Problem Set 1

Alyssa Andrichik

Chapter 2 exercises

1. (a) For this Scenario it would be better for the learning method to be flexible. With such a big sample size and a smaller amount of parameters, there is less of a fear of over-fitting a model since the outliers will be more accounted for. Because the small number of predictors will mean we rely heavily on those parameters.

(b) For this Scenario it would be better for the learning method to be inflexible. With such a small sample size it necessary that there is a large amount of parameters to fit the model well and a flexible model might take into account factors that over-fit the model since there is a not a large enough sample size to be sure that the parameters are relevant.

(c) For this Scenario it would be better for the learning method to be flexible since the outcome is non-linear and the flexible learning methods are more specific and will capture systematic variability better to represent the non-linear relationship.

(d) For this Scenario it would be better for the learning method to be inflexible because the flexible method can capture variables that do not help find the accurate outcome and thus will over-fit the data to be incorrect. We don't want it to be too sensitive to the noise.

2. (a) This is a regression problem because it is focusing solely on numbers. The scenario is most interested in inference because we want to look at the data that exists and not trying to create a prediction. n = 500 p = 4

(b) This is a classification problem because we are looking at 2 category outcomes and not numbers. The scenario is most interested in predicting if a new project will be a success or a failure based on the already formulated model. n = 20 p = 14

(c) This i a regression problem because we are focusing on numbers not categories. The scenario is most interested in predicting the % change in US dollar. n = 52 p = 4

4. (a) 1- Determining if a student passed or failed a class. The predictors would be the scores on different assignments, and the weight of each assignment in relation to the over grade for a class. The response is whether a student passed or failed a class based on how well they did on all their assignments and how those assignments correlate to the weight of the score. The goal would be to predict whether a student passed or failed a class. 2- Determining if a student will get into Reed (yes, no, wait-listed). The predictors would be SAT score, ACT score, essays, GPA, interest in Reed, involvement in extra curricular, and alumni relation. The response would be if a prospie is admitted, wait-listed, or denied. The goal would be to predict whether a student was admitted, wait-listed, or denied. 3- Determining whether a dog should be washed or not. The predictors would be time since last bath, smelliness factor, dirtiness factor, and the weather. The response would be if an owner will wash their dog or not. The goal would be to predict if a dog will be washed or not.

(b) 1- Determining the factors that go into the salary of a Reed professor. The predictors would be num
2- Determining how much money should one allocate for groceries a week. The predictors would be how
3- Determining the amount of money a movie will make the first weekend in the box office. The pred

(c) 1- Education/school classes. Organize students based on their level of intelligence, success in cer
2- Animal kingdoms. Organize species based on bone structure, physical attributes, etc. to organize
3- Hospitals. Organize floors by different area of doctor expertise, health, type of disease.

5. Flexible models can fit many different possible functional forms but needs a greater number of parameters so it is best for it to be under-fitted rather than over-fitted. A less flexible approach is best will less observations because it gives the model less certainty since there is not a lot to go on do to higher variance being more difficult to interpret.

6. Non-parametric methods do not make explicit assumptions about the functional form of f, but rather seeks an estimate of f that gets as close to the data points as possible. Non-parametric methods are better are accurately fitting to a wider range of possible shapes for f. But, non-parametric approaches do need a large number of observations. Parametric methods reduce the problem of estimating f down to one of estimating a set of parameters. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate.

Additional exercises

1. (a) n=10

(b) sorry i do not know what the questions are asking.

(c)

(d)