DaigleHomework3.R

daiglechris

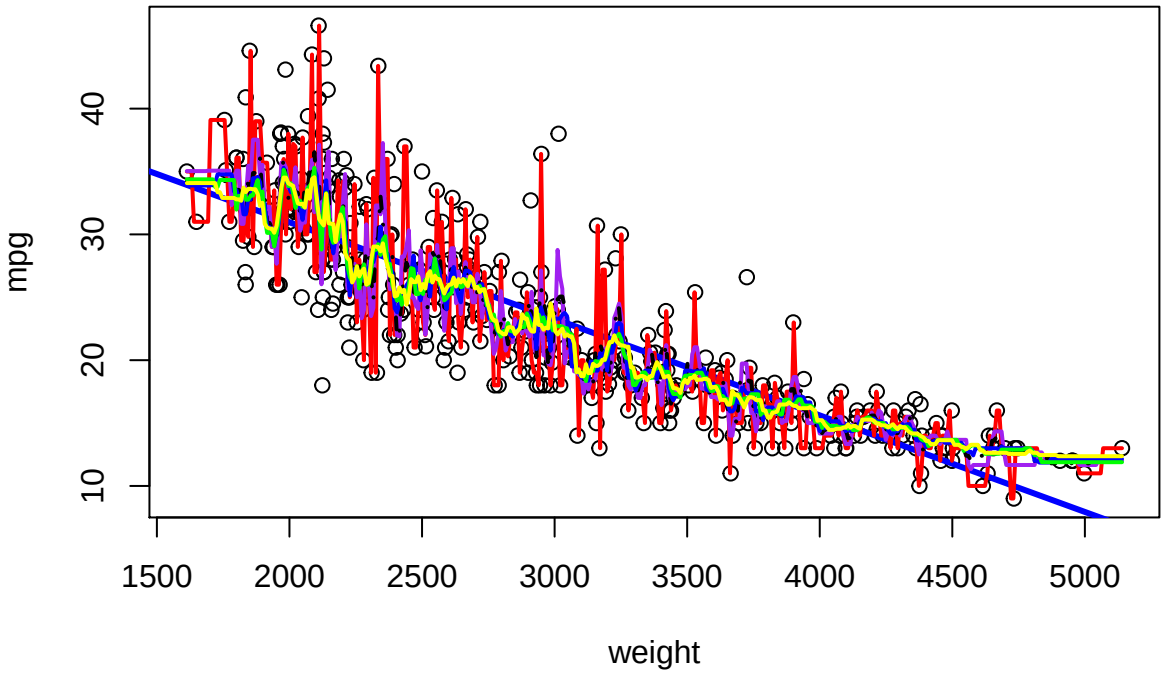
Sun Sep 23 15:00:48 2018

```
# Chris Daigle
# Exercise 3

# Download "Auto.csv" from http://www-bcf.usc.edu/~gareth/ISL/data.html and
# estimate the regression mpg on weight using the KNN method. Draw the
# regression line on the scatterplot and compare it with liner regression line.

rm(list = ls())

setwd('~/Git/MachineLearningAndBigDataWithR')
dataName <- 'Auto.csv'
Adata <- read.csv(dataName, stringsAsFactors = FALSE)

# str(Adata)
# Adata$horsepower <- as.numeric(Adata$horsepower, na.omit = TRUE)
# str(Adata)

mpg <- Adata$mpg
weight <- Adata$weight
plot(weight, mpg)
lm1 <- lm(mpg ~ weight)
summary(lm1)
```

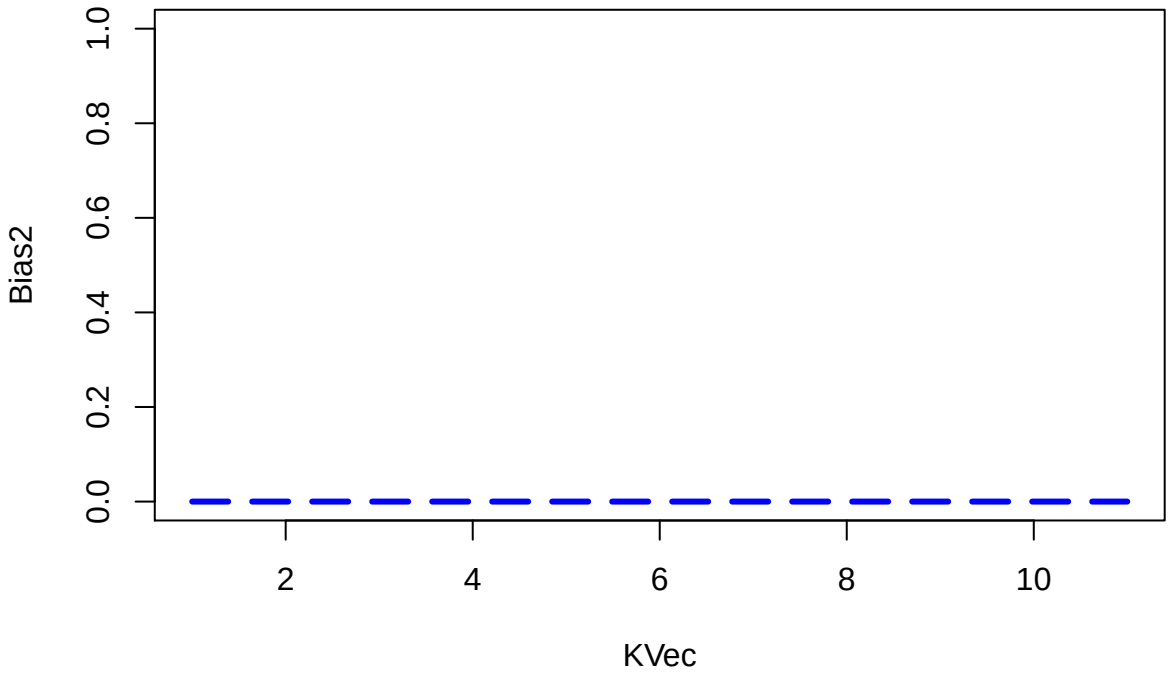
```
##
## Call:
## lm(formula = mpg ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0123  -2.8076  -0.3541   2.1145  16.4802
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.3173992  0.7962915   58.17   <2e-16 ***
## weight      -0.0076766  0.0002578  -29.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.35 on 395 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.691
## F-statistic: 886.6 on 1 and 395 DF,  p-value: < 2.2e-16
```

```
abline(lm1, lwd = 3, col = 'blue')

knn <- function(x0, X, Y, K) {
  x0 <- matrix(rep(x0, length(Y)), byrow = TRUE)
  X <- matrix(X)
  distance <- rowSums((x0 - X) ^ 2)
```

```
  rank <- order(distance)
  Y_K <- Y[rank][1:K]
  mean(Y_K)
}

x <-
  seq.int(
    from = min(weight),
    to = max(weight),
    length.out = length(weight)
  )

fhat <- matrix(rep(NA, length(x) * 6), length(x), 6)

for (j in 1:6) {
  K = 2 * j - 1
  for (i in 1:length(x)) {
    fhat[i, j] <- knn(x[i], weight, mpg, K)
  }
}

lines(x, fhat[, 1], col = 'red', lwd = 2)
lines(x, fhat[, 2], col = 'purple', lwd = 2)
lines(x,
      fhat[, 3],
      col = 'black',
      lwd = 2,
      lty = 10)
lines(x, fhat[, 4], col = 'blue', lwd = 2)
lines(x, fhat[, 5], col = 'green', lwd = 2)
lines(x, fhat[, 6], col = 'yellow', lwd = 2)
```



```r
# Bias Variance Tradeoff ####
# I am still working on this section
B <- matrix(rep(NA, length(x)), length(x), 6)
V <- matrix(rep(NA, length(x)), length(x), 6)

x <-
  seq.int(
    from = min(weight),
    to = max(weight),
    length.out = length(weight)
  )

fhat <- matrix(rep(NA, length(x) * 6), length(x), 6)

for (j in 1:6) {
  K = 2 * j - 1
  for (i in 1:length(x)) {
    fhat[i, j] <- knn(x[i], weight, mpg, K)
    B[i, j] <- knn(x[i], weight, mpg, K) - fhat[i, j]
    V[i, j] <- knn(x[i], weight, mpg, K)
  }
}


Bias <- colMeans(B)
Bias2 <- Bias ^ 2
Var <-
  c(var(V[, 1]), var(V[, 2]), var(V[, 3]), var(V[, 4]), var(V[, 5]), var(V[, 6]))
MSE <- Bias2 + Var

KVec <- 2 * (1:6) - 1
plot(
  KVec,
  Bias2,
  col = 'blue',
  type = 'l',
  lty = 5,
  lwd = 3,
  ylim = c(0, 1)
)
points(
  KVec,
  Var,
  col = 'red',
  type = 'l',
  lty = 10,
  lwd = 3
)
points(KVec,
       MSE,
       type = 'l',
       lty = 1,
       lwd = 3)
```


Bias2
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
KVec