Chapter 2: Exercises

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.
   1. The sample size n is extremely large, and the number of predictors p is small
      1. Better, as there is a large number of observations to train from and a more flexible method would be able to better approximate the data
   2. The number of predictors p is extremely large, and the number of observations n is small
      1. Worse, as it would be prone to overfitting the small number of observations
   3. The relationship between the predictors and response is highly non-linear
      1. Better, as assuming a non-flexible, linear model would introduce a lot of bias relative to the actual relationship
   4. The variance of the error terms, i.e. sigma2 = Var(epsilon), is extremely high
      1. Worse, as a flexible model fits the noise in the error terms and increases variance
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
   1. Inference/regression. n=500, p=3 (profit, # employees, industry)
   2. Prediction/classification. n=20, p=13
   3. Prediction/regression. N=52, p=3
3. Bias-Variance Decomposition
   1. Good
   2. Good
   3. Good
4. Real-life applications of statistical learning
   1. Three real life applications in which classification might be useful
      1. Cancer identification.
         1. Response would be 0|1, depending on the presence of cancer
         2. Predictors would be various biomarkers, scans and physical health data
         3. Prediction problem to better identify when cancer is present
      2. Political district voting results
         1. Response would which political candidate wins election
         2. Predictors would be historical voting records, population stats, demographics, economics, region
         3. Prediction problem to predict winners in a region
            1. Could also be an inference problem, to help understand the factors that influence political victory
      3. Image recognition
         1. Response would be a predicted character
         2. Predictors would be individual pixels of the image
         3. Prediction problem to identify correct characters
   2. Three real life applications in which regression might be useful
      1. Employee Pay
         1. Response would be amount of money an employee is paid
         2. Predictors would be age, sex, job title, years of experience, industry, company history
         3. Prediction problem to predict potential salary
      2. Crime Rate
         1. Response would be percentage of crimes of different types distributed throughout a location, like a city/state
         2. Predictors would be demographic, population, economics, crime history, laws
         3. Inference problem, to better understand what are the factors that drive crime in an area and how lower overall crime rate or crimes of certain kinds
      3. Housing prices
         1. Output variable would the expected cost of a home
         2. Features would be house size, age, configuration, land, location, previous prices, inflation, pricing trends
         3. Prediction problem to predict price of a home to determine whether it is a good idea to sell or buy
   3. Describe three real-life applications in which cluster analysis would be useful
      1. Consumer purchasing trends
         1. Input features would be purchasing habits, shopping habits, item history, economic buying power
         2. Help to understand what the different types of groups are shopping for, learn how to cater to different core desires of these groups
      2. Weather systems on a map
         1. Input features would be weather data, location data
         2. Help to understand weather trends, what regions get what type of precipitation, and how are different regions related to one another
      3. Animal classification
         1. Predictors would be characteristics of the animals themselves
         2. Group animals together into different classes and hierarchies, better see potential relationships between animals
5. What are the advantages and disadvantages of a very flexible (versus less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? The inverse?
   1. Advantages and Disadvantages
      1. A big advantage of a very flexible model is that it has the ability to fit data much more closely than an inflexible model. This will lead to higher accuracy and generally better performance in prediction.
      2. However, the same thing that makes a flexible model good is also what gives it it’s potential disadvantage. Highly flexible models are more prone to overfitting, and so if a model is too flexible, it will closely follow the training data but likely will overestimate the true function, and so will lead to more error and greater inaccuracies
   2. When is more flexible better?
      1. Flexible is better when you have a very non-linear relationship between the predictors and the response. It will model non-linear relationships better, and will generally be better at solving prediction problems, as flexible models tend to have higher accuracy but lower interpretability
   3. When is less flexible better?
      1. Less flexible is better when the relationship between the predictors and the response is very linear. An inflexible model is also generally easier to interpret and understand the features for, so it is better suited to solving inference problems