Solutions to HW3

Shaoyi Zhang

April 30th, 2016

Question 1

```
# set up data frame
setwd("/Users/Shawn/Desktop/PSTAT 231/PSTAT-231/assign3")
spam = read.table("spambase.dat",header=T,sep="")
summary(spam)
```

```
##  word_freq_make   word_freq_address word_freq_all     word_freq_3d
##  Min.   :0.0000   Min.   : 0.000    Min.   :0.0000   Min.   : 0.00000
##  1st Qu.:0.0000   1st Qu.: 0.000    1st Qu.:0.0000   1st Qu.: 0.00000
##  Median :0.0000   Median : 0.000    Median :0.0000   Median : 0.00000
##  Mean   :0.1046   Mean   : 0.213    Mean   :0.2807   Mean   : 0.06542
##  3rd Qu.:0.0000   3rd Qu.: 0.000    3rd Qu.:0.4200   3rd Qu.: 0.00000
##  Max.   :4.5400   Max.   :14.280    Max.   :5.1000   Max.   :42.81000
##  word_freq_our     word_freq_over   word_freq_remove word_freq_internet
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median :0.0000   Median :0.0000   Median : 0.0000
##  Mean   : 0.3122   Mean   :0.0959   Mean   :0.1142   Mean   : 0.1053
##  3rd Qu.: 0.3800   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :10.0000   Max.   :5.8800   Max.   :7.2700   Max.   :11.1100
##  word_freq_order   word_freq_mail    word_freq_receive word_freq_will
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median : 0.0000   Median :0.00000   Median :0.1000
##  Mean   :0.09007   Mean   : 0.2394   Mean   :0.05982   Mean   :0.5417
##  3rd Qu.:0.00000   3rd Qu.: 0.1600   3rd Qu.:0.00000   3rd Qu.:0.8000
##  Max.   :5.26000   Max.   :18.1800   Max.   :2.61000   Max.   :9.6700
##  word_freq_people  word_freq_report   word_freq_addresses
##  Min.   :0.00000   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.0000
##  Median :0.00000   Median : 0.00000   Median :0.0000
##  Mean   :0.09393   Mean   : 0.05863   Mean   :0.0492
##  3rd Qu.:0.00000   3rd Qu.: 0.00000   3rd Qu.:0.0000
##  Max.   :5.55000   Max.   :10.00000   Max.   :4.4100
##  word_freq_free    word_freq_business word_freq_email  word_freq_you
##  Min.   : 0.0000   Min.   :0.0000     Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 0.0000   Median :0.0000     Median :0.0000   Median : 1.310
##  Mean   : 0.2488   Mean   :0.1426     Mean   :0.1847   Mean   : 1.662
##  3rd Qu.: 0.1000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.: 2.640
##  Max.   :20.0000   Max.   :7.1400     Max.   :9.0900   Max.   :18.750
##  word_freq_credit   word_freq_your    word_freq_font    word_freq_000
##  Min.   : 0.00000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.00000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
```

```
##  Median : 0.00000   Median : 0.2200   Median : 0.0000   Median :0.0000
##  Mean   : 0.08558   Mean   : 0.8098   Mean   : 0.1212   Mean   :0.1016
##  3rd Qu.: 0.00000   3rd Qu.: 1.2700   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :18.18000   Max.   :11.1100   Max.   :17.1000   Max.   :5.4500
##  word_freq_money     word_freq_hp     word_freq_hpl     word_freq_george
##  Min.   : 0.00000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.00000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.00000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.09427   Mean   : 0.5495   Mean   : 0.2654   Mean   : 0.7673
##  3rd Qu.: 0.00000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :12.50000   Max.   :20.8300   Max.   :16.6600   Max.   :33.3300
##  word_freq_650    word_freq_lab      word_freq_labs   word_freq_telnet
##  Min.   :0.0000   Min.   : 0.00000   Min.   :0.0000   Min.   : 0.00000
##  1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:0.0000   1st Qu.: 0.00000
##  Median :0.0000   Median : 0.00000   Median :0.0000   Median : 0.00000
##  Mean   :0.1248   Mean   : 0.09892   Mean   :0.1029   Mean   : 0.06475
##  3rd Qu.:0.0000   3rd Qu.: 0.00000   3rd Qu.:0.0000   3rd Qu.: 0.00000
##  Max.   :9.0900   Max.   :14.28000   Max.   :5.8800   Max.   :12.50000
##  word_freq_857     word_freq_data     word_freq_415      word_freq_85
##  Min.   :0.00000   Min.   : 0.00000   Min.   :0.00000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.00000   1st Qu.: 0.0000
##  Median :0.00000   Median : 0.00000   Median :0.00000   Median : 0.0000
##  Mean   :0.04705   Mean   : 0.09723   Mean   :0.04784   Mean   : 0.1054
##  3rd Qu.:0.00000   3rd Qu.: 0.00000   3rd Qu.:0.00000   3rd Qu.: 0.0000
##  Max.   :4.76000   Max.   :18.18000   Max.   :4.76000   Max.   :20.0000
##  word_freq_technology word_freq_1999  word_freq_parts   word_freq_pm
##  Min.   :0.00000      Min.   :0.000   Min.   :0.0000   Min.   : 0.00000
##  1st Qu.:0.00000      1st Qu.:0.000   1st Qu.:0.0000   1st Qu.: 0.00000
##  Median :0.00000      Median :0.000   Median :0.0000   Median : 0.00000
##  Mean   :0.09748      Mean   :0.137   Mean   :0.0132   Mean   : 0.07863
##  3rd Qu.:0.00000      3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.: 0.00000
##  Max.   :7.69000      Max.   :6.890   Max.   :8.3300   Max.   :11.11000
##  word_freq_direct   word_freq_cs     word_freq_meeting word_freq_original
##  Min.   :0.00000   Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median : 0.0000   Median :0.0000
##  Mean   :0.06483   Mean   :0.04367   Mean   : 0.1323   Mean   :0.0461
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :4.76000   Max.   :7.14000   Max.   :14.2800   Max.   :3.5700
##  word_freq_project  word_freq_re     word_freq_edu     word_freq_table
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.000000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.000000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :0.000000
##  Mean   : 0.0792   Mean   : 0.3012   Mean   : 0.1798   Mean   :0.005444
##  3rd Qu.: 0.0000   3rd Qu.: 0.1100   3rd Qu.: 0.0000   3rd Qu.:0.000000
##  Max.   :20.0000   Max.   :21.4200   Max.   :22.0500   Max.   :2.170000
##  word_freq_conference  char_freq_.       char_freq_..1    char_freq_..2
##  Min.   : 0.00000      Min.   :0.00000   Min.   :0.000   Min.   :0.00000
##  1st Qu.: 0.00000      1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000
##  Median : 0.00000      Median :0.00000   Median :0.065   Median :0.00000
##  Mean   : 0.03187      Mean   :0.03857   Mean   :0.139   Mean   :0.01698
##  3rd Qu.: 0.00000      3rd Qu.:0.00000   3rd Qu.:0.188   3rd Qu.:0.00000
##  Max.   :10.00000      Max.   :4.38500   Max.   :9.752   Max.   :4.08100
##  char_freq_..3     char_freq_..4     char_freq_..5
```

```
##  Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 0.00000
##  Median : 0.0000   Median :0.00000   Median : 0.00000
##  Mean   : 0.2691   Mean   :0.07581   Mean   : 0.04424
##  3rd Qu.: 0.3150   3rd Qu.:0.05200   3rd Qu.: 0.00000
##  Max.   :32.4780   Max.   :6.00300   Max.   :19.82900
##  capital_run_length_average capital_run_length_longest
##  Min.   :   1.000           Min.   :   1.00
##  1st Qu.:   1.588           1st Qu.:   6.00
##  Median :   2.276           Median :  15.00
##  Mean   :   5.191           Mean   :  52.17
##  3rd Qu.:   3.706           3rd Qu.:  43.00
##  Max.   :1102.500           Max.   :9989.00
##  capital_run_length_total       y
##  Min.   :    1.0          Min.   :0.000
##  1st Qu.:   35.0          1st Qu.:0.000
##  Median :   95.0          Median :0.000
##  Mean   :  283.3          Mean   :0.394
##  3rd Qu.:  266.0          3rd Qu.:1.000
##  Max.   :15841.0          Max.   :1.000
```

```
spam$y = factor(spam$y,levels=c(0,1),labels=c("good","spam"))

# partition the data set
# train set size = sample size - 1000
# test set size = 1000
train_size <- floor(nrow(spam)-1000)

# set the seed to make your partition reproductible
set.seed(1)
train_index <- sample(seq_len(nrow(spam)), size = train_size)

train <- spam[train_index, ]
test <- spam[-train_index, ]
```

Now, we can start build the decision tree

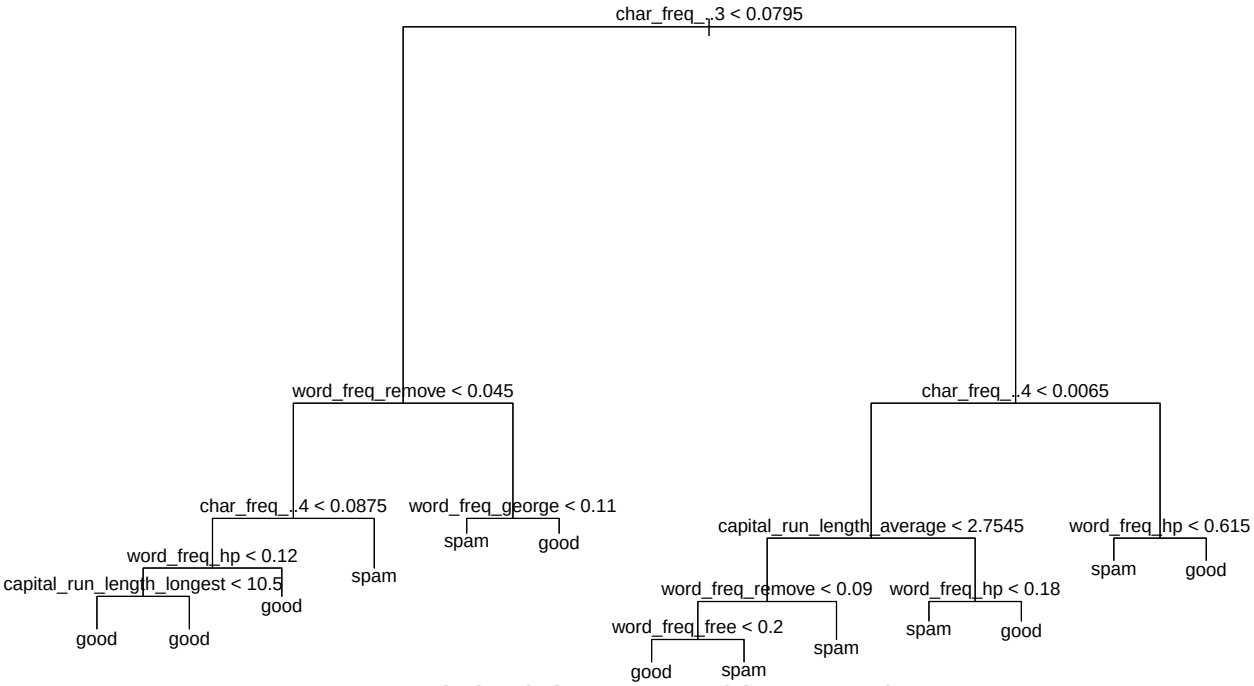
```
require(tree)
```

```
## Loading required package: tree
```

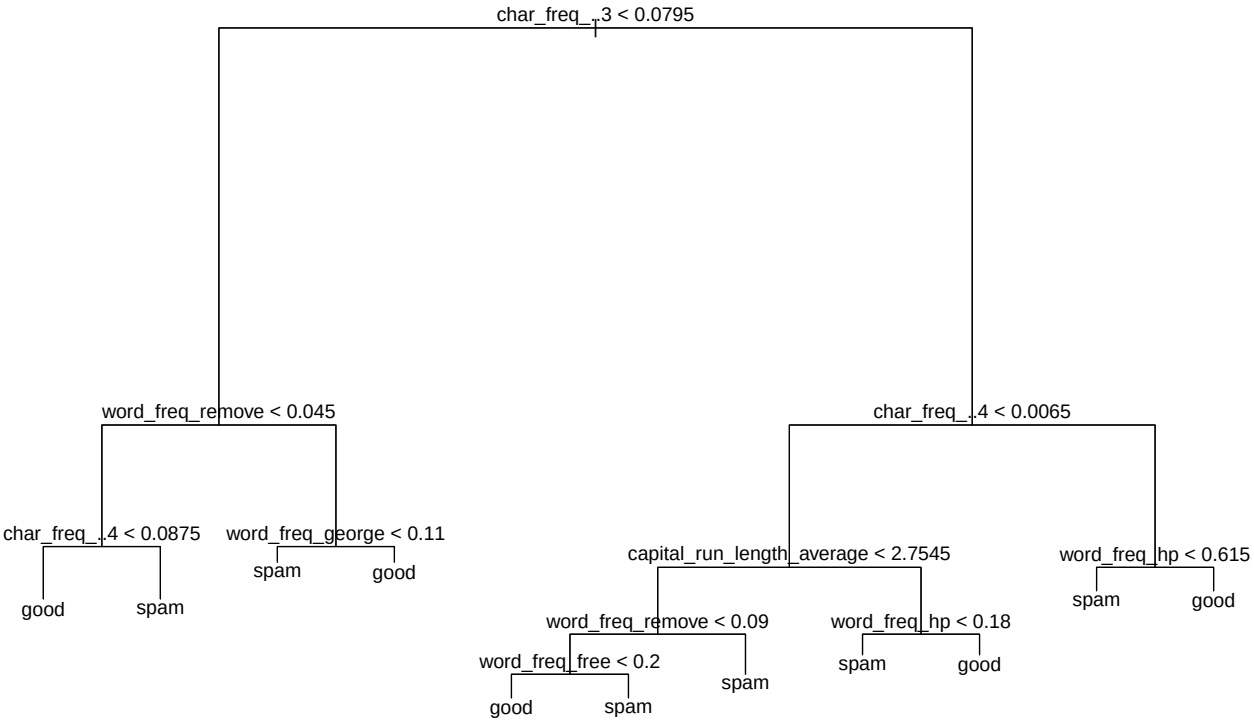
```
spam.tree = tree(y~.,data=train)

cv.spam.treee = cv.tree(spam.tree,FUN=prune.misclass)
prune.spam.tree = prune.misclass(spam.tree,best = 11)
```

unpruned decision tree without option



pruned decision tree without option




```
# make prediction on test set
spam.tree.pred = predict(spam.tree,test,type="class")
conti.table = table(spam.tree.pred,test$y)

prune.pred = predict(prune.spam.tree,test,type="class")
```

```
prune.conti.table = table(prune.pred,test$y)


# construct error rate vector
test.error.rates = vector()
model.index = 0

# compute the test error rate
test.error.rates[model.index] = (conti.table[3] + conti.table[2])/nrow(test)
model.index = model.index + 1
test.error.rates
```

```
## numeric(0)
```

Then, let's try a decision tree with options

```
# decision tree with option
spam.tree.option = tree(y~.,data=train,control=tree.control(nrow(spam),mincut=2,minsize=5,mindev=0.001))

spam.tree.option.pred = predict(spam.tree.option,test,type="class")
conti.table = table(spam.tree.option.pred,test$y)

test.error.rates[model.index] = (conti.table[3] + conti.table[2])/nrow(test)
model.index = model.index + 1
test.error.rates
```

```
## [1] 0.079
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.