Problem Set 1 – MATH243

Theodore Dounias

September 2, 2017

Problems 1-7, Chapter 2, ISLR

1

a) Flexible Model: Here the flexible model's tendency to overfit is mitigated by the existence of a very large amount of data, and the small amount of predictors also means that the variance problem that would normally exist is less strong, therefore we would preffer the model with less bias.

b) Inflexible: A small number of observations in the training dataset means that the flexible model would tend to overfit dramaticaly, causing the model to perform poorly in terms of variance. The very large number of predictors also matches this effect.

c) Flexible: Assuming that the linear model is a relatively inflexible model, we would choose a more flexible model than that. I hesitate to clearly define this as flexible or inflexible, it would generally just be flexible in comparison to the linear.

d) This is irreducible error and therefore irrelevant to what model we choose.

2

a) Regression since we are predicting an arithmetic value, unless we want to divide CEO income in brackets and not values. Inference because we are interested in the relationship that variables have with the quantity we are "predicting". n = 500, p = 3.

b) Classification, since we are modeling a yes/no situation. We are interested in both just predicting whether the product will fail, and potentially inferring what will make it fail so we can fix it. n = 20, p = 13.

c) Regression since we are predicting a clearly arithmetic value. We are interested in prediction. n = number of weeks in a year, p = 3.

4

a)

b) We want to analyze the factors that contribute to alcoholism. This would be an example of inference. The response is whether a person is an alcoholic or not, and predictors might include upbringing and related events, community, income, race, gender etc.

ii) We want to model the conductivity of a diode at different voltages and temperatures. The diode either conducts, or does not, making it an instance of classification. The response is conductivity, predictors include temperature, voltage, materials used, error terms and experimental circumstances etc.

iii) We want to understand the impact of news consumption on voting tendencies. The goal is inferences, the response is each person's vote, and predictors might include how much news a person consumes, where they get it from, how often they do so, etc.

b)

c) We want to predict the change in an athlete's performance after the administration of a particular drug. This is mostly predictive, but we could have use for inference methods if we want to see how/why the drug had its effects. The response is the numerical change in the athlete's performance, and the predictors are whether they are in a control group, diet, body mass and physical characteristics, gender etc.

ii) We want to model the body mass of farmed fish. This would have inference as its goal in order to optimize farming practices. The response is the fishes' body mass, the predictors would be fish type, environment, genetic characteristics, types of feed, seasonal characteristics etc.

iii) We want to predict the price of a certain stock in the long run. This is a predictive model entirely, since all that matters for whether we invest or not is the stock's level. Response is the price in USD, predictors might be exchange rates, general economic variables, type of services the company provides, stock levels of competitors etc.

iv)

v) We want to predict the race of a voter in a specific area, in order to model voting on racial lines. The group is race, predictors are neighborhood, education, income, social media data etc. The goal is prediction, since we are using data we already know is regrettably correlated to race in order to predict it.

vi) Hogwarts sorting hat uses a classification model for prediction purposes. The grouping is one of the four houses, predictors include social status, psychological profile (in quantitative variable form), personal preferences etc.

vii) The manager of Liverpool wants to create a list of football players to fill a midfielder position, and wants to submit a proposal that includes four tiers of candidates based on his preference. He would use classification based on different predictors for each player to divide them into groups.

5

A very flexible approach is useful for eliminating bias, which means we would prefer it when we have a very large amount of data to model on, considering that its disadvantage is mainly that it has high variability, which introduces the risk that the training data set will skew it in a way that renders it useless for predicting responses. Many times when modeling based on a dataset, it is also necessary to increase flexibility if the data does not follow a linear/cubic/quadratic etc path. Also, depending on what we want the model to actually do, our preferences would change; maybe for prediction purposes a more flexible model works best, but if we want to simplify in order to observe the relationship that some variables have a linear model might work better. Generally, the two tradeoffs involved are variability vs bias, and accuracy vs interpretability.

6

There are two key tradeoffs in choosing between a non-parametric and a parametric approach. First, a non-parametric approach needs a large amount of data in order to be functional; something that a parametric approach circumvents by making assumptions about the form that f takes. Second, a parametric approach tends to be more interpretable in the results it offers, meaning that if we want to understand the relationship between variables and predictors it is sometimes preferable to sacrifice accuracy by assuming f's form.

7

a)

Obs.	Euclidean Distance	Colour
1	3	Red
2	2	Red
3	3.16	Red
4	2.24	Green
5	1.41	Green
6	1.73	Red

b) With K = 1, our prediction would be Green, since the nearest point is Green.

c) With K = 3, our prediction would be Red, since in the three closest points one is Green and two are Red.

d)