Lab 9 Classification and Regression Trees (CART)

The goal of this lab is to become familiar with Classification and Regression (CART). CART allows you to determine what variables are important in separating groups. It differs from DA in that it does not have the assumptions of a parametric test and thus provides a more flexible approach to discrimination.

# Set up R session

## Data

Today you will be using the iris data set found in the package datasets in R. To access this data set, simply type:

iris

To learn more about the data set:

?iris

You will use the “ozone.csv” avaliable in Module 9 in CANVAS for regression trees.

oz<-read.csv("../Data/lab\_9/ozone.csv",header=TRUE)

## Download packages

We will be using the following packages:

library(MASS)  
library(rpart)  
library(ade4)  
library(vegan)  
library(rattle)

# CART

## Setting the training and testing data

Like you did for DA, you are going to split our data set into “training” data and “testing” data. This will allow you to test the predictions of the CART model on a “new” data set.

Randomly select 75 samples from the iris data set. Use “set.seed” so we are all working with the same training set:

set.seed(51)  
train <- sample(1:150, 75)

Check the frequency of each species in the training data, to make sure they are relatively proportional to the frequency in the complete data set.

freq<-table(iris$Sp[train])  
freq

##   
## setosa versicolor virginica   
## 27 29 19

## Specifying the model

Next specify the CART model. In the iris data set, Species is the categorical response variables and the four measures of flower morphology are the explanatory variables:

model <- Species ~ .

**“Species ~ .” is the model and the “dot” stands for all of the variables (so you don’t have to type them all)**

## Running the CART algorithm

You will use the rpart function in the *rpart* package to develop CART models.

?rpart

iris\_rpart <- rpart(model, data = iris[train,], method="class", control = rpart.control(minsplit = 10))

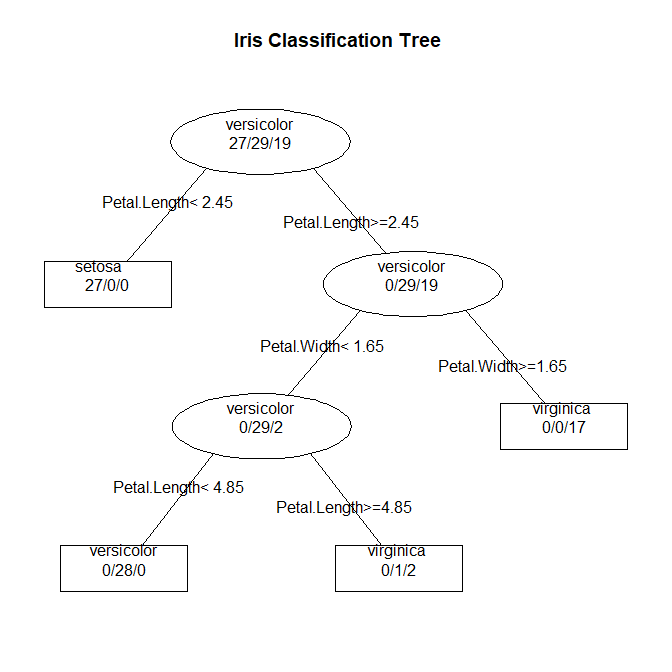
**Look up what rpart.control and it’s parameters do and play with them!**

## Plotting Cart Tree and viewing summary

You will plot your tree using the function post.rpart:

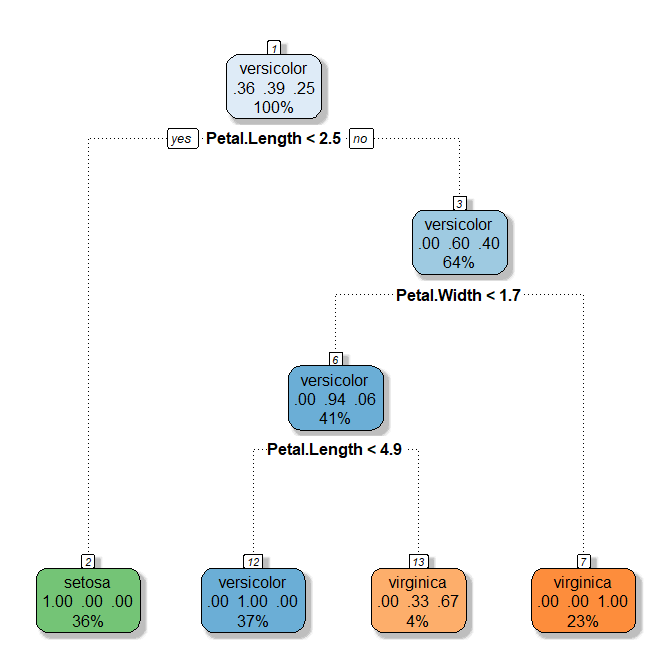
?post.rpart

post(iris\_rpart, file = "", title = " Iris Classification Tree")



Fancier plot!

fancyRpartPlot(iris\_rpart, sub= " " )



Now, look at a node by node summary of the tree and the variable importance:

summary(iris\_rpart)

## Cost-complexity pruning

Now let’s look at the cross-validation results:

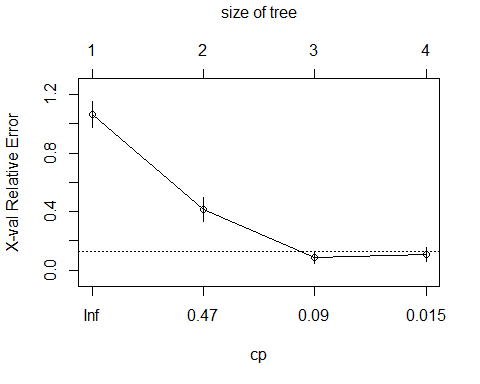
printcp(iris\_rpart)

##   
## Classification tree:  
## rpart(formula = model, data = iris[train, ], method = "class",   
## control = rpart.control(minsplit = 10))  
##   
## Variables actually used in tree construction:  
## [1] Petal.Length Petal.Width   
##   
## Root node error: 46/75 = 0.61333  
##   
## n= 75   
##   
## CP nsplit rel error xerror xstd  
## 1 0.586957 0 1.000000 1.065217 0.089598  
## 2 0.369565 1 0.413043 0.413043 0.081881  
## 3 0.021739 2 0.043478 0.086957 0.042303  
## 4 0.010000 3 0.021739 0.108696 0.046962

* The first column is the complexity parameter (CP) which shows how much each node contributes to the classification rate and thus how misclassification would increase with the removal of a given node.
* The second column shows the split.
* The third column shows the relative error (misclassification rate). This is the proportion of samples misclassified in the “root node” (first node) that are misclassified at any subsequent node. To get the absolute error, multiply these values times the misclassification rate of the first node.
* The fourth column is the estimated error from the cross-validation procedure.
* The fifth column is the standard error from the cross-validation procedure.

Next, plot the results to determine the optimal tree:

plotcp(iris\_rpart)



Remember, you are going to use the 1SE rule of thumb. Select the tree size furthers to the left (i.e. fewest leaves) that is within 1 SE of the minimum estimated error.

How big is you optimal tree?

## Pruning the tree

You will prune the tree using prune.rpart function.

?prune.rpart

First let’s set the cp for the optimal tree size. This extracts the cp value according to the 1-SE rule. *Note you will have to change the column, row designations for other examples*.

cp<-printcp(iris\_rpart)[3,1]

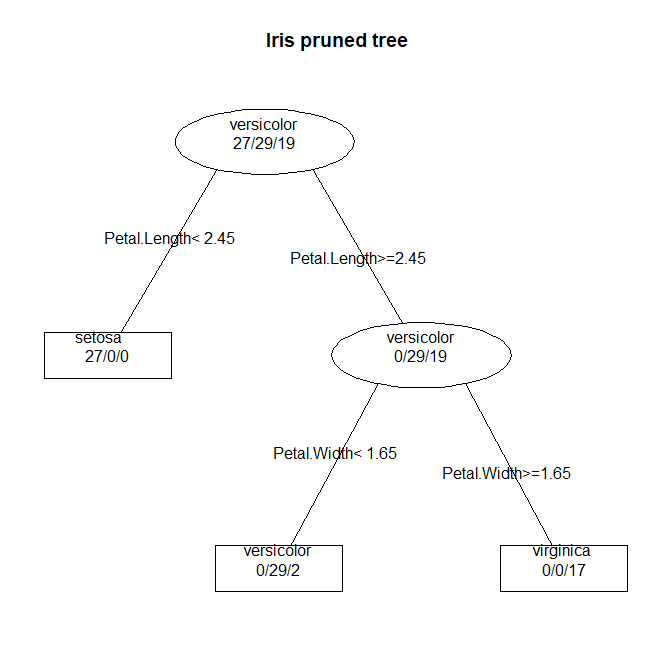
##   
## Classification tree:  
## rpart(formula = model, data = iris[train, ], method = "class",   
## control = rpart.control(minsplit = 10))  
##   
## Variables actually used in tree construction:  
## [1] Petal.Length Petal.Width   
##   
## Root node error: 46/75 = 0.61333  
##   
## n= 75   
##   
## CP nsplit rel error xerror xstd  
## 1 0.586957 0 1.000000 1.065217 0.089598  
## 2 0.369565 1 0.413043 0.413043 0.081881  
## 3 0.021739 2 0.043478 0.086957 0.042303  
## 4 0.010000 3 0.021739 0.108696 0.046962

Now, get out the pruning shears:

iris\_prune <- prune(iris\_rpart, cp = cp)  
print(iris\_prune)

## n= 75   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 75 46 versicolor (0.36000000 0.38666667 0.25333333)   
## 2) Petal.Length< 2.45 27 0 setosa (1.00000000 0.00000000 0.00000000) \*  
## 3) Petal.Length>=2.45 48 19 versicolor (0.00000000 0.60416667 0.39583333)   
## 6) Petal.Width< 1.65 31 2 versicolor (0.00000000 0.93548387 0.06451613) \*  
## 7) Petal.Width>=1.65 17 0 virginica (0.00000000 0.00000000 1.00000000) \*

#And plot the pruned tree:  
post(iris\_prune, file = "", title = " Iris pruned tree")



## Classification accuracy

Now we have our optimum tree. Let’s see how accurately both the unpruned and pruned tree classify the testing data:

#classification matrix  
Ct\_unprune<-table(predict(iris\_rpart, iris[-train,], type = "class"), iris[-train, "Species"])  
Ct\_prune<-table(predict(iris\_prune, iris[-train,], type = "class"), iris[-train, "Species"])  
  
Ct\_unprune

##   
## setosa versicolor virginica  
## setosa 23 0 0  
## versicolor 0 17 0  
## virginica 0 4 31

Ct\_prune

##   
## setosa versicolor virginica  
## setosa 23 0 0  
## versicolor 0 19 2  
## virginica 0 2 29

#classification accuracy  
class\_unprune<-sum(diag(prop.table(Ct\_unprune)))  
class\_prune<-sum(diag(prop.table(Ct\_prune)))  
  
class\_unprune

## [1] 0.9466667

class\_prune

## [1] 0.9466667

**Did pruning help much?**

# Compare with DA

Pull up your code from last week and run a DA for the iris data. Compare classification rates and variable importance.

# Regression Tree

Here we will use regression trees on the “ozone.csv” avaliable in Module 9 in CANVAS. This data set contains information on ozone, radiation, temperature, and wind on 110 days from May to September 1973 in New York. Measurements of daily ozone con- centration (ppb), wind speed (mph), daily maximum temperature (degrees F), and solar radiation (langleys). Here, you will use ozone and the response variable and the remaining variable as predicators (i.e., used to make splits in the tree).

Create a training data set with 60 samples:

train <- sample(1:110, 60)

Specify the model:

model <- ozone ~ .

Utilize rpart to construct the tree:

oz\_rpart <- rpart(model, data = oz[train,],method="anova", control = rpart.control(minsplit = 10))

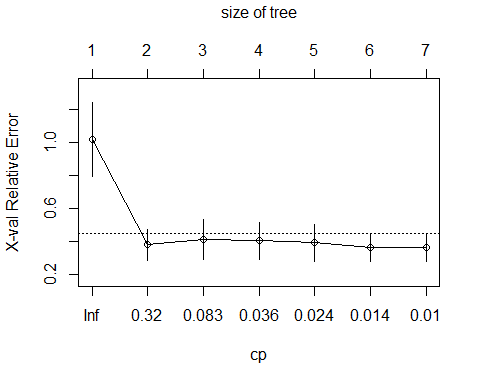
Here we denote the type of tree using method = “anova”, which is used for continues variables.

Now let’s look at the cross-validation results:

printcp(oz\_rpart)

##   
## Regression tree:  
## rpart(formula = model, data = oz[train, ], method = "anova",   
## control = rpart.control(minsplit = 10))  
##   
## Variables actually used in tree construction:  
## [1] radiation temp wind   
##   
## Root node error: 56993/60 = 949.88  
##   
## n= 60   
##   
## CP nsplit rel error xerror xstd  
## 1 0.646046 0 1.000000 1.02048 0.223321  
## 2 0.155438 1 0.353954 0.37812 0.095144  
## 3 0.043976 2 0.198517 0.41128 0.122348  
## 4 0.029682 3 0.154540 0.40140 0.112500  
## 5 0.018615 4 0.124859 0.39032 0.112259  
## 6 0.010583 5 0.106244 0.35884 0.085449  
## 7 0.010000 6 0.095661 0.36316 0.085386

plotcp(oz\_rpart)



cp<-printcp(oz\_rpart)[3,1]

##   
## Regression tree:  
## rpart(formula = model, data = oz[train, ], method = "anova",   
## control = rpart.control(minsplit = 10))  
##   
## Variables actually used in tree construction:  
## [1] radiation temp wind   
##   
## Root node error: 56993/60 = 949.88  
##   
## n= 60   
##   
## CP nsplit rel error xerror xstd  
## 1 0.646046 0 1.000000 1.02048 0.223321  
## 2 0.155438 1 0.353954 0.37812 0.095144  
## 3 0.043976 2 0.198517 0.41128 0.122348  
## 4 0.029682 3 0.154540 0.40140 0.112500  
## 5 0.018615 4 0.124859 0.39032 0.112259  
## 6 0.010583 5 0.106244 0.35884 0.085449  
## 7 0.010000 6 0.095661 0.36316 0.085386

cp

## [1] 0.0439761

Now, get out the pruning shears:

oz\_prune <- prune(oz\_rpart, cp = cp)  
print(oz\_prune)

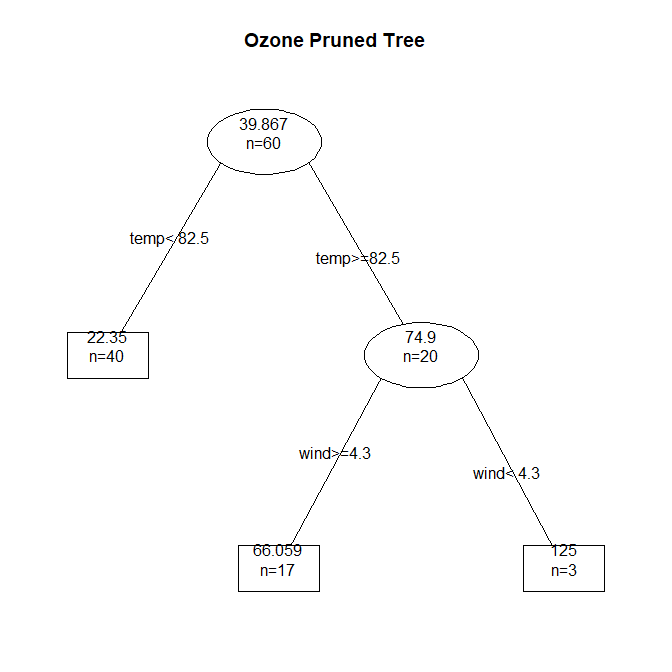
## n= 60   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 60 56992.930 39.86667   
## 2) temp< 82.5 40 6271.100 22.35000 \*  
## 3) temp>=82.5 20 13901.800 74.90000   
## 6) wind>=4.3 17 4884.941 66.05882 \*  
## 7) wind< 4.3 3 158.000 125.00000 \*

summary(oz\_prune)

## Call:  
## rpart(formula = model, data = oz[train, ], method = "anova",   
## control = rpart.control(minsplit = 10))  
## n= 60   
##   
## CP nsplit rel error xerror xstd  
## 1 0.6460456 0 1.0000000 1.0204837 0.22332089  
## 2 0.1554378 1 0.3539544 0.3781170 0.09514399  
## 3 0.0439761 2 0.1985166 0.4112816 0.12234802  
##   
## Variable importance  
## temp wind radiation   
## 59 35 6   
##   
## Node number 1: 60 observations, complexity param=0.6460456  
## mean=39.86667, MSE=949.8822   
## left son=2 (40 obs) right son=3 (20 obs)  
## Primary splits:  
## temp < 82.5 to the left, improve=0.6460456, (0 missing)  
## wind < 6.6 to the right, improve=0.5092466, (0 missing)  
## radiation < 157 to the left, improve=0.2805027, (0 missing)  
## Surrogate splits:  
## wind < 5.7 to the right, agree=0.783, adj=0.35, (0 split)  
## radiation < 157 to the left, agree=0.700, adj=0.10, (0 split)  
##   
## Node number 2: 40 observations  
## mean=22.35, MSE=156.7775   
##   
## Node number 3: 20 observations, complexity param=0.1554378  
## mean=74.9, MSE=695.09   
## left son=6 (17 obs) right son=7 (3 obs)  
## Primary splits:  
## wind < 4.3 to the right, improve=0.63724550, (0 missing)  
## temp < 86.5 to the left, improve=0.24521220, (0 missing)  
## radiation < 272 to the right, improve=0.06123428, (0 missing)  
##   
## Node number 6: 17 observations  
## mean=66.05882, MSE=287.3495   
##   
## Node number 7: 3 observations  
## mean=125, MSE=52.66667

Plot pruned tree:

post(oz\_prune, file = "", title = "Ozone Pruned Tree")

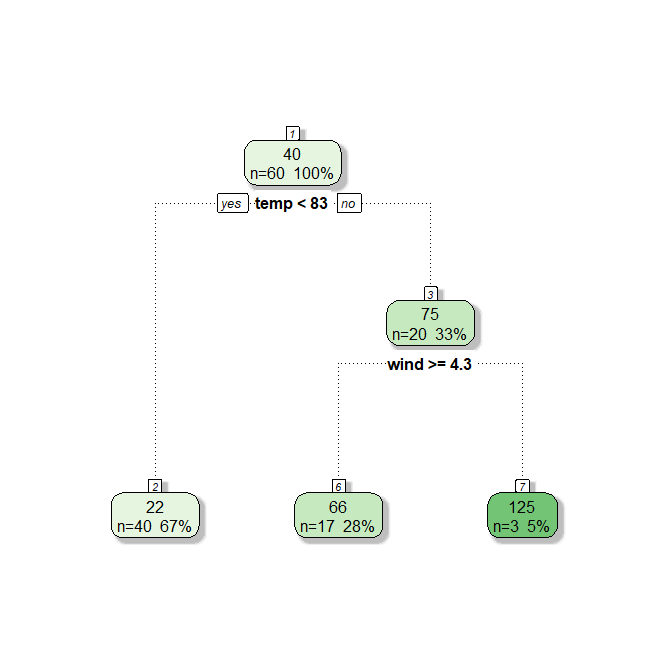


sum(residuals(oz\_prune)^2)

## [1] 11314.04

Fancier plot!

fancyRpartPlot(oz\_prune, sub= " " )



Predict with both trees:

pruned<-predict(oz\_prune, oz[-train,],type = "vector")  
  
full<-predict(oz\_rpart, oz[-train,],type = "vector")