Midterm Exam 2

Read through the entirety of this exam before getting started

Exam Logistics

Reprinted from course website

1. The exam will be based on material from Chapters 2, 3, and 5 of ISLR, along with topics we discussed in class through Friday 10-1.
2. The exam will be made available on GitHub at 5pm PDT on Friday, 10-8. A link to the exam will be posted in the `#announcements` channel on Slack.
3. You must submit your completed exam via pushing any commits to GitHub prior to 9:00am PDT on Monday, 10-11.
4. This is a timed exam. You may take up to 3 hours to work on the exam between Friday and Monday. This time does not need to be spent consecutively. However, if you take a break from the exam, you should not spend the time during that break working on the exam, reviewing notes, or actively thinking about the problems.
5. You are responsible for keeping track of your own time on the exam and will be asked to provide an estimate for the total amount of time you spent.
6. You may freely consult the following references during the exam:

- Your ISLR and Applied Predictive Modeling textbooks
- Any course notes **you** have taken for this class
- Cheatsheats or notes from other classes, provided **you** were the one to create the notes
- Lecture slides on the course website
- Homework problems you have submitted
- Built-in RStudio help files and cheatsheets

7. You may not consult any other resources, including (but not limited to):

- Classmates
- Tutors
- Other faculty
- Other textbooks
- Online help (stackexchange, message boards, slack, etc.)

8. If you run into problems while taking the exam, document the problem in your exam and message me on slack. I will try to respond as soon as I can, but can't guarantee I will available at that moment.

Instructions

Each of the following 4 problems will be worth approximately equal number of points. Compose your answer to each problem between the bars of red stars. Show your work and justify your answers.

Problem 1

Problem 2

Problem 3

The `Auto` data from the `ISLR` package contains fuel efficiency and other data for a group of 392 vehicles. After loading the data using the following code chunk, run `?Auto` to see the data documentation and `View(Auto)` to skim the data. Your goal in this problem is to investigate the relationship between `mpg` and several of the features in the data set.

```
library(ISLR)
data(Auto)
```

a. Create appropriate visualizations comparing `mpg` to `weight`, `cylinders` and `horsepower`. Additionally, create graphics showing the individual distribution of each predictor and the response.

b. Suppose you are interested in estimating the standard deviation of the parameter on `cylinders` in a multiple regression model predicting `mpg` based on `weight`, `cylinders` and `horsepower`. Write an algorithm to create a 1000 bootstrap samples of the `cylinders` estimate.

c. Plot the distribution of your bootstrap estimates, compute the standard error of the statistic, and find the 95% confidence interval for the value of the parameter.

d. Fit a multiple regression model predicting `mpg` based on `weight`, `cylinders` and `horsepower`. How does the standard error listed in the `summary` table compare to the standard error you computed in the previous part?

e. Fit 6 more linear models, one for each subset of predictors from the list of: `weight`, `cylinders`, `horsepower`.

f. Use 5-fold cross-validation to compute error rates for each of the 7 models in parts d and e. Which model has the lowest CV error?

Problem 4

While I was commons dining the other day, I saw the following display:

```
include_graphics("img/pumpkin_comp.jpg")
```

Your task to describe a process for building a model to win this competition, using only the photo below (for reference, the Bon Appetit floor tiles ares 12" x 12"). You cannot directly weigh the pumpkin in this picture, but you may assume you can go to the store to measure and weigh other pumpkins. *(Of course, I don't actually expect you to build a model for this midterm and/or to actually go to the store to weigh pumpkins, although you are welcome to do so after the exam in order to actually win the competition!)*

```
include_graphics("img/pumpkin.jpg")
```

a. Is the goal of this task regression or classification? What is the response variable for this problem?

b. Write down several predictors (at least 5, with at least one quantitative and one categorical predictor) that you could use to make your prediction. These predictors need to be ones that you can find the value of based on the picture alone. Which of these predictors are quantitative and which are categorical? For example, one predictor that you **cannot** use is *odor* (on a scale from 0 = fresh to 5 = rancid), since you can't discern this from the picture.

c. For each of the predictors you identified in the previous part, describe the approximate relationship between that predictor and the response. For example, using the *odor* predictor, we might guess that stinky pumpkins weigh less, since some mass is lost due water evaporation during the rotting process.

d. Describe any correlations you might expect to find between any of the predictors you listed in part (b).

e. What is one *interaction* effect you may expect to see between predictors? Explain.

f. What is one predictor with a *non-linear* relationship to the response. Explain.

g. Suppose you can go to the store and measure data on 10 pumpkins. Describe how you could use information from this data estimate the variability in MSE for your model. Be sure to include specific terminology from our course.

h. Describe how you could use the data from the store pumpkins in order to compare the quality of several candidate models. Be sure to include specific terminology from our course.

i. Suppose each of the 19 students in our class (and me) independently build a model and submit a prediction based on the model. We agree to split the winnings equally if any prediction wins. Do you think we would be better off if we all collected data from the same 10 store pumpkins, or if we each

collected data on different sets of 10 pumpkins each (assume we are not able to pool our data together to get a set of 200 pumpkins). Explain your reasoning using terminology from the course.