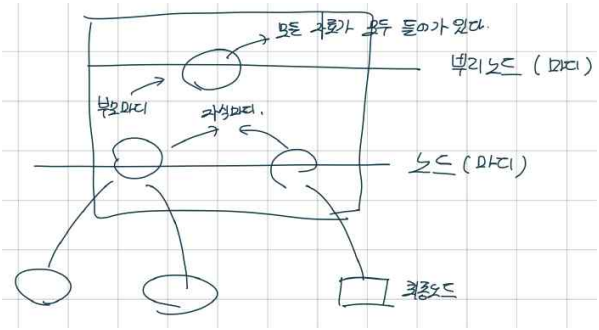
```
rm(list=ls())
##### Chapter 4 의사결정나무


### 4.1 서론
# 의사결정나무 또는 나무모형은 의사결정규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다.
# 상위 노드에서의 (분류변수, 분류 기준값)은 이 기준에 의해 분기되는 하위 노드에서
# 노드(집단) 내에서는 동질성이, 노드(집단) 간에는 이질성이 가장 커지도록 선택된다.
# 나무 모형의 크기는 과대 적합(또는 과소 적합)되지 않도록 합리적 기준에 의해 적당히 조절되어야 한다.
```

4.2 의사결정나무



뿌리노드 : 맨 위 노드라고 하며 분류 대상이 되는 모든 자료 집단을 포함한다.
상위 노드가 하위 노드로 분기될 때, 상위 노드를 부모 노드라고 하고, 하위 노드를 자식 노드라고 한다.
최종노드 : 더이상 분기되지 않는 노드를 최종노드라고 한다.
가지분할(split)은 나무의 가지를 생성하는 과정을
가지치기(pruning)는 생성된 가지를 잘라내어 모형을 단순화하는 과정을 말한다.

의사결정나무는 분류나무(목표변수가 이산형인 경우)와 회귀나무(목표변수가 연속형인 경우)로 나뉜다.

목표변수가 이산형인 분류나무
분류나무의 가지분할을 할 때, 분류(기준)변수와 분류 기준값의 선택방법으로
카이제곱통계량의 p값, 지니 지수, 엔트로피 지수 등이 사용된다.
카이제곱통계량의 p값은 그 값이 작을수록 자식 노드 간의 이질성이 큼을 나타내며
자식 노드에서의 지니 지수나 엔트로피 지수는 그 값이 작을수록 자식 노드 간의 이질성이 큼을 의미한다.

지니 지수의 값이 클수록 (자식 노드 안에서) 이질적이며 순수도가 낮다고 할 수 있다.
따라서 이 값들이 가장 작아지는 방향으로 가지분할을 수행하게 된다.
A B C A B B A D : 높은 이질성, 낮은 순수도
A A A A A B A A : 낮은 이질성, 높은 순수도

불확실성 측도인 지니 지수와 엔트로피 지수에 대한 정의는 다음과 같다. 불확실성이 커지면 안 좋은 것이며, 작아지면 가지가 잘 나뉜다.
```

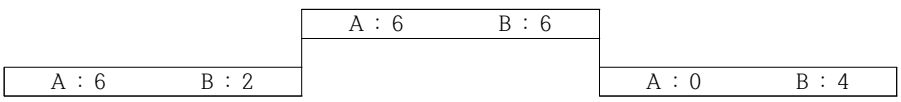
지니지수 : (부모노드) $G_k = 1 - \sum_i^c p_i^2$, $c$ : 종류의 개수 ,

(자식노드) $\sum_{k=1}^{d} G_k R_k$ , $R_k$는 가중값 : $\frac{\text{해당되는 자식노드의 자료 총 개수}}{\text{부모노드의 자료 총 개수}}$

엔트로피 지수 : $E = -\sum_i^c p_i \log_2 p_i$ , $0 \le E \le 1$

위에 식에서 c는 목표변수의 범주의 수이다.

## 참고: 지니 지수와 카이제곱 통계량의 계산




● 엔트로피 지수

분기전(부모노드) : $-[\frac{6}{12}\log_2(\frac{6}{12}) + \frac{6}{12}\log_2(\frac{6}{12})] = -(-0.5-0.5) = 1$

분기후(자식노드) : $-[\frac{6}{8}\log_2(\frac{6}{8}) + \frac{2}{8}\log_2(\frac{2}{8})] \times \frac{8}{12} - [\frac{0}{4}\log_2(\frac{0}{4}) + \frac{4}{4}\log_2(\frac{4}{4})] \times \frac{4}{12}$

$= -(-0.3112781-0.5) \times \frac{8}{12} - (0+0) \times \frac{4}{12} = 0.5408521$
```

● 지니지수

분기 전(부모노드) : $1-(\frac{6}{12})^2-(\frac{6}{12})^2=0.5$

분기 후(자식노드) : $[1-(\frac{6}{8})^2-(\frac{2}{8})^2]\times\frac{8}{12}+[1-(\frac{0}{4})^2-(\frac{4}{4})^2]\times\frac{4}{12}=\frac{1}{8}$ (분기 전에 비해 감소함)

● 카이제곱 통계량

귀무가설(두 개의 노드(Right, Lift) 간에 A와 B 구성 비율이 동질적이다.)에 대한 기각이 좋다는 가정하에 진행.
기각한다는 것은 구성 비율이 다르다는 것이고, 구성 비율이 다르게 분류하는 것이 궁극적이 목표이기 때문이다.
()는 귀무가설 하에 기대빈도이다.

	A	B	total
Left	6, (기대빈도 : $\frac{8\times6}{12}=4$)	2, (기대빈도 : $\frac{8\times6}{12}=4$)	8
Right	0, (기대빈도 : $\frac{4\times6}{12}=2$)	4, (기대빈도 : $\frac{4\times6}{12}=2$)	4
total	6	6	12

카이제곱 통계량 : $\chi^2=\sum_{i=1}^{k}\frac{(빈도-기대빈도)^2}{기대빈도}=\frac{(6-4)^2}{4}+\frac{(2-4)^2}{4}+\frac{(0-2)^2}{2}+\frac{(4-2)^2}{2}=6$

목표변수가 연속형 자료인 회귀나무
회귀나무의 경우에는 분류변수와 분류 기준값의 선택방법으로 F-통계량의 P-값, 분산의 감소량 등이 사용된다.
F-통계량은 일원배치법에서의 검정통계량으로, 그 값이 클수록 오차변동에 비해 처리변동이 크다는 것을 의미하며,
자식 노드(처리들) 간의 이질적임을 의미하므로, 이 값이 커지는(p값이 작아지는)방향으로 가지분할을 수행(자식노드를 생성)하게 된다.
분산의 감소량도 이 값이 최대화 되는 방향으로 가지분할을 수행하게 된다.

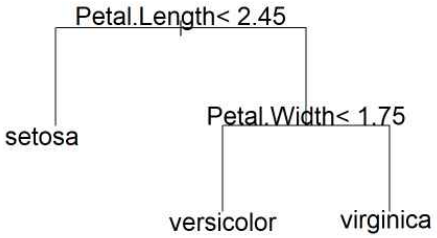
의사결정나무의 분석과정은 다음과 같다.
단계 1 : 목표변수와 관계가 있는 설명변수들의 선택
단계 2 : 분석목적과 자료의 구조에 따라 적절한 분리기준과 정지규칙을 정하여 의사결정나무의 생성
단계 3 : 부적절한 나뭇가지는 제거 : 가지치기
단계 4 : 이익(gain), 위험(risk), 비용(cost) 등을 고려하여 모형평가
단계 5 : 분류 및 예측 수행

의사결정나무의 주료 알고리즘

	이산형 목표변수	연속형 목표변수
CHAID(다지분할)	카이제곱통계량	ANOVA F-통계량
CART(이진분할)	지니 지수	분산 감소량
C4.5	엔트로피 지수	.

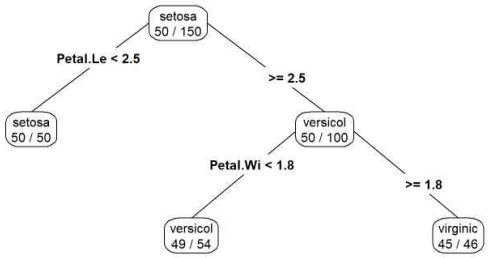
예제 1
{rpart}패키지의 rpart()함수를 사용하여 분석 수행(rpart : recursive partitioning and regression tree의 약어)
```
library(rpart)
c <- rpart(Species ~., data=iris) # rpart를 사용하여 가지분할 수행
c # 가지분할의 결과가 저장됨
ls(c) # 가지분할이 수행된 data에서 설정할 수 있는 매서드를 출력
```

plot.rpart 함수의 옵션 설명
compress : 출력 시 노드의 배치에 관한 사항. 만약 T이면 더 압축된 의사결정나무 그림을 출력, 디폴트는 F로 설정
margin : 나무 출력 시 주변 여백 값, 너무 작은 경우 분할 규칙이 잘리는 경우도 있음. 디폴트는 0으로 설정
```
plot(c, compress=T, margin=0.3) # 틀 그리기
text(c, cex=1.5) # 각 텍스드 입력
```



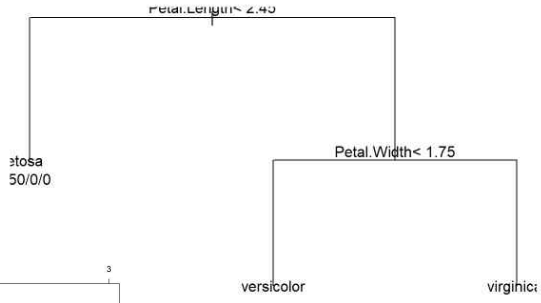
predict() 함수 (type=class)를 이용하여 새로운 자료에 대해 예측을 수행한다.
가지분할을 했을 때, 각 변수가 해당하는 분류를 나타냄
```
head(predict(c, newdata=iris, type="class")) # Levels는 분류 기준을 말한다.
tail(predict(c, newdata=iris, type="class")) # Levels는 분류 기준을 말한다.
```

{rpart.plot}패키지의 prp함수를 사용하여 의사결정나무 모형을 시각화를 해본다.
# 앞선 {rpart}plot() 함수보다 보기 편한 시각화를 사용할 수 있다.
# prp함수의 옵션 설명
# type : 출력 형식 지정
# 0이면 중간노드를 그리지 않고,
# 1이면 중간노드를 도형으로 표현하고 노드를 분할하는 조건을 노드 위쪽에 출력
# 2이면 중간노드를 그리면서 노드를 분할하는 조건을 노드 아래쪽에 출력
# 3이면 중간노드를 그리지 않으면서 왼쪽으로 분할하는 조건과 오른쪽으로 분할하는 조건을 모두 출력
# 4이면 중간 노드를 그리면서 왼쪽으로 분할하는 조건과 오른쪽으로 분할하는 조건을 모두 출력함
# extra : 중간노드 내의 출력 형식을 지정
# 0인 경우 분류나무이면 다수 점주를 출력하고 회귀나무이면 노드의 목표변수 평균을 출력
# 1인 경우 0 옵션의 결과에 추가하여 분류나무이면 범주별 관찰값 개수 출력, 회귀나무이면 노드의 총 관찰값 개수 출력
# 2인 경우 분류나무 인 경우 다수 범주의 관찰값 개수와 총 관찰값 개수를 출력함. 회귀나무는 해당사항 없음.
install.packages("rpart.plot")
library(rpart.plot)
prp(c, type=4, extra =2)
# 두 가지의 기준을 만족하는 노드에서 두 번째 최종노드는 해당 개체 54개 중 versicolor가 49개임을 나타내고,
# 이후 새로운 데이터가 이 노드로 분류된다면 versicolor 라고 판단한다.

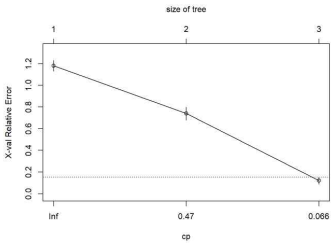


가지치기의 결정 및 시기
사전 가지치기 : 노드 수를 몇 개로 할건지 등을 정해서 분석 진행
사후 가지치기 : cp($cptable 사용) 등,
ls(c) # 가지분할이 수행된 data에서 설정할 수 있는 매서드를 출력

cp를 이용한 사후 가지치기 하는 방법
$cptable는 트리의 크기가 클 때(가지의 수가 많을 때) 비용-복잡도 모수를 제공하며, 교차타당성오차를 함께 제공한다.
교차타당성오차(xerror)를 사용해 가지치기와 트리의 최대 크기를 조절하기 위한 옵션으로 사용
주어진 자료는 트리의 크기(가지의 수)가 충분히 크지 않아서, cp로 가지치기를 하기 전과 후의 그래프가 같다.
만약 트리의 크기가 크면, cp로 가지치기 했을 때 모양을 달라질것이다.
c$cptable
opt <- which.min(c$cptable[,"xerror"]) # 교차타당성 오차(xerror)가 가장 작은 번호를 반환 : 3
cp <- c$cptable[opt, "CP"] # 교차타당성 오차(xerror)가 가장 작은 번호의 cp값을 저장
prune.c <- prune(c, cp = cp) # cp값에 따른 가지분할을 진행한다.
plot(prune.c) # 틀 그리기
text(prune.c, use.n=T) # 각 노드의 텍스트 입력



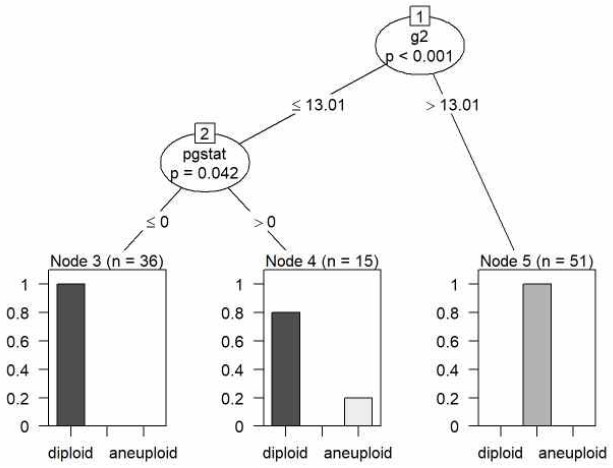
{rpart}의 패키지 plotcp 함수를 사용하여 cp 의 값을 시각화
plotcp(c)



예제 2
데이터 선언
# {party}의 패키지의 ctree()함수를 사용하여 분석 수행
install.packages("party")
library(party)
library(caret)
data(stagec) # 146명의 전립선 암 환자의 자료이다.
str(stagec) # 데이터 요약, 8개 변수(pgtime, pgstat, age, eet, g2, grade, gleason, ploidy)와 146개 데이터

결측값을 제거하는 과정(g2, gleason, eet의 결측값 제거)
stagec1<- subset(stagec, !is.na(g2))
stagec2<- subset(stagec1, !is.na(gleason))
stagec3<- subset(stagec2, !is.na(eet))
str(stagec3) # 8개 변수와 134개 데이터 (12개의 결측값 제거)

훈련용(training) 자료와 검증용(test)자료로 나누기 위해 7:3로 나눈다.
set.seed(1234) # 분석의 재현성을 만족시키기 위해 시드 고정
ind <- sample(2, nrow(stagec3), replace=TRUE, prob=c(0.7, 0.3))
trainData <- stagec3[ind==1,] # n=102개
testData <- stagec3[ind==2,] # n=32개



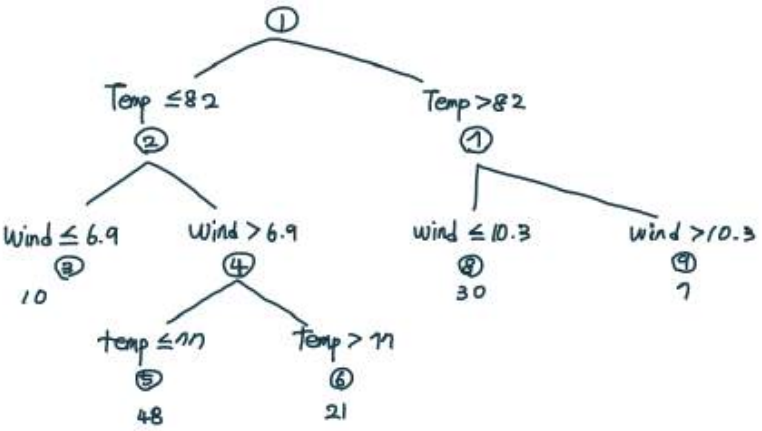
{party}의 패키지의 ctree()함수를 사용하여 가지분할 진행
# ctree()함수는 사전 가지치기가 디폴트이다.
tree <- ctree(ploidy ~ ., data=trainData) # ploidy : 반응변수
tree # 가지분할 결과가 저장됨
plot(tree) # 최종노드의 막대 그래프는 반응변수의 각 범주별 비율을 나타낸다

predict() 함수를 통해 검증용 자료에 대해 적합모형을 적용(test 데이터의 수 : 32개, 데이터 수가 작다.)
# 위에서 생성된 의사결정나무에 새로운 데이터를 적용하여 예측한다.
# diploid, tetraploid는 각자로 잘 분류되었지만, aneuploid는 자료의 수가 매우 작아서 diploid, tetraploid로 잘못 분류되었다.
testPred = predict(tree, newdata=testData) # 의사결정나무모형을 출력
table(testPred, testData$ploidy) # ploidy : 반응변수, 각 변수들이 자신의 어떤 노드로 분류가 되었는지 table로 출력

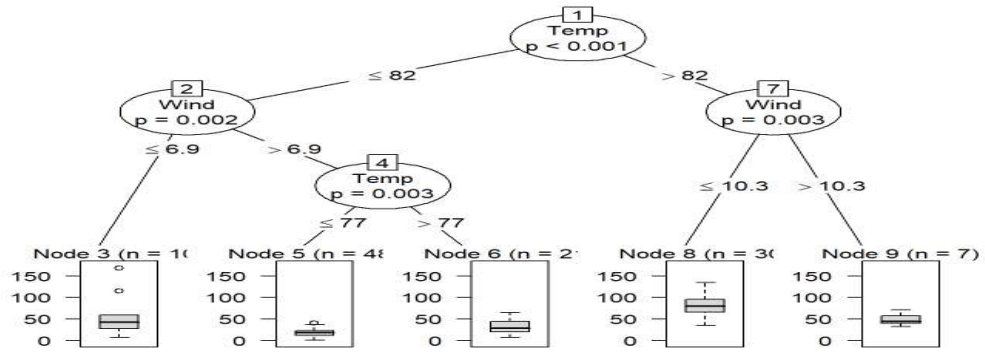
예제 3
# {party}의 패키지의 ctree()함수를 사용하여 분석진행
# airquality 자료에 대해 반응변수가 Ozone로 분류를 한다. 따라서 Ozone이 결측값인 자료를 제외한 데이터로 분석을 진행
airq <- subset(airquality, !is.na(Ozone)) # 결측치 제거
head(airq) # 결측치 제거한 데이터 요약

{party}의 패키지의 ctree()함수를 사용하여 가지분할 진행
# criterion = 1-p 이 일정값 이상이면 분할 진행
airct <- ctree(Ozone ~ ., data=airq)
airct

결과값
```
## Response:  Ozone
## Inputs:  Solar.R, Wind, Temp, Month, Day
## Number of observations:  116
##
## 1) Temp <= 82; criterion = 1, statistic = 56.086
##    2) Wind <= 6.9; criterion = 0.998, statistic = 12.969
##       3)*  weights = 10
##    2) Wind > 6.9
##       4) Temp <= 77; criterion = 0.997, statistic = 11.599
##          5)*  weights = 48
##       4) Temp > 77
##          6)*  weights = 21
## 1) Temp > 82
##    7) Wind <= 10.3; criterion = 0.997, statistic = 11.712
##       8)*  weights = 30
##    7) Wind > 10.3
##       9)*  weights = 7
```



# 가지분할의 결과를 시각화
plot(airct)



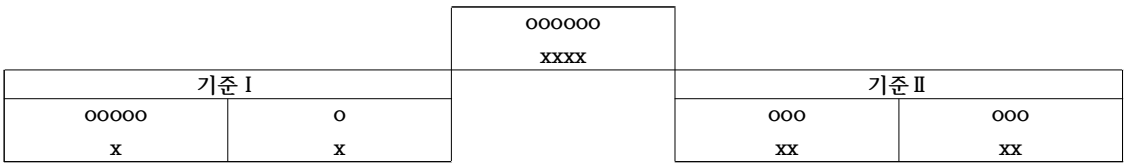
predict() 함수를 통해 새로운 자료에 대해 예측을 진행한다.
연속형 반응변수에 대한 예측값은 최종노드에 속한 자료들의 평균값이 제공
type = "node" 옵션을 이용하여 자료가 속하는 해당 최종노드의 번호를 출력
head(predict(airct, data=airq)) # 출력 : [1,] 18.47917 [2,] 18.47917 [3,] 18.47917 [4,] 18.47917 。。。 [6,] 18.47917
predict(airct, data=airq, type="node") # 출력 : [1] 5 5 5 5 5 5 5 5 3 5 5 5 。。。8 6 9 8 8 8 。。。5 3 5 5 5

예측값을 이용하여 평균제곱오차를 구할 수 있다.
# 다른 방식으로 고른 결과들을 비교할 때 MSE를 사용하기 위해 구함.
mean((airq$Ozone - predict(airct))^2)

의사나무결정의 장단점
장점
구조가 단순하여 해석에 용이하고,
유용한 입력변수의 파악과 예측변수 간의 상호작용 및 비선형성을 고려하여 분석이 수행되며,
선형성, 정규성, 등분산성 등의 수학적 가정이 불필요한 비모수적 모형이다.

단점
분류의 기준값의 경계선 근방의 자료값에 대해서는 오차가 클 수 있으며(비연속성),
로지스틱 회귀의 오즈비와 같이 각 예측변수의 효과를 파악하기 어려우며
새로운 자료에 대한 예측이 불안정할 수 있다.

추가 계산 예제
cf) 정보이득(Information Gain : I.A) : 정보 이득이 클수록 좋은 지표이다.
I.A(S.A) = E(S) - E(A) = (가지분할 전의 불확실성) - (가지분할 후의 불확실성)



$$G = 1-(\frac{6}{10})^2-(\frac{4}{10})^2 = 0.48$$

$$G_1 = [1-(\frac{5}{6})^2-(\frac{1}{6})^2]\times\frac{6}{10}+[1-(\frac{1}{4})^2-(\frac{3}{4})^2]\times\frac{4}{10} = [0.2778]\times\frac{6}{10}+[0.375]\times\frac{4}{10} = 0.31668$$

$$G_2 = [1-(\frac{3}{5})^2-(\frac{2}{5})^2]\times\frac{1}{2}+[1-(\frac{3}{5})^2-(\frac{2}{5})^2]\times\frac{1}{2} = [0.48]\times\frac{1}{2}+[0.48]\times\frac{1}{2} = 0.48$$

$$I.A_1 = 0.48-0.316 = 0.164(\text{채택})$$

$$I.A_2 = 0.48-0.48 = 0$$

$$E = -[\frac{6}{10}\log_2\frac{6}{10}+\frac{4}{10}\log\frac{4}{10}] = -(-0.4421794-0.5287712) = 0.9709506$$

$$E_1 = -[\frac{5}{6}\log_2\frac{5}{6}+\frac{1}{6}\log\frac{1}{6}]\times\frac{6}{10}-[\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log\frac{3}{4}]\times\frac{4}{10}$$

$$= -(-0.2191953-0.4308271)\times\frac{6}{10}-(-0.5-0.3112781)\times\frac{4}{10} = 0.6500224\times0.6+0.8112781\times0.4 = 0.7145247$$

$$E_2 = -[\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log\frac{2}{5}]\times\frac{5}{10}-[\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log\frac{2}{5}]\times\frac{5}{10} = -[\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log\frac{2}{5}] = -(-0.4421794-0.5287712) = 0.9709506$$

$$I.E_1 = 0.9709506-0.7145247 = 0.2564259(\text{채택})$$

$$I.E_2 = 0.9709506-0.9709506 = 0$$