26 BE 7024 & 26 PH 7024

Computational Statistics

Autumn, 2014

MB Rao

Homework Sheet No. 5 Due Date: November 25, 2014 Maximum Points: 30

Classification Trees

Work with the Wheat data.

1. Build a classification (decision) tree. 3 points

> setwd('/Users/aishwaryakulkarni/Documents/Comp\_Stats')

> getwd()

[1] "/Users/aishwaryakulkarni/Documents/Comp\_Stats"

> Wheat <- read.delim("wheat.txt")

> head(Wheat)

class density hardness size weight moisture type

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy

> Wheat1 <- rpart(type ~ ., data = Wheat)

Error: could not find function "rpart"

> install.packages("rpart")

trying URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/rpart\_4.1-8.tgz'

Content type 'application/x-gzip' length 888653 bytes (867 Kb)

opened URL

==================================================

downloaded 867 Kb

The downloaded binary packages are in

/var/folders/sy/bj43l9rj7t9gz45yx9h\_hcbr0000gn/T//Rtmp0ZIrIb/downloaded\_packages

> loadModule(rpart)

Error: could not find function "loadModule"

> module load(rpart)

Error: unexpected symbol in "module load"

> module load("rpart")

Error: unexpected symbol in "module load"

> module load rpart

Error: unexpected symbol in "module load"

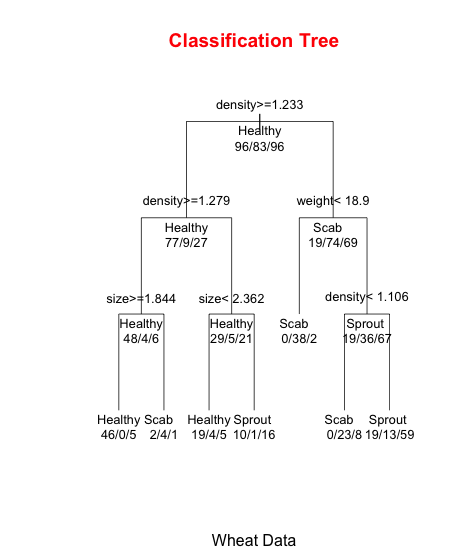
> library("rpart")

> Wheat1 <- rpart(type ~ ., data = Wheat)

> plot(Wheat1, uniform = T, margin = 0.1)

> text(Wheat1, all = T, use.n = T, cex = 0.8)

> title(main = "Classification Tree", col.main = "red", sub = "Wheat Data")



1. Provide a tight verbal description of how the patterns are identified in the decision tree. 5 points

7 terminal nodes, originally 96 H,83 Sp,96 Sc, best predictor is density and best cut point is 1.233 and hence the first split point is at this value of the predictor.

1. Carry out the command ‘predict’ for the entire data. Show the first six predictions. 3 points

Wheat2 <- predict(Wheat1, newdata = Wheat)

> head(Wheat2)

Healthy Scab Sprout

1 0.9019608 0.00000000 0.09803922

2 0.9019608 0.00000000 0.09803922

3 0.3703704 0.03703704 0.59259259

4 0.9019608 0.00000000 0.09803922

5 0.6785714 0.14285714 0.17857143

6 0.9019608 0.00000000 0.09803922

1. Build a confusion matrix and interpret it. 5 points

> Wheat$Prediction<-Wheat4

> head(Wheat)

class density hardness size weight moisture type Prediction

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy Healthy

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy Healthy

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy Sprout

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy Healthy

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy Healthy

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy Healthy

> kyphosis$Disagree <- ifelse(Wheat$type != Wheat$Prediction, 1, 0)

Error in `$<-.data.frame`(`\*tmp\*`, "Disagree", value = c(0, 0, 1, 0, 0, :

replacement has 275 rows, data has 81

> head(Wheat)

class density hardness size weight moisture type Prediction

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy Healthy

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy Healthy

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy Sprout

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy Healthy

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy Healthy

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy Healthy

> kyphosis$Disagree <- ifelse(Wheat$type == Wheat$Prediction, 0, 1)

Error in `$<-.data.frame`(`\*tmp\*`, "Disagree", value = c(0, 0, 1, 0, 0, :

replacement has 275 rows, data has 81

> Wheat$Disagree <- ifelse(Wheat$type != Wheat$Prediction, 1, 0)

> head(Wheat)

class density hardness size weight moisture type Prediction Disagree

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy Healthy 0

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy Healthy 0

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy Sprout 1

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy Healthy 0

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy Healthy 0

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy Healthy 0

1. Calculate the misclassification rate. 2 points

(70/275)\*100

[1] 25.45455

1. You have already fit a multinomial logistic regression model for the data. What were the significant predictors? Compare and contrast the multinomial model and classification tree. 5 points

The misclassification rate for multinomial model is

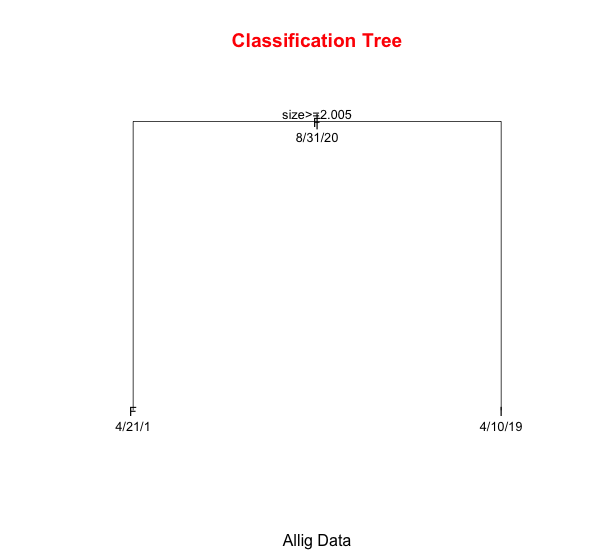
> (78/275)\*100

[1] 28.36364

density , hardness and weight are significant predictors.

1. Look at the alligator data from one of the previous lecture notes. There is only one predictor. The response variable is ternary. Would a classification tree work? Explore. 7 points

It is not possible to construct a decision tree where the response variable is ternary using only single predictor.



It is ideally possible to construct a decision tree from a single predictor with a response variable, which is ternary, except in this case the ‘Other’ category is not important enough and the size is small such that it is not assigned an individual split in this case.