Homework6_Hwang

Charles Hwang

4/13/2022

Charles Hwang

Dr. Perry

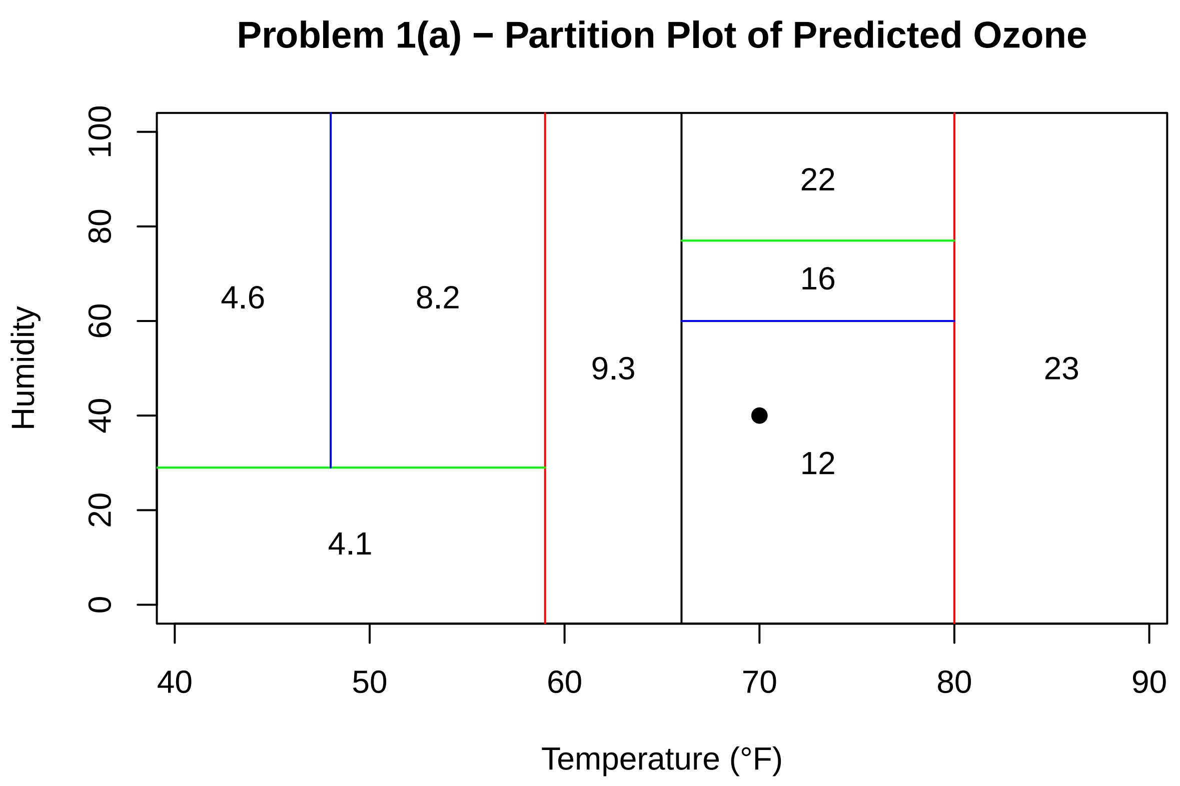
STAT 451-001

13 April 2022

Problem 1

Problem 1(a)

```
rm(list=ls())
plot(70,40,xlim=c(41,89),ylim=c(0,100),pch=19,xlab="Temperature (°F)",ylab="Humidity",main="Problem 1(a)
abline(v=66)                                                   # Branch 1
abline(v=c(59,80),col="red")                                   # Branch 2
segments(c(0,66),c(29,77),c(59,80),c(29,77),col="green")  # Branch 3
segments(c(48,66),c(29,60),c(48,80),c(110,60),col="blue") # Branch 4
text(49,13,"4.1")
text(43.5,65,"4.6")
text(53.5,65,"8.2")
text(62.5,50,"9.3")
text(73,30,"12")
text(73,69,"16")
text(73,90,"22")
text(85.5,50,"23")
```



Problem 1(b)

We can see from the partition plot in Problem 1(a) that the predicted ozone concentration is **12** for a temperature of 70°F and humidity of 40.

Problem 2

```
bl<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 451 - Nonparametric Statistical Methods/BankLoa
bl$Personal.Loan<-as.factor(bl$Personal.Loan)
library(rpart)
library(tree)
set.seed(1304,sample.kind="Rounding")                              # Problem 2(a)
s<-sample(1:nrow(bl),nrow(bl)*.7)
train<-bl[s,]
test<-bl[-s,]
rpart<-rpart(Personal.Loan~.,data=train,method="class")            # Problem 2(b)
summary(rpart)
```

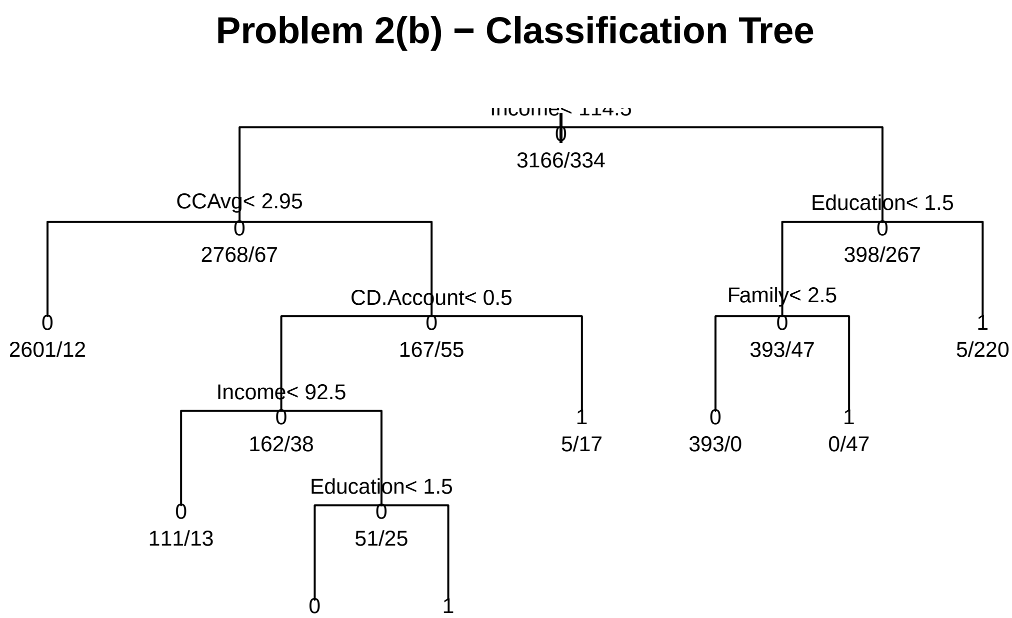
```
## Call:
## rpart(formula = Personal.Loan ~ ., data = train, method = "class")
##   n= 3500
##
##           CP nsplit rel error    xerror       xstd
## 1 0.32185629      0 1.0000000 1.0000000 0.05204131
## 2 0.14071856      2 0.3562874 0.4491018 0.03587464
## 3 0.01796407      3 0.2155689 0.2514970 0.02710929
## 4 0.01497006      5 0.1796407 0.2275449 0.02581625
## 5 0.01000000      7 0.1497006 0.2215569 0.02548174
##
```

```
## Variable importance
##  Education     Income     Family      CCAvg CD.Account   Mortgage
##         33         23         20         13          6          3
##
## Node number 1: 3500 observations,    complexity param=0.3218563
##   predicted class=0  expected loss=0.09542857  P(node) =1
##     class counts:  3166   334
##    probabilities: 0.905 0.095
##   left son=2 (2835 obs) right son=3 (665 obs)
##   Primary splits:
##       Income     < 114.5    to the left,  improve=153.82360, (0 missing)
##       CCAvg      < 2.95     to the left,  improve=105.84010, (0 missing)
##       CD.Account < 0.5      to the left,  improve= 57.10933, (0 missing)
##       Mortgage   < 280.5    to the left,  improve= 25.42160, (0 missing)
##       Education  < 1.5      to the left,  improve= 12.84149, (0 missing)
##   Surrogate splits:
##       CCAvg    < 4.05     to the left,  agree=0.885, adj=0.392, (0 split)
##       Mortgage < 336.5    to the left,  agree=0.827, adj=0.090, (0 split)
##
## Node number 2: 2835 observations,    complexity param=0.01796407
##   predicted class=0  expected loss=0.02363316  P(node) =0.81
##     class counts:  2768    67
##    probabilities: 0.976 0.024
##   left son=4 (2613 obs) right son=5 (222 obs)
##   Primary splits:
##       CCAvg              < 2.95     to the left,  improve=24.195630, (0 missing)
##       Income             < 92.5     to the left,  improve=13.533180, (0 missing)
##       CD.Account         < 0.5      to the left,  improve= 4.859985, (0 missing)
##       Mortgage           < 298      to the left,  improve= 1.997850, (0 missing)
##       Securities.Account < 0.5      to the left,  improve= 0.243782, (0 missing)
##
## Node number 3: 665 observations,    complexity param=0.3218563
##   predicted class=0  expected loss=0.4015038  P(node) =0.19
##     class counts:   398   267
##    probabilities: 0.598 0.402
##   left son=6 (440 obs) right son=7 (225 obs)
##   Primary splits:
##       Education  < 1.5      to the left,  improve=225.860100, (0 missing)
##       Family     < 2.5      to the left,  improve=135.018000, (0 missing)
##       CD.Account < 0.5      to the left,  improve= 39.164190, (0 missing)
##       CCAvg      < 6.633333 to the right, improve= 11.196190, (0 missing)
##       Income     < 156.5    to the left,  improve=  6.314526, (0 missing)
##   Surrogate splits:
##       Family     < 2.5      to the left,  agree=0.749, adj=0.258, (0 split)
##       CD.Account < 0.5      to the left,  agree=0.707, adj=0.133, (0 split)
##       CCAvg      < 8.9      to the left,  agree=0.671, adj=0.027, (0 split)
##       Mortgage   < 529.5    to the left,  agree=0.669, adj=0.022, (0 split)
##       Income     < 116.5    to the right, agree=0.666, adj=0.013, (0 split)
##
## Node number 4: 2613 observations
##   predicted class=0  expected loss=0.004592423  P(node) =0.7465714
##     class counts:  2601    12
##    probabilities: 0.995 0.005
##
```

```
## Node number 5: 222 observations,    complexity param=0.01796407
##   predicted class=0  expected loss=0.2477477  P(node) =0.06342857
##     class counts:   167    55
##    probabilities: 0.752 0.248
##   left son=10 (200 obs) right son=11 (22 obs)
##   Primary splits:
##       CD.Account < 0.5      to the left,  improve=13.460480, (0 missing)
##       Income     < 90.5     to the left,  improve= 8.149924, (0 missing)
##       Education  < 1.5      to the left,  improve= 3.252252, (0 missing)
##       Family     < 2.5      to the left,  improve= 2.247064, (0 missing)
##       Experience < 36.5     to the left,  improve= 1.541951, (0 missing)
##   Surrogate splits:
##       Age < 64.5     to the left,  agree=0.905, adj=0.045, (0 split)
##
## Node number 6: 440 observations,    complexity param=0.1407186
##   predicted class=0  expected loss=0.1068182  P(node) =0.1257143
##     class counts:   393    47
##    probabilities: 0.893 0.107
##   left son=12 (393 obs) right son=13 (47 obs)
##   Primary splits:
##       Family     < 2.5      to the left,  improve=83.959090, (0 missing)
##       CD.Account < 0.5      to the left,  improve=10.459300, (0 missing)
##       CCAvg      < 6.633333 to the right, improve= 2.762938, (0 missing)
##       Mortgage   < 189      to the left,  improve= 2.104018, (0 missing)
##       ZIP.Code   < 95057    to the left,  improve= 1.110451, (0 missing)
##   Surrogate splits:
##       Mortgage   < 566      to the left,  agree=0.895, adj=0.021, (0 split)
##       CD.Account < 0.5      to the left,  agree=0.895, adj=0.021, (0 split)
##
## Node number 7: 225 observations
##   predicted class=1  expected loss=0.02222222  P(node) =0.06428571
##     class counts:     5   220
##    probabilities: 0.022 0.978
##
## Node number 10: 200 observations,    complexity param=0.01497006
##   predicted class=0  expected loss=0.19  P(node) =0.05714286
##     class counts:   162    38
##    probabilities: 0.810 0.190
##   left son=20 (124 obs) right son=21 (76 obs)
##   Primary splits:
##       Income    < 92.5     to the left,  improve=4.733175, (0 missing)
##       Education < 1.5      to the left,  improve=3.849829, (0 missing)
##       Family    < 2.5      to the left,  improve=1.605397, (0 missing)
##       Age       < 29.5     to the right, improve=1.601667, (0 missing)
##       Online    < 0.5      to the right, improve=1.600584, (0 missing)
##   Surrogate splits:
##       CCAvg      < 4.05     to the left,  agree=0.700, adj=0.211, (0 split)
##       ZIP.Code   < 90718.5  to the right, agree=0.665, adj=0.118, (0 split)
##       Mortgage   < 248.5    to the left,  agree=0.660, adj=0.105, (0 split)
##       Age        < 58.5     to the left,  agree=0.650, adj=0.079, (0 split)
##       Experience < 32.5     to the left,  agree=0.645, adj=0.066, (0 split)
##
## Node number 11: 22 observations
##   predicted class=1  expected loss=0.2272727  P(node) =0.006285714
```

```
##     class counts:     5    17
##    probabilities: 0.227 0.773
##
## Node number 12: 393 observations
##   predicted class=0  expected loss=0  P(node) =0.1122857
##     class counts:   393     0
##    probabilities: 1.000 0.000
##
## Node number 13: 47 observations
##   predicted class=1  expected loss=0  P(node) =0.01342857
##     class counts:     0    47
##    probabilities: 0.000 1.000
##
## Node number 20: 124 observations
##   predicted class=0  expected loss=0.1048387  P(node) =0.03542857
##     class counts:   111    13
##    probabilities: 0.895 0.105
##
## Node number 21: 76 observations,    complexity param=0.01497006
##   predicted class=0  expected loss=0.3289474  P(node) =0.02171429
##     class counts:    51    25
##    probabilities: 0.671 0.329
##   left son=42 (42 obs) right son=43 (34 obs)
##   Primary splits:
##       Education < 1.5      to the left,  improve=12.451790, (0 missing)
##       Family    < 2.5      to the left,  improve= 8.343401, (0 missing)
##       CCAvg     < 4.25     to the right, improve= 3.219531, (0 missing)
##       Income    < 104.5    to the right, improve= 2.860755, (0 missing)
##       Online    < 0.5      to the right, improve= 1.558187, (0 missing)
##   Surrogate splits:
##       Family < 2.5      to the left,  agree=0.711, adj=0.353, (0 split)
##       Online < 0.5      to the right, agree=0.658, adj=0.235, (0 split)
##       Age    < 60.5     to the left,  agree=0.645, adj=0.206, (0 split)
##       Income < 102.5    to the right, agree=0.645, adj=0.206, (0 split)
##       CCAvg  < 4.25     to the right, agree=0.645, adj=0.206, (0 split)
##
## Node number 42: 42 observations
##   predicted class=0  expected loss=0.07142857  P(node) =0.012
##     class counts:    39     3
##    probabilities: 0.929 0.071
##
## Node number 43: 34 observations
##   predicted class=1  expected loss=0.3529412  P(node) =0.009714286
##     class counts:    12    22
##    probabilities: 0.353 0.647
```

```r
plot(rpart,uniform=TRUE,main="Problem 2(b) - Classification Tree")
text(rpart,use.n=TRUE,all=TRUE,cex=0.7)
```




```
mean(predict(rpart,newdata=train,type="class")==train$Personal.Loan) # Problem 2(c)
```

```
## [1] 0.9857143
```

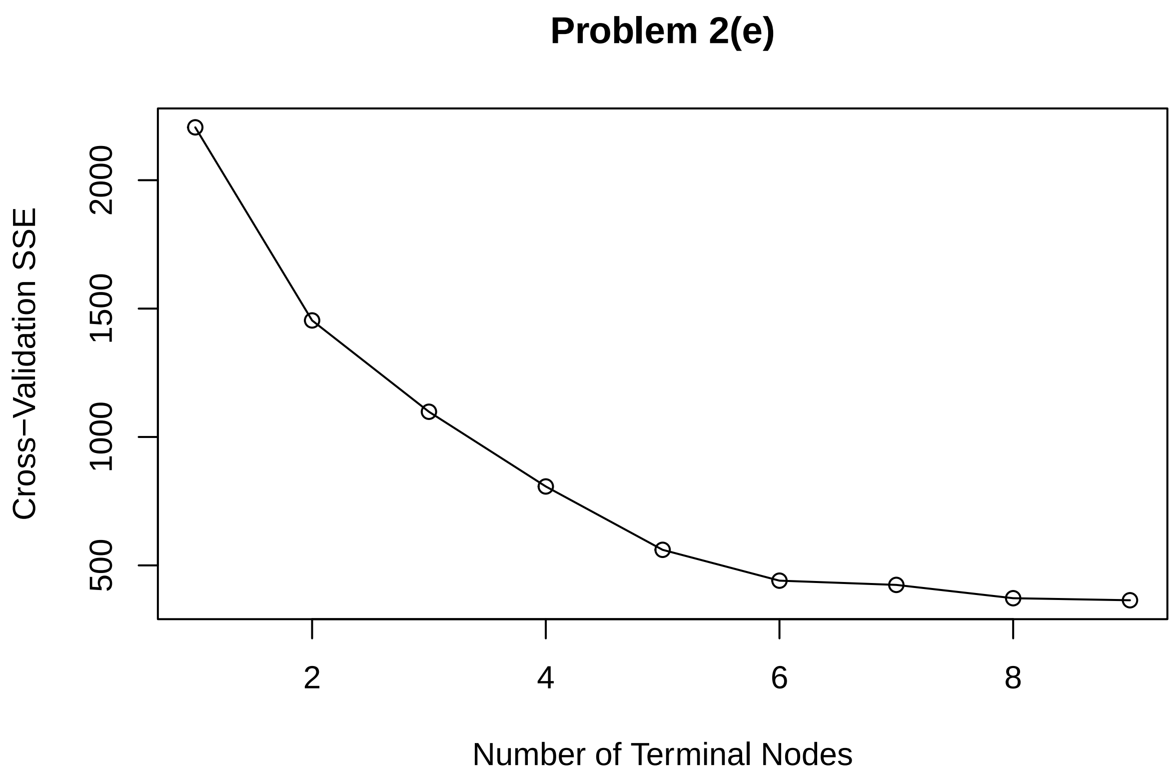
```
# We can see the accuracy of this initial model is approximately 98.57143 percent
# and that there are eight terminal nodes included in the classification tree.
mean(predict(rpart,newdata=test,type="class")==test$Personal.Loan)   # Problem 2(d)
```

```
## [1] 0.99
```

```
table(Predicted=predict(rpart,newdata=test,type="class"),Actual=test$Personal.Loan)
```

```
##          Actual
## Predicted    0    1
##         0 1349   10
##         1    5  136
```

```
cv<-cv.tree(tree(Personal.Loan~.,data=train))                          # Problem 2(e)
plot(cv$size,cv$dev,type="o",xlab="Number of Terminal Nodes",ylab="Cross-Validation SSE",main="Problem 2
```




```
# I believe the optimal number of nodes is 5. After pruning the tree to 5, 6, and 8 nodes,
# I saw there was very little loss of accuracy when using 5 nodes compared to 6 or 8.
ptree<-prune.tree(tree(Personal.Loan~.,data=train),best=5)          # Problem 2(f)
mean(predict(ptree,newdata=train,type="class")==train$Personal.Loan)
```

```
## [1] 0.9754286
```

```
mean(predict(ptree,newdata=test,type="class")==test$Personal.Loan)  # Problem 2(g)
```

```
## [1] 0.978
```

```
table(Predicted=predict(ptree,newdata=test,type="class"),Actual=test$Personal.Loan)
```

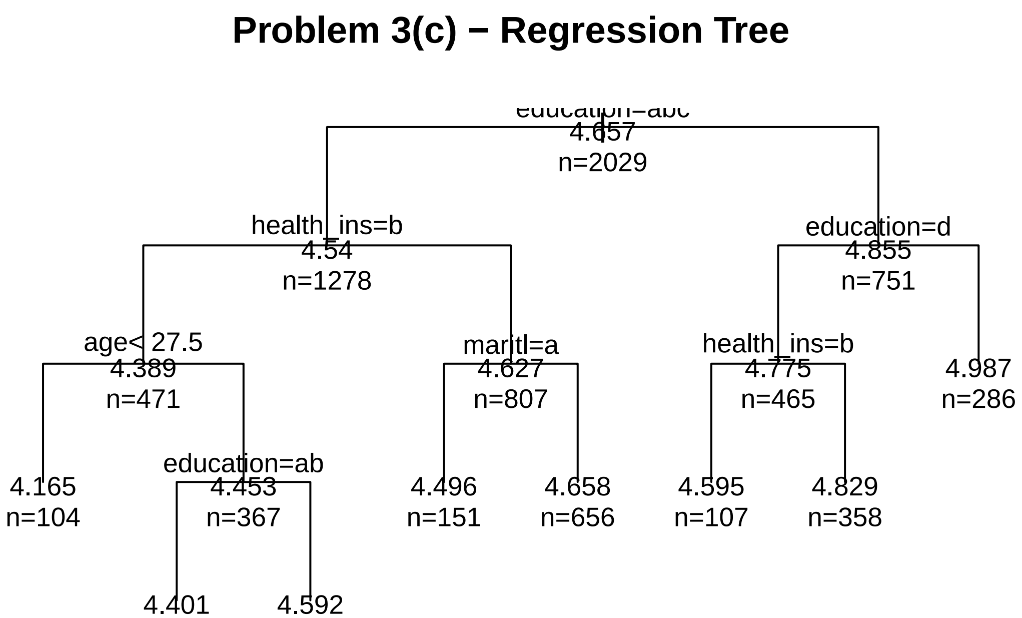
```
##          Actual
## Predicted    0    1
##         0 1345   24
##         1    9  122
```

We can see the accuracy of this pruned tree is approximately 97.8 percent, which indicates this model seems to be predicting the outcome variable ("Personal.Loan") quite well.

Problem 3

```
w<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 451 - Nonparametric Statistical Methods/wage.cs
w$maritl<-as.factor(w$maritl)          # Problem 3(a)
w$race<-as.factor(w$race)
w$education<-as.factor(w$education)
w$region<-as.factor(w$region)
w$jobclass<-as.factor(w$jobclass)
w$health<-as.factor(w$health)
w$health_ins<-as.factor(w$health_ins)
```

```r
set.seed(1304,sample.kind="Rounding") # Problem 3(b)
ws<-sample(1:nrow(w),nrow(w)*.7)
wtrain<-w[ws,]
wtest<-w[-ws,]
wtree<-rpart(logwage~.,data=wtrain)    # Problem 3(c)
plot(wtree,uniform=TRUE,main="Problem 3(c) - Regression Tree")
text(wtree,use.n=TRUE,all=TRUE,cex=0.8)
```




```r
mean((predict(wtree,newdata=wtrain)-wtrain$logwage)^2)
```

```
## [1] 0.08582914
```

```r
mean((predict(wtree,newdata=wtest)-wtest$logwage)^2)
```

```
## [1] 0.08074707
```

```r
wp<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 451 - Nonparametric Statistical Methods/wagepre
wp$maritl<-as.factor(wp$maritl)        # Problem 3(d)
wp$race<-as.factor(wp$race)
wp$education<-as.factor(wp$education)
wp$region<-as.factor(wp$region)
wp$jobclass<-as.factor(wp$jobclass)
wp$health<-as.factor(wp$health)
wp$health_ins<-as.factor(wp$health_ins)
library(car)
Export(as.data.frame(predict(wtree,newdata=wp)),"Charles Hwang WagePredictions.csv")
```