Lab 6 - Classification and Regression Trees

Loading Packages

Run the code chunk below to load packages needed for this lab.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ISLR)
library(rpart)
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

This lab is adapted from Exercise 8.9 in ISLR.

Orange Juice Purchases

The `OJ` data frame that comes with the ISLR package "contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded."

```
head(OJ)
```

```
##   Purchase WeekofPurchase StoreID PriceCH PriceMM DiscCH DiscMM SpecialCH
## 1       CH            237       1    1.75    1.99   0.00    0.0         0
## 2       CH            239       1    1.75    1.99   0.00    0.3         0
## 3       CH            245       1    1.86    2.09   0.17    0.0         0
## 4       MM            227       1    1.69    1.69   0.00    0.0         0
## 5       CH            228       7    1.69    1.69   0.00    0.0         0
## 6       CH            230       7    1.69    1.99   0.00    0.0         0
##   SpecialMM  LoyalCH SalePriceMM SalePriceCH PriceDiff Store7 PctDiscMM
## 1         0 0.500000        1.99        1.75      0.24     No  0.000000
## 2         1 0.600000        1.69        1.75     -0.06     No  0.150754
## 3         0 0.680000        2.09        1.69      0.40     No  0.000000
## 4         0 0.400000        1.69        1.69      0.00     No  0.000000
## 5         0 0.956535        1.69        1.69      0.00    Yes  0.000000
## 6         1 0.965228        1.99        1.69      0.30    Yes  0.000000
##   PctDiscCH ListPriceDiff STORE
## 1  0.000000          0.24     1
## 2  0.000000          0.24     1
## 3  0.091398          0.23     1
## 4  0.000000          0.00     1
## 5  0.000000          0.00     0
## 6  0.000000          0.30     0
```

```
dim(OJ)
```

```
## [1] 1070   18
```

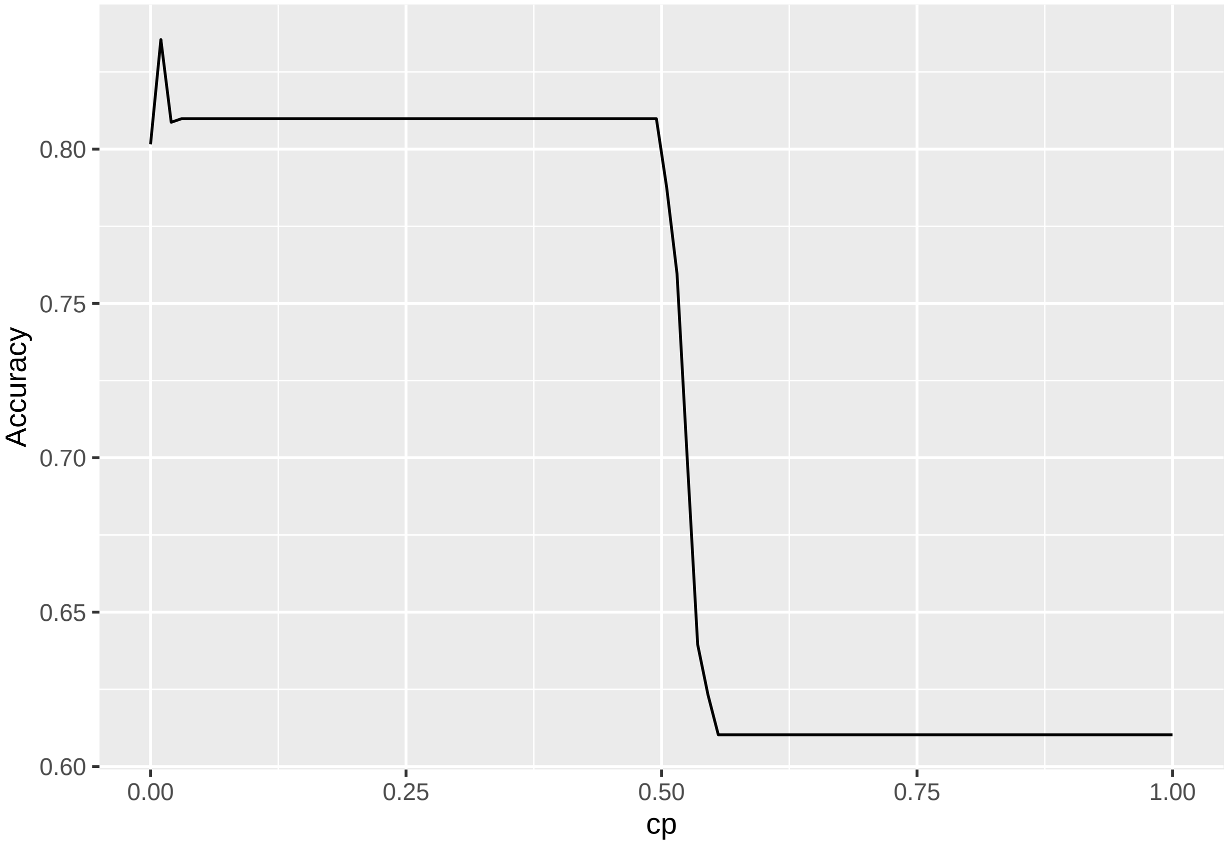
```
set.seed(71490)
train_inds <- caret::createDataPartition(OJ$Purchase, p = 0.8)
OJ_train <- OJ %>% dplyr::slice(train_inds[[1]])
OJ_test <- OJ %>% dplyr::slice(-train_inds[[1]])
```

The `Purchase` variable specifies which brand of orange juice the customer purchased (CH for Citrus Hill or MM for Minute Maid). See the help file (`?OJ`) for descriptions of the other variables.

Problem 1: Fit a classification tree to the training data with `Purchase` as the response. Use cross-validation to select the penalty parameter `cp`. You can have the train function do cross-validation for you, you don't need to implement cross-validation yourself. Make a plot of classification accuracy vs. cp.

```
tree_fit <- train(
  Purchase ~ .,
  data = OJ_train,
  method = "rpart",
  trControl = trainControl(method = "cv"),
  tuneGrid = data.frame(cp = seq(from = 0, to = 1, length = 100))
)
```

```
ggplot(data = tree_fit$results, mapping = aes(x = cp, y = Accuracy)) +
  geom_line()
```



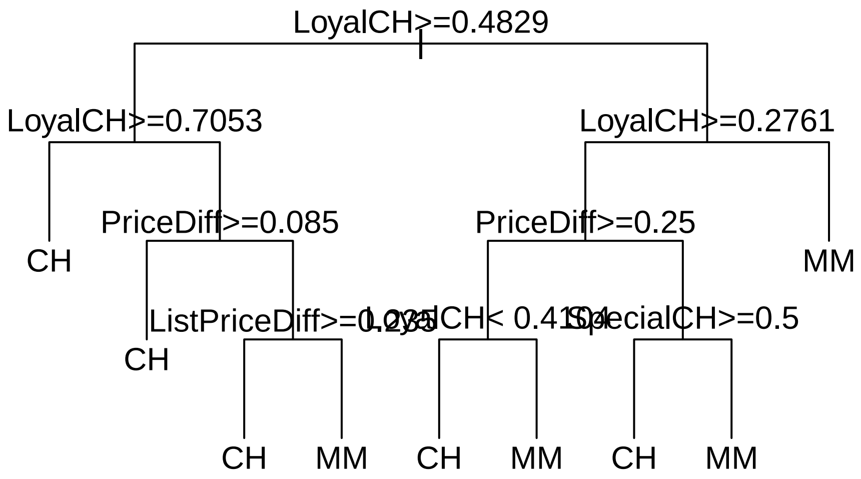
Problem 2: What is the predicted value for a customer with the characteristics given in the data frame below? You should know how to solve this problem using the `predict` function and by tracing through a picture of the estimated tree.

```
test_customer_problem3 <- data.frame(
  WeekofPurchase = 245,
  StoreID = 7,
  PriceCH = 1.90,
  PriceMM = 2.09,
  DiscCH = 0,
  DiscMM = 0,
  SpecialCH = 0,
  SpecialMM = 0,
  LoyalCH = 0.58031,
  SalePriceMM = 2.09,
  SalePriceCH = 1.9,
  PriceDiff = 0.19,
  Store7 = factor("Yes", levels = c("No", "Yes")),
  PctDiscMM = 0,
  PctDiscCH = 0,
  ListPriceDiff = 0.19,
  STORE = 0)
```

```
predict(tree_fit, newdata = test_customer_problem3)
```

```
## [1] CH
## Levels: CH MM
```

```
plot(tree_fit$finalModel, margin = 0.1, uniform = TRUE)
text(tree_fit$finalModel)
```



In the first split, the value of LoyalCH for our customer is 0.58031. This is greater than 0.4829, so we go to the left branch.

In the second split, the value of LoyalCH for our customer is 0.58031. This is not greater than 0.7053, so we go to the right branch.

In the third split, the value of PriceDiff for our customer is 0.19. This is greater than 0.085, so we go to the left branch. We have reached a leaf of the tree, so our prediction is the value listed at that leaf: "CH".

Problem 3: Find the test set error rate for the classification tree.

```
oj_preds <- predict(tree_fit, newdata = OJ_test)

mean(oj_preds != OJ_test$Purchase)
```

```
## [1] 0.2018779
```

Our test set classification error rate is 0.202.