데이터마이닝(DataMining)

Chapter 4.1. 의사결정나무

- 의사결정나무(decision tree) 또는 나무 모형(tree model)은 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법

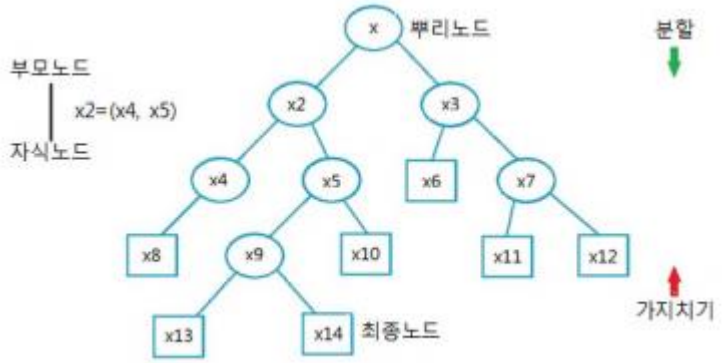
- 상위 노드로부터 하위노드로 트리구조를 형성하는 매 단계마다 분류변수와 분류 기준값의 선택이 중요

- 상위노드에서의 (분류변수, 분류 기준값)은 이 기준에 의해 분기되는 하위노드에서 노드(집단) 내에서는 동질성이, 노드(집단) 간에는 이질성이 가장 커지도록 선택

- 나무 모형의 크기는 과대적합(또는 과소적합) 되지 않도록 합리적 기준에 의해 적당히 조절

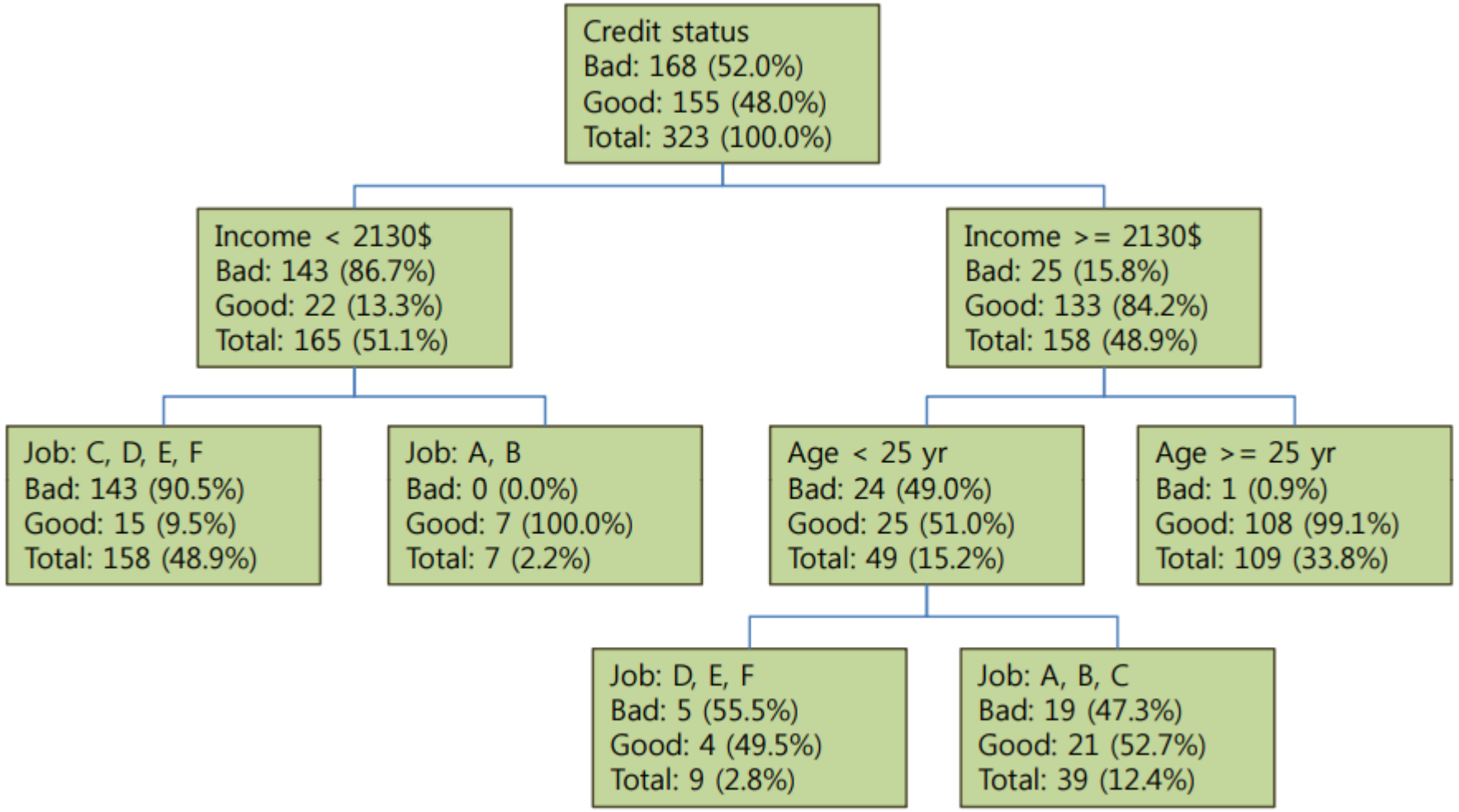
의사결정나무

- 맨 위의 마디를 뿌리노드(root node)라 하며, 이는 분류(또는 예측) 대상이 되는 모든 자료 집단을 포함
- 상위 마디를 부모마디(parent node)라 하고, 하위 마디를 자식마디(child node)라 하며, 더 이상 분기되지 않는 마디를 최종노드(terminal node)
- 가지분할(split)은 나무의 가지를 생성하는 과정을, 가지치기(pruning)는 생성된 가지를 잘라내 어 모형을 단순화하는 과정



의사결정나무

- 신용자료에 대한 의사결정나무 예



의사결정나무

- 의사결정나무는 목표변수가 이산형인 경우의 분류나무(classification tree)와 목표변수가 연속형인 경우의 회귀나무(regression tree)로 구분

- 목표변수가 이산형인 분류나무의 경우 상위노드에서 가지분할을 수행할 때, 분류(기준)변수와 분류 기준값의 선택 방법으로 카이제곱 통계량(Chi-square statistic)의 p-값, 지니 지수(Gini index), 엔트로피 지수(entropy index) 등이 사용

- 선택된 기준에 의해 분할이 일어날 때, 카이제곱통계량의 p-값은 그 값이 작을수록 자식노드 간의 이질성이 큼을 나타내며, 자식노드에서의 지니 지수나 엔트로피 지수는 그 값이 클수록 자식노드 내의 이질성이 큼을 의미한다.

- 값들이 가장 작아지는 방향으로 가지분할을 수행

의사결정나무

- 불확실성 측도(uncertainty measure)인 지니 지수와 엔트로피 지수의 값의 범위는 다르나, 해석은 그 크기에 따라 유사
- m개의 개체가 속하는 노드 A에 대한 엔트로피

$Entropy(A) = -\sum_k p_k \log_2 p_k$, $p_k = k$범주에 속하는 개체의 비율

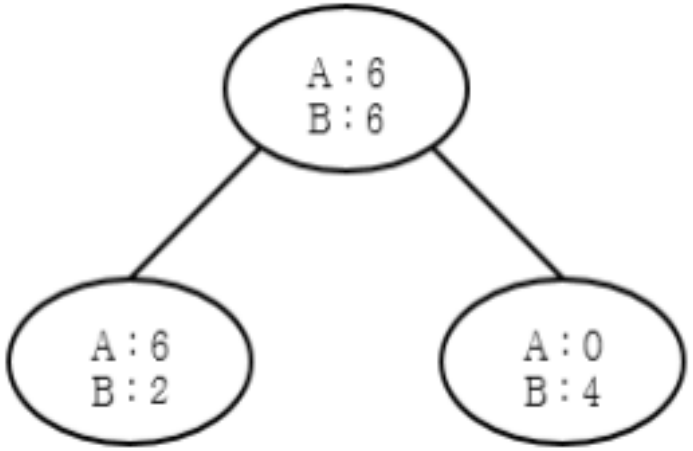
- A노드를 분할 A_1, A_2노드로 분할한다면

$Entropy(A) = \sum_{i=1}^{2} p_{A_i} \left(-\sum_k p_{A_{ik}} \log_2 p_{A_{ik}}\right), p_{A_{ik}} = A_i$노드의 k범주에 속하는 개체의 비율

의사결정나무

-분기 전 $Entropy = -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12} = 1$

-분기 후 $Entropy = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right)\frac{8}{12} - \left(\frac{0}{4}\log_2\frac{0}{4} + \frac{4}{4}\log_2\frac{4}{4}\right)\frac{4}{12} = 0.5408521$



의사결정나무

- 지니계수 $G = \sum_{i=1}^{2} p_{A_i}(1 - \sum_k p_{A_{ik}}^2)$

-분기 전 $Entropy = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$

-분기 후 $Entropy = \left[1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2\right] \times \frac{2}{3} + \left[1 - \left(\frac{0}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right] \times \frac{1}{3} = \frac{1}{8}$

- 카이제곱통계량

-$\chi^2 = \frac{(6-4)^2}{4} + \frac{(2-4)^2}{4} + \frac{(0-2)^2}{2} + \frac{(4-2)^2}{2} = 6$

	A	B
Left	6(4)	2(4)
Right	0(2)	4(2)

의사결정나무

높은 이질성 ⇔ 낮은 순수도

$G=1-(3/8)^2-(3/8)^2-(1/8)^2-(1/8)^2=0.69$

낮은 이질성 ⇔ 높은 순수도

$G=1-(7/8)^2-(1/8)^2=0.24$

의사결정나무

- 목표변수가 연속형인 회귀나무의 경우에는 분류변수와 분류 기준값의 선택방법으로 F-통계량 의 F-값, 분산의 감소량 등이 사용

- F-통계량은 일원배치법에서의 검정통계량으로 그 값이 클수록 오차의 변동에 비해 처리(treatment)의 변동이 크다는 것을 의미하며, 이는 자식노드(처리들) 간이 이질적임을 의미하므로 이 값이 커지는(p-값은 작아지는) 방향으로 가지분할을 수행(자식노드를 생성)

- 분산의 감소량(variance reduction)도 이 값이 최대화 되는 방향으로 가지분할을 수행

의사결정나무

- 의사결정나무의 분석과정
  - [1단계] : 목표변수와 관계가 있는 설명변수들의 선택
  - [2단계] : 분석목적과 자료의 구조에 따라 적절한 분리기준과 정지규칙을 정하여 의사결정 나무의 생성
  - [3단계] : 부적절한 나뭇가지는 제거: 가지치기
  - [4단계] : 이익(gain), 위험(risk), 비용(cost) 등을 고려하여 모형평가(교차타당성 방법도 사용됨)
  - [5단계] : 분류(classification) 및 예측(prediction) 수행

의사결정나무

- 의사결정나무분석을 위한 알고리즘에는 CHAID(Kass, 1980), CART(Breiman 등, 1984), ID3(Quinlan, 1986), C4.5(Quinlan, 1993), C5.0(Quinlan, 1998) 등과 이들의 장점을 결합한 다양한 알고리즘이 있음

	이산형 목표변수	연속형 목표변수
CHAID(다지분할)	카이제곱 통계량	ANOVA F-통계량
CART(이진분할)	지니지수	분산감소량
C4.5	엔트로피지수	·

의사결정나무

- CART(classification and regression trees)
  - 가장 널리 사용되는 알고리즘으로 이진분리(binary split)
  - 분류나무: 지니지수, 회귀나무: 분산
  - 입력변수들의 선형결합들중에서 최적의 분리를 찾을 수도 있음
- C4.5와 C5.0
  - 다지분리(multiple split)가 가능
  - 엔트로피지수를 사용
- CHAID(chi-squared automatic interaction detection)
  - 가지치기를 하지 않고 조기 종료(early stopping)에 의해 적당한 크기에서 나무모형의 성장을 중지
  - 입력변수는 범주형
  - 카이제곱 통계량을 사용

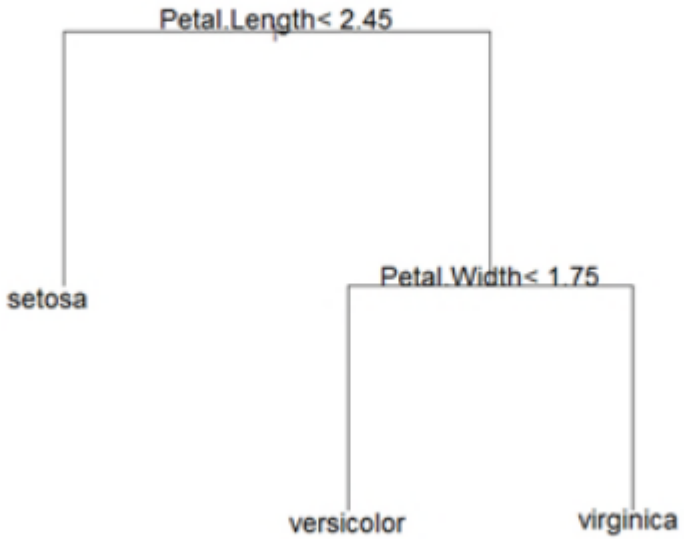
의사결정나무

- R 패키지 {rpart}의 rpart() 함수를 이용하여 의사결정나무 분석을 수행
- recursive partitioning and regression tree

```
> library(rpart)
> c <- rpart(Species ~., data=iris)
> c
n= 150
 node), split, n, loss, yval, (yprob)
           * denotes terminal node
 1) root 150 100 setosa (0.3333 0.3333 0.3333)
   2) Petal.Length< 2.45 50 0 setosa (1.0000 0.0000 0.0000) *
   3) Petal.Length>=2.45 100 50 versicolor (0.0000 0.5000 0.5000)
     6) Petal.Width< 1.75 54 5 versicolor (0.0000 0.9074 0.0925) *
     7) Petal.Width>=1.75 46 1 virginica (0.0000 0.0217 0.9782) *
```

의사결정나무

```
> plot(c, compress=T, margin=0.3)
> text(c, cex=1.5)
```



의사결정나무

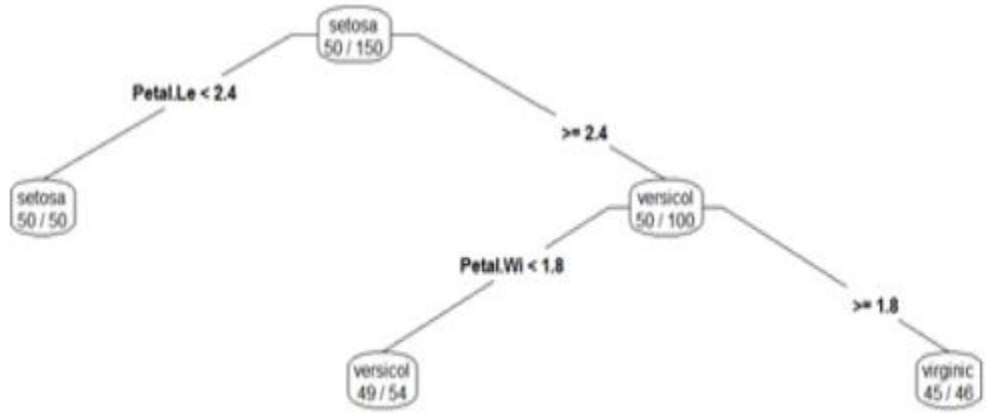
- predict() 함수를 이용하여 새로운 자료에 대해 예측을 수행

```
> head(predict(c, newdata=iris, type="class"))
     1          2          3          4          5          6
setosa     setosa     setosa     setosa     setosa     setosa
Levels: setosa versicolor virginica
```

```
> tail(predict(c, newdata=iris, type="class"))
      145       146       147       148       149       150
virginica virginica virginica virginica virginica virginica
Levels: setosa versicolor virginica
```

의사결정나무

```
> library(rpart.plot)
> prp(c, type=4, extra =2)
```



- 두 조건(Petal.Length>=2.4 와 Petal.Width<1.8) 을 만족하는 노드에서 49/54는 이 노드에 속하는 해당 개체가 54개이며 이 가운데 versicolor가 49임을 나타냄
- 따라서 이 노드에 해당되는 새로운 자료는 versicolor로 분류

```
> ls(c)
 [1] "call"    "control"   "cptable"
 [4] "frame"   "functions" "method"
 [7] "numresp" "ordered"   "parms"
[10] "splits"  "terms"     "variable.importance"
[13] "where"   "y"
```

- $cptable은 트리의 크기에 따른 비용-복잡도 모수(cost-complexity parameter)를 제공하며, 교차타당성오차(cross-validation error)를 함께 제공
- 이 값들은 prune() 또는 rpart.control() 함수에서 가지치기(pruning)와 트리의 최대 크기(maximum size)를 조절하기위한 옵션으로 사용

의사결정나무

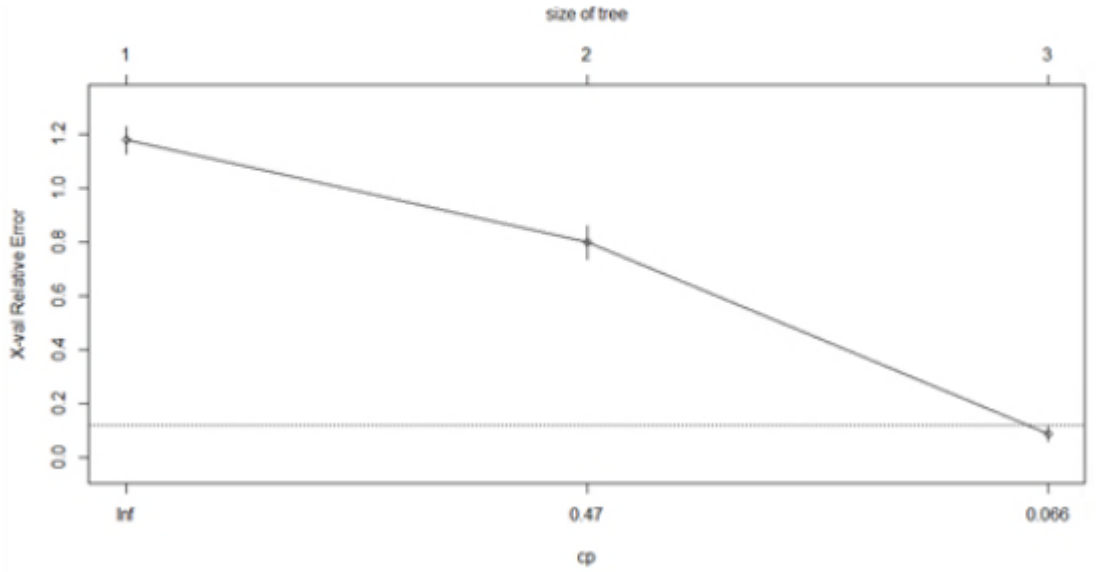
- 교차타당성오차를 최소로 하는 트리를 형성하는 과정

```
> c$cptable
CP nsplit  rel error xerror          xstd
 1   0.50    0  1.00   1.18 0.05017303393
 2   0.44    1  0.50   0.80 0.06110100927
 3   0.01    2  0.06   0.09 0.02908607914
```

```
> opt <- which.min(c$cptable[,"xerror"])
> cp <- c$cptable[opt, "CP"]
> prune.c <- prune(c, cp = cp)
> plot(prune.c)
> text(prune.c, use.n=T)
```

의사결정나무

```
> plotcp(c)
```



의사결정나무

- ctree는 “conditional inference tree”의 약어
- 146명의 전립선 암 환자의 자료(stagec)
- 7개의 예측변수를 이용하 여 범주형의 반응변수(ploidy)를 예측(또는 분류)

```
> library(party)
> data(stagec)    # stagec는 {rpart}에서 제공함
> str(stagec)
 'data.frame':  146 obs. of 8 variables:
 $ pgtime    : num 6.1 9.4 5.2 3.2 1.9 4.8 5.8 7.3 3.7 15.9 ...
 $ pgstat    : int 0 0 1 1 1 0 0 0 1 0 ...
 $ age       : int 64 62 59 62 64 69 75 71 73 64 ...
 $ eet       : int 2 1 2 2 2 1 2 2 2 2 ...
 $ g2        : num 10.26 NA 9.99 3.57 22.56 ...
 $ grade     : int 2 3 3 2 4 3 2 3 3 3 ...
 $ gleason   : int 4 8 7 4 8 7 NA 7 6 7 ...
 $ ploidy    : Factor w/ 3 levels "diploid","tetraploid",..: 1 3 1 1 2 1 ...
```

의사결정나무

- 결측값을 제거

```
> stagec1<- subset(stagec, !is.na(g2))
> stagec2<- subset(stagec1, !is.na(gleason))
> stagec3<- subset(stagec2, !is.na(eet))
> str(stagec3)
```

- 결측값이 제거된 134개의 자료를 이용하여 모형을 적합
- 모형구축을 위한 훈련용 자료 (training data)와 모형의 성능을 검증하기위한 검증용 자료 (test data)를 70%와 30%로 구성

```
> set.seed(1234)
> ind <- sample(2, nrow(stagec3), replace=TRUE, prob=c(0.7, 0.3))
```

의사결정나무

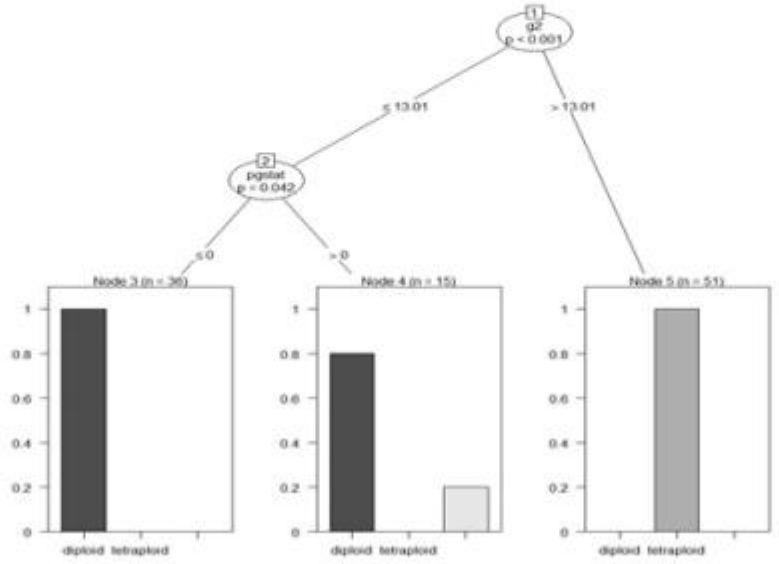
```
> ind
[1] 1 1 1 1 2 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 2 1 2 2 1 1 1
1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 2 2 1 1 1 1 2
1 1 1
[70] 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2
1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 2 2 1 1 2 2 2 2 2 1 1 1 1 1 1 2 1 1 1
```

```
> trainData <- stagec3[ind==1, ]    # n=102개
> testData <- stagec3[ind==2, ]    # n=32개
```

의사결정나무

- 훈련용 자료(n=20)에 대해 ctree()를 적용한 결과

```
> tree <- ctree(ploidy ~ ., data=trainData)
> tree
> plot(tree)
```



- 최종노드의 막대그래프(barplot)는 반응변수(ploidy)의 각 범주별 비율을 나타냄

의사결정나무

```
> testPred = predict(tree, newdata=testData)
> table(testPred, testData$ploidy)
  testPred diploid tetraploid aneuploid
   diploid      17          0         1
tetraploid       0         13         1
 aneuploidy      0          0         0
```

의사결정나무

- ctree() 함수를 이용하여 반응변수가 연속형인 경우 의사결정나무(회귀나무) 를 통한 예측을 수행
- airquality 자료에 대해 의사결정나무모형을 적합

```
> airq <- subset(airquality, !is.na(Ozone))
> head(airq)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
6    28      NA 14.9   66     5   6
7    23     299  8.6   65     5   7
```

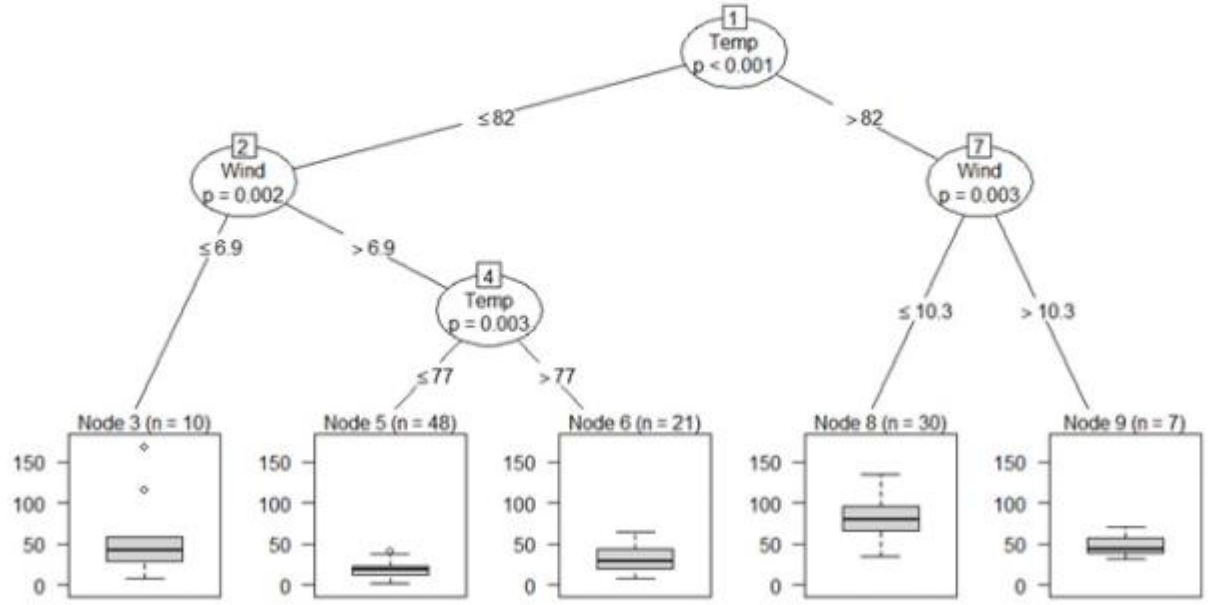
의사결정나무

```
> airct <- ctree(Ozone ~ ., data=airq)
> Airct

       Conditional inference tree with 5 terminal nodes
Response: Ozone
Inputs: Solar.R, Wind, Temp, Month, Day
Number of observations: 116
 1) Temp <= 82; criterion = 1, statistic = 56.086  # criterion=1-p값으로 기준값 이상이면 분할 수행
   2) Wind <= 6.9; criterion = 0.998, statistic = 12.969  # 예측변수와 반응변수 간의 독립성 검정
     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
```

의사결정나무

```
> plot(airct)
```



의사결정나무

- 연속형 반응변수에 대한 예측값은 최종노드에 속한 자료들의 평균값이 제공

```
> head(predict(airct, data=airq))
         Ozone
 [1,] 18.47917
 [2,] 18.47917
 [3,] 18.47917
 [4,] 18.47917
 [5,] 18.47917
 [6,] 18.47917
```

- 자료가 속하는 해당 최종노드의 번호를 출력하고 싶을 때는 type="node" 옵션을 사용

```
> predict(airct, data=airq, type="node")    # > where(airct)의 결과
와 동일
 [1] 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 3 5 6 9 9 6
```

의사결정나무

- 예측값을 이용하여 평균제곱오차 구하기

```
> mean((airq$Ozone - predict(airct))^2)
[1] 403.6668
```

의사결정나무

- 장점
  - 구조가 단순하여 해석이 용이
  - 유용한 입력변수의 파악과 예측변수간의 상호작용 및 비선형성을 고려하여 분석이 수행
  - 선형성, 정규성, 등분산성 등의 수학적 가정이 불필요한 비모수적 모형
  - 연속형 변수와 범주형 변수를 모두 취급할 수 있음

- 단점
  - 분류 기준값의 경계선 근방의 자료 값에 대해서는 오차가 클 수 있음(비연속성)
  - 로지스틱 회귀와 같이 각 예측변수의 효과를 파악하기 어려움
  - 새로운 자료에 대한 예측이 불안정할 수 있음
  - 일반적으로 너무 복잡한 나무모형은 예측력이 떨어지고 해석도 어렵고 계산량이 많을 수 있음
  - 분산이 매우 큰 불안정한 방법