14강

분류분석 (2)

서강대학교 경영학부 이윤동교수

목 차

1 나무 방법

나무 방법

종속변수 : $y \in \{y_{(0)}, \ldots, y_{(K)}\}$ **독립변수 :** $x = (x_1, x_2, \ldots, x_p)'$

훈련자료 : $y_i \in \{y_{(0)}, \ldots, y_{(K)}\}$ $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$ $i = 1, \ldots, n$

$$R_m = \Pi_{j=1}^{p}\{x_j | l_j < x_j < u_j\} \quad m = 1, \ldots, M \qquad R = \cup_{m=1}^{M} R_m$$

회귀나무 : $\hat{f}(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$ $c_m = average\{y_i | x_{i\cdot} \in R_m\}$

분류나무 : $\hat{k} = \arg\max_k p_{mk}$ $p_{mk} = \frac{1}{n_m} \sum_{x_{i\cdot} \in R_m} I(y_i = y_{(k)})$

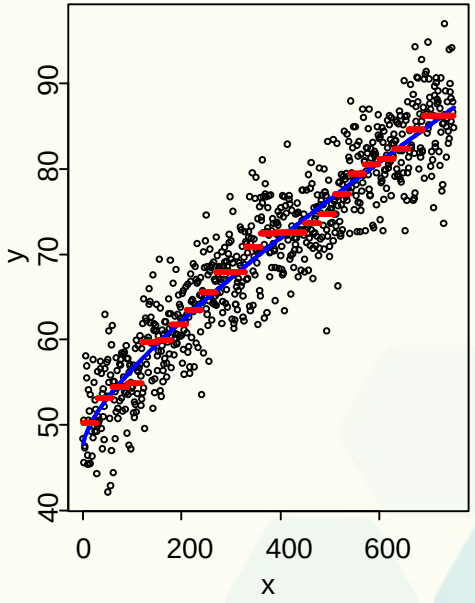
회귀나무 방법의 예

생후 경과 일수(x) 따른 신생아의 신장(y) 예측 모형

회귀분석 : 곡선 혹은 직선 추정

회귀분석 : x와 y 사이의 관계,
즉 곡선 혹은 직선을 구하여 예측.

회귀나무 방법 : 적당한 시간 단위별
(예 : 월별)로 신장의 평균을 구하여 예측.



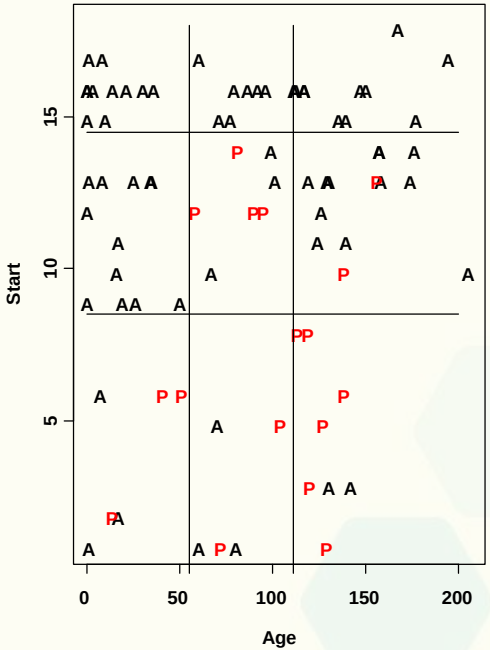
분류나무 방법의 예

환자 연령(Age) 질병 발생 부위(Start) 등 기타 여러 원인변수에 따라
후유증의 존재 여부(A 혹은 P)를 분류.

분류나무 방법 :

(1) 여러 변수 중 적절한 변수 선택
(2) 선택된 각 변수들을 구간 분할

특정 구간 분할된 영역 내에서의 후유증 존재 여부가 동일해지도록 하여,
각 구간을 후유증 유무 기준으로 판단.



불순도 척도

불순도척도(impurity measure) : 분할된 영역 R_m 내에서 y 값의 다양성을 말하는 척도. 일종의 적합결여도.

회귀나무 : $I_m = \sum_{x_i \in R_m} (y_i - c_m)^2$

분류나무 :

$I_m = 1 - p_{m\hat{k}}$ (오분류 오차)

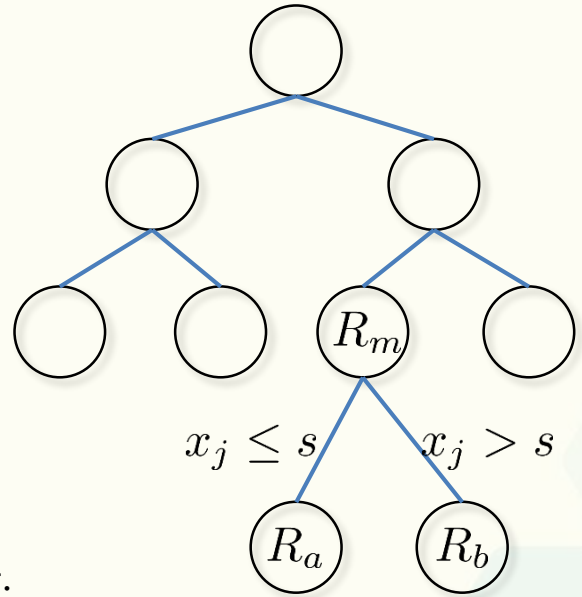
$I_m = \sum_{k=0}^{K} p_{m\hat{k}}(1 - p_{m\hat{k}})$ (지니계수)

$I_m = -\sum_{k=0}^{K} p_{m\hat{k}} \log p_{m\hat{k}}$ (크로스 엔트로피, 이탈도)

가지분할 방법

영역분할 과정의 말단노드 R_m 을 다시 분할하기 위하여 분할변수 x_j 를 선택하고, 분할값 s 를 선택하여, $x_j \leq s$ 인 조건과 $x_j > s$ 인 조건을 부과하여 새로운 말단노드 R_a 와 R_b 로 분할한다.

이때 불순도척도 $I_a + I_b$ 를 최소화하는 x_j 와 s 를 선택한다.



가지치기

복잡도기준(complexity criterion) : 어떤 나무 T 의 말단노드의 개수를 M 이라고 할 때, 적당한 복잡도모수 α 에 대하여,

cp

$$c_\alpha(T) = \sum_{m=1}^{M} I_m + \alpha M$$

을 말하는 것으로, AIC 와 같은 일종의 모형선택 기준이다.

- 불순도척도가 비슷한 경우에 말단노드의 수가 적을수록 더 낮은 복잡도기준을 갖게 된다.
- 복잡도기준을 최소화 하는 방법으로 **말단노드의 수가 가능한 적은 모형을 선택**한다.

2 나무방법의 적용 예

R의 rpart 함수

> library(rpart)

rpart(formula, data, ...)

기타 주요 전달인자 :

- method : 분류는 "class", 회귀는 "anova", "poisson", "exp"
- parms, control : 나무 방법의 기타 상세 사항 설정을 위한 선택사항.
- cp : 복잡도모수. rpart.control 모수 중 하나. **기정값은 0.01**

회귀나무의 예 : 붓꽃자료

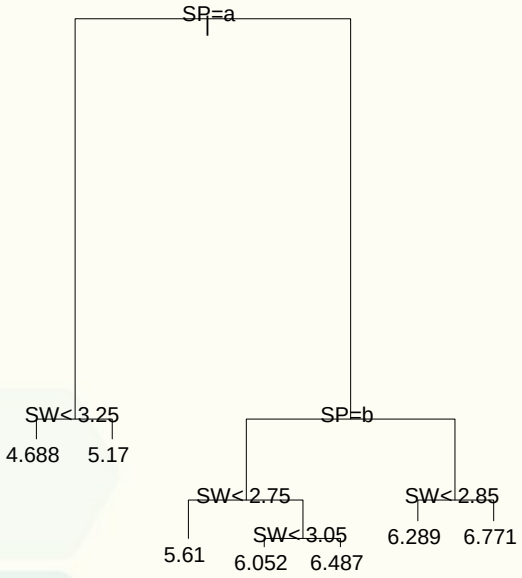
종속변수 : 꽃받침의 길이(**SL**) **독립변수** : 붓꽃의 종류(SP), 꽃받침의 폭(SW)

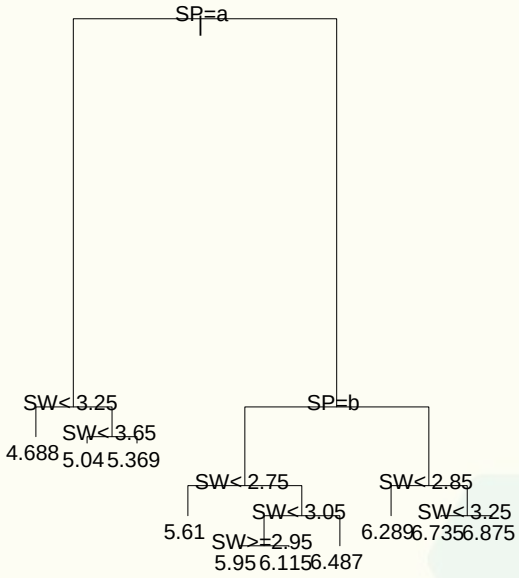
```
> library(rpart)

> xrp<-expand.grid(SP=unique(iris$SP), SW=swq)
> iris7t2<-rpart(SL~SP+SW, cp=1/1000, data=iris)
> ( iris7t1<-rpart(SL~SP+SW, data=iris) )  # cp=0.01
n= 150
node), split, n, deviance, yval  * denotes terminal node
 1) root 150      102.1683000 5.843333
   2) SP=st 50    6.0882000     5.006000
     4) SW< 3.25 17      0.8176471    4.688235 *
     5) SW>=3.25 33     2.6696970    5.169697 *
   3) SP=vc,vg 100    43.4956000      6.262000
      ...
     7) SP=vg 50   19.8128000 6.588000
      14) SW< 2.85 19      9.0178950    6.289474 *
      15) SW>=2.85 31     8.0638710    6.770968 *
```

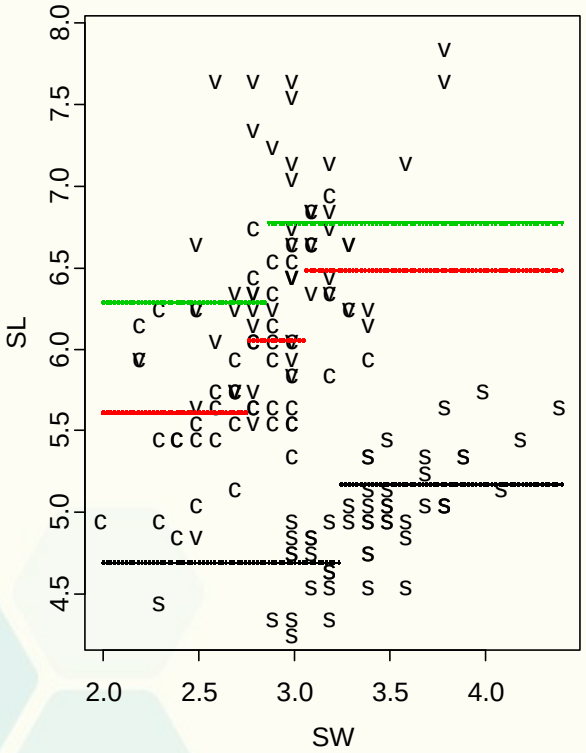
회귀나무의 예 : 붓꽃자료

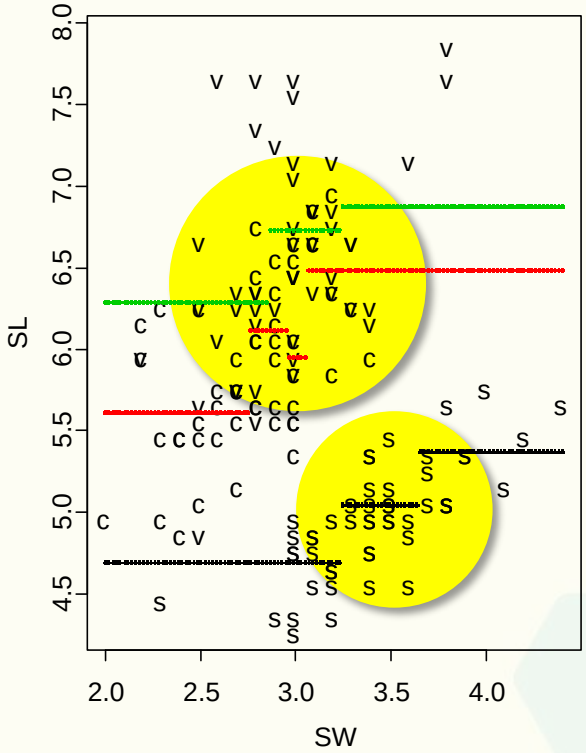
```
> plot(iris7t1); text(iris7t1)   # cp=0.01
> plot(iris7t2); text(iris7t2)   # cp=0.001
```





회귀나무의 예 : 붓꽃자료


SL
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


SL
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

분류나무의 예 : 척추후만증자료

```
> library(rpart)
> rbind(head(kyphosis),tail(kyphosis))
   Kyphosis Age Number Start
1    absent  71      3     5
2    absent 158      3    14
....
80  present  42      7     6
81   absent  36      4    13
```

Kyphosis	수술 후에 기형 증상이 있는지 (present) 하는지 없는지(absent)를 나타내는 요인 변수
Age	월 단위 나이
Number	문제가 된 척추의 개수
Start	문제가 된 맨 윗 쪽 척추번호

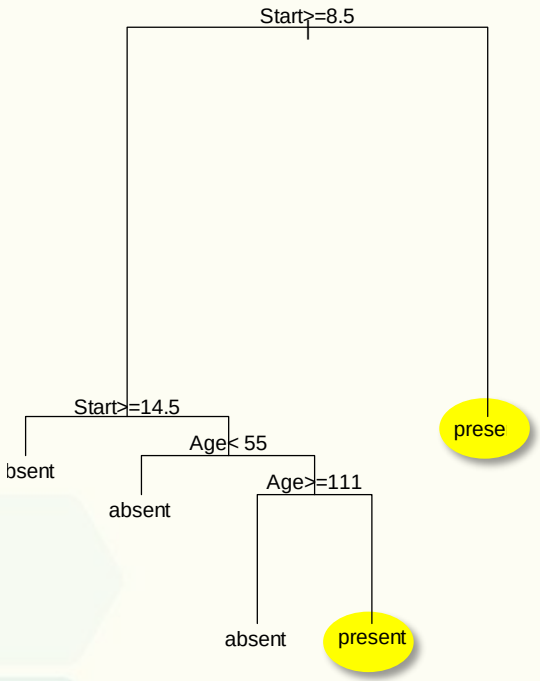
분류나무의 예 : 척추후만증자료

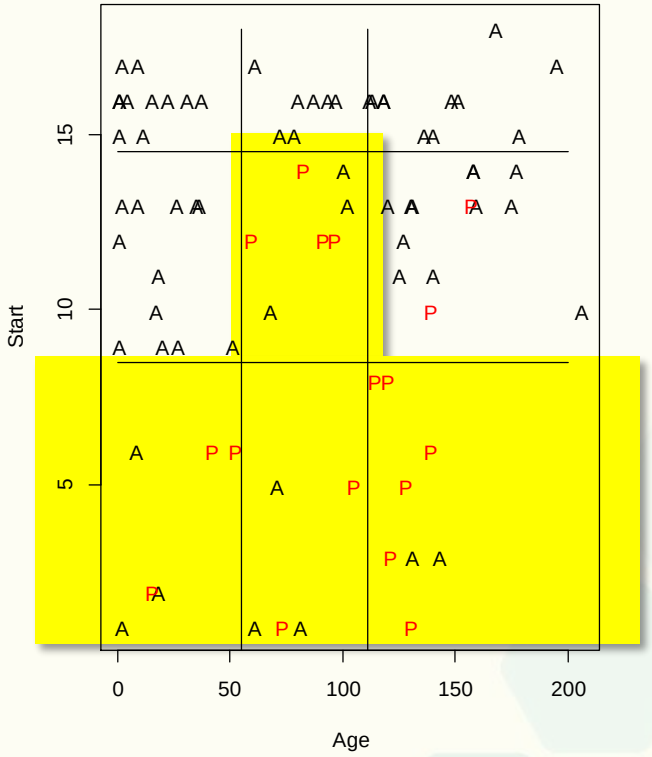
```
> ( fit <- rpart(Kyphosis ~ Age + Number + Start, data = kyphosis) )
n= 81

node), split, n, loss, yval, (yprob)      * denotes terminal node

 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=111 14  2 absent (0.85714286 0.14285714) *
        23) Age< 111 7  3 present (0.42857143 0.57142857) *
   3) Start< 8.5 19  8 present (0.42105263 0.57894737) *
```

분류나무의 예 : 척추후만증자료





분류나무의 예 : 붓꽃자료 – 2변수

종속변수 : 붓꽃의 종류(**SP**), **독립변수** : 꽃받침의 길이(SL), 꽃받침의 폭(SW)

```
> ( iris7t3<-rpart(SP~SW+SL, data=iris) )

n= 150
node), split, n, loss, yval, (yprob)  * denotes terminal node

 1) root 150 100 st (0.33333333 0.33333333 0.33333333)
   2) SL< 5.45 52   7 st (0.86538462 0.11538462 0.01923077)
     4) SW>=2.8 45   1 st (0.97777778 0.02222222 0.00000000) *
     5) SW< 2.8 7   2 vc (0.14285714 0.71428571 0.14285714) *
   3) SL>=5.45 98  49 vg (0.05102041 0.44897959 0.50000000)
     6) SL< 6.15 43  15 vc (0.11627907 0.65116279 0.23255814)
      12) SW>=3.1 7   2 st (0.71428571 0.28571429 0.00000000) *
      13) SW< 3.1 36  10 vc (0.00000000 0.72222222 0.27777778) *
     7) SL>=6.15 55  16 vg (0.00000000 0.29090909 0.70909091) *
```

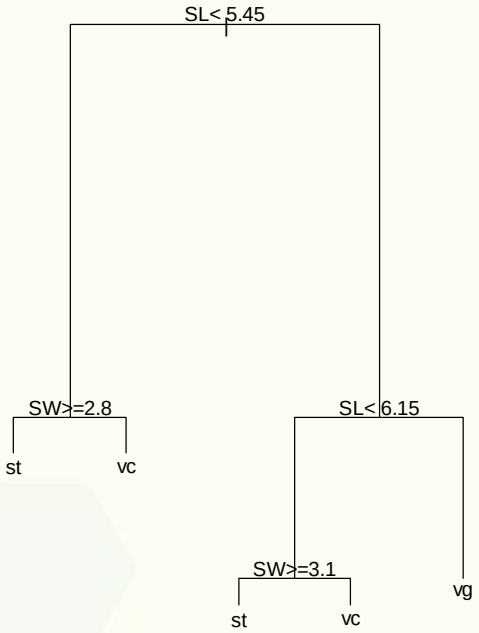
분류나무의 예 : 붓꽃자료 – 2변수

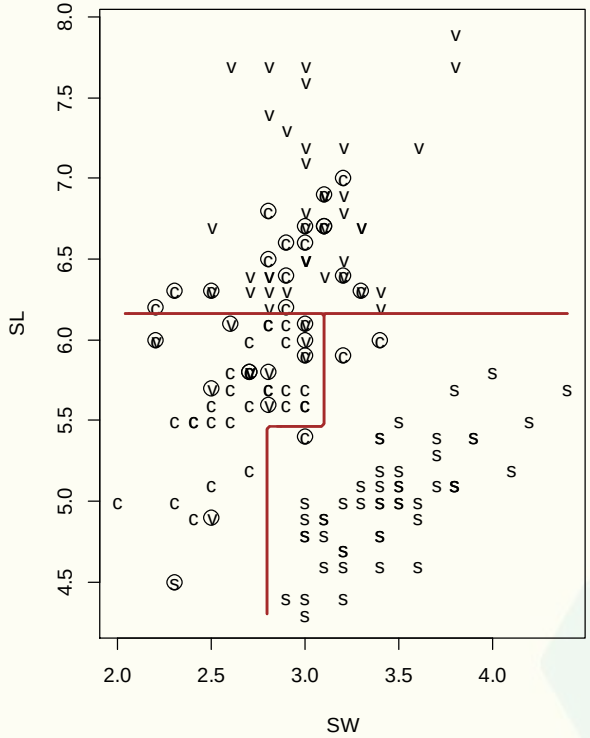
```
> plot(iris7t3); text(iris7t3)

> with(iris, plot(SW,SL, pch=c('s','c','v')[SP] ))
> PSP=unique(iris$SP)[max.col(predict(iris7t3))]
> with(iris[PSP!=iris$SP,], points(SW, SL, cex=1.8 ))

> prd01<- class.ind(max.col(predict(iris7t3, xr)))
> apply(prd01, 2, function(x) contour(swq,slq, matrix(x,120,),
        levels=0.5, lwd=2, col="brown", labels="", add=TRUE))
```

분류나무의 예 : 붓꽃자료 – 2변수





분류나무의 예 : 붓꽃자료 – 4변수

종속변수 : 붓꽃의 종류(**SP**),

독립변수 : 꽃받침의 길이(**SL**)와 폭(**SW**); 꽃잎의 길이(**PL**)와 폭(**PW**)

```
> ( iris7t4<-rpart(SP~., data=iris) )

n= 150

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 150 100 st (0.33333333 0.33333333 0.33333333)
  2) PL< 2.45 50   0 st (1.00000000 0.00000000 0.00000000) *
  3) PL>=2.45 100  50 vc (0.00000000 0.50000000 0.50000000)
    6) PW< 1.75 54   5 vc (0.00000000 0.90740741 0.09259259) *
    7) PW>=1.75 46   1 vg (0.00000000 0.02173913 0.97826087) *
```

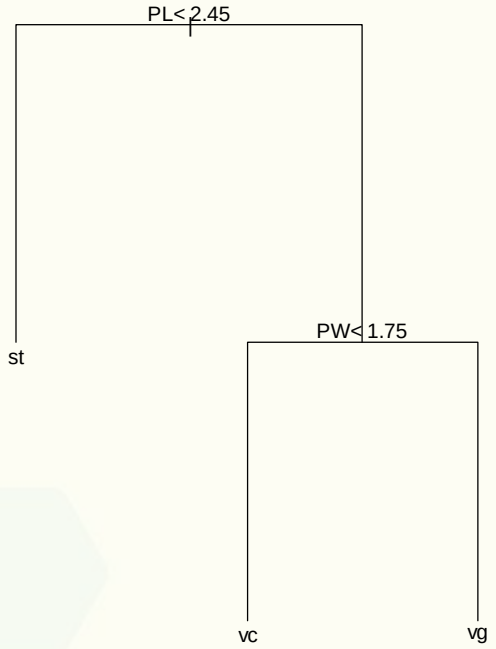
분류나무의 예 : 붓꽃자료 – 4변수

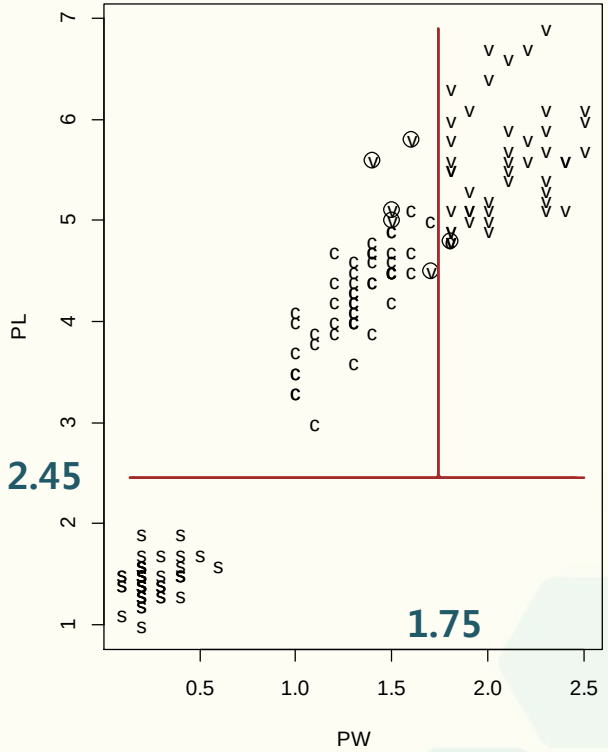
```
> plot(iris7t4); text(iris7t4)

> with(iris, plot(PW,PL, pch=c('s','c','v')[SP] ))

> with(iris[PSP!=iris$SP,], points(PW, PL, cex=1.8 ))
> prd01<- class.ind(max.col(predict(iris7t4, xr2)))
> apply(prd01, 2, function(x) contour(pwq,plq, matrix(x,120,),
+       levels=0.5, lwd=2, col="brown", labels="", add=TRUE))
```

분류나무의 예 : 붓꽃자료 – 4변수





분류나무의 예 : 붓꽃자료 – 4변수

```
> PSP=unique(iris$SP)[max.col(predict(iris7t4))]
> table(iris$SP,PSP)
    PSP
     st vc vg
  st 50  0  0
  vc  0 49  1
  vg  0  5 45
```

〈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

3 서포트벡터 기계

서포트벡터 기계

서포트벡터(support vector) : 받침점, 지탱점

서포트벡터 기계(**S**upport **V**ector **M**achine) : 자료전체를 이용한 분류방법이 아닌 각 그룹의 경계 획정에 영향을 미치는 서포트 벡터들을 이용한 분류방법

- 서포트벡터 기계:

$$f(x) = h(x)'\beta + \beta_0$$

$$x = (x_1, \ldots, x_p)'$$

$$h(x) = (h_1(x), \ldots, h_q(x))'$$

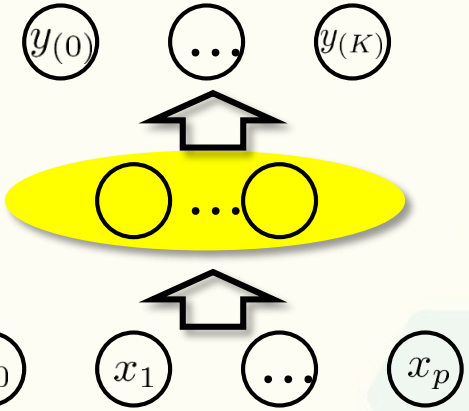
모형의 단순화

$$Min\ \|\beta\|^2$$

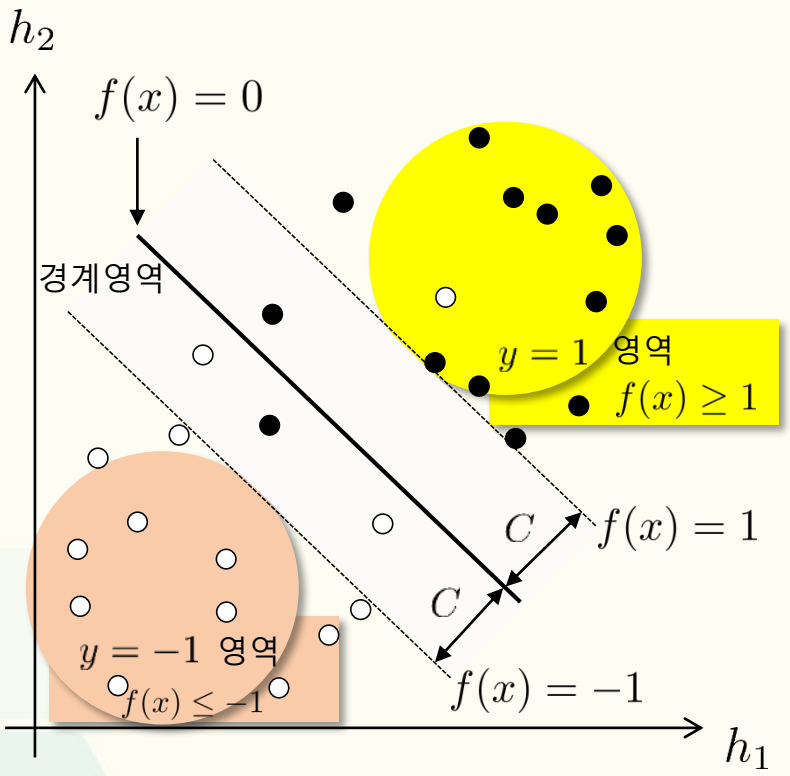
$$f(x) = \beta_1 h_1 + \beta_2 h_2 + \cdots + \beta_q h_q + \beta_0$$

$$h(x) = (h_1(x), \ldots, h_q(x))'$$

$$x = (x_1, \ldots, x_p)'$$



경계영역폭의 최대화



$$f(x) = \beta_1 h_1 + \beta_2 h_2 + \beta_0$$

$$C = \frac{1}{\sqrt{\beta_1^2 + \beta_2^2}}$$

$$Min \ \|\beta\|^2$$

편차변수

	$h'\beta + \beta_0 \leq -1$	$-1 < h'\beta + \beta_0 < 1$	$h'\beta + \beta_0 \geq 1$
$y = -1$	○	△	X
$y = 1$	X	△	○

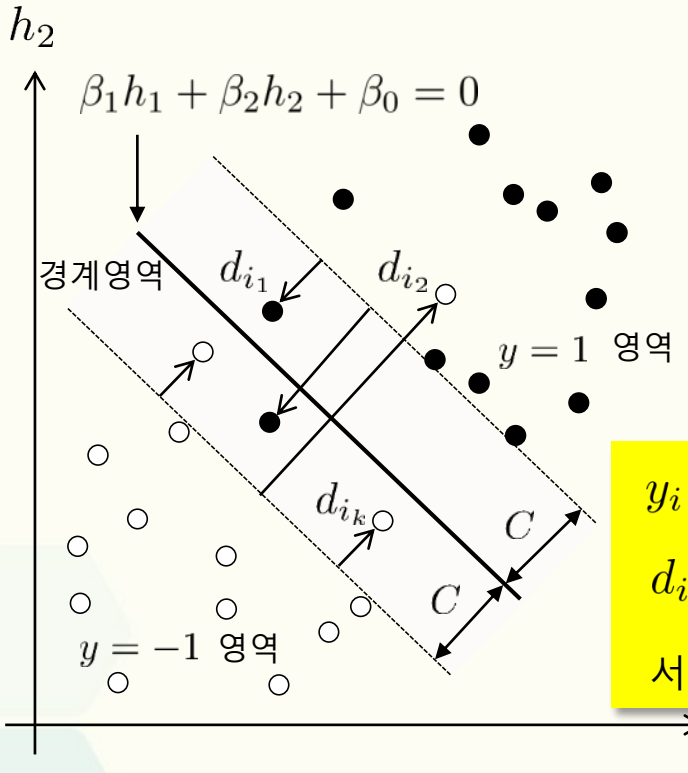
$$f(x_{i\cdot}) = h(x_{i\cdot})'\beta + \beta_0$$

바람직한 경우 : $y_i\, f(x_{i\cdot}) \geq 1$

일반적 표현 : $y_i\, f(x_{i\cdot}) + d_i - l_i = 1 \quad d_i,\ l_i \geq 0$

(동일한 표현) $y_i\, f(x_{i\cdot}) + d_i \geq 1 \quad d_i \geq 0$

편차변수의 의미



$$y_i\, f(x_{i\cdot}) + d_i \;\geq\; 1$$

$$d_i \geq 0$$

$$d_i \;=\; [1 - y_i\, f(x_{i\cdot})]_+$$

$$y_i\, f(x_{i\cdot}) + d_i - l_i \;=\; 1$$

$$d_i,\; l_i \geq 0$$

서포트벡터: $l_i = 0$ 인 경우

서포트벡터 최적화 문제

$$f(x) = h(x)'\beta + \beta_0$$

$$Min_{\beta,\beta_0} \quad \lambda \|\beta\|^2 + \sum_i [1 - y_i f(x_{i.})]_+$$

cost

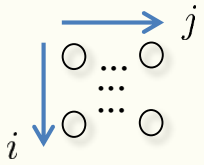
$$Min_{\beta,\beta_0} \quad \frac{1}{2} \|\beta\|^2 + c \sum_i [1 - y_i f(x_{i.})]_+$$

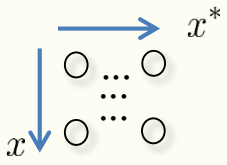
$$Min_{\beta,\beta_0} \quad \sum_i [1 - y_i f(x_{i.})]_+ \quad s.t. \quad \|\beta\|^2 \leq d_o$$

$$Min_{\beta,\beta_0} \quad \|\beta\|^2 \quad s.t. \quad \sum_i [1 - y_i f(x_{i.})]_+ \leq r_o$$

재생커널

재생커널 힐버트공간 (**R**eproducing **K**ernel **H**ilbert **S**pace) :

$A = A(i,j)$ 

$K = K(x, x^*)$ 

- 사영행렬 $A = H(H'H)^{-1}H'$: 대칭성, 비음정성, 멱등성

$$y \in ColSP(H) \quad \Rightarrow \quad Ay = y$$

- 재생커널 K : 대칭성, 비음정성, 멱등성

$$h \in RKHS(K) \quad \Rightarrow \quad K * h = h \quad \textbf{(재생성)}$$

$$h = h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i.)$$

RKHS

$$h \in RKHS(K) \qquad h(\cdot) = \sum_{i=1}^{n} \alpha_i \, K(\cdot, x_i.)$$

$$f(x) = h(x)'\beta + \beta_0 \quad = \sum_{i=1}^{n} \alpha_i \, K(x, x_i.) \, y_i + \beta_0$$

$$K(x, x^*) = < x, x^* > \qquad \text{(linear)}$$

$$K(x, x^*) = (c_0 + \gamma < x, x^* >)^d \qquad \text{(polynomial)}$$

$$K(x, x^*) = \exp\{c_0 - \gamma \, \|x - x^*\|^2\} \qquad \text{(radial)}$$

$$K(x, x^*) = \tanh\{c_0 + \gamma < x, x^* > \} \qquad \text{(sigmoid)}$$

4 서포트벡터 기계의 적용 예

R의 svm 함수

> library(e1071)

svm(formula, data = NULL, type="C-classification", kernel="radial", cost=1, gamma=1, coef0=0, degree=3, ...)

기타 주요 전달인자 :

type : 사용목적 "C-classification", "nu-classification", "nu-regression", "one-classification", "eps-regression".

kernel : 커널종류, "radial"(기정값), "linear", "polynimial", "sigmoid"

$$K(x, x^*) = (c_0 + \gamma < x, x^* >)^d$$

coef0 gamma degree

R의 svm 함수

```
> library(e1071)
> iris9s2a<- svm(SP ~ SW+SL , data =iris, cost = 100, gamma = 1)
> tune(svm, SP~ SW+SL, data = iris,
            ranges = list(gamma = 2^(-4:4), cost = 2^(-4:4) ),
            tunecontrol = tune.control(sampling = "fix")        )
```

...

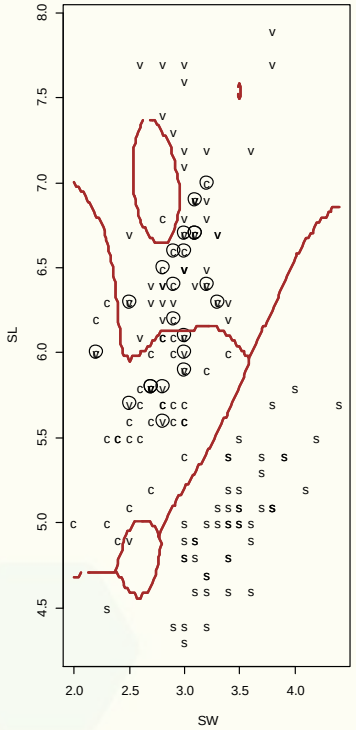
- best parameters:

gamma	cost
0.125	1

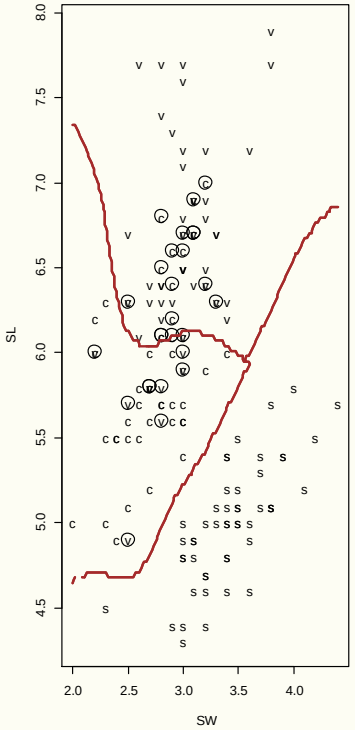
- best performance: 0.2

```
> iris9s2b<- svm(SP ~ SW+SL , data =iris, cost =4, gamma = 1)
> iris9s2c<- svm(SP ~ SW+SL , data =iris, cost = 1, gamma = 0.1)
> par(mfrow=c(1,3))
> plot.iris.mnom(iris9s2a)
> plot.iris.mnom(iris9s2b)
> plot.iris.mnom(iris9s2c)
```

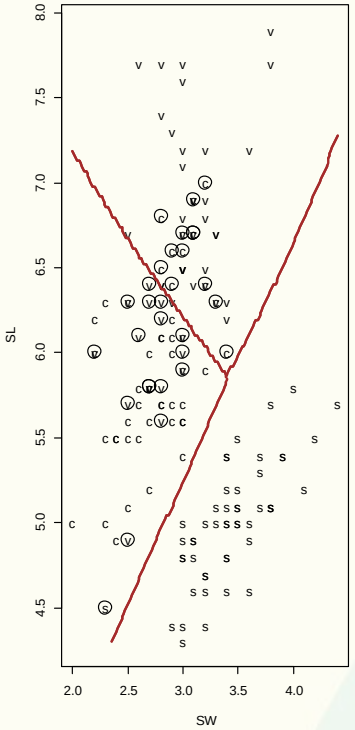
파라메터의 영향



cost=100, gamma=1



cost=4, gamma=1



cost=1, gamma=0.1

$$Min_{\beta,\beta_0} \quad \frac{1}{2}\|\beta\|^2 + c\sum_i d_i$$

$$K(x, x^*) = \exp\{c_0 - \gamma\|x - x^*\|^2\}$$

서포트벡터 기계 : 붓꽃자료 결과 (2변수)

〈표 7.5〉 붓꽃자료에 대한 서포트벡터 기계를 이용한 분류 결과

	Case I			Case II			Case III		
	st	vc	vg	st	vc	vg	st	vc	vg
st	50	0	0	50	1	0	50	0	0
vc	0	37	13	0	33	17	0	37	13
vg	0	9	41	0	10	40	0	16	34

〈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

서포트벡터 기계 : 붓꽃자료 4변수

```
> tune(svm, SP~., data = iris,
              ranges = list( gamma = 2^(-4:4), cost = 2^(3:7)),
              tunecontrol = tune.control(sampling = "fix")        )
> iris9s4<- svm(SP ~ ., data =iris, cost = 8, gamma =0.2)

> table(iris$SP, predict(iris9s2a))
> table(iris$SP, predict(iris9s2b))
> table(iris$SP, predict(iris9s2c))
> table(iris$SP, predict(iris9s4))
     st vc vg
  st 50  0  0
  vc  0 48  2
  vg  0  0 50
```

LDA, QDA & LRA : 4변수

```
> table(iris$SP, predict(iris9s2a))
> table(iris$SP, predict(iris9s2b))
> table(iris$SP, predict(iris9s2c))
> table(iris$SP, predict(iris9s4))
    st vc vg
st 50  0  0
vc  0 48  2
vg  0  0 50
```

〈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

다음시간 안내

총정리