# Homework 7

## Instructions

**Due: 1:35pm on Wednesday, November 11th**

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 6.3*

Suppose we are estimating the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{subject to } \sum_{i=1}^{p}|\beta_i| \leq s$$

for a particular value of $s$. For parts (a) throguh (e), select exactly one option from the list of i. through v. Justify your answer.

a. As we increase s from 0, the training RSS will…
b. As we increase s from 0, the test RSS will…
c. As we increase s from 0, the variance will..
d. As we increase s from 0, the (squared) bias will…
e. As we increase s from 0, the irreducible error will…

**Options:**

i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in an inverted U shape.
iii. Steadily increase
iv. Steadily decrease.
v. Remain constant.

## Problem 2

*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 cutpoints, 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 digitial figure using a digitial drawing application of your choice and include the resulting image suing `include_graphics()`.
3. Fabricate an appropriate data set in R and use the `tree` package to create appropriate visualizations of this data within R.

# Applied

## Problem 3

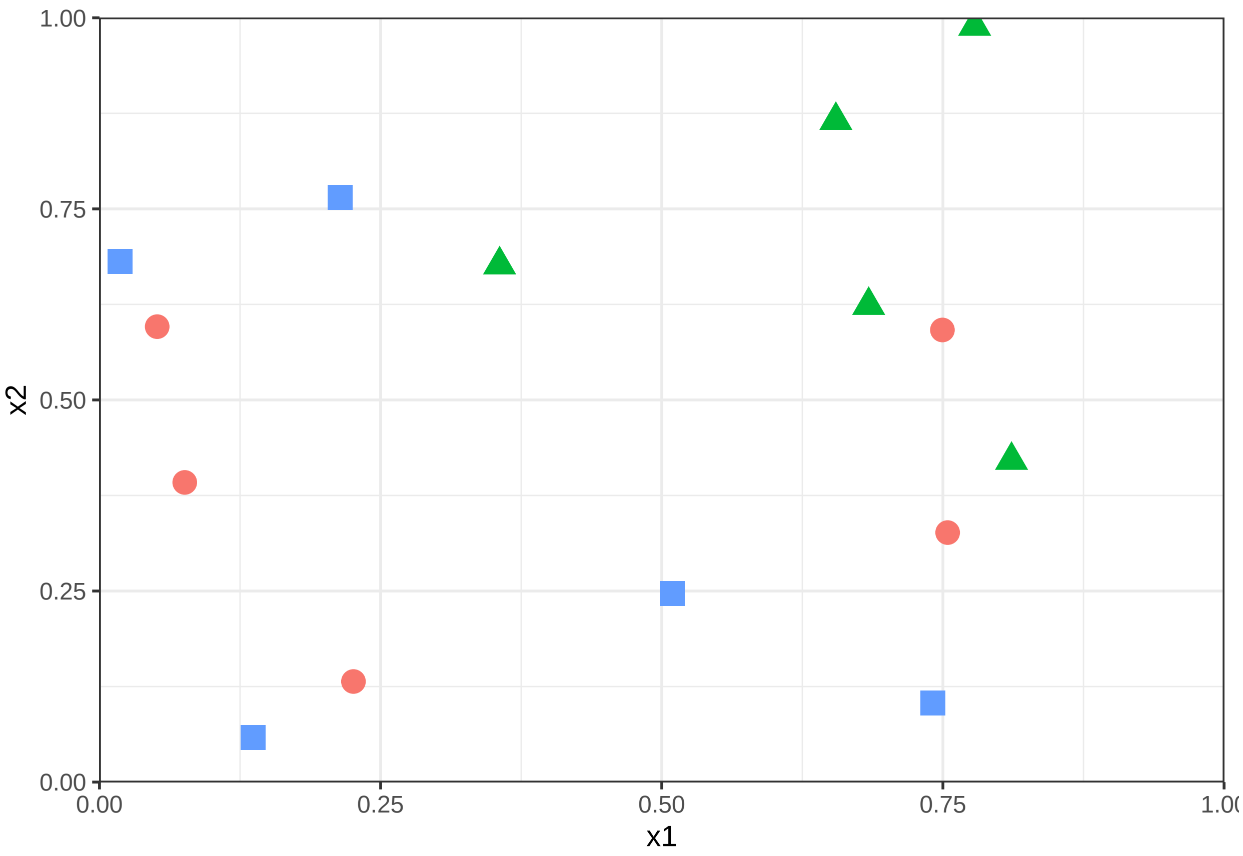
In class, we estimated by eye the first split in a classification tree for the `shapes` data set constructed and plotted using the code chunk below. Now we'llcheck to see if our graphical intuition agrees with that of the full classification tree algorithm.

```r
set.seed(75)

n <- 16
x1 <- runif(n)
x2 <- runif(n)
group <- as.factor(sample(1:3, n, replace = TRUE))
levels(group) <- c("circle", "triangle", "square")
shapes <- data.frame(x1, x2, group)
shapes[1, 2] <- .765 # tweaks to make a more interesting configuration
shapes[9, 1] <- .741
shapes <- shapes[-7, ]



library(ggplot2)
ggplot(shapes, aes(x = x1, y = x2, col = group, shape = group)) +
  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)) +
  scale_color_discrete(guide = FALSE) +
  scale_shape_discrete(guide = FALSE) +
  theme_bw()
```



a. Use the `tree` package in R to fit a full unpruned tree to this data set, making splits based on the *Gini index*. Plot the resulting tree.

b. The two most common splits that we saw in class were a horizontal split around $X_2 \approx 0.50$ and a vertical split around $X_1 \approx 0.30$. Was either of these the first split decided upon by your classification tree?

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

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

e. Now refit the tree based on the *deviance* as the splitting criterion (you set this as an argument to the `tree()` function). Plot the resulting tree. Why does this tree differ from the tree fit based on the Gini Index? Note that the deviance is defined for the classification setting as:

$$-2 \sum_m \sum_k n_{mk} \log \hat{p}_{mk}$$

---

---

## Problem 4

The following `crime_trn` data presents the amount of violent crime in a community along with other characteristics of those communities. In this problem, you will construct a regression tree model predicting crime based on these features using a training data set with information on over 100 variables recorded for 800 communities. Approximately 400 communities have been set aside as test data. Further information on this data can be found in the `data description` file in the `data` folder of the assignment repo.

```
crime_trn<-read_csv("data/crime.csv")
```

a. Review the data description and write down at least 5 predictors that you expect to have a strong association with violent crime. Briefly (1 sentence for each) explain why you theorize these variables may have strong association.

b. Create pairwise scatterplots and compute correlations for the predictors you identified above and the response variable. Do these relationships look strong/weak? Linear/non-linear? Does it seem like a transformation would be useful?

c. Create a multiple regression model for violent crime based on these predictors. Report the model `summary` along with the quartet of diagnostic plots. What are some problems indicated by the diagnostic plots?

d. Fit a regression tree to this data using the default splitting criteria (here, the deviance is essentially the RSS). Next, perform cost-complexity pruning and generate a plot showing the relationship between tree size and deviance to demonstrate the size of the best tree. Finally, construct the tree diagram for this best tree.

e. Run the following code chunk to load test data. Use your regression tree to compute the MSE for the test data set. How does it compare to the test MSE for your regression model in part c?

```
crime_tst<-read_csv("data/crime_tst.csv")
```