Cross Validation Homework (HW5)

Kunaal Raghav

Install Packages

```r
#install.packages("class")
#install.packages("boot")
library( class)
library(boot)
```

```
Warning: package 'boot' was built under R version 4.5.2
```

Part 1: Predicting a Grade

Subsection C:

```r
x <- c(90, 88, 83, 78, 85, 84)
g <- factor(c("A", "A", "A", "B", "B", "B"))
pred <- knn.cv(train = data.frame(x), cl = g, k = 3)
pred
```

```
[1] B B B B A A
Levels: A B
```

```r
mean(pred != g)
```
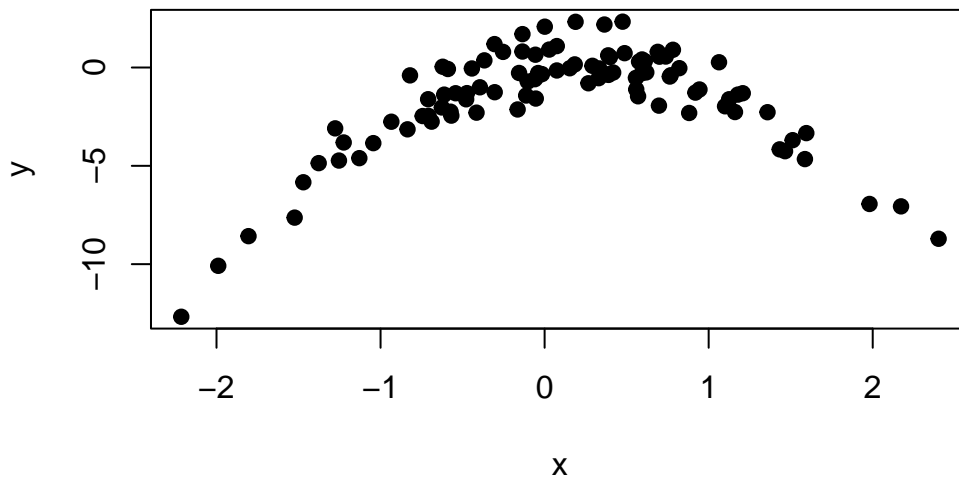
```
[1] 0.8333333
```

Part 2: Excersize 5.4.8

Subsection A

```
y <- x - 2*x^2 + rnorm(100)
```

Warning in x - 2 * x^2 + rnorm(100): longer object length is not a multiple of
shorter object length

```
set.seed(1)
x <- rnorm(100)
y <- x - 2*x^2 + rnorm(100)
sim.df <- data.frame(x, y)

plot(x, y, pch = 19)
```


Note that the first line of code can be rewritten as

$$Y = X - 2X^2 + \varepsilon, \varepsilon \sim N(0, 1)$$

n is the 00 observation, while p is the one predictor (x)

2

The plot is nonlinear , dominated by the quadratic term, with a downward opening parabola with noise

Subsection B

```
set.seed(1)
x <- rnorm(100)
y <- x - 2*x^2 + rnorm(100)
sim.df <- data.frame(x, y)

#Model 1: linear
glm1 <- glm(y ~ x, data = sim.df)
set.seed(12)
cv1 <- cv.glm(sim.df, glm1)$delta[1]


# Model 2: quadratic
glm2 <- glm(y ~ x + I(x^2), data = sim.df)
set.seed(12)
cv2 <- cv.glm(sim.df, glm2)$delta[1]


# Model 3: cubic
glm3 <- glm(y ~ x + I(x^2) + I(x^3), data = sim.df)
set.seed(12)
cv3 <- cv.glm(sim.df, glm3)$delta[1]


# Model 4: quartic
glm4 <- glm(y ~ x + I(x^2) + I(x^3) + I(x^4), data = sim.df)
set.seed(12)
cv4 <- cv.glm(sim.df, glm4)$delta[1]

cv1; cv2; cv3; cv4
```

[1] 7.288162

[1] 0.9374236

[1] 0.9566218

[1] 0.9539049

3

Part C

The quadratic model (Model 2) has the smallest LOOCV Error, as the tru model is quadratic