Introduction to Statistical Learning June 12, 2017

Chapter 2 Exercises

Name: roninlaw
Email: roninlaw

2

(a) Regression, $n = 500$, $p = 4$, Inference

(b) Classification, $n = 20$, $p = 13$, Prediction

(c) Regression, $n = 52$, $p = 4$, Prediction

4

(a) (1) *Application*: Whether someone will go to college or not.
Predictors: High school ranking, Family Support Network.
PorI: Prediction

(2) *Application*: Whether someone will go to jail or not in their lifetime.
Predictors: Personality, Support Network
PorI: Prediction

(3) *Application*: College Major
Predictors: High school courses taken, Desired salary, Interests
PorI: Prediction

(b) (1) *Application*: Average income in neighborhood @ age 30
Predictors Education Level, Career Choice
PorI: Prediction

(2) *Application*: Stock price of Company X
Predictors News articles about Company X, Value of stock from past year
PorI: Prediction

(3) *Application*: Percent on probability exam
Predictors Number of hours studied, Previous exposure to material
PorI: Prediction

(c) (1) *Application* Grouping people based on projected salary

(2) *Application* Grouping people based on musical taste

(3) *Application*: Grouping mammals based on characteristics

5

An advantage of flexible models is they can find more complex patterns. A disadvantage of flexible models is they come with the inherent risk of overfitting. A more flexible approach would be preferred when dealing with prediction problems, while a less flexible approach would be preferred when dealing with inference problems.

6

Parametric models make assumptions about the functional form of f, and then try to solve for those parameters, while non-parametric methods make no assumptions about functional form. The big disadvantage to using non-parametric methods is that they require a lot more data points.

7

Let $x_0 = (0, 0, 0)$, and x_i denote the ith observation for $i \in [5]$

(a) $d(x_0, x_1) = \sqrt{9} = 3$
$d(x_0, x_2) = \sqrt{4} = 2$
$d(x_0, x_3) = \sqrt{10}$
$d(x_0, x_4) = \sqrt{5}$
$d(x_0, x_5) = \sqrt{10}$
$d(x_0, x_6) = \sqrt{3}$

(b) The closest point is green. Therefore for $K = 1$ we predict green.

(c) Two of the three closest points are red. Therefore for $K = 3$ we predict red.

(d) Large, because as K increases the decision boundary for kNN becomes more linear.

8

(a)
```
College = read.csv("College.csv", header = T, na.strings = "?")
```

(b)
```
rownames(College) = College[,1]
College = College[, -1]
fix(College)
```

(c)
```
summary(College) # part (i)
```