HW1

Yigao Li

January 28, 2018

2.4 - 4

(a)

1. Museums want to know about their visitors including their attitudes, preferences, experiences and etc. Predictors could be visiting purpose, previous knowledge about exhibitions, visitor's companions (friends, family or individuals), job backgrounds and so on. Response would be whether they will come back visit museum again. The goal is inference because it is discussing a future event.
2. Object detection. Response is the object itself and predictors are features of objects and pixels in pictures. The goal is prediction because computers recognize objects.
3. Predicting whether a person will fail to pay his loan. Response is either success or failure. Predictors can be many factors such as income, martial status, age, social credits and so on. The goal, apparently, is prediction since we are predicting a person's credit of paying loan.

(b)

1. New-born baby weight. Response is baby's weight, and predictors are mostly related to mothers: smoke and alcohol history, mother's weight and age, boy or girl, gestation period and many others. The goal is prediction, we are predicting new-born baby weights.
2. Stock price prediction. Response is stock price change in a certain day. Predictors can be company's headline news, stock index behavior and industry trend. The goal is prediction because we do not know how the price will change before any further information.
3. Sports game results. This is interesting for people who gamble in sport games. Responses can be final result, scores, score differences, first event occurance and so on. Predictors are factors that can be used from game history of teams on both sides. The goal is also prediction.

(c)

1. Shoppers preference. Data are items in their shopping bags.
2. Museum visitors. Cluster audience into groups and discover what people like to see in their museums.
3. Product quality. Group products based on quality with various factors. Furthermore, retailers are able to set prices for different groups to expand sales.

2.4 - 7

(a)

```r
x1 <- c(0, 2, 0, 0, -1, 1)
x2 <- c(3, 0, 1, 1, 0, 1)
x3 <- c(0, 0, 3, 2, 1, 1)
y <- c("Red", "Red", "Red", "Green", "Green", "Red")
data <- data.frame(x1, x2, x3, y)
for (i in 1:6){
  obs <- c(data$x1[i], data$x2[i], data$x3[i])
  cat("Euclidean distance between observation", i, "and the test point is",
      dist(rbind(obs, c(0,0,0))), "\n")
}
```

```
## Euclidean distance between observation 1 and the test point is 3
## Euclidean distance between observation 2 and the test point is 2
## Euclidean distance between observation 3 and the test point is 3.162278
## Euclidean distance between observation 4 and the test point is 2.236068
## Euclidean distance between observation 5 and the test point is 1.414214
## Euclidean distance between observation 6 and the test point is 1.732051
```

(b)

If $k = 1$, since point $(0, 0, 0)$ is closest to observation 5 which is green, prediction for $X_1 = X_2 = X_3 = 0$ is Green as well.

(c)

If $k = 3$, since 3 closest observations from $(0, 0, 0)$ is 2, 5 and 6, in which 2 of them are red and 1 is green, prediction for $X_1 = X_2 = X_3 = 0$ is Red.

(d)

Because Bayes decision boundary is non-linear, we would expect K to be small because small K means more flexible and the boundary will be less linear.

2.4 - 9

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.3
```

```r
auto <- Auto
auto <- auto[complete.cases(auto),]
```

(a)

“mpg”, “cylinders”, “displacement”, “horsepower”, “weight”, “acceleration”, “year” are quantitative.
“origin” is qualitative.

(b)

```
## [1] "mpg"
## [1]  9.0 46.6
## [1] "cylinders"
## [1] 3 8
## [1] "displacement"
## [1]  68 455
## [1] "horsepower"
## [1]  46 230
## [1] "weight"
## [1] 1613 5140
## [1] "acceleration"
## [1]  8.0 24.8
## [1] "year"
## [1] 70 82
```

(c)

```
## [1] "mpg"
## [1] "mean"
## [1] 23.44592
## [1] "standard deviation"
## [1] 7.805007
## [1] "cylinders"
## [1] "mean"
## [1] 5.471939
## [1] "standard deviation"
## [1] 1.705783
## [1] "displacement"
## [1] "mean"
## [1] 194.412
## [1] "standard deviation"
## [1] 104.644
## [1] "horsepower"
## [1] "mean"
## [1] 104.4694
## [1] "standard deviation"

## [1] 38.49116

## [1] "weight"

## [1] "mean"

## [1] 2977.584

## [1] "standard deviation"

## [1] 849.4026

## [1] "acceleration"

## [1] "mean"

## [1] 15.54133

## [1] "standard deviation"

## [1] 2.758864

## [1] "year"

## [1] "mean"

## [1] 75.97959

## [1] "standard deviation"

## [1] 3.683737
```

(d)

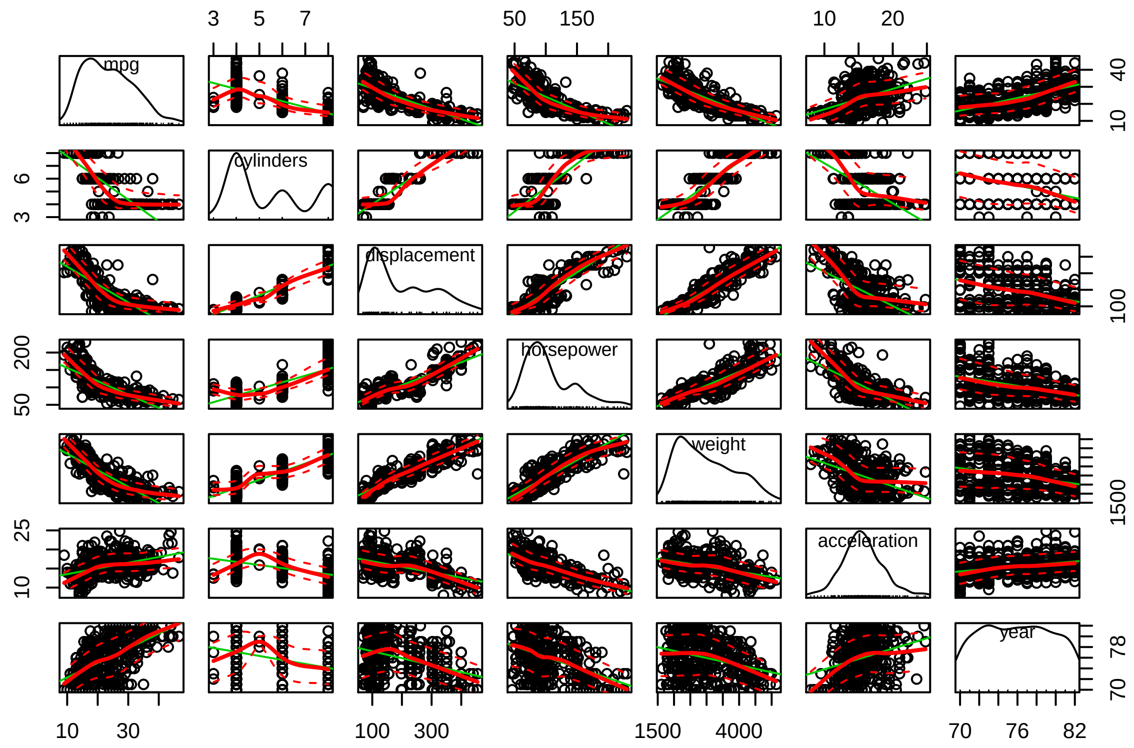
```
## [1] "mpg"

## [1] "range"

## [1] 11.0 46.6

## [1] "mean"

## [1] 24.40443

## [1] "standard deviation"

## [1] 7.867283

## [1] "cylinders"

## [1] "range"

## [1] 3 8

## [1] "mean"

## [1] 5.373418

## [1] "standard deviation"

## [1] 1.654179

## [1] "displacement"

## [1] "range"

## [1]  68 455
## [1] "mean"
## [1] 187.2405
## [1] "standard deviation"
## [1] 99.67837
## [1] "horsepower"
## [1] "range"
## [1]  46 230
## [1] "mean"
## [1] 100.7215
## [1] "standard deviation"
## [1] 35.70885
## [1] "weight"
## [1] "range"
## [1] 1649 4997
## [1] "mean"
## [1] 2935.972
## [1] "standard deviation"
## [1] 811.3002
## [1] "acceleration"
## [1] "range"
## [1]  8.5 24.8
## [1] "mean"
## [1] 15.7269
## [1] "standard deviation"
## [1] 2.693721
## [1] "year"
## [1] "range"
## [1] 70 82
## [1] "mean"
## [1] 77.14557
## [1] "standard deviation"
## [1] 3.106217
```

(e)

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```r
scatterplotMatrix(~mpg+cylinders+displacement+horsepower+weight+acceleration+year, data = auto)
```



From the scatter matrix above, we can see there are some obvious linear relations between some predictors such as:
(mpg, cylinders),
(mpg, displacement),
(mpg, horsepower),
(mpg, weight),
(mpg, year),
(cylinders, displacment),
(cylinders, weight),
(displacement, horsepower),
(displacement, acceleration),
(horsepower, weight),
(horsepower, acceleration),
(horsepower, year),
(weight, acceleration),
(acceleration, year).

(f)

From last part, mpg has linear relations with cylinders, displacement, horsepower, weight and year.

```
model.1 <- lm(mpg ~ cylinders + displacement + horsepower + weight + year, data = auto)
summary(model.1)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8714 -2.3852 -0.0895  2.0971 14.4267
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.278e+01  4.274e+00  -2.990  0.00297 **
## cylinders    -3.437e-01  3.316e-01  -1.037  0.30058
## displacement  6.996e-03  7.310e-03   0.957  0.33908
## horsepower   -7.715e-03  1.070e-02  -0.721  0.47149
## weight       -6.524e-03  5.866e-04 -11.122  < 2e-16 ***
## year          7.499e-01  5.244e-02  14.302  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.434 on 386 degrees of freedom
## Multiple R-squared:  0.8089, Adjusted R-squared:  0.8064
## F-statistic: 326.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

Summary of above linear model suggests that these variables are useful in predicting mpg because its F-statistics is large.

Exploration of the Bias-Variance Tradeoff

(a)

10 simulations with degree = 1. Residual SSEs are: 128.77
49.87
102.12
55.36
99.86
76.22
151.02
123.05
78.25
89.27
Average residual SSE is 95.379
Approximate range of the highest order coefficient is from -3 to -6

(b)

10 simulations with degree = 3. Residual SSEs are:
33.93
55.58
65.92
56.21
66.87
45.95
54.35
75.04
68.7
38.64
Average residual SSE is 56.119
The largest range of coefficient is of order 1. Range approximately from -5 to 9.

(c)

10 simulations with degree = 15. Residual SSEs are:
11.33
3.01
0.16
29.23
31.68
2.29
6.97
26.76
33.05
12.13
Average residual SSE is 15.661 The largest range of coefficient is of order 6. Range approximately from -10^5 to 8×10^5.

(d)

Simulation results in previous parts illustrate the fact of bias-variance trade-off. When we are increasing model complexity, model curve is more flexible and model residual sum of squared error gets smaller. But at the same time, coefficients of polynomials are larger and less stable. They can range extremely wide.

(e)

I would choose to use model complexity 2. From the plot, the true curve seems to be a concave parabola. 2^{nd} order polynomial model is the most appropriate to estimate. Furthermore, for most of the time, when I choose model complexity to be 3, coefficient of 3^{rd} order is very small, close to 0. Thus, I believe that 3^{rd} order term is not necessary in this model.