Homework 1

Instructions

Due: 1:35pm on Wednesday, September 16th

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

For each of parts (a) through (d), indicate whether we would generally expected the performance of a complex statistical learning method to be better or worse than a low complexity method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors is small.

(b) The number of predictors is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error term $\text{Var}(\epsilon)$ is extremely high.

Problem 2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide the sample size n and the number of predictors p.

(a) We collect a set of data on the top 500 firms in the U.S. For each firm we record profit, number of employees, industry and CEO salary. We are interested in understanding which factors effect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a *success* or *failure*. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week, we record the % change in the USD/Euro, the % change in the US market, the % change in teh British market, and the % change in the German market.

Problem 3

The following problem asks you to think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which *classification* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answers.

(b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answers.

Problem 4

What are the advantages and disadvantages of a very high complexity model (versus a lower complexity model)? Under what circumstances might a high complexity approach be preferred to a lower complexity one? When might the low complexity approach be preferred?

Applied

Problem 5

To begin, load in the `Boston` data set. The `Boston` data set is part of the `MASS` library in R.

```
library(MASS)
```

Now the data set is contained in the object `Boston`. Read about the data set. By running the following code chunk. Note that the code options include `echo = F` so that the code chunk isn't printed in the .pdf output, and include `eval = F` so that the code is not run when knitting to .pdf.

(a) How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Make some (2-3) pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(d) Are there any suburbs of Boston that appear to have particularly high crime rates? Tax rate? Pupil-teacher ratios? Comment on the range of each predictor.

(e) How many of the suburbs in this data set bound the Charles river?

(f) What is the median pupil-teacher ratio among the towns in this data set?

(g) If you want to build a model to predict the average value of a home based on the other variables, what is your output/response? What is your input?