Lab 7: Decision Trees

PSTAT 131/231, Fall 2023

Learning Objectives

- Fit decision tree models using package `tree` and `base` - `tree()` and `summary()` - `predict()` and `table()` - `cv.tree()` and `prune.tree()`
- Decision trees visualization

---

1. Install packages and import dataset

We are going to use the dataset `Carseats` in the package `ISLR` and various tree-fitting functions in `tree`. As we have seen in previous labs, `Carseats` is a simulated data set containing sales of child car seats at 400 different stores on 11 features. The features include: `Sales`, `CompPrice`, `Income`, `Advertising`, `Population`, `Price`, `ShelveLoc`, `Age`, `Education`, `Urban` and `US`. Among all the variables, `ShelveLoc`, `Urban` and `US` are categorical and the rest are continuous.

Notice that originally `Sales` is a continuous variable. Just as in Lab 2, we create a new binary variable `High` using `Sales`:

$$\text{High} = \begin{cases} \text{No}, & \text{if Sales} \leq \text{median(Sales)} \\ \text{Yes}, & \text{if Sales} > \text{median(Sales)} \end{cases}$$

Our goal is to investigate how other features (`CompPrice`, `Income`, `Advertising`, `Population`, `Price`, `ShelveLoc`, `Age`, `Education`, `Urban` and `US`) influence whether the unit sales at each location is high or not. In other words, we look for the relationship between the binary response `High` and all variables but `Sales`.

We first load in the data, and the required packages needed for using decision trees:

```r
##install.packages("ISLR")
##install.packages("tree")
##install.packages('maptree')

# Load libraries
library(ISLR)
library(tree)
library(maptree)

# Utility library
library(dplyr)
```

Using `mutate()` and `ifelse()` to create the binary response variable `High`, then check the structure of resulting data frame with the following codes:

```r
# Create data frame with the oringinal eleven variables and High
Carseats = Carseats %>%
    mutate(High=as.factor(ifelse(Sales <= median(Sales), "No", "Yes")))
```

```r
# Check the structure of above data frame we just created
glimpse(Carseats)
```

```
## Rows: 400
## Columns: 12
## $ Sales       <dbl> 9.50, 11.22, 10.06, 7.40, 4.15, 10.81, 6.63, 11.85, 6.54, ~
## $ CompPrice   <dbl> 138, 111, 113, 117, 141, 124, 115, 136, 132, 132, 121, 117~
## $ Income      <dbl> 73, 48, 35, 100, 64, 113, 105, 81, 110, 113, 78, 94, 35, 2~
## $ Advertising <dbl> 11, 16, 10, 4, 3, 13, 0, 15, 0, 0, 9, 4, 2, 11, 11, 5, 0, ~
## $ Population  <dbl> 276, 260, 269, 466, 340, 501, 45, 425, 108, 131, 150, 503,~
## $ Price       <dbl> 120, 83, 80, 97, 128, 72, 108, 120, 124, 124, 100, 94, 136~
## $ ShelveLoc   <fct> Bad, Good, Medium, Medium, Bad, Bad, Medium, Good, Medium,~
## $ Age         <dbl> 42, 65, 59, 55, 38, 78, 71, 67, 76, 76, 26, 50, 62, 53, 52~
## $ Education   <dbl> 17, 10, 12, 14, 13, 16, 15, 10, 10, 17, 10, 13, 18, 18, 18~
## $ Urban       <fct> Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Ye~
## $ US          <fct> Yes, Yes, Yes, Yes, No, Yes, No, Yes, No, Yes, Yes, Yes, N~
## $ High        <fct> Yes, Yes, Yes, No, No, Yes, No, Yes, No, No, Yes, Yes, No,~
```

2. A decision tree trained with the entire dataset

Based on the data frame `Carseats` with `High`, we will build a classification tree model, in which `High` will be the response (dependent variable), and the rest 10 features, excluding `Sales`, will be the predictors (independent variables). The classification tree model can be built with function `tree()` in the package `tree`. (Yeah, they share the same name!)

(a). Fit and summarize the tree

- `tree()` can be used to fit both classification and regression tree models. A regression tree is very similar to a classification tree, except that it is used to predict a quantitative response rather than a qualitative one. In this lab, we will focus on classification trees. We put the response variable on the left of tilde, explanatory variables on the right of tilde; the dot is merely an economical way to represent "everything else but High"[1]. Recall that this syntax is exactly the same as we use in fitting `lm()`, `glm()`, but different from `glmnet()`.

  ```r
  tree.carseats = tree(High ~.-Sales, data = Carseats)
  ```

- `summary()` is a generic function used to produce result summaries of various model fitting functions. When we call the `summary` of a tree, we will have the following reported:
  - *Classification tree*: displays the model and the dataset
  - *Variables ... construction*: variables that are truly useful to construct the tree
  - *Number ... nodes*: the number of leaf node, which is a node that has no child nodes. Let's denote this quantity as T_0 for further reference
  - *Residual mean deviance*: is simply the deviance divided by $n - T_0$, which in this case is $400 - 23 = 377$
  - *Misclassification error rate*: is the number of wrong predictions divided by the number of total predictions (on the input dataset). This is really the training error rate.

  ```r
  summary(tree.carseats)
  ```

  ```
  ##
  ## Classification tree:
  ## tree(formula = High ~ . - Sales, data = Carseats)
  ## Variables actually used in tree construction:
  ## [1] "ShelveLoc"   "Price"       "Advertising" "CompPrice"   "Age"
  ## [6] "Population"  "Income"
  ```

[1] Note: The reason why we have to exclude `Sales` from the explanatory variables is that the response (`High`) is derived from it.

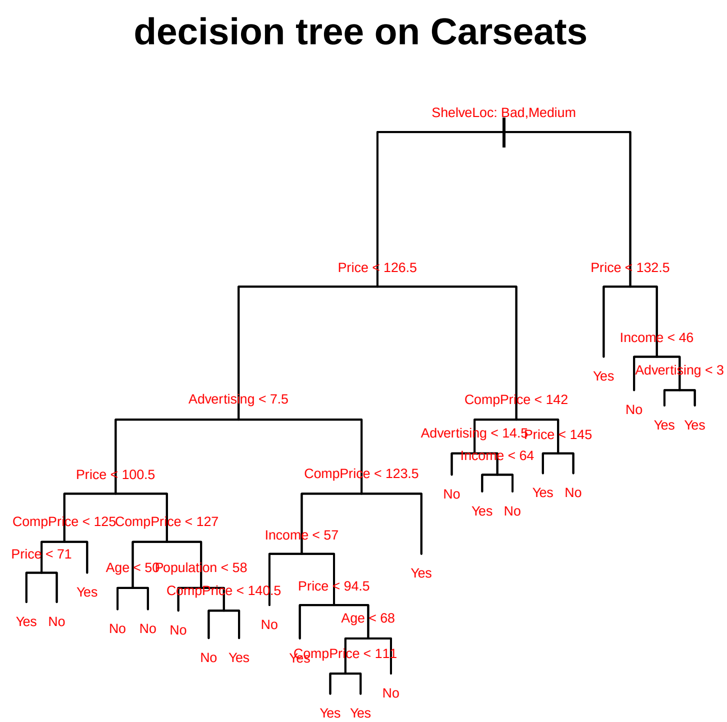
```
## Number of terminal nodes:  23
## Residual mean deviance:  0.4945 = 186.4 / 377
## Misclassification error rate: 0.115 = 46 / 400
```

(b). Visualize the tree

There are essentially two ways of visualizing a tree fitted from `tree` function call.

- The built-in function `plot` and `text` in the `tree` package, demonstrated as follows:

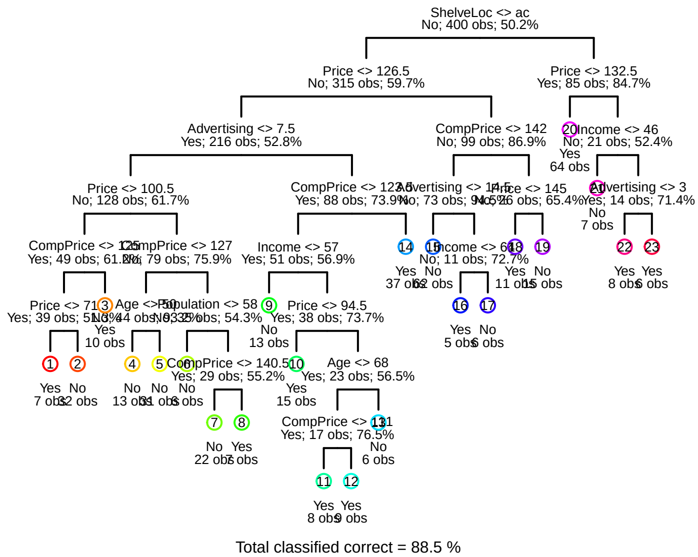
```r
plot(tree.carseats)
text(tree.carseats, pretty = 0, cex = .4, col = "red")
title("decision tree on Carseats", cex = 0.8)
```



Note that `text()` is to display the node labels. The argument `pretty=0` instructs R to include the category names for any qualitative predictors, rather than simply displaying a letter for each category. The function `title()` is to display the theme of the plot. `cex` controls the size of labels in the plots.

- Alternatively, `draw.tree()` in the `maptree` package is helpful for visualizing the structure

```r
draw.tree(tree.carseats, nodeinfo=TRUE, cex = 0.4)
```



3. A decision tree trained with training/test split

In order to properly evaluate the performance of a classification tree, we should estimate the **test error rate** rather than simply compute the training error rate. Therefore, as what we have been doing in this course, we split all observations into a **training set** and a **test set**, build the tree using the training set, and evaluate the model's performance on the test set.

(a). Split the data into a training set and a test set

We sample 75% of observations as the training set and the rest 25% as the test set.

```r
# Set random seed for results being reproducible
set.seed(3)
# Get dimension of dataset
dim(Carseats)
```

```
## [1] 400  12
```

```r
# Sample 75% of observations as the training set
train = sample(nrow(Carseats), 0.75*nrow(Carseats))
Carseats.train = Carseats[train, ]
# The rest 25% as the test set
```

```r
Carseats.test = Carseats[-train,]

# For later convenience in coding, we create High.test, which is the true labels of the
# test cases
High.test = Carseats.test$High
```

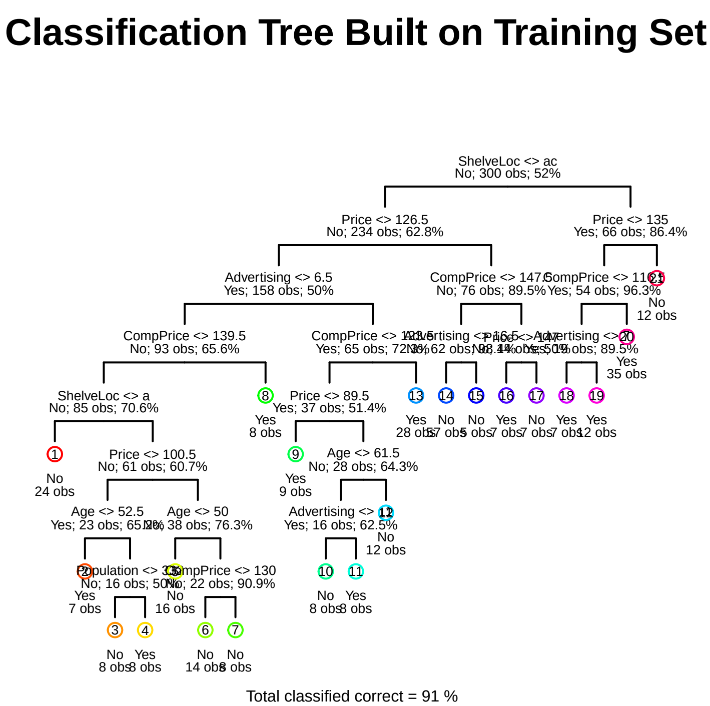
(b). Fit the tree on training set and compute test error rate

- `tree()` can be used to grow the tree as we discussed in the previous section.
- `predict()` is helpful to predict the response (`High`) on the test set. In the case of a classification tree, specifying `type="class"` instructs R to return the actual class predictions instead of probabilities.

  As discussed earlier, we build the model on the training set and predict the labels for `High` on the test set:

```r
# Fit model on training set
tree.carseats = tree(High~.-Sales, data = Carseats.train)

# Plot the tree
draw.tree(tree.carseats, nodeinfo=TRUE, cex = 0.4)
title("Classification Tree Built on Training Set")
```




```r
# Predict on test set
tree.pred = predict(tree.carseats, Carseats.test, type="class")
```

```
tree.pred
```

```
##   [1] No  Yes Yes No  Yes Yes No  Yes No  No  Yes No  Yes No  No  No  No  No
##  [19] Yes Yes No  No  Yes No  Yes Yes No  No  No  No  No  Yes No  No  No  No
##  [37] Yes Yes Yes Yes No  No  No  No  Yes Yes Yes No  No  No  No  No  No  Yes
##  [55] Yes No  Yes No  Yes No  Yes No  Yes No  No  Yes Yes Yes No  Yes Yes No
##  [73] No  No  No  No  No  No  No  No  Yes No  Yes No  No  Yes Yes No  Yes No
##  [91] No  Yes No  No  Yes Yes No  No  Yes Yes
## Levels: No Yes
```

- To calculate the test error rate, we can construct a confusion matrix and use the counter diagonal sum divided by the total counts.

```
# Obtain confusion matrix
error = table(tree.pred, High.test)
error
```

```
##          High.test
## tree.pred No Yes
##       No  39  20
##       Yes  6  35
```

```
# Test accuracy rate
sum(diag(error))/sum(error)
```

```
## [1] 0.74
```

```
# Test error rate (Classification Error)
1-sum(diag(error))/sum(error)
```

```
## [1] 0.26
```

This approach leads to correct predictions for 74% of the locations in the test set. In other words, the test error rate is 26%.

Note that this is really equivalant to

```
mean(tree.pred != High.test)
```

```
## [1] 0.26
```

4. Prune the tree using `cv.tree()` and `prune.misclass()`

Next, we consider whether pruning the tree might lead to a lower test error. To do so, primarily we have to decide what the best size of the tree should be, then we can trim the tree to this pre-determined size.

(a). Determine the best size

By 'best' size, for example, if we use classification error rate to guide the pruning process, we mean the number of terminal nodes which corresponds to the **smallest** classification error. There are multiple ways of pruning trees in R. Here we focus on a k-fold cross-validation approach.

- `cv.tree()` performs k-fold Cross-validation in order to determine the optimal level of tree complexity; cost-complexity pruning is used in order to select a sequence of trees for consideration. The argument `FUN=prune.misclass` is to indicate that misclassification error should guide the Cross-validation and pruning process, rather than the default deviance in the `cv.tree()` function. `K=10` instructs R to use a 10-fold Cross-validation in order to find the best size. The `cv.tree()` function reports the number of terminal nodes of each tree considered, as well as the corresponding error rate and the value of the cost-complexity parameter `k` used.

```r
# Set random seed
set.seed(3)

# K-Fold cross validation
cv = cv.tree(tree.carseats, FUN=prune.misclass, K=10)
# Print out cv
cv$size
```

```
##  [1] 21 16 14 12 10  8  5  4  2  1
```

```r
cv$dev
```

```
##  [1]  60  60  78  78  75  76  78  74  97 144
```

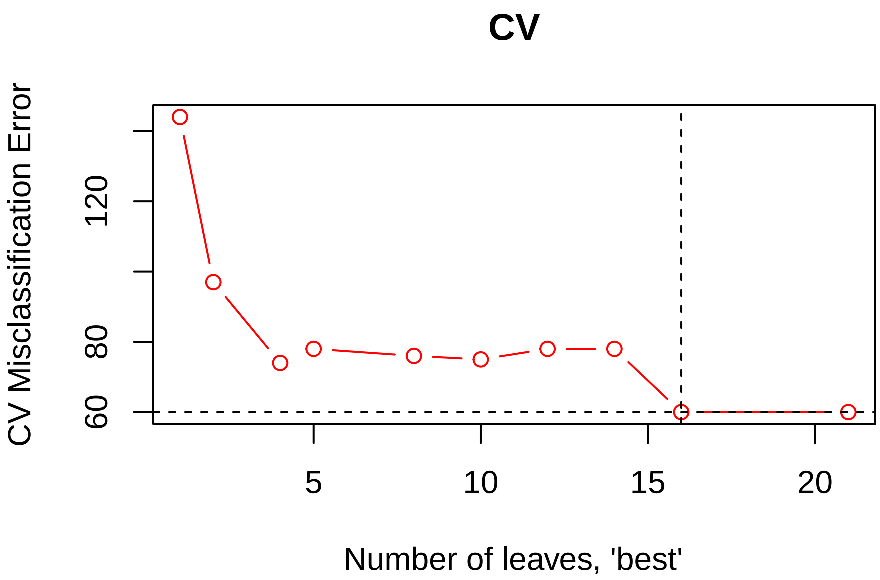
```r
# Best size
# note that there is a tie
# tree of size 21 and size 16 both have the minimum CV estimate of test error rate
# we prefer the tree with smaller size
best.cv = min(cv$size[cv$dev == min(cv$dev)])
best.cv
```

```
## [1] 16
```

Note that, despite the name, `$dev` is the Cross-validation error instead of deviance. The tree with 16 terminal nodes results in the lowest error.

(b). Error vs. Best Size plot

```r
# Plot size vs. cross-validation error rate
plot(cv$size , cv$dev, type="b",
     xlab = "Number of leaves, \'best\'", ylab = "CV Misclassification Error",
     col = "red", main="CV")
abline(v=best.cv, lty=2)
# Add lines to identify complexity parameter
min.error = which.min(cv$dev) # Get minimum error index
abline(h = cv$dev[min.error],lty = 2)
```



(c) Prune the tree and visualize it

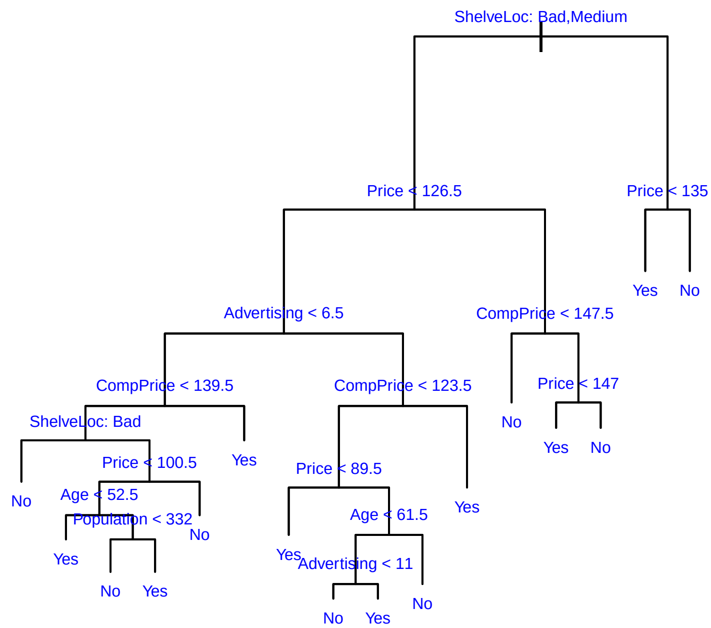
- Note that `cv.tree()` only helps determine the best tree complexity. We then use `prune.misclass` to prune a tree in order to have a tree with targeted best number of terminal nodes.

  Let's trim `tree.carseats` to have 16 nodes. This number was determined by `cv.tree()`.

```r
# Prune tree.carseats
pt.cv = prune.misclass (tree.carseats, best=best.cv)

# Plot pruned tree
plot(pt.cv)
text(pt.cv, pretty=0, col = "blue", cex = .5)
title("Pruned tree of size 16")
```

Pruned tree of size 16



(d) Calculate respective test error rate for model `pt.cv`

Recall that in (3b), we built `tree.carseats` on the training set and obtained the test error rate. In (4c), we trimmed the tree in using CV to get `pt.cv`, thus we want to see if the trimmed tree is better than `tree.carseats`, judged by the test error rate. Let's predict the labels for `High` on test set for two models and construct confusion matrices.

```r
# Predict on test set
pred.pt.cv = predict(pt.cv, Carseats.test, type="class")
# Obtain confusion matrix
err.pt.cv = table(pred.pt.cv, High.test)
err.pt.cv
```

```
##              High.test
## pred.pt.cv No Yes
##         No   39  20
##         Yes   6  35
```

```r
# Test accuracy rate
sum(diag(err.pt.cv))/sum(err.pt.cv)
```

```
## [1] 0.74
```

```r
# Test error rate (Classification Error)
1-sum(diag(err.pt.cv))/sum(err.pt.cv)
```

```
## [1] 0.26
```

The test error rate for `pt.cv` is 0.26, which is identical to (3b). Basically that means we get a simpler tree for free (without any cost in prediction error rate) by pruning. We would thus prefer the pruned tree.

Your turn

Using the original tree `tree.carseats`, perform 5-fold Cross-validation to determine the best size of the tree:

```
# Codes start here
```

Calculate the test error rate:

```
# Codes start here:

# Test set is Carseats.test
```

Credit: Adopted from *An Introduction to Statistical Learning* by *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani*