Introduction to classification and regression trees, random forests and model-based recursive partitioning in R

Day 1: Single trees

Exercise 3: Illustrating variable selection bias

Using function `rnorm()`, generate three standard normally distributed (that is, $\mu = 0$ and $\sigma = 1$) variables `x1`, `x2` and `y`, each with 100 observations. Round the values of `x2` to 0 decimal places. Note that `x1` now has much more levels than `x2`. Fit two trees that aim to predict `y` from `x1` and `x2`: one using the `tree()` function from the `tree` package and one using the `ctree()` function from the `partykit` package.

Did the algorithms create spurious splits? If so, were `x1` and `x2` equally often selected for splitting, or was there a selection bias towards `x1`? Repeat the sampling and fitting process ten times. Do you see the same pattern every time?

```
library(tree)
library(partykit)
```

```
## Loading required package: grid
```

```
x1 <- rnorm(100)
x2 <- round(rnorm(100))
y <- rnorm(100)
CART <- tree(y ~ x1 + x2)
print(CART)
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 100 86.780  0.08028
##     2) x1 < -1.56598 8  4.646 -0.88570 *
##     3) x1 > -1.56598 92 74.020  0.16430
##       6) x1 < -0.708266 14 10.050  0.51970
##        12) x1 < -0.991703 9  5.738  0.16800 *
##        13) x1 > -0.991703 5  1.195  1.15300 *
##       7) x1 > -0.708266 78 61.880  0.10050
##        14) x2 < 0.5 55 41.080  0.20510
##          28) x1 < 1.13675 46 35.320  0.27400
##            56) x1 < 0.397348 25 27.320  0.13980
##             112) x1 < -0.2002 15 17.630  0.37380 *
##             113) x1 > -0.2002 10  7.644 -0.21110 *
##            57) x1 > 0.397348 21  7.013  0.43380
##             114) x1 < 0.431076 5  1.032  1.00600 *
##             115) x1 > 0.431076 16  3.834  0.25500
##               230) x1 < 0.679734 10  1.759  0.06703 *
##               231) x1 > 0.679734 6  1.133  0.56830 *
##          29) x1 > 1.13675 9  4.419 -0.14730 *
##        15) x2 > 0.5 23 18.760 -0.14960
##          30) x1 < -0.00499258 6  3.351 -0.54480 *
##          31) x1 > -0.00499258 17 14.140 -0.01019 *
```

```
ct <- ctree(y ~ x1 + x2)
print(ct)
```

```
##
```

```
## Model formula:
## y ~ x1 + x2
##
## Fitted party:
## [1] root: 0.080 (n = 100, err = 86.8)
##
## Number of inner nodes:    0
## Number of terminal nodes: 1
```

```r
# The CART tree (fit with the tree function) often creates spurious splits,
# mostly using variable x1. CART shows vairable selection bias: Both x1 andx2
# are in fact not associated with the outcome, but x1 gets picked for splitting
# much more often than x2. The ctree does not (or rarely) create spurious splits,
# as none fo the predictor variables are significantly associated with the
# outcome.
#
# Note that if we would prune the CART trees created above, maybe (hopefully)
# all splits would be removed, as there really is no association between the
# x variables and y in this example.
```

Exercise 4: Revisiting carseat sales

```r
# Code from exercise 2:
data("Carseats", package = "ISLR")
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising    
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000  
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000  
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000  
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635  
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000  
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000  
##    Population        Price        ShelveLoc        Age       
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00  
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75  
##  Median :272.0   Median :117.0   Medium:219   Median :54.50  
##  Mean   :264.8   Mean   :115.8                Mean   :53.32  
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00  
##  Max.   :509.0   Max.   :191.0                Max.   :80.00  
##    Education    Urban       US     
##  Min.   :10.0   No :118   No :142  
##  1st Qu.:12.0   Yes:282   Yes:258  
##  Median :14.0                      
##  Mean   :13.9                      
##  3rd Qu.:16.0                      
##  Max.   :18.0
```

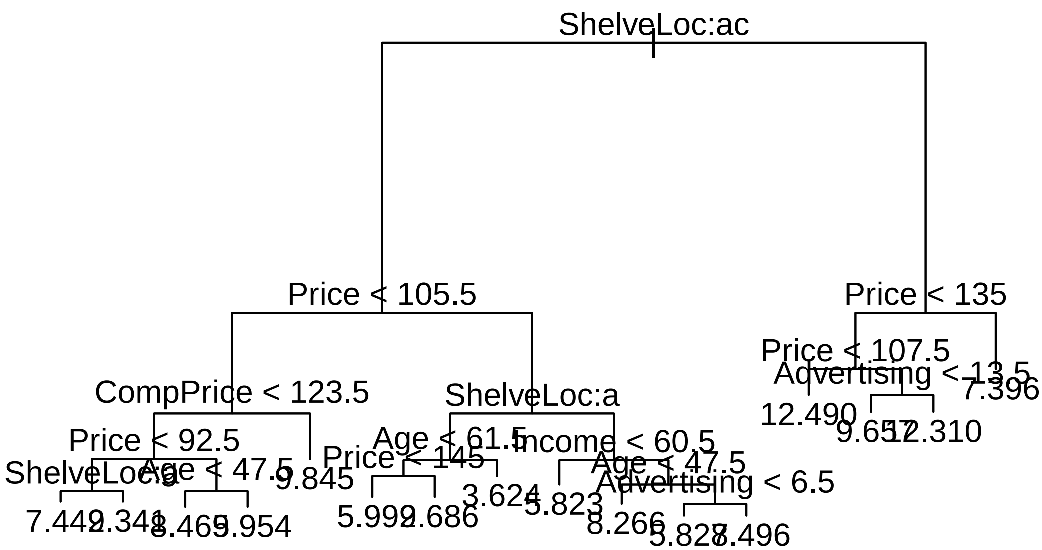
```r
?ISLR::Carseats
```

```
## starting httpd help server ...
```

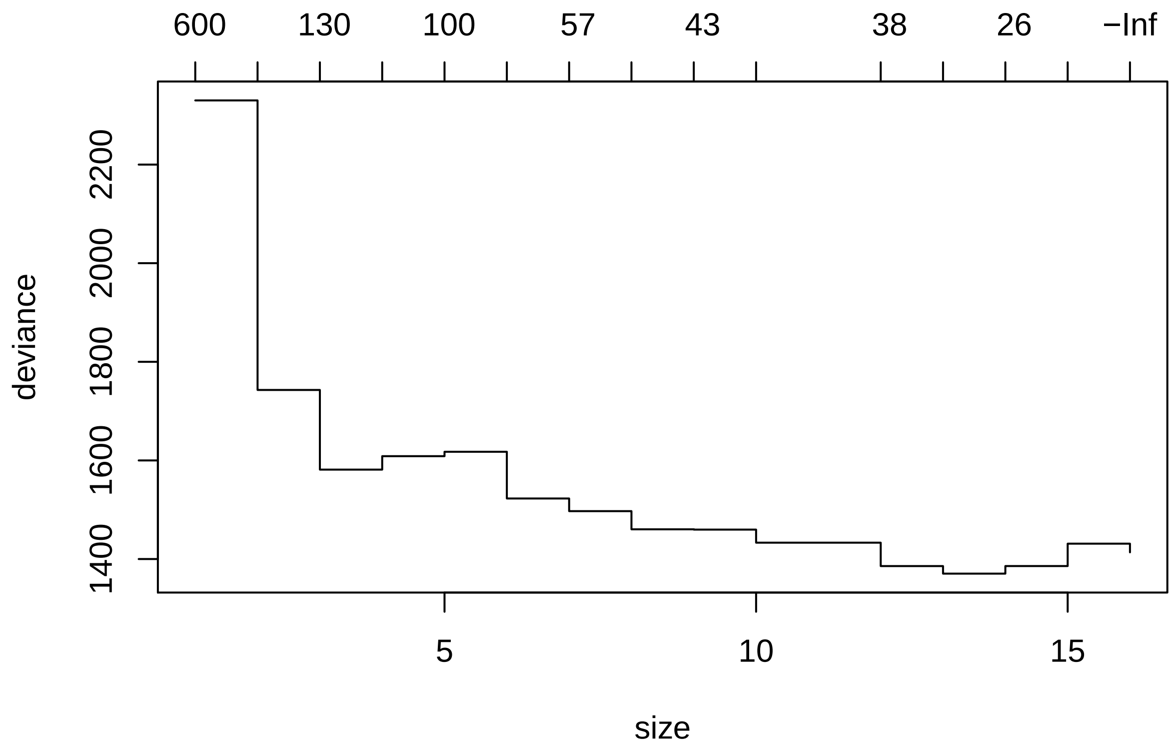
```
##  done
```

```r
set.seed(42)
train <- sample(1:400, 300)
```

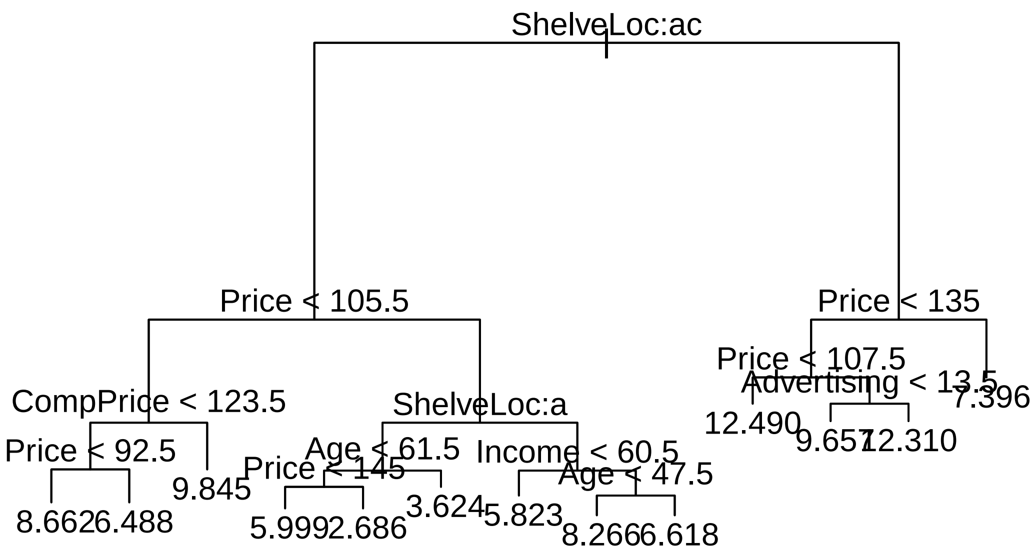
```
CART.cs <- tree(Sales ~ ., data = Carseats[train,])
plot(CART.cs)
text(CART.cs)
```




```
CART.cs.cv <- cv.tree(CART.cs)
plot(CART.cs.cv)
```




```r
CART.cs.pruned <- prune.tree(CART.cs, best = 13)
plot(CART.cs.pruned)
text(CART.cs.pruned)
```




```r
# Shelve location is the most important predictor (followed by Price)
CART_preds_pruned <- predict(CART.cs.pruned, newdata =  Carseats[-train,])
CART_preds <- predict(CART.cs, newdata =  Carseats[-train,])
cor(cbind(Carseats[-train, "Sales"], CART_preds, CART_preds_pruned))
```

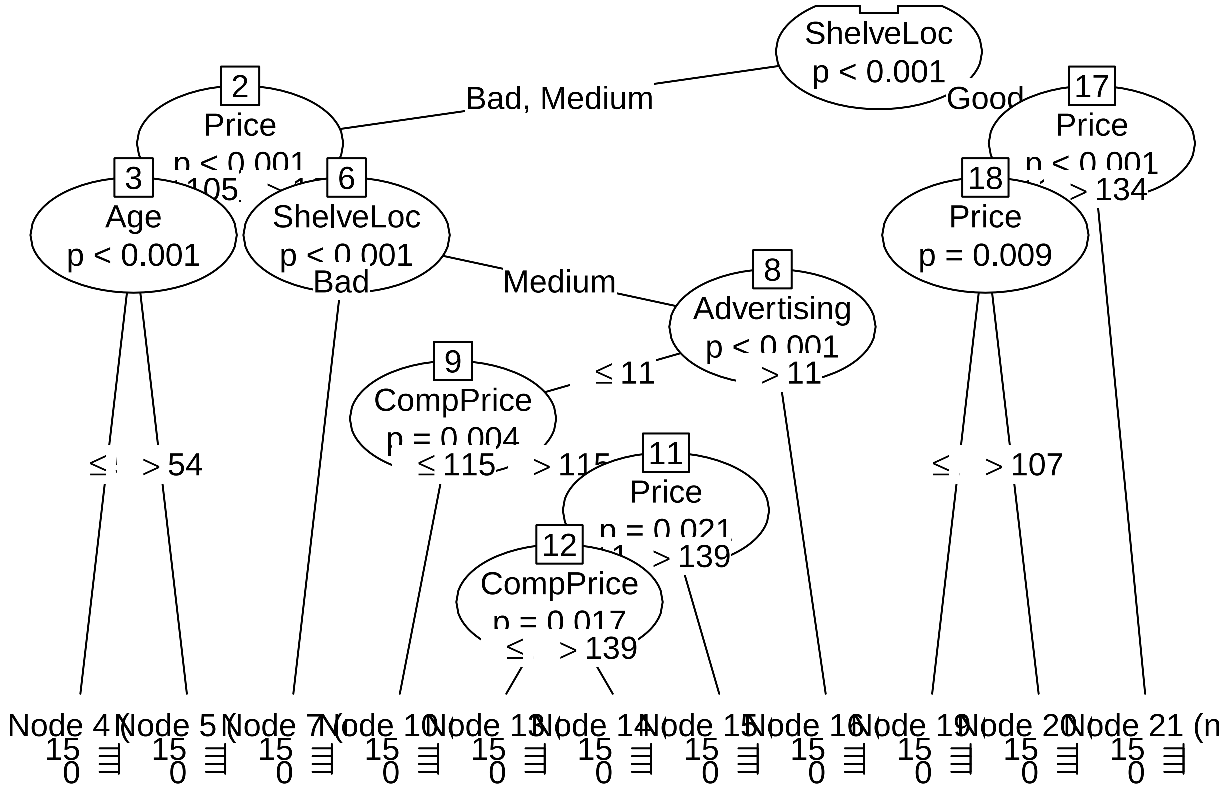
```
##                             CART_preds CART_preds_pruned
##                   1.0000000  0.7169186         0.6830204
## CART_preds        0.7169186  1.0000000         0.9762475
## CART_preds_pruned 0.6830204  0.9762475         1.0000000
```

```r
# Pruning reduced the accuracy on test data (in this case, should not always be the case).
```

a) Using the 'ctree()' function, fit a conditional inference tree to the training observations of the Carseats data from Exercise 2.

b) plot the tree. Is the ctree the same as the pruned CART tree from Exercise 2? Do the two trees agree on the most important predictor of carseat sales?

```r
ctree.cs <- ctree(Sales ~ ., data = Carseats[train,])
plot(ctree.cs)
```




```
# The two trees are not exactly equal, but very similar. Both seem to agree that
# Shelve location is the most important predictor (followed by Price)
```

c) Using the `predict()` function, generate predictions for the test observations from. For the ctree, calculate the correlation with the true car seat sales in the test data. Does the ctree predict better than the (un)pruned CART trees?

```
ctree_preds <- predict(ctree.cs, newdata =  Carseats[-train,])
cor(cbind(Carseats[-train, "Sales"], ctree_preds, CART_preds, CART_preds_pruned))
```

```
##                              ctree_preds CART_preds CART_preds_pruned
##                    1.0000000   0.7131133  0.7169186         0.6830204
## ctree_preds        0.7131133   1.0000000  0.8432362         0.8351069
## CART_preds         0.7169186   0.8432362  1.0000000         0.9762475
## CART_preds_pruned  0.6830204   0.8351069  0.9762475         1.0000000
```

```
# The ctree predicts better than both the pruned and unpruned CART trees.
```

Exercise 5: Predicting Glaucoma

Load the `Glaucoma` dataset from the package `TH.data`:

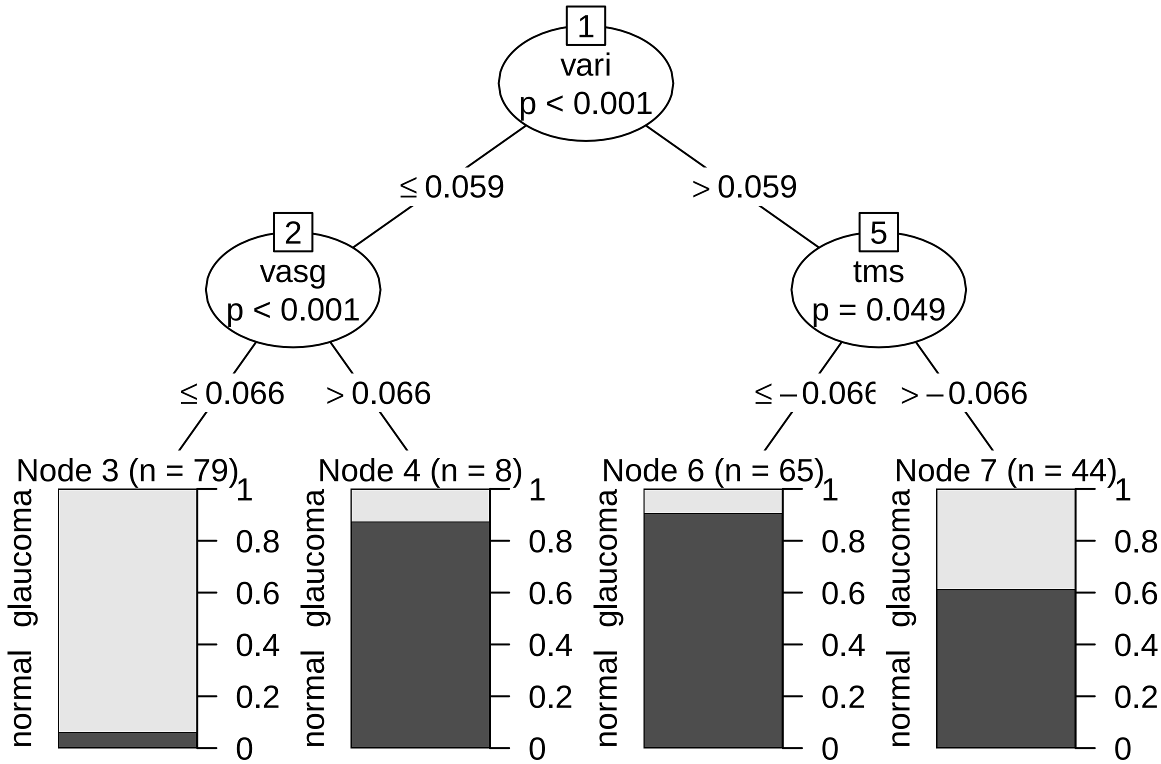
```
data("GlaucomaM", package = "TH.data")
?TH.data::GlaucomaM
```

The data set contains 62 continuous predictor variables. The response is the binary factor `GlaucomaM$Class`.

a) Fit a classification tree with the `ctree` function.

b) Plot the tree. Does it indicate main and/or interaction effects? Of which variables?

```r
ct <- ctree(Class ~ ., data = GlaucomaM)
plot(ct)
```




```r
# There seems to be a main effect of vari, but we should be careful in
# interpreting this in the presence of interactions. There seems to be an
# interaction between vari and vasg, and vari and tms: with lower values of
# vari and vasg, the probability of having glaucoma is very high. With higher
# values of vari and tms, the prboability of having glaucoma is somewhat
# increased.
```