Chapter 9 –
Classification and Regression Trees

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Iris - 1

```
library(party)
str(iris)
#Compactly Display the Structure of an Arbitrary R Object
ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7,
0.3))
#Assign 1 or 2 to ind randomly with Prob(1)=0.7 and
Prob(2)=0.3
trainData <- iris[ind==1,]
testData <- iris[ind==2,]


myFormula <- Species ~ Sepal.Length + Sepal.Width +
Petal.Length + Petal.Width
iris_ctree <- ctree(myFormula, data=trainData)
# check the prediction
table(predict(iris_ctree), trainData$Species)
```

Iris - 2

```
print(iris_ctree)

plot(iris_ctree)

plot(iris_ctree, type="simple")

# predict on test data


testPred <- predict(iris_ctree, newdata = testData)

table(testPred, testData$Species)
```

Bodyfat - 1

```
library(TH.data)
library(rpart)
data("bodyfat", package="TH.data")
dim(bodyfat)
bodyfat[1:5,]

set.seed(1234)
ind <- sample(2, nrow(bodyfat), replace=TRUE, prob=c(0.7,
0.3))
bodyfat.train <- bodyfat[ind==1,]
bodyfat.test <- bodyfat[ind==2,]
```

Bodyfat - 2

```
# train a decision tree
myFormula <- DEXfat ~ age + waistcirc + hipcirc +
elbowbreadth + kneebreadth
bodyfat_rpart <- rpart(myFormula, data = bodyfat.train,
                       control = rpart.control(minsplit = 10))
#minsplit: the minimum number of observations that
must exist in a node in order for a split to be attempted.
attributes(bodyfat_rpart)

?rpart.object
print(bodyfat_rpart$cptable)
#a matrix of information on the optimal prunings based
on a complexity parameter.
```

Bodyfat - 3

```
print(bodyfat_rpart)
plot(bodyfat_rpart)
text(bodyfat_rpart, use.n=T)
#以下对决策树进行剪枝。
opt <- which.min(bodyfat_rpart$cptable[,"xerror"])
#通过cptable观察哪个树的xerror值最小。
cp <- bodyfat_rpart$cptable[opt, "CP"]
#选择xerror最小的树，找到它所对应的cp值。
bodyfat_prune <- prune(bodyfat_rpart, cp = cp)
#根据上面得到的cp值进行剪枝。
print(bodyfat_prune)
#剪枝之后得到的决策树是xerror值最小的决策树。
plot(bodyfat_prune)
text(bodyfat_prune, use.n=T)
```