DATA 605 Assignment 12

Bin Lin

2017-4-29

Using the stats and boot libraries in R perform a cross-validation experiment to observe the bias variance tradeoff. You'll use the auto data set from previous assignments. This dataset has 392 observations across 5 variables. We want to fit a polynomial model of various degrees using the glm function in R and then measure the cross validation error using cv.glm function.

Fit various polynomial models to compute mpg as a function of the other four variables acceleration, weight, horsepower, and displacement using glm function. For example:

glm.fit=glm(mpg~poly(disp+hp+wt+acc,2), data=auto)

cv.err5[2]=cv.glm(auto,glm.fit,K=5)$delta[1]

will fit a 2nd degree polynomial function between mpg and the remaining 4 variables and perform 5 iterations of cross-validations. This result will be stored in a cv.err5 array. cv.glm returns the estimated cross validation error and its adjusted value in a variable called delta. Please see the help on cv.glm to see more information. Once you have fit the various polynomials from degree 1 to 8, you can plot the cross-validation error function as

degree=1:8

plot(degree,cv.err5,type='b')

For you assignment, please create an R-markdown document where you load the auto data set, perform the polynomial fit and then plot the resulting 5 fold cross validation curve. Your output should show the characteristic U-shape illustrating the tradeoff between bias and variance.

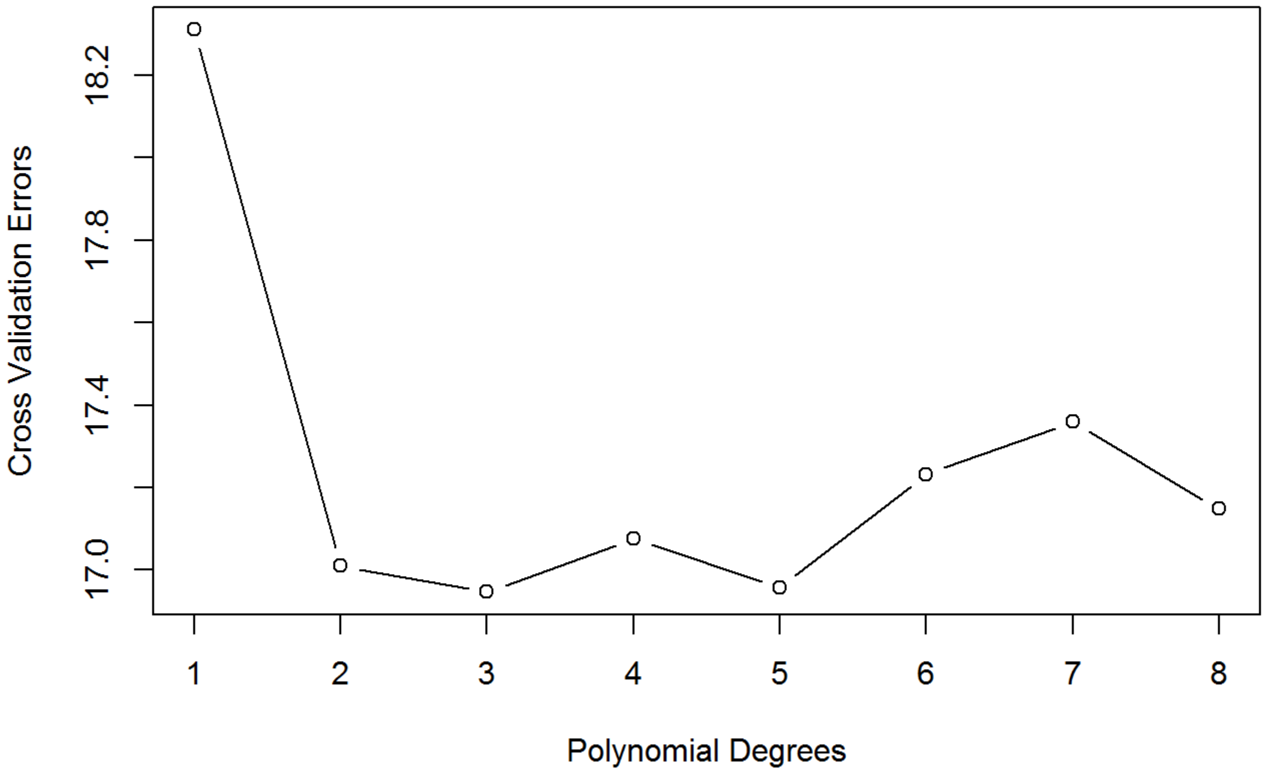
```
library(stats)
library(boot)
auto <- read.table("C:/Users/blin261/Desktop/DATA605/assign11/auto-mpg.data", stringsAsFactors =
FALSE)
colnames(auto) <- c("displacement", "horsepower", "weight", "acceleration", "mpg")
head(auto)
```

```
##   displacement horsepower weight acceleration mpg
## 1          307        130   3504         12.0  18
## 2          350        165   3693         11.5  15
## 3          318        150   3436         11.0  18
## 4          304        150   3433         12.0  16
## 5          302        140   3449         10.5  17
## 6          429        198   4341         10.0  15
```

```
cv.err5 <- c()

set.seed(888)
#Getting cross validation errors for each degree polynomial functions and save it as cv.err
for (i in 1:8)
{
  glm.fit <- glm(mpg ~ poly(displacement + horsepower + weight + acceleration, i), data = auto)
  cv.err5[i] = cv.glm(auto, glm.fit, K=5)$delta[1]
}

cv.err5
```

```
## [1] 18.31018 17.01014 16.94730 17.07645 16.95848 17.23192 17.35926 17.14822
```

```
degree=1:8

#Built a data frame contains the degree of polynomial functions and its corresponding cv errors.
polynomial_cv <- data.frame(degree, cv.err5)
plot(degree, cv.err5, type='b', xlab = "Polynomial Degrees", ylab = "Cross Validation Errors")
```



```
#Getting the minimum cv errors.
min_cv_error <- min(polynomial_cv$cv.err5)
min_cv_error
```

```
## [1] 16.9473
```

```
#Getting the degree of polynomial function which generates lowest cv error.
polynomial_cv$degree[polynomial_cv$cv.err5 == min_cv_error]
```

```
## [1] 3
```