Homework 1

1. (2.4-1, p.52) 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.

A flexible statistical learning method would have better performance. It can improve the accuracy of the model.

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

A flexible statistical learning method will be worse than an inflexible method. A flexible might bring a higher variance and might overfit the data. Also, the model might be hard to interpret.

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

A flexible statistical learning method will be better. It allows the liner model to extend to various non-linear models.

(d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

A flexible statistical learning method will be worse. It would produce a model with very high variance.

2. (2.4-2, p.52) 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.

- It is a regression problem.
- We are most interested in inference.
- $n = 500$
  $p = 3$

(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.

- It is a classification problem.
- We are most interested in prediction.
- $n = 20$
  $p = 13$

(c) We are interesting in predicting the % change in the US dollar 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 dollar, the the % change in the British market, and the German market.

- It is a regression problem.
- We are most interested in prediction.
- $n = 52$
  $p = 2$

3. (2.4-4, p. 53) You will now think of some real-life applications for statistical learning. In each example, describe the response and the predictors and state the goal - inference or prediction.

(a) Describe two real-life applications in which classification might be useful.

(i) To predict Reddit users' decision to subscribe the Premium service

- The response: subscription (yes/no)
- The predictors: income/ Reddit age/ total number of upvotes and downvotes/ number of posts/ number of comments
- Its goal: prediction

(ii) To predict whether a student will be admitted to become a AU graduate student

- The response: admission (yes/no)
- The predictors: GPA, work experience, GRE
- Its goal: prediction

(b) Describe two real-life applications in which regression might be useful.

(i) To estimate the house price in DC

- The response: house price
- The predictors: zipcode, house age, lot area, number of bedrooms, number of bathrooms, number of floors
- Its goal: Inference

(ii) To predict a student's grade on a quiz

- The response: grade
- The predictors: length of sleep, length of time spent on the quiz, average grade of his/her homeworks
- Its goal: prediction

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

(i) To recommen advertisement to Reddit users

- The response: type of advertisement

- The predictors: users' subscribed communities, recently visited communities, content of posts, saved posts, upvoted posts
- Its goal: maybe prediction

(ii) Object recognition in digital image

- The response: what object it is
- The predictors: pixels
- Its goal: maybe prediction