Intro to Statistical Learning

Chapter 2 Solutions

Conceptual

1.

A) Flexible method would be better because the flexible method would fit the data closer. This is true with a large sample size as an inflexible method would be very difficult to fit the data.

B) Inflexible method is better because a flexible method would overfit the small number of observations.

C) Flexible method is better because fitting the data of a non-linear relationship is easier than an inflexible method.

D) Inflexible method is better because a flexible method would fit the data to the error terms and increase the variance.

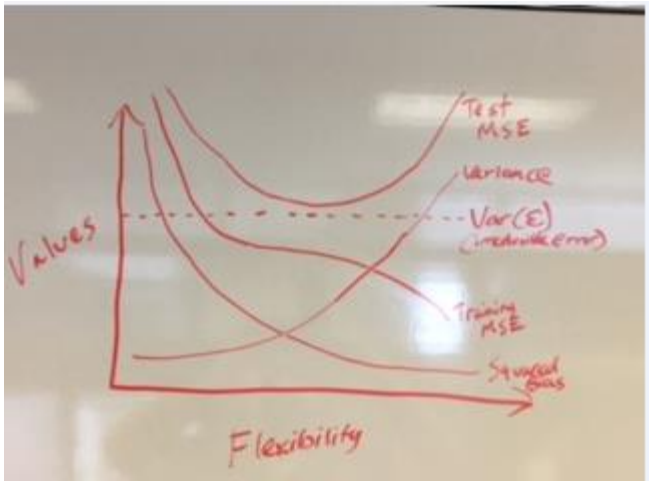
2.

A) Regression, Inference, n=500 and p=3

B) Classification, Prediction, n=20 and p=13

C) Regression, Prediction, n=52 and p=3

3.

A) 

B) Training MSE and Test MSE, see page 31, Figure 2.9; Squared Bias, Variance, and Var (ε) see page 36, Figure 2.12

The training MSE declines monotonically as flexibility increases because as flexibility increases the f curve fits the observed data more closely. Test MSE intially declines as flexibility increases but starts to level off and then starts to increase again. This happens because when a f curve yields a small training MSE but a large test MSE the data is overfitted. The squared bias decreases monotonically and the variance increases monotonically. As flexible methods are used, the variance will increase and the bias will decrease. Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. If the curve fits the observations very closely, changing any point may cause $\hat{f}$ to change considerably, and so will result in some variance. Bias refers to the error that is introduced by approximating a real-life problem by a much simpler model, so if we use a very simple model (linear regression) it is unlikely that any real-life problem has such a simple linear relationship, and so performing linear regression will result in some bias in the estimate of f. The irreducible error is a constant so it is a parallel line, this curve lies below the test MSE curve because the expected test MSE will always be greater the Var(ε).

4.

A) Classification 1: Will Checkpoint Strikeforce program in Maryland decrease alcohol related fatalities? (Response: Alcohol related fatalities increase or decrease from the program, Predictors: Budget allocation for the program, Number of law enforcement, Number of Checkpoints, Goal: Prediction). Classification 2: Should a person be accepted for a housing loan? (Response: Accepted/Not Accepted, Predictors: Income, Credit History, Length of Employment, Education, Debt, Goal: Prediction). Classification 3: Is a marketing campaign going to increase or decrease revenue? (Response: Increase/Decrease of revenue, Predictors: Budget for campaign, Number of email campaigns, Number of direct mail campaigns, Number of calls, Search Engine Optimization (Clicks, Impressions), Goal: Prediction).

B) Regression 1: What offensive statistics affect how baseball teams spend money on hitters? (Response: Total salary per year of baseball team, Predictors: Home runs, RBI, Stolen Bases, On Base Percentage, On Base Plus Slugging, Walks, Goal: Inference). Regression 2: GDP Growth in the United States (Response: What is GDP in US to be predicted to be in 20 years?, Predictors: Population, Life Expectancy, Income per capita, Government Spending, Tax Revenue, Private Sector Production, Goal: Inference). Regression 3: Does gun control legislation decrease gun related deaths? (Response: Deaths from guns, Predictors: Gun legislation bills, Gun free zones, Economic Status, Location, Law enforcement officers, Goal: Inference).

C) Cluster 1: SAT students at a school are placed into classes based on results from practice test. (Response: Classes split into math/verbal with 700-800 score, 600-700 score, etc. Predictors: Grade level, verbal courses completed previously, math courses completed, Grade point average, Goal: Prediction). Cluster 2: Customer segmentation for marketing. (Response: Customer segments to predict buying patterns, Predictors: Department Sales, Number of visits to store, Response to marketing campaigns, Total Sales, Goal: Prediction). Cluster 3: Cluster of cities in Texas that have high sports ratings (Response: Cities in Texas, Predictors: Sports ratings, Number of televisions, Income, Time spent watching sports, Goal: Prediction).

5. The advantages of a flexible vs an inflexible approach is that flexible models have a better fit for non-linear models and decreases bias. Disadvantages of a flexible vs inflexible approach is that flexible models can overfit errors too closely and increases the variance. A flexible approach is preferred if prediction is the desired result and not inference. An inflexible approach is preferred if inference is the desired result.

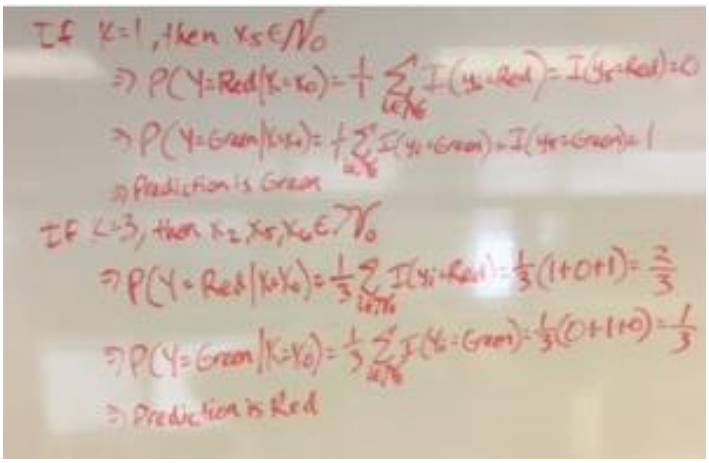
6. A parametric model reduces the problem of estimating f down to one of estimating a set of parameters while non-parametric models do not make explicit assumoptions about the functional form of f. The advantages of a parametric approach is that simplifying the model to a few parameters and not as many observations required helps with interpretability compared with a non-parametric approach. The disadvantages of the parametric approach is that non-parametric approaches are more accurate while parametric can overfit the data. Also parametric methods do not handle more observations as efficiently.

7.

A)

Test point $X_1=X_2=X_3=0$

Obs	X_1	X_2	X_3	Y	Distance
1	0	3	0	Red	$\sqrt{9}=3$
2	2	0	0	Red	$\sqrt{4}=2$
3	0	1	3	Red	$\sqrt{10}=3.16$
4	0	1	2	Green	$\sqrt{5}=2.23$
5	-1	0	1	Green	$\sqrt{2}=1.41$
6	1	1	1	Red	$\sqrt{3}=1.73$

B) and C) 

D) As K increases, the Bayes decision boundary becomes more linear. So the best values for K to have a nonlinear Bayes decision boundary is to choose a value for K that is small.

Applied

8.

A), B), C) See Ch 2 ISL Problem Set.R

9.

A), B), C), D), E), F) See Ch 2 ISL Problem Set.R

E) There seems to be more miles per gallon on a 4 cylinder vehicle than the others. Displacement, horsepower and weight are negatively correlated with mpg. Miles per gallon increases over time.

F) Yes, cylinders, horsepower, year, and origin were used as predictors from the scatterplot matrix. Displacement and weight are highly correlated with horsepower and with each other.

10.

A), B), C), D), E), F), G), H) See Ch 2 ISL Problem Set.R

C) There may be a relationship between crim and nox, rm, age, dis, lstat, and medv.