Rahul Narvekar | ECON 573 | Problem Set 1

Part 1: Ex 1, 2, 4, 5, 6 from Chapter 2 of ISL.

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
(a) The sample size n is extremely large, and the number of predictors p is small.

- In this situation a flexible method will be better because it will more close fit the data and since the sample size is large

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

- In this situation a flexible method will be worse, as it would not fit well to smaller number of observations

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

- A flexible method will fit the data better

(d) The variance of the error terms, i.e. σ2 = Var(ǫ), is extremely high.

- Since the error term has a lot of variance, a flexible method would be worse than an inflexible one

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.
(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- n = 500
- p = 3
- regression
- inference

(b) We are considering launching a new product and wish to know whether it will be a success or a 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.

- n = 20
- p = 14
- classification
- prediction

(c) We are interest in predicting the ed % 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.

- n = 52, for 52 weeks in a year
- p = 4
- regression
- prediction

4. You will now 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 answer.

- Will an online training model for onboarding new employees work?
    - Response: Training works or doesn't
    - Predictors:
        - Num of modules
        - Length
        - Assessments in the training
    - Prediction
- Should Toyota create a new crossover for the north America market?
    - Response: Should or shouldn't create the crossover
    - Predictors
        - Costs
        - Sale Price
        - MPG
        - Competitors

    - Prediction
- Should Apple create a new MacBook with a screen an inch bigger than the current one?
    - Response: should or shouldn't create the MacBook
    - Predictors:
        - Price
        - Manufacturing costs
        - Distribution
        - Competitors
    - Prediction of whether the product should be created or not

(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 answer.

- How much will a car be worth n years from now?
    - Response: an estimations of the cars price
    - Predictors
        - Miles
        - Mpg
        - Body condition
        - Previous owners
        - Accidents
    - Prediction of car price
- How does college years of education compare with salary in first job
    - Response: salary amount after education amount
    - predictors
        - Years of education
        - Certifications
        - Prior experience or internships
    - Prediction of salary
- How much will a plant grow in a certain amount of days?
    - Response: heigh of the plant
    - Predictors
        - Fertilizer
        - Water
        - Sunlight
        - Soil conditions
    - Prediction of height

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

- Separating groups of people into tax brackets based on their income
    - Response: groups of tax brackets
    - Predictors
        - Salary
        - Number of children
        - Marital status
    - Prediction of tax bracket
- Separating students based on their grade
    - Response: group of students based on letter grade
    - Predictors:
        - Grade
    - Prediction of letter grade
- Separating cars into budget, mistier, and luxury brands
    - Response: classification of cars by quality
    - Predictors:
        - Types of materials
        - Brand prestige
        - Car horsepower and torque
    - Prediction of quality

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

- Advantages
    - Good for large population sizes
    - If the model is non-linear
- Disadvantages
    - If there are a lot of predictors, variance could be very high
    - Could lead to an overfit of the data
- You should use a flexible approach when there are not a lot of predictors and the general population size is large

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

- Parametric
    - Takes an assumption about the functional form
    - Once a model is found the data is fit
    - Advantages
        - Very simple to estimate the model
        - Don’t need as large of a sample size
    - Disadvantages
        - Maybe using the wrong model to estimate
        - An overfit may occur if an overly flexible model is used
- Non-Parametric
    - Do not make explicit assumptions about the functional form
    - Estimate the model to get as close the data points as possible