# 데이터과학을 위한 R프로그래밍

## 11주차. 의사결정나무와 랜덤포레스트

**이혜선 교수**
포항공과대학교 산업경영공학과

# 목차

## 11주차. 의사결정나무와 랜덤포레스트

11주차

2차시

# 의사결정나무 II

POSTECH

# 의사결정나무 – rpart 패키지

## 의사결정나무 실행 패키지: rpart 패키지(tree패키지 외 사용)

```
# lec11_2_tree.R
# Decision tree
# use package rpart and party

# other package for tree
install.packages("rpart")
install.packages("party")
library(rpart)
library(party)

#package for confusion matrix
#install.packages("caret")
library(caret)
```

```
#decision tree : use rpart package
help("rpart")
```

rpart {rpart} R Documentation

### Recursive Partitioning and Regression Trees

**Description**

Fit a `rpart` model

**Usage**

```
rpart(formula, data, weights, subset, na.action = na.rpart, method,
      model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)
```

**Arguments**

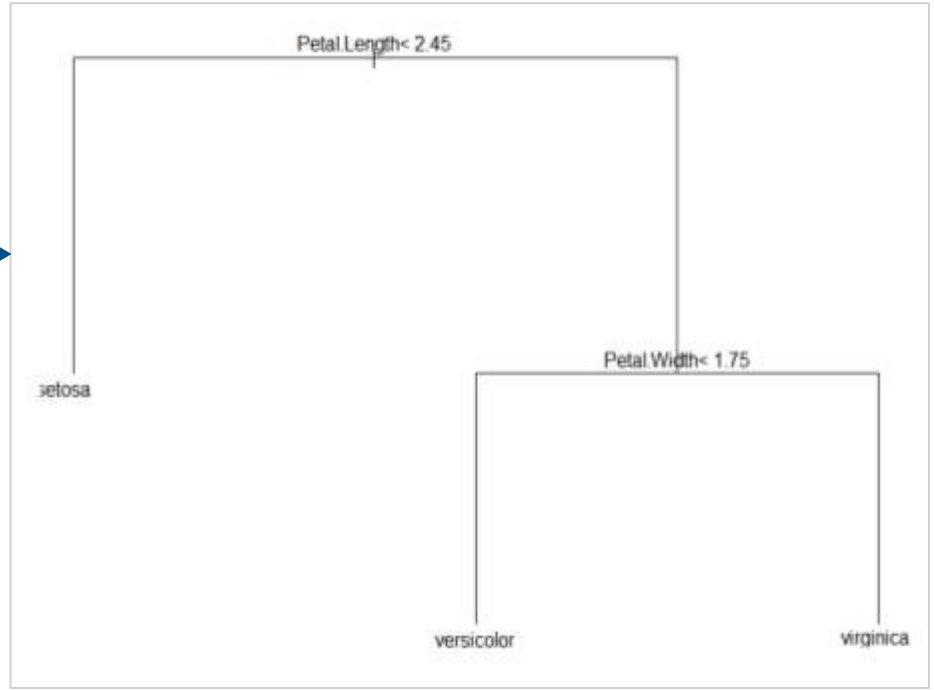
| | |
|---|---|
| `formula` | a formula, with a response but no interaction terms. If this a a data frame, that is taken as the model frame (see `model.frame`) . |
| `data` | an optional data frame in which to interpret the variables named in the formula. |
| `weights` | optional case weights. |
| `subset` | optional expression saying that only a subset of the rows of the data should be used in the fit. |

POSTECH

# 의사결정나무 – rpart 패키지

- 의사결정나무 함수 : rpart(종속변수~x1+x2+x3+x4, data= )

```
cl1<-rpart(Species~., data=train)
plot(cl1)
text(cl1, cex=1)
```




rpart 함수는 가지치기한 결과
→ 데이터에 따라 추가적인
가지치기가 필요할 수도 있음

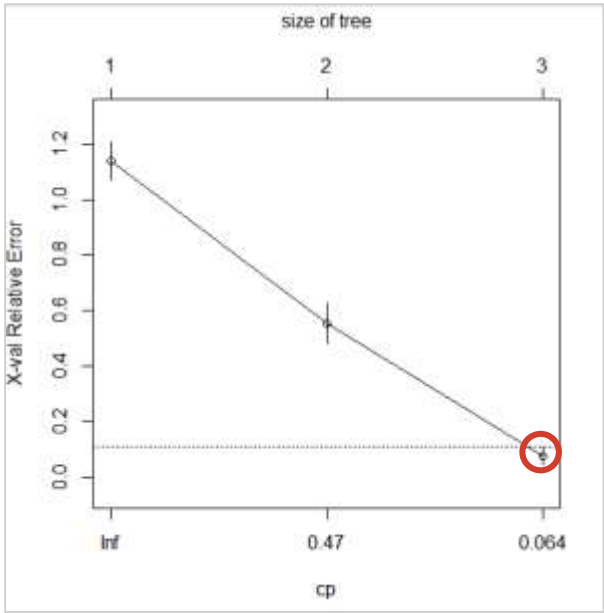
POSTECH

# 의사결정나무 – rpart 패키지

- rpart 패키지에서의 최적 트리모형 (cp(complexity parameter) 비교)
- printcp에서 xerror(cross validation error)의 값이 최소가 되는 트리를 선택

```
#pruning (cross-validation)-rpart
printcp(cl1)
plotcp(cl1)
help(printcp)
```

```
> printcp(cl1)

Classification tree:
rpart(formula = Species ~ ., data = train)

Variables actually used in tree construction:
[1] Petal.Length Petal.Width

Root node error: 65/100 = 0.65

n= 100

        CP nsplit rel error   xerror      xstd
1 0.52308      0   1.000000 1.138462 0.067482
2 0.41538      1   0.476923 0.553846 0.073846
3 0.01000      2   0.061538 0.076923 0.033530
```



POSTECH

# 의사결정나무 – rpart 패키지

## rpart 결과에서 복잡도계수에 기반한 최적 가지치기

```
#pruning (cross-validation)-rpart
printcp(cl1)
plotcp(cl1)
help(printcp)
```

cp(complexity parameter)

printcp {rpart} R Documentation

### Displays CP table for Fitted Rpart Object

**Description**

Displays the `cp` table for fitted `rpart` object.

**Usage**

```
printcp(x, digits = getOption("digits") - 2)
```

**Arguments**

| | |
|---|---|
| `x` | fitted model object of class `"rpart"`. This is assumed to be the result of some function that produces an object with the same named components as that returned by the `rpart` function. |
| `digits` | the number of digits of numbers to print. |

**Details**

Prints a table of optimal prunings based on a complexity parameter.

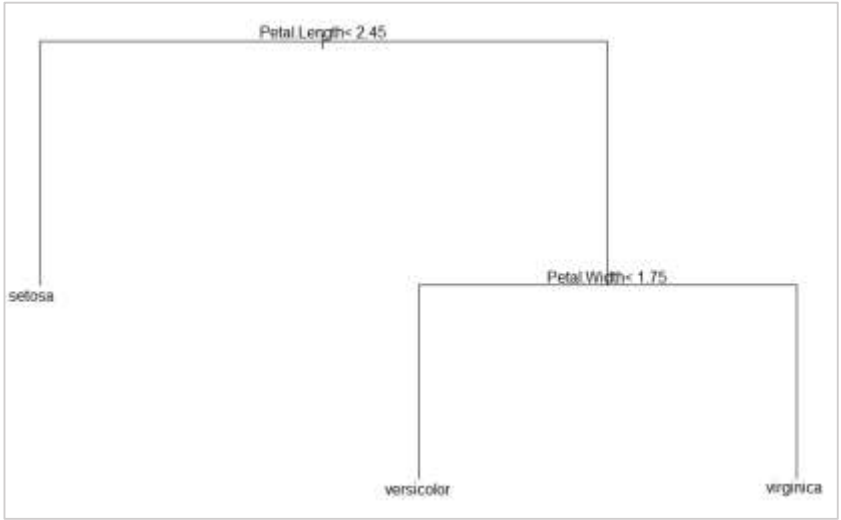
POSTECH

# 의사결정나무 – rpart 패키지

## rpart를 사용한 최종 tree모형

cp(complexity parameter)

```
#final tree model -rpart
pcl1<-prune(cl1, cp=cl1$cptable[which.min(cl1$cptable[,"xerror"]),"CP"])
plot(pcl1)
text(pcl1)
```




rpart를 이용한 최종 tree모형

POSTECH

# 의사결정나무 – rpart 패키지

- 의사결정나무 결과 정확도 : test data에 대한 정확도

```
#measure accuracy -rpart
pred2<- predict(pcl1,test, type='class')
confusionMatrix(pred2,test$Species)
```

```
> confusionMatrix(pred2,test$Species)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         16          0         0
  versicolor      0         15         2
  virginica       0          0        17

Overall Statistics

               Accuracy : 0.96
```

POSTECH

## 분류모형의 평가척도

```
Overall Statistics

               Accuracy : 0.96
                 95% CI : (0.8629, 0.9951)
    No Information Rate : 0.38
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.94

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor
Sensitivity                   1.00            1.0000
Specificity                   1.00            0.9429
Pos Pred Value                1.00            0.8824
Neg Pred Value                1.00            1.0000
Prevalence                    0.32            0.3000
Detection Rate                0.32            0.3000
Detection Prevalence          0.32            0.3400
Balanced Accuracy             1.00            0.9714
                     Class: virginica
Sensitivity                    0.8947
Specificity                    1.0000
```

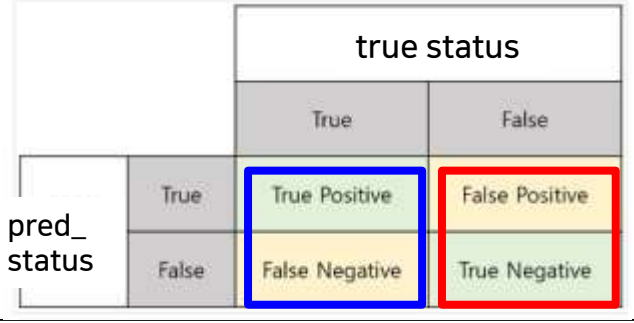
Accuracy

Sensitivity

Specificity

Sensitivity : 실제로 True인것을 True로 예측한 비율 = True Positive / (True Positive, False Negative)

Specificity : 실제로 False를 False로 예측한 비율= True Negative / (False Positive, True Negative)



POSTECH

# 의사결정나무 – party 패키지

## 의사결정나무 실행 패키지 : party 패키지(tree패키지 외 사용)

```
help(ctree)
```

Conditional Inference Trees {party} R Documentation

### Conditional Inference Trees

**Description**

Recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables in a conditional inference framework.

**Usage**

```
ctree(formula, data, subset = NULL, weights = NULL,
      controls = ctree_control(), xtrafo = ptrafo, ytrafo = ptrafo,
      scores = NULL)
```

**Arguments**

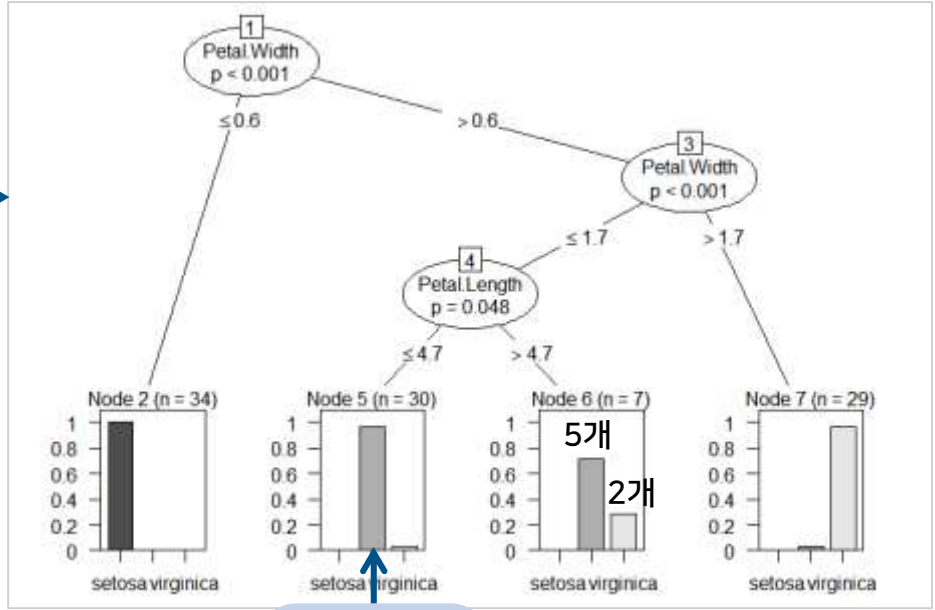
| | |
|---|---|
| `formula` | a symbolic description of the model to be fit. Note that symbols like `:` and `-` will not work and the tree will make use of all variables listed on the rhs of `formula`. |
| `data` | a data frame containing the variables in the model. |
| `subset` | an optional vector specifying a subset of observations to be used in the fitting process. |
| `weights` | an optional vector of weights to be used in the fitting process. Only non-negative integer valued weights are allowed. |

POSTECH

# 의사결정나무 – party 패키지

- 의사결정나무 함수 : ctree(종속변수~x1+x2+x3+x4, data= )

```
partymod<-ctree(Species~.,data=train)
plot(partymod)
partymod
```



POSTECH

# 의사결정나무 – party 패키지

## party 패키지를 이용한 결과

```
partymod<-ctree(Species~.,data=train)
plot(partymod)
partymod
```

```
> partymod

        Conditional inference tree with 4 terminal nodes

Response:  Species
Inputs:  Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Number of observations:  100

1) Petal.Width <= 0.6; criterion = 1, statistic = 92.056
  2)*  weights = 34
1) Petal.Width > 0.6
  3) Petal.Width <= 1.7; criterion = 1, statistic = 45.613
    4) Petal.Length <= 4.7; criterion = 0.952, statistic = 6.27
      5)*  weights = 30
    4) Petal.Length > 4.7
      6)*  weights = 7
  3) Petal.Width > 1.7
    7)*  weights = 29
```

POSTECH

# 의사결정나무 – party 패키지

- 의사결정나무결과 정확도 : test data에 대한 정확도

```
#measuring accuracy(party)
partypred<-predict(partymod,test)
confusionMatrix(partypred,test$Species)
```

```
> confusionMatrix(partypred,test$Species)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         16          0         0
  versicolor      0         15         2
  virginica       0          0        17

Overall Statistics

               Accuracy : 0.96
```