교

	LDA			QDA			LRA		
	st	vc	cg	st	vc	vg	st	vc	vg
st	49	1	0	49	1	0	50	0	0
vc	0	36	14	0	37	13	0	38	12
vg	1	15	35	0	16	34	0	13	37

LDA, QDA & LRA : 4변수

```
> iris4<-lda(SP~. , data=iris)           #  LDA
> iris4q<-qda(SP~. , data=iris)          #  QDA
> iris4m<-multinom(SP~., data=iris)      # Logistic

> table(iris$SP, predict(iris4)$class )
> table(iris$SP, predict(iris4q)$class )
> table(iris$SP, predict(iris4m) )
```

〈표 7.2〉 네 변수를 이용한 붓꽃자료 LDA, QDA, LRA 적용 결과

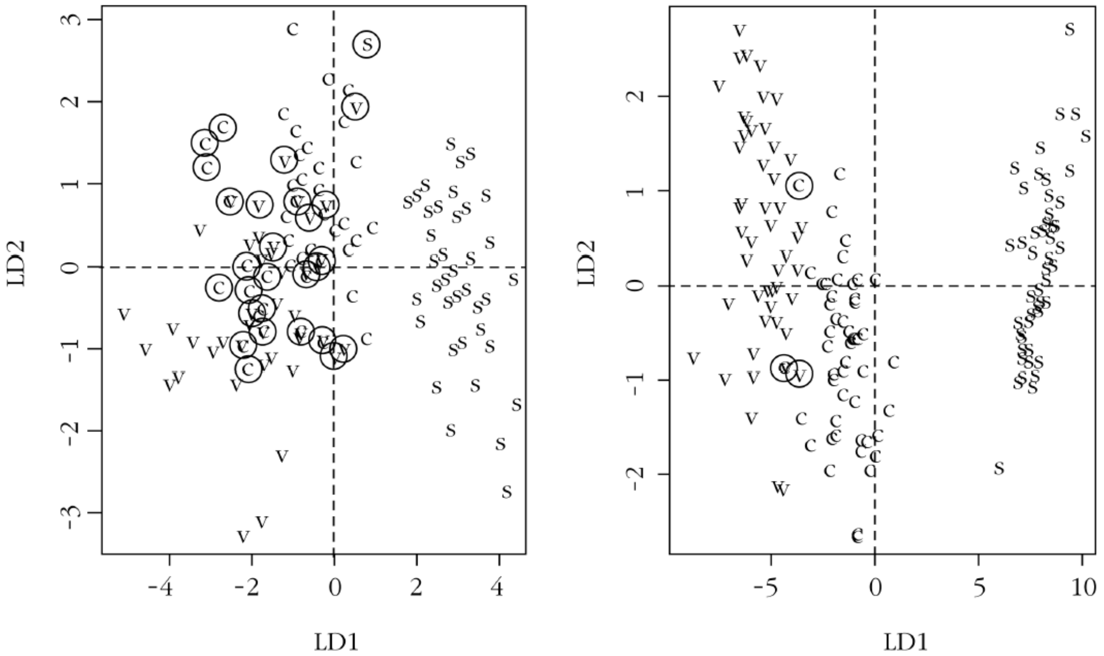
	LDA			QDA			LRA		
	st	vc	vg	st	vc	vg	st	vc	vg
st	50	0	0	50	1	0	50	0	0
vc	0	48	2	0	48	2	0	49	1
vg	0	1	49	0	1	49	0	1	49

결과 : LDA (2) & LDA (4)

```
> lda(SP ~ SW + SL, data = iris)

Prior probabilities of groups:
       st        vc        vg 
0.3333333 0.3333333 0.3333333 

Group means:
      SW    SL
st 3.428 5.006
vc 2.770 5.936
vg 2.974 6.588

Coefficients of linear discriminants:
         LD1        LD2
SW  2.768109 -2.0960764
SL -2.141178 -0.8152721

Proportion of trace:
   LD1    LD2 
0.9628 0.0372 
```

```
> lda(SP~. ,data=iris)

Prior probabilities of groups:
       st        vc        vg 
0.3333333 0.3333333 0.3333333 

Group means:
      SL    SW    PL    PW
st 5.006 3.428 1.462 0.246
vc 5.936 2.770 4.260 1.326
vg 6.588 2.974 5.552 2.026

Coefficients of linear discriminants:
          LD1         LD2
SL  0.8293776  0.02410215
SW  1.5344731  2.16452123
PL -2.2012117 -0.93192121
PW -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2 
0.9912 0.0088 
```

LDA : 2변수, 4변수



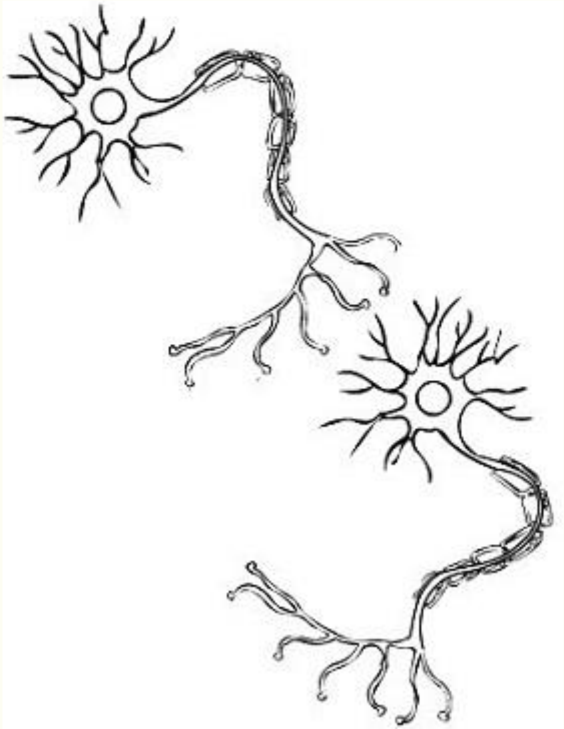
[그림 7.4] 붓꽃자료에 대한 LDA 결과 비교
(좌측 (SW, SL) 사용, 우측 (SW, SL, PW, PL) 사용 경우
잘못 분류된 결과들에 동그라미(O) 표시가 되어 있다)

```
> par(mfrow=c(1,2))

> plot.iris.discriminant(iris3a)
> plot.iris.discriminant(iris4)
```

plot.iris.discriminant() 함수는 부록에 수록되어 있음

3 신경망 분류방법

neuron & triggering



LRA의 그림표현 : 4변수

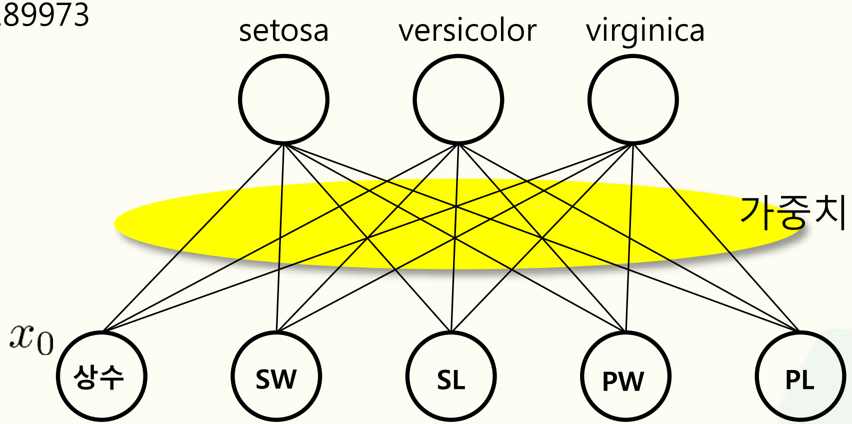
> multinom(SP ~ . , data = iris)

Coefficients:

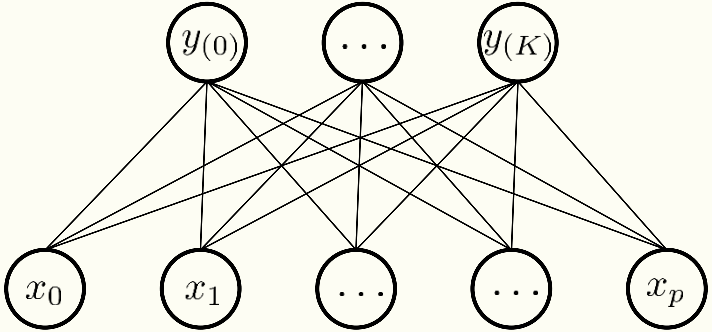
	(Intercept)	SL	SW	PL	PW
st	0	0	0	0	0
vc	18.69	-5.46	-8.70	14.24	-3.10
vg	-23.84	-7.92	-15.37	23.66	15.14

가중치 (weight : Wts)

Residual Deviance: 11.89973
AIC: 31.89973



LRA의 수리적 표현



$$y = (y_{(0)}, y_{(1)}, \ldots, y_{(K)})' \sim Multinom(1, p_0, p_1, \ldots, p_K)$$

$$x = (1, x_1, \ldots, x_p)' \qquad \beta_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})'$$

$$p_k = \frac{\exp(\beta_k' x)}{\sum_{k=0}^{K} \exp(\beta_k' x)} \qquad k = 0, 1, \ldots, K$$

Softmax 함수

은닉변수층이 있는 모형의 예

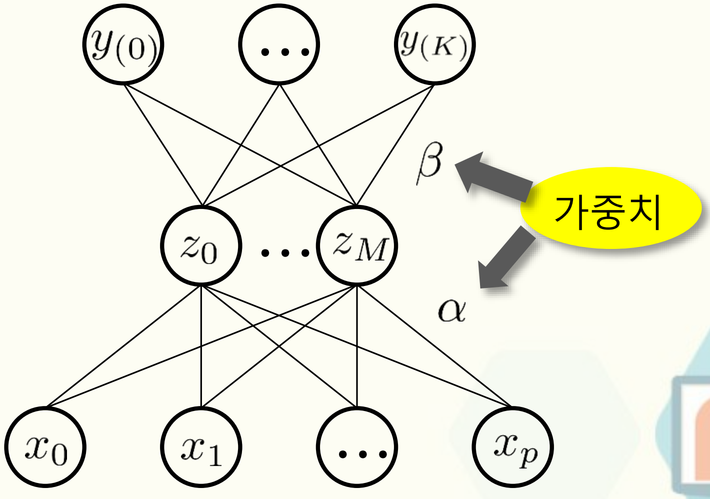
$$y = (y_{(0)}, y_{(1)}, \ldots, y_{(K)})' \sim Multinom(1, p_0, p_1, \ldots, p_K)$$

$$z_m \sim Bin(1, q_m) \qquad m = 0, 1, \ldots, M$$

$$p_k = \frac{\exp(\beta_k' q)}{\sum_{k=0}^{K} \exp(\beta_k' q)} \qquad k = 0, 1, \ldots, K$$

$$q_m = \frac{1}{1 + \exp(\alpha_m' x)} \qquad m = 0, 1, \ldots, M$$

Sigmoid 함수



모형적합 방법

$$R(\alpha,\beta) + \lambda J(\alpha,\beta)$$

$$y_i \in \{y_{(0)}, \ldots, y_{(K)}\} \qquad x_{i\cdot} = (x_{i0}, \ldots, x_{ip})' \qquad i = 1, \ldots, n$$

$$y_{i(k)} = I(y_i = y_{(k)})$$

적합결여도 :

$$R(\alpha,\beta) = \sum_{i=1}^{n}\sum_{k=0}^{K}(y_i - f_k(x_{i\cdot}))^2$$ 잔차제곱합(회귀목적)

$$R(\alpha,\beta) = -\sum_{i=1}^{n}\sum_{k=0}^{K} y_{i(k)} \log p_k(x_{i\cdot})$$ 엔트로피(분류목적)

거칠기 벌칙 : (모형의 단순성 추구)

$$J(\alpha,\beta) = \sum_{k,m}\beta_{km}^2 + \sum_{mi}\alpha_{mi}^2$$ 제곱합벌칙

$$J(\alpha,\beta) = \sum_{k,m}\frac{\beta_{km}^2}{1+\beta_{km}^2} + \sum_{mi}\frac{\alpha_{mi}^2}{1+\alpha_{mi}^2}$$ 소거벌칙

4 신경망 분류방법의 적용 예

R의 nnet 함수

```
> library(nnet)
```

```
nnet(formula, data, ...)

nnet(x, y,...)
```

기타 주요 전달인자 :

- size : 히든노드의 수
- Wts : 추정해야 할 각 weight 의 초기값.
- rang : 초기 weight 값이 없는 경우 배정할 확률난수 범위
- decay : 모형적합에서 사용할 거칠기 벌칙항에 대한 가중값
- maxit : 최적화 과정의 최대 허용 반복 횟수
- weights : 자료의 가중치

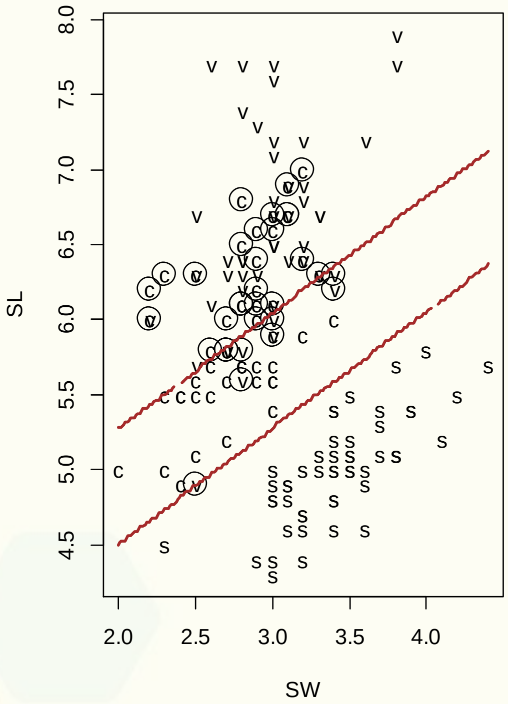
class.ind & max.col

```
> ( smp<- iris$SP[sample(1:150,9)] )
[1] vg st vg vc vc vg vg st st
Levels: st vc vg
> as.integer(smp)
[1] 3 1 3 2 2 3 3 1 1
> class.ind(smp)
      st vc vg
 [1,]  0  0  1
 [2,]  1  0  0
 [3,]  0  0  1
 [4,]  0  1  0
 ...
 [9,]  1  0  0
> max.col(class.ind(smp))
[1] 3 1 3 2 2 3 3 1 1
```

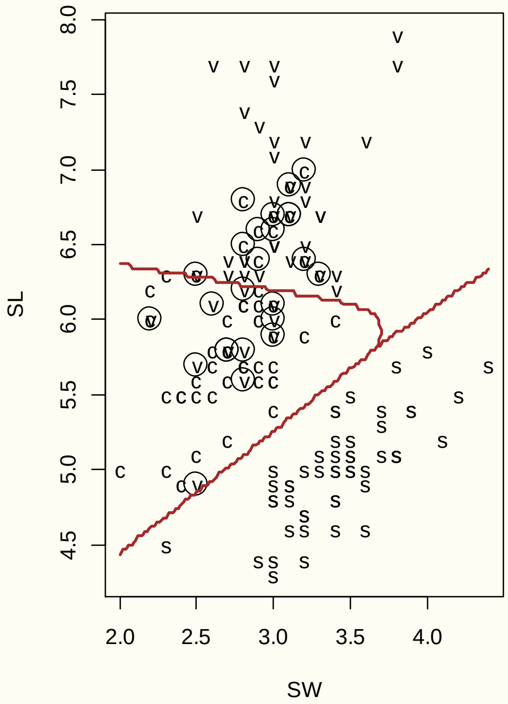
신경망

```
> irx <- iris[,-5]
> spc <- class.ind( iris$SP )
> irx3n1 <- nnet(spc~ SW+SL , data=irx, size = 1, rang = 0.1, decay = 5e-4, maxit = 300 )

> irx3n2 <- nnet(spc~ SW+SL , data=irx, size = 2, rang = 0.1, decay = 5e-4, maxit = 300 )
# weights:  15
initial  value 113.276176
iter  10 value 84.446863
iter 240 value 37.693508
final  value 37.692786
converged


> par(mfrow=c(1,2))
> plot.iris.mnom(irx3n1)
> plot.iris.mnom(irx3n2)
```

히든노드의 수에 따른 신경망 적합결과



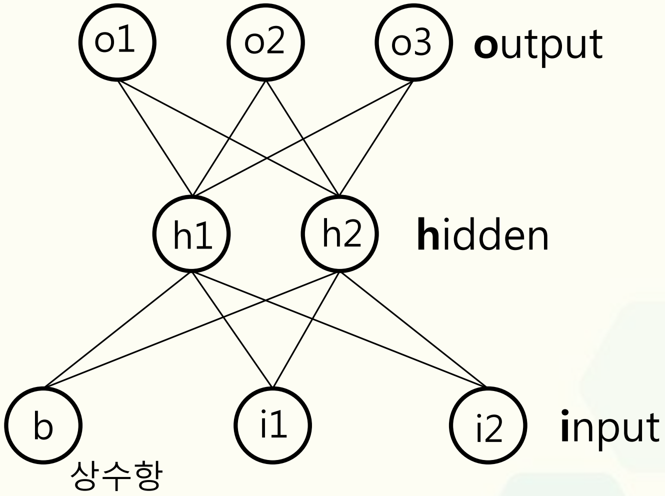
히든노드 수 1개



히든노드 수 2개

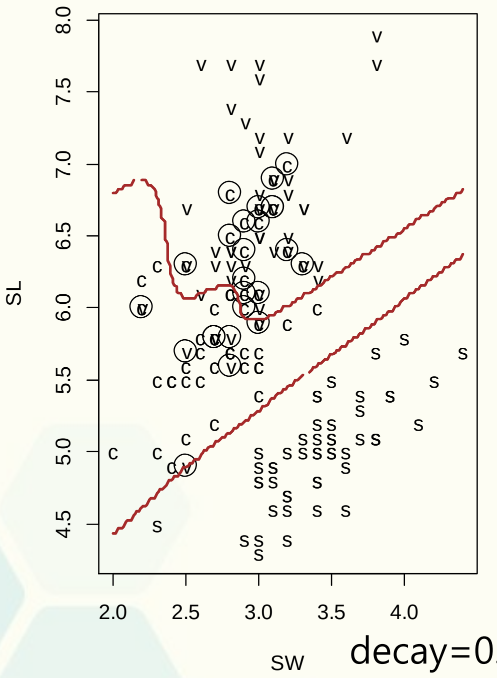
가중치의 추정

```
> summary(irx3n2)

a 2-2-3 network with 15 weights
options were - decay=5e-04

 b->h1   i1->h1    i2->h1
24.77    6.25      -8.41
 b->h2   i1->h2    i2->h2
  6.93   -0.18     -1.15
 b->o1   h1->o1    h2->o1
-4.05    13.11     -3.04
 b->o2   h1->o2    h2->o2
-2.85    -11.26     8.82
 b->o3   h1->o3    h2->o3
  2.73   -6.21     -8.26
```

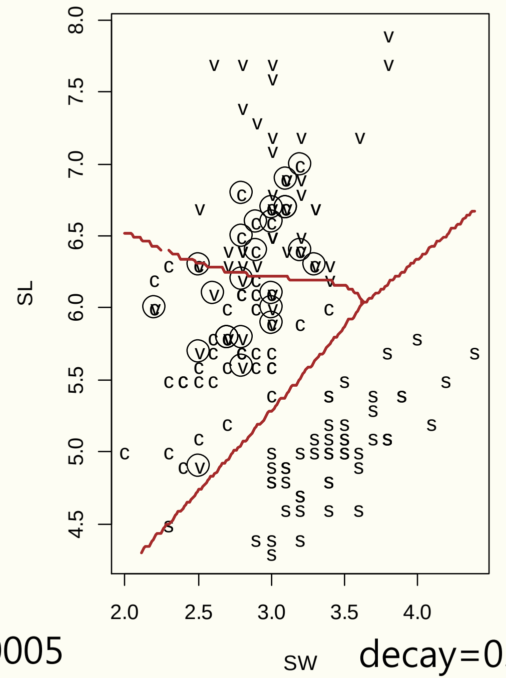


과적합과 벌칙항

> irx3n5 <- nnet(spc~ SW+SL , data=irx, **size = 5**, rang = 0.1, **decay = 5e-4**, maxit = 300)
> irx3n6 <- nnet(spc~ SW+SL , data=irx, **size = 5**, rang = 0.1, **decay = 0.01**, maxit = 300)



decay=0.0005



decay=0.01

히든노드 수 5개

$$R(\alpha, \beta) + \lambda J(\alpha, \beta)$$

decay

다음시간 안내

분류분석 (2)