Homework 8

Instructions

Due: 5pm on Wednesday, November 10th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 8.1

Draw an example (of your own invention) of a partition of two-dimensional feature space that could result from recursive binary splitting. Your example should contain at least 6 regions. Draw the decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions, the splitting points, and so forth.

There are a number of ways to add your "drawing" to your .Rmd file. You could...

1. Draw the figure by hand, take a picture/scan the figure, and then include in your .Rmd file using `include_graphics(...)`
2. Create a digital figure using a digital drawing application of your choice and include the resulting image suing `include_graphics()`.
3. Use `ggplot2` and `geom_segment` to manually create a partition plot.

Applied

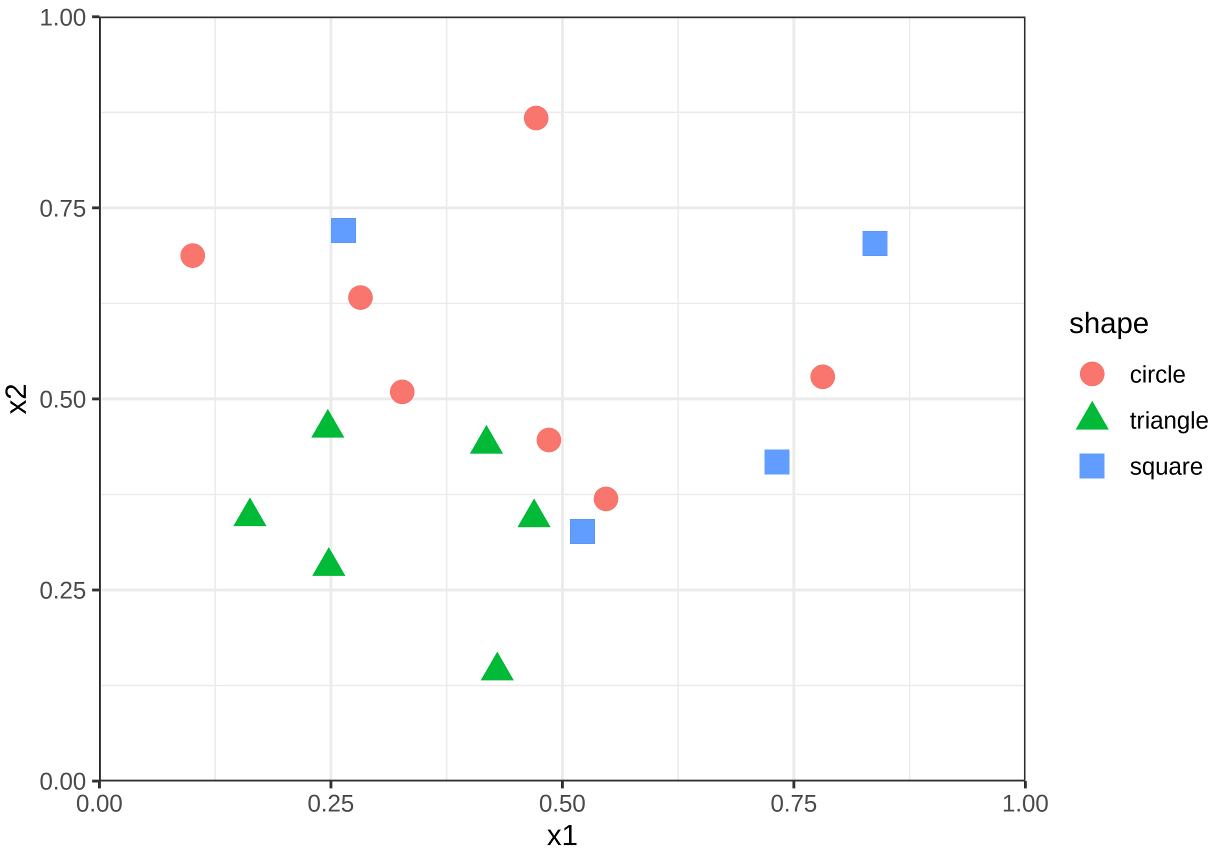
Problem 2

In Friday's class, we estimated by eye the first split in a classification tree for the `shapes` data set constructed and plotted using the code chunk below. Now we'll check to see if our graphical intuition agrees with that of the full classification tree algorithm.

```r
set.seed(100)
n <- 6
circles <- data.frame(
  shape = "circle",
  x1 = runif(n+1, 0.05,.95),
  x2 = runif(n+1, .25,.95 )
)
triangles <- data.frame(
  shape = "triangle",
  x1 = runif(n, 0.05,.6),
  x2 = runif(n,  0.05,.6)
)
squares <- data.frame(
  shape = "square",
  x1 = runif(n-2, 0.0,.95),
  x2 = runif(n-2, 0.1,.75)
)
shapes <- rbind(circles, triangles, squares) %>% mutate(shape = as.factor(shape))


g<- ggplot(shapes, aes(x = x1, y = x2, col = shape, shape = shape)) +
  geom_point(size = 4) +
  scale_x_continuous(expand = c(0, 0) , limits = c(0, 1)) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 1)) +
  theme_bw()

g
```



a. Use the `rpart` package in R to fit a tree **of depth 2 or 3** to this data set, making splits based on the *Gini index* (the default for `rpart`). *You will need to change the minimum split to something more appropriate, since the default is 20 will produce a tree with no splits.* Plot the resulting tree.

b. Use geom_segment in ggplot2 to recreate the plot above with partition boundaries (the ggplot above is saved as the object g. You can recreate the plot in your answer just by adding multiple geom_segment() layers to g graphic, i.e. g + geom_segment()...)

c. Two commonly suggested splits for this data are to make a horizontal split around $X_2 \approx 0.5$ and a vertical split around $X_1 \approx 0.5$. Was either of these the first split decided upon by your classification tree?

d. What is the benefit of the second split in the tree?

e. Which class would this model predict for the new observation with $X_1 = 0.21, X_2 = 0.56$?

f. Which classes are easiest to predict based your tree? Which classes are hard to predict or distinguish between?

Problem 3

Based on ISLR Exercise 8.9

This problem uses the OJ (as in Orange Juice, not the 1990s murder trial defendant) data set from the ISLR2 package. If you haven't used ISLR2 before, you may need to first install the package by running the code install.packages("ISLR2").

Load the data by running the following code chunk:

```
#If you haven't used ISLR2 before, you may need to first install
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 3.6.2
```

```
data(OJ)
```

The data set contains information for 18 variables on 1070 purchases of either Citrus Hill or Minute Maid Orange Juice.

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

b. Fit a tree to the training data, with Purchase as the response and all other variables as predictors.

c. Create a plot of the tree.

d. What is the training error rate? How many terminal nodes does the tree have?

e. Give a 2 - 3 sentence description of the decision rules used to predict whether a customer purchases Citrus Hill or Minute Maid. Your description should be aimed at a non-statistical audience.

f. Make predictions on the test data and produce a confusion matrix of the results. What is the test error rate?

g. Plot the cross-validated error rates for the tree, and determine the optimal size for the tree.

h. Create a pruned tree corresponding to the size you selected in the previous part. Plot the pruned tree.

i. Compare the training error rates between the pruned and unpruned trees. Which is higher? Explain why this occurs.

j. Compare the test error rates between the pruned and unpruned trees. Which is higher. Explain why this occurs.

Probelm 4

For this exercise, we will use the `College` data from the ISLR package. Familiarize yourself with this dataset before performing analysis. We will attempt to predict the `Outstate` variable, which indicates the college's out-of-state tutition rate in 1995. Load the data with the following code:

```
library(ISLR2)
data("College")
```

a. Split the data into a training and test set.
b. Create the following **four** models on the training data and tune relevant hyperparameters using cross-validation.

- The full linear model
- A well-tuned LASSO model
- A well-tuned KNN model
- A well-tuned decision tree

c. Calculate the test RMSE for each model and compare. Which model performs best on test data? What features seem most important for determining out-of-state tuition?