FIT5197_ass3_wk11

Chuangfu Xie, 27771539

28/05/2018

First load tree package:

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 3.4.4
```

Then, we load the fire and preprocess for tree:

```
train <- read.csv('./diabetes_train.csv')
test <- read.csv('./diabetes_test.csv')
train$Y <- as.factor(train$Y > 200)
test$Y <- as.factor(test$Y > 200)
tree_diabetes <- tree(Y~., train)
summary(tree_diabetes)
```

```
##
## Classification tree:
## tree(formula = Y ~ ., data = train)
## Variables actually used in tree construction:
## [1] "S5"  "BP"  "BMI" "S1"  "S6"
## Number of terminal nodes:  22
## Residual mean deviance:  0.4364 = 144.9 / 332
## Misclassification error rate: 0.0904 = 32 / 354
```

Take a peek at the tree:

```
tree_diabetes
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##   1) root 354 405.600 FALSE ( 0.74011 0.25989 )
##     2) S5 < 4.88275 241 173.200 FALSE ( 0.88382 0.11618 )
##       4) BP < 102.5 199  85.080 FALSE ( 0.94472 0.05528 )
##         8) BMI < 25.75 144  29.160 FALSE ( 0.97917 0.02083 )
##          16) S1 < 184.5 91   0.000 FALSE ( 1.00000 0.00000 ) *
##          17) S1 > 184.5 53  23.060 FALSE ( 0.94340 0.05660 )
##            34) S5 < 4.3882 27   0.000 FALSE ( 1.00000 0.00000 ) *
##            35) S5 > 4.3882 26  18.600 FALSE ( 0.88462 0.11538 )
##              70) S6 < 88.5 11  12.890 FALSE ( 0.72727 0.27273 ) *
##              71) S6 > 88.5 15   0.000 FALSE ( 1.00000 0.00000 ) *
##         9) BMI > 25.75 55  45.620 FALSE ( 0.85455 0.14545 )
##          18) BMI < 29.1 30  32.600 FALSE ( 0.76667 0.23333 ) *
##          19) BMI > 29.1 25   8.397 FALSE ( 0.96000 0.04000 ) *
##       5) BP > 102.5 42  56.690 FALSE ( 0.59524 0.40476 )
##        10) BMI < 26.9 25  21.980 FALSE ( 0.84000 0.16000 ) *
##        11) BMI > 26.9 17  18.550 TRUE ( 0.23529 0.76471 )
##          22) S5 < 4.42395 8  11.090 FALSE ( 0.50000 0.50000 ) *
##          23) S5 > 4.42395 9   0.000 TRUE ( 0.00000 1.00000 ) *
##     3) S5 > 4.88275 113 154.700 TRUE ( 0.43363 0.56637 )
##       6) BMI < 31.55 78 106.800 FALSE ( 0.56410 0.43590 )
##        12) BMI < 24.25 13   7.051 FALSE ( 0.92308 0.07692 ) *
##        13) BMI > 24.25 65  90.090 TRUE ( 0.49231 0.50769 )
##          26) S5 < 5.3822 54  73.670 FALSE ( 0.57407 0.42593 )
##            52) S6 < 99.5 43  55.620 FALSE ( 0.65116 0.34884 )
##             104) BMI < 29.45 35  47.800 FALSE ( 0.57143 0.42857 )
##               208) S6 < 86.5 7   5.742 TRUE ( 0.14286 0.85714 ) *
##               209) S6 > 86.5 28  35.160 FALSE ( 0.67857 0.32143 )
##                 418) BMI < 28.35 23  24.080 FALSE ( 0.78261 0.21739 )
##                   836) S5 < 5.0138 12  16.300 FALSE ( 0.58333 0.41667 ) *
##                   837) S5 > 5.0138 11   0.000 FALSE ( 1.00000 0.00000 ) *
##                 419) BMI > 28.35 5   5.004 TRUE ( 0.20000 0.80000 ) *
##             105) BMI > 29.45 8   0.000 FALSE ( 1.00000 0.00000 ) *
##            53) S6 > 99.5 11  12.890 TRUE ( 0.27273 0.72727 )
##             106) S5 < 5.0464 5   6.730 FALSE ( 0.60000 0.40000 ) *
##             107) S5 > 5.0464 6   0.000 TRUE ( 0.00000 1.00000 ) *
##          27) S5 > 5.3822 11   6.702 TRUE ( 0.09091 0.90909 ) *
##       7) BMI > 31.55 35  28.710 TRUE ( 0.14286 0.85714 )
##        14) BP < 109.5 17  20.600 TRUE ( 0.29412 0.70588 )
##          28) S6 < 103.5 11  15.160 TRUE ( 0.45455 0.54545 )
##            56) S1 < 195 6   5.407 TRUE ( 0.16667 0.83333 ) *
##            57) S1 > 195 5   5.004 FALSE ( 0.80000 0.20000 ) *
##          29) S6 > 103.5 6   0.000 TRUE ( 0.00000 1.00000 ) *
##        15) BP > 109.5 18   0.000 TRUE ( 0.00000 1.00000 ) *
```

```
train$S5_cnt = as.factor(train$S5<4.88275)
train$BP_cnt = as.factor(train$BP<102.5)
train$BMI_cnt = as.factor(train$BMI<31.55)
attach(train)
table(S5_cnt, Y)
```

```
##        Y
## S5_cnt  FALSE TRUE
##   FALSE    49   64
##   TRUE    213   28
```

```
table(BP_cnt, Y)
```

```
##        Y
## BP_cnt  FALSE TRUE
##   FALSE    45   53
##   TRUE    217   39
```

```
table(BMI_cnt, Y)
```

```
##         Y
## BMI_cnt FALSE TRUE
##   FALSE    15   37
##   TRUE    247   55
```

From above, we can compute their entropy:

$$H(Y|S5>4.88275)=-\frac{49}{113}log_2(\frac{49}{113})-\frac{64}{113}log_2(\frac{64}{113})=0.6899765$$

$$H(Y|S5<4.88275)=-\frac{213}{241}log_2(\frac{213}{241})-\frac{28}{241}log_2(\frac{64}{113})=0.5182873$$

$$p(S5>4.88275)=\frac{113}{354}=0.319209$$

$$p(S5<4.88275)=\frac{241}{354}=0.680791$$

$$H(S5)=p(S5>4.88275)\cdot H(Y|S5>4.88275)+p(S5<4.88275)\cdot H(Y|S5<4.88275)=0.667985$$

$$H(Y|BP>102.5)=-\frac{45}{98}log_2(\frac{45}{98})-\frac{53}{98}log_2(\frac{53}{98})=0.9951877$$

$$H(Y|BP<102.5)=-\frac{217}{256}log_2(\frac{217}{256})-\frac{39}{256}log_2(\frac{39}{256})=0.6156746$$

$$p(BP>102.5)=\frac{98}{354}=0.2768362$$

$$p(BP<102.5)=\frac{256}{354}=0.7231638$$

$$H(BP)=p(BP>102.55)\cdot H(Y|BP>102.5)+p(BP<102.5)\cdot H(Y|BP<102.5)=0.7207375$$

$$H(Y|BMI>31.55)=-\frac{15}{52}log_2(\frac{15}{52})-\frac{37}{52}log_2(\frac{37}{52})=0.8667256$$

$$H(Y|BMI<31.55)=-\frac{247}{302}log_2(\frac{247}{302})-\frac{55}{302}log_2(\frac{55}{302})=0.6846912$$

$$p(BMI>31.55)=\frac{52}{354}=0.1468927$$

$$p(BMI<31.55)=\frac{302}{354}=0.8531073$$

$$H(BP)=p(BMI>31.55)\cdot H(Y|BMI>31.55)+p(BMI<31.55)\cdot H(Y|BMI<31.55)=0.7114307$$

Hence, the best test to use at the root of the tree is S5.