Stat 154 HW7

Mokssh Surve

4/1/2020

```
library('SimDesign')
```

Q4) kNN Regression

a) Function

```
kNNr <- function(k, z, x, y){

  yhat <- c()
  for (i in 1:length(z)){
    ix <- sort(abs(x-z[i]), index.return=T)$ix
    ## Note: ix = order(abs(x-z[i])) would also work
    ix <- ix[1:k]
    ynn <- y[ix]
    yhat <- c(yhat, mean(ynn))
  }
  return(yhat)
}
```

b) Running the experiment

```
yhat_tbl <- c()
z <- c(-1,0,1)
for (i in 1:100){

  set.seed(i)

  # (i) generating data set
  x <- runif(-2,2, n=30)
  e <- rnorm(length(x), mean=0, sd=1)
  f <- function(x){ return(x^3 - 3*x)}
  y <- f(x) + e

  # (ii) run kNN regression
  yhat_temp <- kNNr(k=4, z, x, y)
  yhat_tbl <- rbind(yhat_tbl, yhat_temp)
```

```
}

colnames(yhat_tbl) <- c('neg_one', 'zero', 'one')
yhat_tbl <- as.data.frame(yhat_tbl)

# Squared Bias
bias_sq__n1 <- (mean(yhat_tbl$neg_one)-2)^2
bias_sq_0 <- (mean(yhat_tbl$zero)-0)^2
bias_sq_1 <- (mean(yhat_tbl$one)+2)^2

# Variance
var_n1 <- var(yhat_tbl$neg_one)
var_0 <- var(yhat_tbl$zero)
var_1 <- var(yhat_tbl$one)


k4_sq_bias <- rbind(bias_sq__n1, bias_sq_0, bias_sq_1)
k4_var <- rbind(var_n1, var_0, var_1)

k4_sq_bias
```

```
##                     [,1]
## bias_sq__n1 2.056543e-02
## bias_sq_0   8.876057e-07
## bias_sq_1   2.311360e-02
```

```
k4_var
```

```
##             [,1]
## var_n1 0.1946582
## var_0  0.2905230
## var_1  0.2509568
```

It should be noted that the bias squared for the kNNr estimate at **point = 0**.

c) Repeat for k=1 & k=100

```
## k=1
yhat_tbl <- c()
z <- c(-1,0,1)
for (i in 1:100){

  set.seed(i)

  # (i) generating data set
  x <- runif(-2,2, n=30)
  e <- rnorm(length(x), mean=0, sd=1)
  f <- function(x){ return(x^3 - 3*x)}
  y <- f(x) + e
```

```
  # (ii) run kNN regression
  yhat_temp <- kNNr(k=1, z, x, y)
  yhat_tbl <- rbind(yhat_tbl, yhat_temp)

}

colnames(yhat_tbl) <- c('neg_one', 'zero', 'one')
yhat_tbl <- as.data.frame(yhat_tbl)

# Squared Bias
bias_sq__n1 <- (mean(yhat_tbl$neg_one)-2)^2
bias_sq_0 <- (mean(yhat_tbl$zero)-0)^2
bias_sq_1 <- (mean(yhat_tbl$one)+2)^2

# Variance
var_n1 <- var(yhat_tbl$neg_one)
var_0 <- var(yhat_tbl$zero)
var_1 <- var(yhat_tbl$one)


k1_sq_bias <- rbind(bias_sq__n1, bias_sq_0, bias_sq_1)
k1_var <- rbind(var_n1, var_0, var_1)

k1_sq_bias
```

```
##                      [,1]
## bias_sq__n1 2.049416e-04
## bias_sq_0   2.248254e-05
## bias_sq_1   1.211484e-02
```

```
k1_var
```

```
##             [,1]
## var_n1 0.8850729
## var_0  0.8937837
## var_1  1.0285162
```

```
## k=10
yhat_tbl <- c()
z <- c(-1,0,1)
for (i in 1:100){

  set.seed(i)
  # (i) generating data set
  x <- runif(-2,2, n=30)
  e <- rnorm(length(x), mean=0, sd=1)
  f <- function(x){ return(x^3 - 3*x)}
  y <- f(x) + e

  # (ii) run kNN regression
  yhat_temp <- kNNr(k=10, z, x, y)
  yhat_tbl <- rbind(yhat_tbl, yhat_temp)
```

```
}

colnames(yhat_tbl) <- c('neg_one', 'zero', 'one')
yhat_tbl <- as.data.frame(yhat_tbl)

# Squared Bias
bias_sq__n1 <- (mean(yhat_tbl$neg_one)-2)^2
bias_sq_0 <- (mean(yhat_tbl$zero)-0)^2
bias_sq_1 <- (mean(yhat_tbl$one)+2)^2

# Variance
var_n1 <- var(yhat_tbl$neg_one)
var_0 <- var(yhat_tbl$zero)
var_1 <- var(yhat_tbl$one)


k10_sq_bias <- rbind(bias_sq__n1, bias_sq_0, bias_sq_1)
k10_var <- rbind(var_n1, var_0, var_1)

k10_sq_bias
```

```
##                    [,1]
## bias_sq__n1 0.276086429
## bias_sq_0   0.005184387
## bias_sq_1   0.290785863
```

```
k10_var
```

```
##             [,1]
## var_n1 0.1364805
## var_0  0.2332195
## var_1  0.2193662
```

It is seen that as k increases, the kNN bias increases, and the variance decreases.
To summarise:

- **k increase => bias increase**
- **k increase => variance decrease**