Problem Set 1

Alice Chang

Chapter 2 exercises

Exercise 1

a) Better: In covering a broader range of functional forms and shapes for f, a flexible statistical learning method would better fit the data with a larger number of observations (n).
b) Worse: Flexible methods often do not relate response variables to a small set of predictors or estimate f through a few parameters. As a result, the model is less interpretable as it will be difficult to see how any individual predictor relates to the reponse variable, and there is a greater possibility of overfitting the data.
c) Better: More flexible methods allow for consideration of non-linear relationships between X and Y. As flexibility increases, the bias – included in the test MSE – generally decreases as well when the true f is non-linear.
d) Worse: The risk of overfitting the data with a flexible method is much higher in this case. Generally, while using a more flexible method may decrease bias, it will also increase variance even further, so that the rate of increase in variance is likely to be higher than the decrease in bias.

Exercise 2

a) Regression (quantitative/continuous response variable: CEO salary) Inference (how CEO salary affected by firm features) n = 500 (firms) p = 3 (profit, number of employees, industry)
b) Classification (qualitative response variable: success/failure) Prediction (predict success/failure of new product) n = 20 (products) p = 13 (price, marketing budget, competition price + ten variables)
c) Regression (quantitative/continous reponse variable: % change in USD/EURO) Prediction (predict % change in USD/Euro) n = 52 (weeks in 2012) p = 3 (% change in US market, % change in British market, % change in German market)

Exercise 4

a) 1. Consider whether the family income of a college student impacts whether or not they graduate from the college. The goal of this study is inference because we are observing the impact of family income on whether or not a student graduates. Response: Graduates/Does not graduate Predictors: Family income of student

2. Consider whether the size of regressive payments/monetary penalties paid by released offenders impacts whether or not they return to jail within five years of their release. The goal of this study is inference because we are observing the potential relationship between monetary penalties and the likelihood someone returns to prison. Response: Returns/Does not return to prison within five eyars of release Predictor: Total size of regressive payments of released offenders (Fees and Fines)
3. Consider the potential impact of the number of hours a high school student works a job outside of school on whether or not they drop out. The goal of this study is inference because we are evaluating the statistical significance of the effect of the amount of time a student spends working a differen job on their likelihood of dropping out. Response: Drops out/Does not drop out Predictor: Amount of time spent working a job outside of school

b) 1. Consider the extent that the monetary penalties paid by offenders in the criminal justice affects the size of their individual debt.The goal of this study is inference because we are studying how

regressive payments affects debt. Response = Size of individual debt of offenders Predictor = Total amount of regressive payments (fines, fees, bail bonds)

2. Consider the relationship between the number of Title I schools in a county and the school funding per student in the county. The goal of this study is inference because we examining whether or not there is a statisically significant relationship between the number of Title I schools and funding per student. Response = School funding per student in a county Predictor = The number of Title I eligible schoools in a county
3. Consider the impact of yearly fluctuations in temperature on the yearly growth rate of trees in a five year study of a forest region in the Midwest. The goal of this study is predictive because we want to predict the future impact of temperature fluctuations on tree growth rates in the Midweset during global warming. Response = Yearly growth rate of sample of trees Predictor: Yearly increase/decrease in temperature in region

c) 1. Consider clustering most frequent keyword user searches on new search engine by different topics (Entertainment, Politics, Academics) to create a more user-friendly sites, perhaps creating links/separate search tools for each category
2. Consider clustering illnesses/health issues relevant to kids/young adults by type to determine which units should be covered in health class.
3. Consider clustering public transit as well as companies providing transit by their different modes (eg. bus, subway, etc) to make navigating transit more accessible and understandible in urban areas.

Exercise 5

The advantages of a very flexible approach include its potential to more accurately fit a broader range of possible shapes for f. More flexilbe approaches also decrease bias and facilitate the consideration of non-linear relationships between predictors and the response variable.

A more flexible approach might be preferred when the sole goal of accurate prediction allows for the deemphasis on interpretatblity.

The disadvantages of a flexible approach is the possibility of producing complex estimates of f that make it difficult to interpret relationships between individual explanatory variables and the response variable. Furthernore, flexible methods increase variance and the potential of overfitting the data.

A less flexible approach might be preferred when the goal of inference increases the need for model interpretability.

Exercise 6

A parametric approach is a model-based approach that makes an explicit assumption about the functional form/shape of f and subsequently uses the training data to fit the model. In assuming a parametric form for f, the approach simplifies the estimation of f to the estimation of a number of parameters.

In contrast, a non-parametric approach does not make any explicit assumptions about the functional form of f, and subsequently seeks to obtain an f nearest to the data points.

The advantages of a parametric approach in regression/classification is its simplification of the estimation of f to the estimation of a set of parameters. Moreover, as fewer observations are required in this approach, the task of estimating f is much easier.

The disadvantages of a parametric approach in regression/classification is that the final model is more likely to be inaccurate and more misaligned with the true f. Using more flexible models to address this problem may also result in overfitting the data.

Additional exercises