Classification Trees

Ex. MLB

```
library(ISLR)
names(Hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "
##  [6] "Walks"     "Years"     "CAtBat"    "CHits"     "
## [11] "CRuns"     "CRBI"      "CWalks"    "League"    "
## [16] "PutOuts"   "Assists"   "Errors"    "Salary"    "
```

Can we predict the league that a player is in based on his other variables?

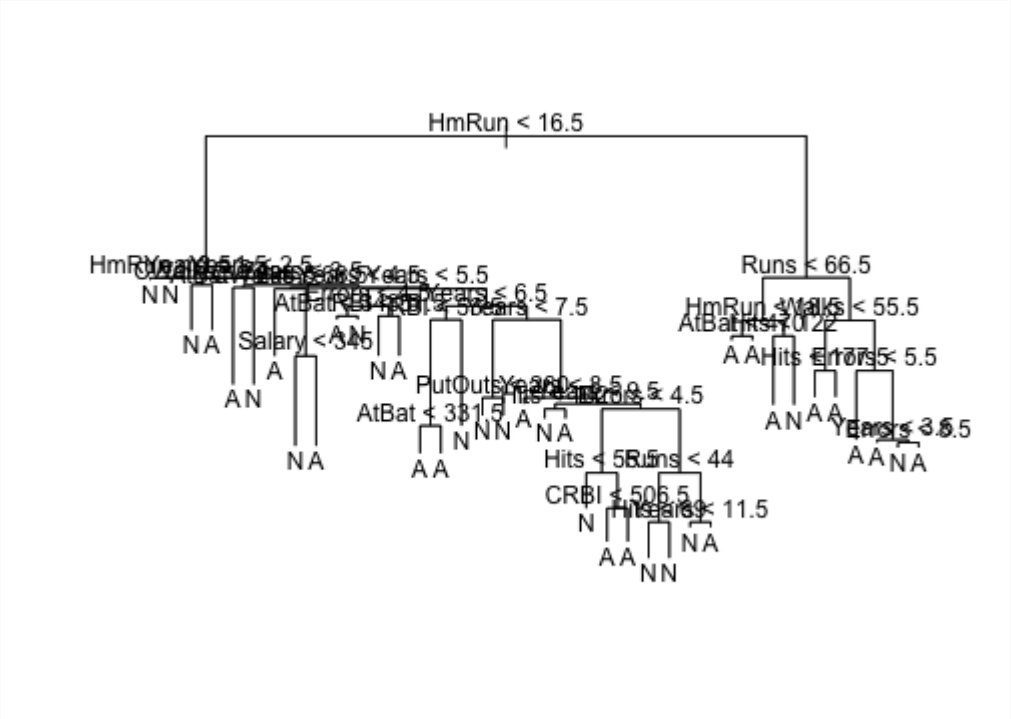
Fitting a tree

```
library(tree)
t1 <- tree(League ~. - NewLeague,
           data = Hitters, split = "gini")
class(t1)
```

```
## [1] "tree"
```

Default stopping rule: stop splitting when terminal nodes get too small.

Fitting a tree, cont.



Too much?

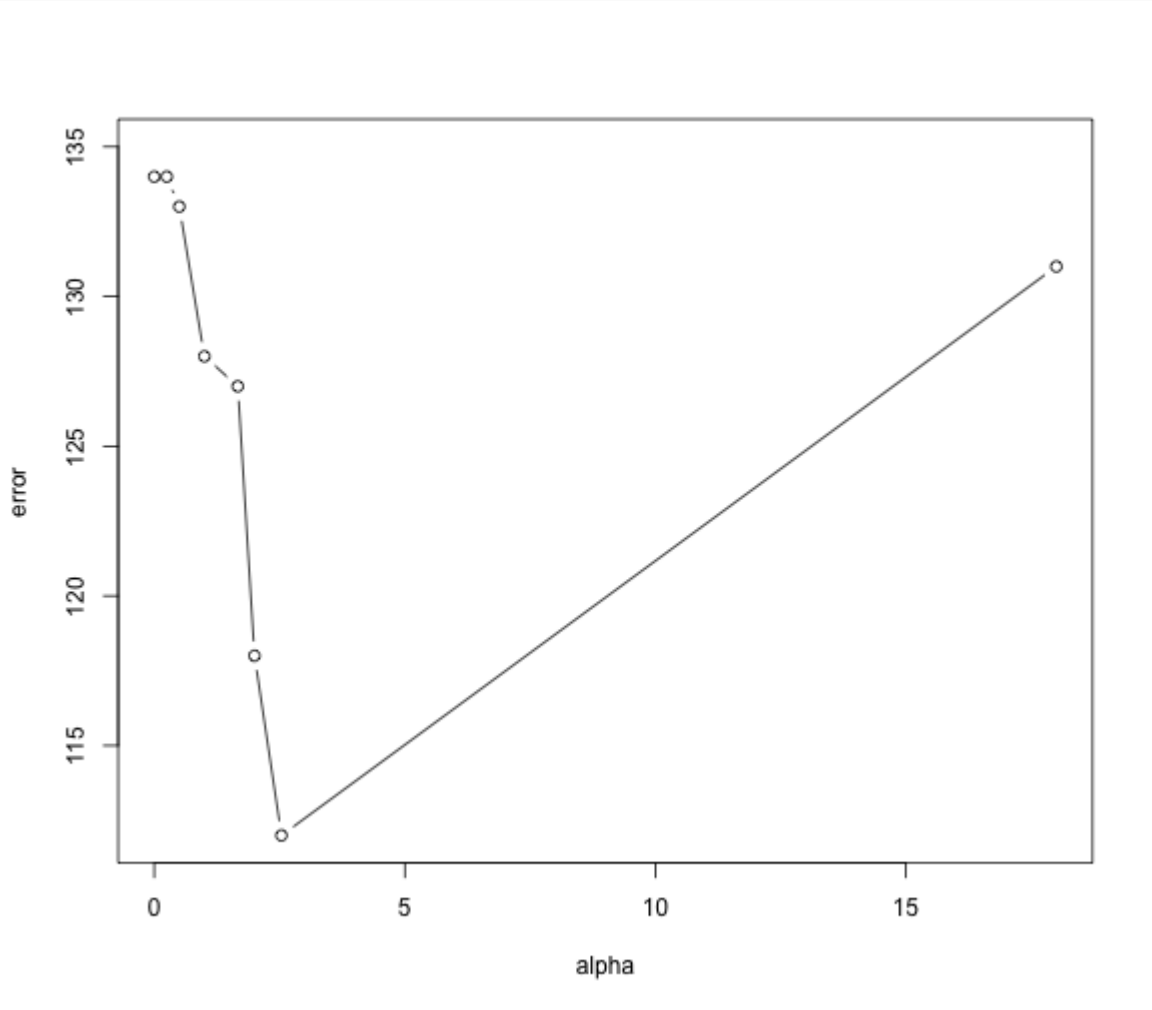
Cost-complexity pruning

Assess the performance of many trees with size indexed by α via 10-fold cross-validation on misclassification rate.

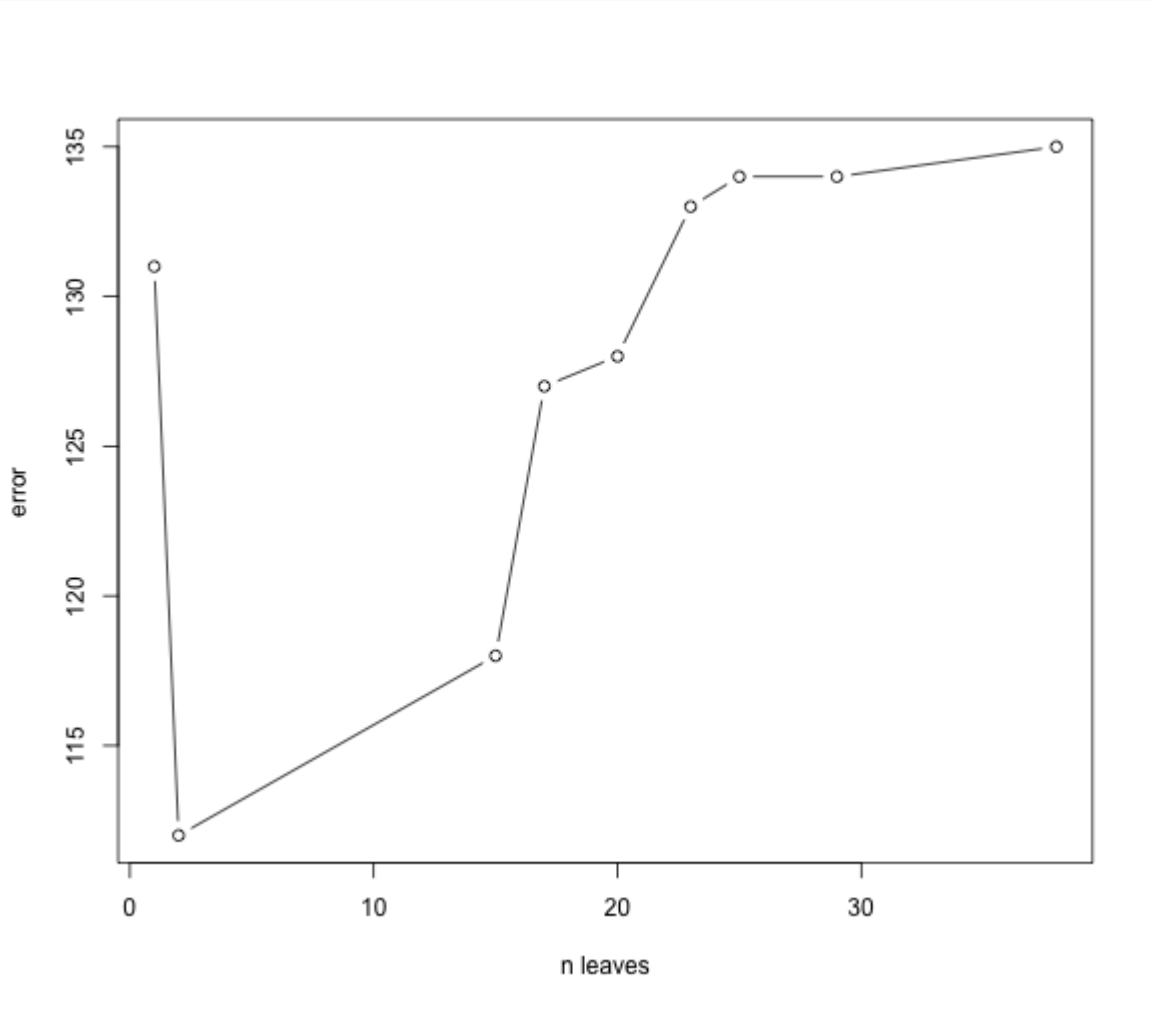
```
set.seed(40)
t1cv <- cv.tree(t1, FUN = prune.misclass)
t1cv
```

```
## $size
## [1] 38 29 25 23 20 17 15  2  1
##
## $dev
## [1] 135 134 134 133 128 127 118 112 131
##
## $k
## [1]      -Inf  0.000000  0.250000  0.500000  1.000000
## [8]  2.538462 18.000000
##
## $method
```

Alpha vs Error



Size vs Error

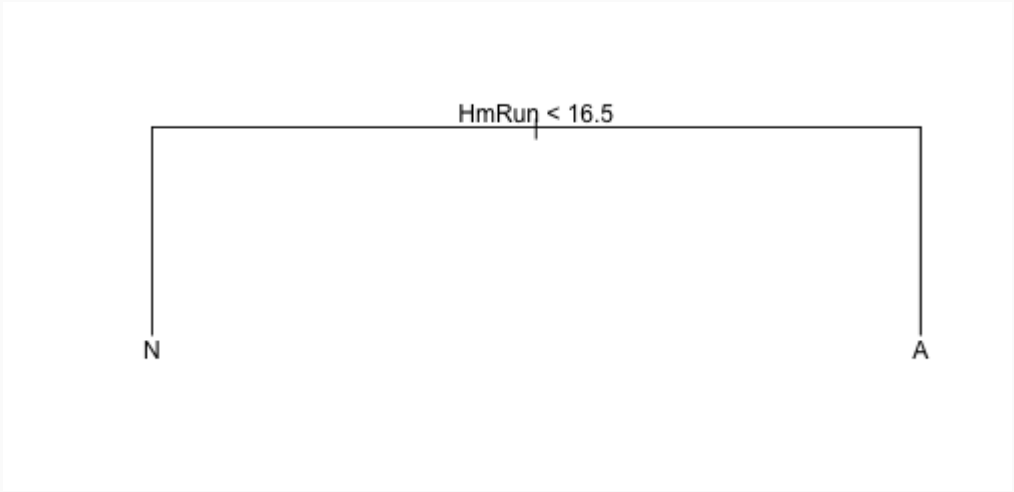


Prune the tree

```
t1cv$size[which.min(t1cv$dev)]
```

```
## [1] 2
```

```
t1prune <- prune.misclass(t1, best = 2)
```



Activity 4: Off in the distance

Return to your Lab 4, where you fit Logistic and an LDA model for the civil wars data set. In a new .Rmd file, add a new classification tree that has been pruned back.

1. What is the training misclassification rate? (code for creating the confusion matrix can be found on p. 327 of your book)
2. How does this rate compare to the other classification models that you used in Lab 4? Why do you think this is?