# Chapter 2 solution

1. (a) better - a more flexible approach will fit the data closer and with the

large sample size a better fit than an inflexible approach would be obtained

(b) worse - a flexible method would overfit the small number of observations

(c) better - with more degrees of freedom, a flexible model would obtain a

better fit

(d) worse - flexible methods fit to the noise in the error terms and increase variance

2. (a) regression. inference. quantitative output of CEO salary based on CEO

firm's features.

n - 500 firms in the US

p - profit, number of employees, industry

(b) classification. prediction. predicting new product's success or failure.

n - 20 similar products previously launched

p - price charged, marketing budget, comp. price, ten other variables

(c) regression. prediction. quantitative output of % change

n - 52 weeks of 2012 weekly data

p - % change in US market, % change in British market, % change in German market

3. (a) See 3a.jpg.

(b)

all 5 lines >= 0

i. (squared) bias - decreases monotonically because increases in flexibility

yield a closer fit

ii. variance - increases monotonically because increases in flexibility yield

overfit

iii. training error - decreases monotonically because increases in flexibility

yield a closer fit

iv. test error - concave up curve because increase in flexibility yields a closer

fit before it overfits

v. Bayes (irreducible) error - defines the lower limit, the test error is bounded

below by the irreducible error due to variance in the error (epsilon) in the output

values (0 <= value). When the training error is lower than the irreducible error,

overfitting has taken place.

The Bayes error rate is defined for classification problems and is determined by

the ratio of data points which lie at the 'wrong' side of the decision boundary,

(0 <= value < 1).

4. (a) i. stock market price direction, prediction, response: up, down,

input: yesterday's price movement % change, two previous day price movement %

change, etc.

ii. illness classification, inference, response: ill, healthy, input: resting

heart rate, resting breath rate, mile run time

iii. car part replacement, prediction, response: needs to be replace, good,

input: age of part, mileage used for, current amperage

(b) i. CEO salary. inference. predictors: age, industry experience, industry,

years of education. response: salary.

ii. car part replacement. inference. response: life of car part. predictors: age

of part, mileage used for, current amperage.

iii. illness classification, prediction, response: age of death,

input: current age, gender, resting heart rate, resting breath rate, mile run

time.

(c) i. cancer type clustering. diagnose cancer types more accurately.

ii. Netflix movie recommendations. recommend movies based on users who have

watched and rated similar movies.

iii. marketing survey. clustering of demographics for a product(s) to see which

clusters of consumers buy which products.

5. The advantages for a very flexible approach for regression or classification

are obtaining a better fit for non-linear models, decreasing bias.

The disadvantages for a very flexible approach for regression or classification

are requires estimating a greater number of parameters, follow the noise too

closely (overfit), increasing variance.

A more flexible approach would be preferred to a less flexible approach when we

are interested in prediction and not the interpretability of the results.

A less flexible approach would be preferred to a more flexible approach when we

are interested in inference and the interpretability of the results.

6. A parametric approach reduces the problem of estimating f down to one of

estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a functional form for f and so

requires a very large number of observations to accurately estimate f.

The advantages of a parametric approach to regression or classification are the

simplifying of modeling f to a few parameters and not as many observations are

required compared to a non-parametric approach.

The disadvantages of a parametric approach to regression or classification

are a potential to inaccurately estimate f if the form of f assumed is wrong or

to overfit the observations if more flexible models are used.

7.

(a) Obs. X1 X2 X3 Distance(0, 0, 0) Y

---------------------------------------------

1 0 3 0 3 Red

2 2 0 0 2 Red

3 0 1 3 sqrt(10) ~ 3.2 Red

4 0 1 2 sqrt(5) ~ 2.2 Green

5 -1 0 1 sqrt(2) ~ 1.4 Green

6 1 1 1 sqrt(3) ~ 1.7 Red

(b) Green. Observation #5 is the closest neighbor for K = 1.

(c) Red. Observations #2, 5, 6 are the closest neighbors for K = 3. 2 is Red,

5 is Green, and 6 is Red.

(d) Small. A small K would be flexible for a non-linear decision boundary,

whereas a large K would try to fit a more linear boundary because it takes more

points into consideration.

# 8. (a)

college = read.csv("../data/College.csv")

# 8. (b)

fix(college)

rownames(college) = college[,1]

college = college[,-1]

fix(college)

# 8. (c)

# i.

summary(college)

# ii.

pairs(college[,1:10])

# iii.

plot(college$Private, college$Outstate)

# iv.

Elite = rep("No", nrow(college))

Elite[college$Top10perc>50] = "Yes"

Elite = as.factor(Elite)

college = data.frame(college, Elite)

summary(college$Elite)

plot(college$Elite, college$Outstate)

# v.

par(mfrow=c(2,2))

hist(college$Apps)

hist(college$perc.alumni, col=2)

hist(college$S.F.Ratio, col=3, breaks=10)

hist(college$Expend, breaks=100)

# vi.

par(mfrow=c(1,1))

plot(college$Outstate, college$Grad.Rate)

# High tuition correlates to high graduation rate.

plot(college$Accept / college$Apps, college$S.F.Ratio)

# Colleges with low acceptance rate tend to have low S:F ratio.

plot(college$Top10perc, college$Grad.Rate)

# Colleges with the most students from top 10% perc don't necessarily have

# the highest graduation rate. Also, rate > 100 is erroneous!

# 9.

Auto = read.csv("../data/Auto.csv", header=T, na.strings="?")

Auto = na.omit(Auto)

dim(Auto)

summary(Auto)

# (a)

# quantitative: mpg, cylinders, displacement, horsepower, weight,

# acceleration, year

# qualitative: name, origin

# (b)

# apply the range function to the first seven columns of Auto

sapply(Auto[, 1:7], range)

# mpg cylinders displacement horsepower weight acceleration year

# [1,] 9.0 3 68 46 1613 8.0 70

# [2,] 46.6 8 455 230 5140 24.8 82

# (c)

sapply(Auto[, 1:7], mean)

# mpg cylinders displacement horsepower weight acceleration

# 23.445918 5.471939 194.411990 104.469388 2977.584184 15.541327

# year

# 75.979592

sapply(Auto[, 1:7], sd)

# mpg cylinders displacement horsepower weight acceleration

# 7.805007 1.705783 104.644004 38.491160 849.402560 2.758864

# year

# 3.683737

# (d)

newAuto = Auto[-(10:85),]

dim(newAuto) == dim(Auto) - c(76,0)

newAuto[9,] == Auto[9,]

newAuto[10,] == Auto[86,]

sapply(newAuto[, 1:7], range)

# mpg cylinders displacement horsepower weight acceleration year

# [1,] 11.0 3 68 46 1649 8.5 70

# [2,] 46.6 8 455 230 4997 24.8 82

sapply(newAuto[, 1:7], mean)

# mpg cylinders displacement horsepower weight acceleration

# 24.404430 5.373418 187.240506 100.721519 2935.971519 15.726899

# year

# 77.145570

sapply(newAuto[, 1:7], sd)

# mpg cylinders displacement horsepower weight acceleration

# 7.867283 1.654179 99.678367 35.708853 811.300208 2.693721

# year

# 3.106217

# (e)

pairs(Auto)

plot(Auto$mpg, Auto$weight)

# Heavier weight correlates with lower mpg.

plot(Auto$mpg, Auto$cylinders)

# More cylinders, less mpg.

plot(Auto$mpg, Auto$year)

# Cars become more efficient over time.

# (f)

pairs(Auto)

# See descriptions of plots in (e).

# All of the predictors show some correlation with mpg. The name predictor has

# too little observations per name though, so using this as a predictor is

# likely to result in overfitting the data and will not generalize well.

# 10.

# (a)

library(MASS)

?Boston

dim(Boston)

# 506 rows, 14 columns

# 14 features, 506 housing values in Boston suburbs

# (b)

pairs(Boston)

# X correlates with: a, b, c

# crim: age, dis, rad, tax, ptratio

# zn: indus, nox, age, lstat

# indus: age, dis

# nox: age, dis

# dis: lstat

# lstat: medv

# (c)

plot(Boston$age, Boston$crim)

# Older homes, more crime

plot(Boston$dis, Boston$crim)

# Closer to work-area, more crime

plot(Boston$rad, Boston$crim)

# Higher index of accessibility to radial highways, more crime

plot(Boston$tax, Boston$crim)

# Higher tax rate, more crime

plot(Boston$ptratio, Boston$crim)

# Higher pupil:teacher ratio, more crime

# (d)

par(mfrow=c(1,3))

hist(Boston$crim[Boston$crim>1], breaks=25)

# most cities have low crime rates, but there is a long tail: 18 suburbs appear

# to have a crime rate > 20, reaching to above 80

hist(Boston$tax, breaks=25)

# there is a large divide between suburbs with low tax rates and a peak at 660-680

hist(Boston$ptratio, breaks=25)

# a skew towards high ratios, but no particularly high ratios

# (e)

dim(subset(Boston, chas == 1))

# 35 suburbs

# (f)

median(Boston$ptratio)

# 19.05

# (g)

> t(subset(Boston, medv == min(Boston$medv)))

# 399 406

# crim 38.3518 67.9208 above 3rd quartile

# zn 0.0000 0.0000 at min

# indus 18.1000 18.1000 at 3rd quartile

# chas 0.0000 0.0000 not bounded by river

# nox 0.6930 0.6930 above 3rd quartile

# rm 5.4530 5.6830 below 1st quartile

# age 100.0000 100.0000 at max

# dis 1.4896 1.4254 below 1st quartile

# rad 24.0000 24.0000 at max

# tax 666.0000 666.0000 at 3rd quartile

# ptratio 20.2000 20.2000 at 3rd quartile

# black 396.9000 384.9700 at max; above 1st quartile

# lstat 30.5900 22.9800 above 3rd quartile

# medv 5.0000 5.0000 at min

summary(Boston)

# Not the best place to live, but certainly not the worst.

# (h)

dim(subset(Boston, rm > 7))

# 64

dim(subset(Boston, rm > 8))

# 13

summary(subset(Boston, rm > 8))

summary(Boston)

# relatively lower crime (comparing range), lower lstat (comparing range)