Simple KNN

Michael Schniepp

April 13, 2016

(a)

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.3
```

```
library(ISLR)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
auto.df <- Auto

for(i in 1:length(auto.df$mpg)){
  if(auto.df$mpg[i] < median(auto.df$mpg)){
    auto.df$mpg01[i] <- 0
  } else{
    auto.df$mpg01[i] <- 1
  }
}

auto.df$mpg01 <- as.factor(auto.df$mpg01)
```

(b)

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.2.4
```

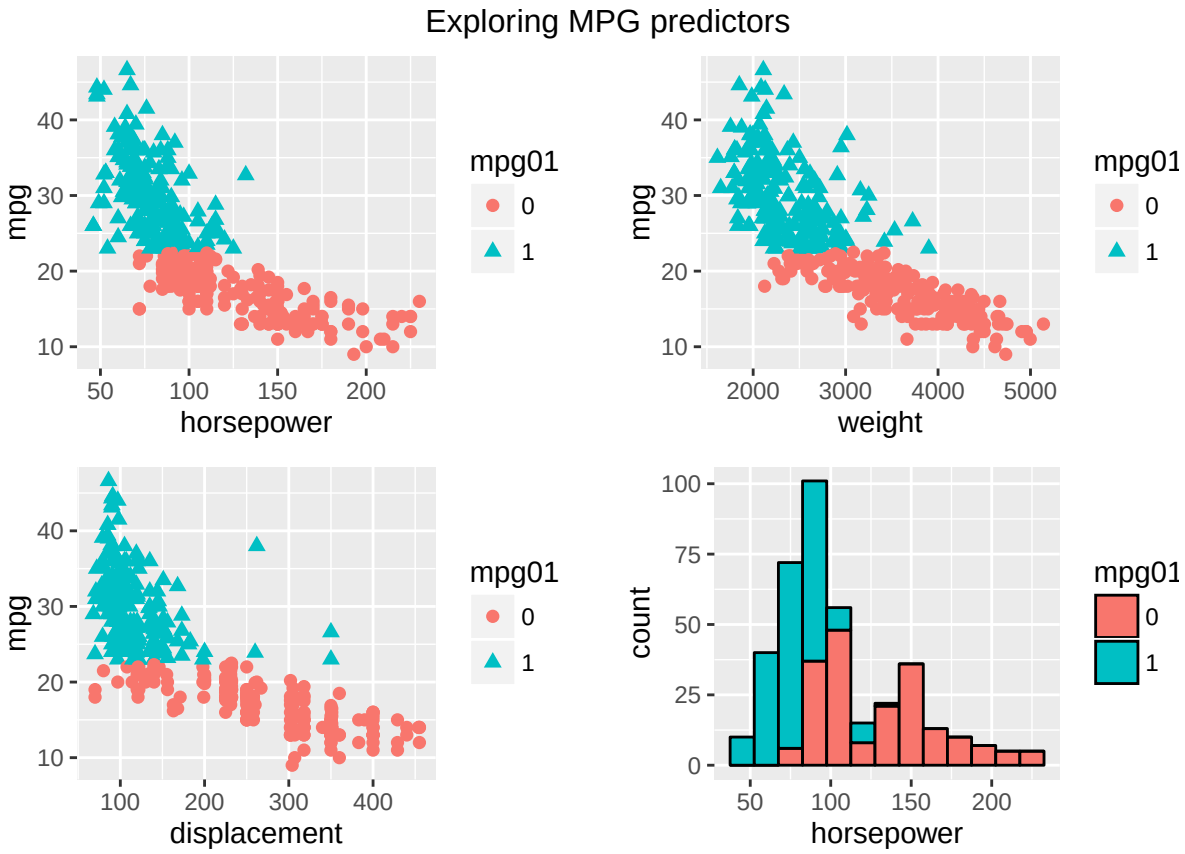
```
mpgGG1 <- ggplot(auto.df, aes(x=horsepower, y=mpg, color=mpg01, shape=mpg01)) + geom_point(size=2)

mpgGG2 <- ggplot(auto.df, aes(x=weight, y=mpg, color=mpg01, shape=mpg01)) + geom_point(size=2)

mpgGG3 <- ggplot(auto.df, aes(x=displacement, y=mpg, color=mpg01, shape=mpg01)) + geom_point(size=2)

auto.df$cylinders <- as.factor(auto.df$cylinders)
mpgBox <- ggplot(auto.df) + geom_histogram(binwidth=15,color='black',aes(horsepower, fill=mpg01))

grid.arrange(mpgGG1, mpgGG2, mpgGG3, mpgBox, nrow=2,ncol=2,  top = "Exploring MPG predictors")
```



We can see a clear divide in the data, incidcating high horsepower, high weight, and high displacement lead to a low mpg. This is a perfect candidate for a K-NN classification.

(c)

```
library(data.table)
auto.df <- as.data.table(auto.df)

smp_size <- floor(0.75 * nrow(auto.df))
set.seed(123)
train_ind <- sample(seq_len(nrow(auto.df)), size = smp_size)
train.set <- auto.df[train_ind, ]
test.set <- auto.df[-train_ind, ]
```

(d)

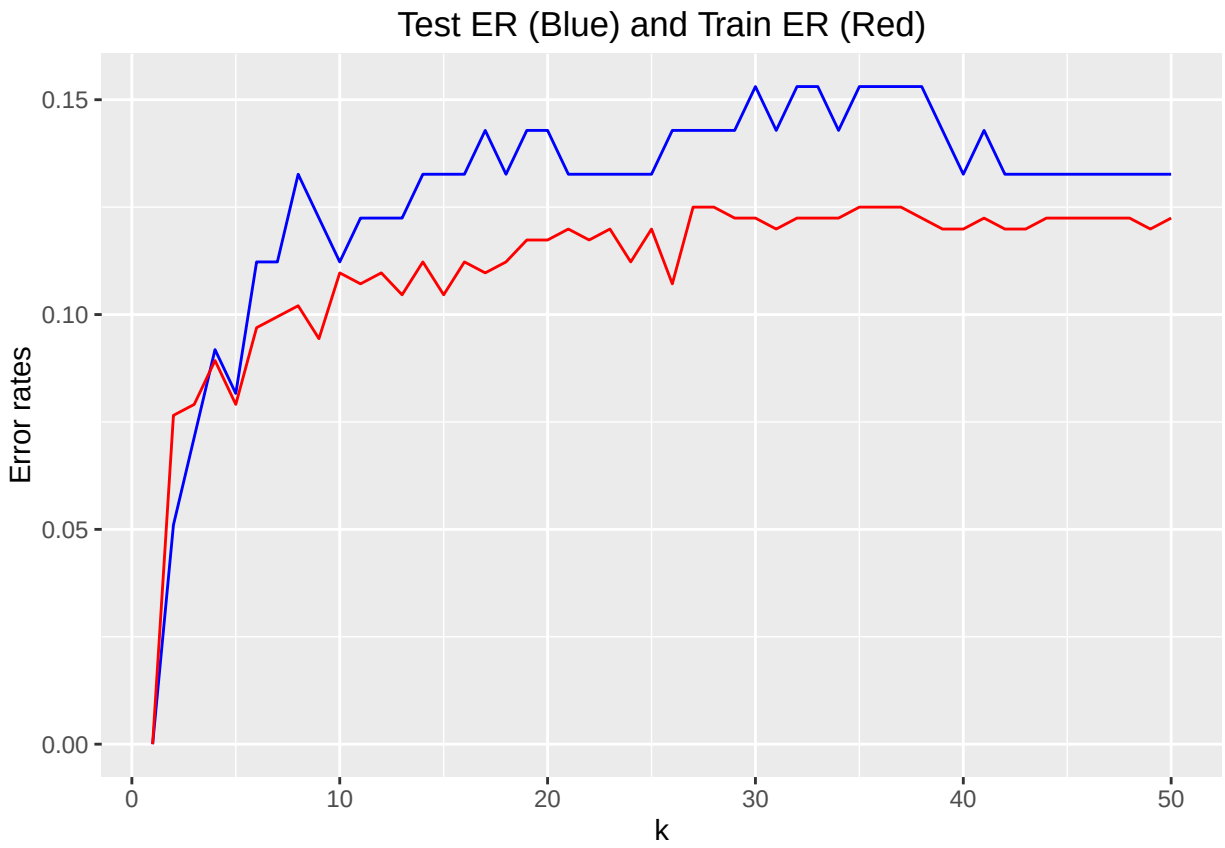
```
Xtrain <- auto.df[,list(weight,horsepower,displacement)]
Xtest <- test.set[,list(weight,horsepower,displacement)]
Ytrain <- auto.df[,mpg01]
library(class)
Ytest <- test.set[,mpg01]
p.YTest = NULL
test.error.rate = NULL
for(i in 1:50){
  set.seed(1)
  p.Ytest = knn(Xtrain,Xtest,Ytrain,k=i)
```

```
  test.error.rate[i] = mean(Ytest != p.Ytest)
}

p.YTrain = NULL
train.error.rate = NULL
for(i in 1:50){
  set.seed(1)
  p.Ytrain = knn(Xtrain,Xtrain,Ytrain,k=i)
  train.error.rate[i] = mean(Ytrain != p.Ytrain)
}


Error.rates<-data.table("k"=1:50, "Test.error.rate"=test.error.rate,"Train.error.rate"=train.error.rate)

gg4<-ggplot(Error.rates)+geom_line(aes(x=k,y=Test.error.rate), color="Blue")+geom_line(aes(x=k,y=Train.e
gg4
```



I used the weight, displacement, and horsepower variables because they showed the strongest relationships to determining the mpg of the car, as was demonstrated in the scatterplots above. The test errors range from about 0 to 15% but the best number of k would be about 5.