

Homework 1

Instructions

Due: 5:00pm on Wednesday, September 15th

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 expect 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 the 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.
 - (c) Describe three real-life applications in which *cluster analysis* might be useful.
-
-

Problem 4

The table below provides a training data set containing six observations, three predictors, one categorical response variable, and one quantitative response variable.

Obs.	X_1	X_2	X_3	Y_1	Y_2
1	0	3	0	Red	5
2	2	0	0	Red	3
3	0	1	3	Red	1
4	0	1	2	Green	2
5	-1	0	1	Green	3
6	1	1	1	Red	4

Suppose we wish to use the data set to make a prediction for Y_1 and for Y_2 when $X_1 = X_2 = X_3 = 0$ using K -nearest neighbors.

- a. Compute the Euclidean distance between each observation and the test point $X_1 = X_2 = X_3 = 0$.
 - b. What are the predictions for Y_1 and for Y_2 when $K = 1$? Explain.
 - c. What are the predictions for Y_1 and for Y_2 when $K = 3$? Explain.
 - d. What is the prediction for Y_1 and for Y_2 when $K = 6$? Explain.
 - e. If the Bayes decision boundary in this problem is highly non-linear, would we expect our predictions to be most accurate when K is large or when K is small? Explain.
-
-

Applied

Problem 5

The `hw1_p5.csv` file contains 20 test data points for a predictor X and a quantitative response Y .

Three models were fit on a separate training data set consisting of 60 observations; a linear model, a quadratic model, and a septic model (i.e. polynomial of degree 7). The predictions made from these models on the test data are included in `hw1_p5.csv` as well.

- Load the `hw1_p5` data set using the `read_csv(file = "...")` function (where ... should be replaced with the file path from your project directory to the csv file).
- Plot the test data, along with color-coded curves for each model (hint: you can use `geom_line` to create a curve in `ggplot2` which interpolates between points in a data frame).
- Based on inspection of the graph, which model seems to best fit the test data?
- Calculate the MSE for each of the three models. (Use R to assist with calculation, don't do this by hand.)
- Which model had the highest test MSE? Which had the lowest?
- Which model do you expect had the highest MSE on the training set? Which had the lowest?
- Suppose the true relationship between X and Y is quadratic. Which model do you think would be most accurate?

Problem 6

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.

- How many rows are in this data set? How many columns? What do the rows and columns represent?
- Make some (2-3) pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.
- Are any of the predictors associated with per capita crime rate? If so, explain the relationship.
- 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.
- How many of the suburbs in this data set bound the Charles river?
- What is the median pupil-teacher ratio among the towns in this data set?
- 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?
