Predictive Modeling
Lesson 1
k-NN

Kevin Zollicoffer

09/15/2013

Introduction

The RStudio project files and accompanying artifacts, including the tex file that created this PDF, are publicly available on GitHub
`https://github.com/zollie/PASS-PredictiveModeling-knn`

Data Setup

I took the Excel spreadsheet and saved it as a CSV for easy import into R

```
> mowers <- read.csv("~/R/PASS/PredictiveModeling/knn/Mowers.csv")
> head(mowers)
```

```
  Observation Income...000.s. Lot.Size..000.s.sq..ft..
1           1            60.0                     18.4
2           2            85.5                     16.8
3           3            64.8                     21.6
4           4            61.5                     20.8
5           5            87.0                     23.6
6           6           110.1                     19.2
  Owners...1..Non.owners...2  X X.1
1                          1 NA  NA
2                          1 NA  NA
3                          1 NA  NA
4                          1 NA  NA
5                          1 NA  NA
6                          1 NA  NA
```

There were some errant extra columns on the end so I cleaned these up

```
> mowers$X <- NULL
> mowers$X.1 <- NULL
```

Partitioning

Next, partition the mowers data into 60% Train and 40% Test sets. I set the RNG seed for reproducibility

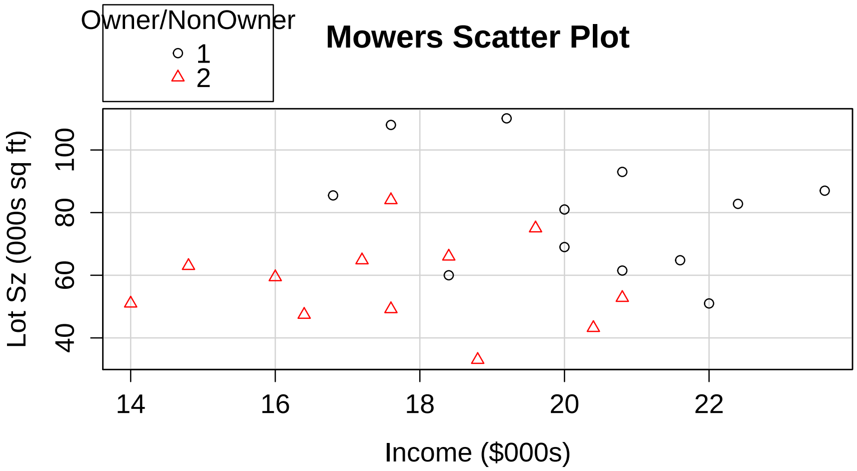
```
> set.seed(21275)
> ind <- sample(2, nrow(mowers), replace=TRUE, prob=c(0.6, 0.4))
> train <- mowers[ind==1,]
> test <- mowers[ind==2,]
> head(train)

  Observation Income...000.s. Lot.Size..000.s.sq..ft..
1           1            60.0                     18.4
2           2            85.5                     16.8
3           3            64.8                     21.6
5           5            87.0                     23.6
6           6           110.1                     19.2
8           8            82.8                     22.4
  Owners...1..Non.owners...2
1                          1
2                          1
3                          1
5                          1
6                          1
8                          1

> head(test)

   Observation Income...000.s. Lot.Size..000.s.sq..ft..
4            4            61.5                     20.8
7            7           108.0                     17.6
10          10            93.0                     20.8
12          12            81.0                     20.0
13          13            75.0                     19.6
15          15            64.8                     17.2
   Owners...1..Non.owners...2
4                           1
7                           1
10                          1
12                          1
13                          2
15                          2
```

Visualization

To better understand the data, scatterplots were created.

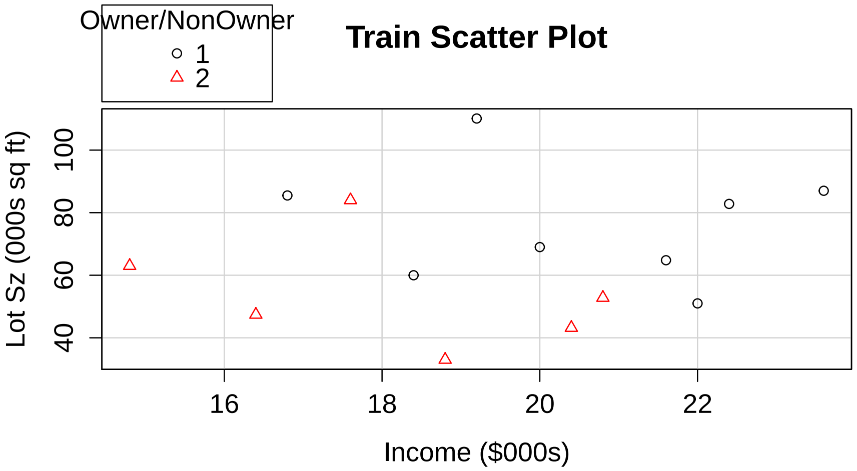
mowers scatterplot

```
> library(car)
> scatterplot(mowers[,2] ~ mowers[,3] | mowers[,4],
+               data=mowers, smoother=FALSE, reg.line=FALSE, xlab="Income ($000s)",
+               ylab="Lot Sz (000s sq ft)", main="Mowers Scatter Plot",
+               legend.title="Owner/NonOwner")
```



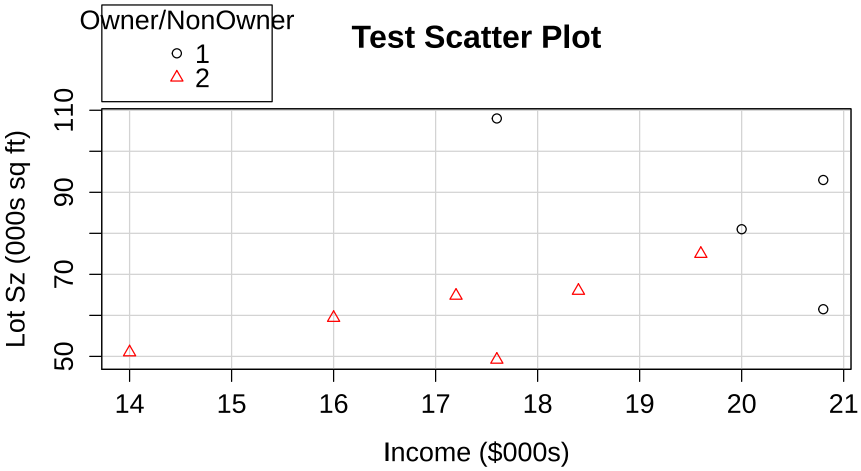
train scatterplot

```
> scatterplot(train[,2] ~ train[,3] | train[,4],
+               data=train, smoother=FALSE, reg.line=FALSE, xlab="Income ($000s)",
+               ylab="Lot Sz (000s sq ft)", main="Train Scatter Plot",
+               legend.title="Owner/NonOwner")
```



test scatterplot

```
> scatterplot(test[,2] ~ test[,3] | test[,4],
+             data=test, smoother=FALSE, reg.line=FALSE, xlab="Income ($000s)",
+             ylab="Lot Sz (000s sq ft)", main="Test Scatter Plot",
+             legend.title="Owner/NonOwner")
```



k-Nearest Neighbor

We need to reference the FNN library that contains the knn function

```
> library(FNN)
```

Factoring the categories

We have to tell the *knn* function what the real categories of the train data are

```
> levels <- factor(train[,4], labels=c("Owner", "NonOwner"))
> levels

 [1] Owner    Owner    Owner    Owner    Owner    Owner    Owner    Owner   
 [9] NonOwner NonOwner NonOwner NonOwner NonOwner NonOwner
Levels: Owner NonOwner
```

We also record the categories of the test data for determining classification error

```
> testLevels <- factor(test[,4], labels=c("Owner", "NonOwner"))
> testLevels

 [1] Owner    Owner    Owner    Owner    NonOwner NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
Levels: Owner NonOwner
```

k-NN for 1:nrow(train)

Below is a control loop to run knn() for k = {1,2,3,4,5,6,7,8,9,10,11,12,13,14}. The last statement of each loop prints the classification error rate for the current value of k. This will be summarized when answering the lesson questions in the next section.

```
> n <- nrow(train)
> z <- nrow(test)
> knn.err <- numeric(z)
> for(k in 1:n) {
+     cat("\n\nPerforming k-NN with k=",k, sep="")
+     cat("\n", "------------------------------\n")
+     pred <- knn(train, test, cl=levels, k=k, prob=TRUE)
+     print(pred)
+     knn.err[k] <- mean(as.integer(factor(pred,
+                                           levels=c("Owner", "NonOwner"),
+                                           ordered=TRUE)) != as.integer(testLevels))
+     cat("Error Rate is ", knn.err[k])
+ }

Performing k-NN with k=1
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
```

```
 [1] 1 1 1 1 1 1 1 1 1 1
attr(,"nn.index")
      [,1]
 [1,]    3
 [2,]    5
 [3,]    4
 [4,]    6
 [5,]    7
 [6,]    7
 [7,]   12
 [8,]   14
 [9,]   14
[10,]   12
attr(,"nn.dist")
          [,1]
 [1,] 3.539774
 [2,] 2.823119
 [3,] 8.296987
 [4,] 5.000000
 [5,] 7.291090
 [6,] 7.904429
 [7,] 3.698648
 [8,] 6.276942
 [9,] 6.161169
[10,] 4.766550
Levels: NonOwner Owner
Error Rate is  0.2

Performing k-NN with k=2
 ------------------------------
 [1] Owner    Owner    Owner    NonOwner Owner    NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 1.0 1.0 1.0 0.5 1.0 0.5 1.0 1.0 0.5 1.0
attr(,"nn.index")
      [,1] [,2]
 [1,]    3    1
 [2,]    5    4
 [3,]    4    6
 [4,]    6   11
 [5,]    7    6
 [6,]    7   14
 [7,]   12    9
 [8,]   14    9
 [9,]   14    7
[10,]   12    9
```

```
attr(,"nn.dist")
          [,1]      [,2]
 [1,] 3.539774  4.124318
 [2,] 2.823119 21.931712
 [3,] 8.296987 10.516653
 [4,] 5.000000  6.384356
 [5,] 7.291090  9.730365
 [6,] 7.904429  9.486833
 [7,] 3.698648  6.260990
 [8,] 6.276942  9.570789
 [9,] 6.161169 11.556816
[10,] 4.766550 11.422784
Levels: NonOwner Owner
Error Rate is  0.2

Performing k-NN with k=3
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 1.0000000 1.0000000 1.0000000 0.6666667 0.6666667 0.6666667 1.0000000
 [8] 1.0000000 0.6666667 1.0000000
attr(,"nn.index")
      [,1] [,2] [,3]
 [1,]    3    1    7
 [2,]    5    4    2
 [3,]    4    6    2
 [4,]    6   11    4
 [5,]    7    6   11
 [6,]    7   14    9
 [7,]   12    9   10
 [8,]   14    9   12
 [9,]   14    7    9
[10,]   12    9   14
attr(,"nn.dist")
          [,1]      [,2]      [,3]
 [1,] 3.539774  4.124318  9.049309
 [2,] 2.823119 21.931712 23.062741
 [3,] 8.296987 10.516653 11.672618
 [4,] 5.000000  6.384356  9.897474
 [5,] 7.291090  9.730365 10.049876
 [6,] 7.904429  9.486833 12.568214
 [7,] 3.698648  6.260990  6.916647
 [8,] 6.276942  9.570789 12.172099
 [9,] 6.161169 11.556816 14.696938
[10,] 4.766550 11.422784 12.068140
```

```
Levels: NonOwner Owner
Error Rate is  0.1

Performing k-NN with k=4
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 1.00 1.00 0.75 0.75 0.75 0.50 0.75 0.75 0.50 1.00
attr(,"nn.index")
      [,1] [,2] [,3] [,4]
 [1,]    3    1    7    8
 [2,]    5    4    2    6
 [3,]    4    6    2   11
 [4,]    6   11    4    2
 [5,]    7    6   11    3
 [6,]    7   14    9    3
 [7,]   12    9   10    8
 [8,]   14    9   12    8
 [9,]   14    7    9    3
[10,]   12    9   14   10
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]
 [1,] 3.539774  4.124318  9.049309 12.676356
 [2,] 2.823119 21.931712 23.062741 25.672553
 [3,] 8.296987 10.516653 11.672618 11.884444
 [4,] 5.000000  6.384356  9.897474 11.423222
 [5,] 7.291090  9.730365 10.049876 14.458216
 [6,] 7.904429  9.486833 12.568214 12.820296
 [7,] 3.698648  6.260990  6.916647  8.520563
 [8,] 6.276942  9.570789 12.172099 13.098091
 [9,] 6.161169 11.556816 14.696938 17.368938
[10,] 4.766550 11.422784 12.068140 12.280065
Levels: NonOwner Owner
Error Rate is  0.1

Performing k-NN with k=5
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] Owner    NonOwner
attr(,"prob")
 [1] 0.8 0.8 0.8 0.8 0.8 0.6 0.8 0.6 0.6 0.8
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5]
 [1,]    3    1    7    8    9
 [2,]    5    4    2    6   11
 [3,]    4    6    2   11    5
 [4,]    6   11    4    2    7
 [5,]    7    6   11    3    4
 [6,]    7   14    9    3    1
 [7,]   12    9   10    8   14
 [8,]   14    9   12    8    7
 [9,]   14    7    9    3    8
[10,]   12    9   14   10    8
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683
Levels: NonOwner Owner
Error Rate is  0.3

Performing k-NN with k=6
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 0.6666667 0.8333333 0.8333333 0.8333333 0.8333333 0.6666667 0.8333333
 [8] 0.6666667 0.5000000 0.8333333
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]    3    1    7    8    9   14
 [2,]    5    4    2    6   11    7
 [3,]    4    6    2   11    5    7
 [4,]    6   11    4    2    7    3
 [5,]    7    6   11    3    4    2
 [6,]    7   14    9    3    1    8
 [7,]   12    9   10    8   14   13
 [8,]   14    9   12    8    7   10
 [9,]   14    7    9    3    8   11
[10,]   12    9   14   10    8   13
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583
Levels: NonOwner Owner
Error Rate is  0.2

Performing k-NN with k=7
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 0.7142857 0.8571429 0.8571429 0.7142857 0.7142857 0.5714286 0.7142857
 [8] 0.5714286 0.5714286 0.7142857
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
 [1,]    3    1    7    8    9   14    6
 [2,]    5    4    2    6   11    7    3
 [3,]    4    6    2   11    5    7    3
 [4,]    6   11    4    2    7    3   14
 [5,]    7    6   11    3    4    2   14
 [6,]    7   14    9    3    1    8   12
 [7,]   12    9   10    8   14   13    1
 [8,]   14    9   12    8    7   10    3
 [9,]   14    7    9    3    8   11   12
[10,]   12    9   14   10    8   13    7
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
Levels: NonOwner Owner
Error Rate is  0.2

Performing k-NN with k=8
```

```
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
attr(,"prob")
 [1] 0.625 0.750 0.750 0.750 0.750 0.500 0.625 0.500 0.500 0.625
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
 [1,]    3    1    7    8    9   14    6   10
 [2,]    5    4    2    6   11    7    3   14
 [3,]    4    6    2   11    5    7    3   14
 [4,]    6   11    4    2    7    3   14    1
 [5,]    7    6   11    3    4    2   14    1
 [6,]    7   14    9    3    1    8   12   11
 [7,]   12    9   10    8   14   13    1    7
 [8,]   14    9   12    8    7   10    3    1
 [9,]   14    7    9    3    8   11   12    1
[10,]   12    9   14   10    8   13    7    1
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
          [,8]
 [1,] 21.91004
 [2,] 48.19585
 [3,] 33.66007
 [4,] 23.76047
 [5,] 19.27278
 [6,] 19.30803
 [7,] 21.90434
 [8,] 18.19670
 [9,] 19.94994
[10,] 24.19421
Levels: NonOwner Owner
Error Rate is  0.1

Performing k-NN with k=9
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] Owner    NonOwner
attr(,"prob")
 [1] 0.5555556 0.7777778 0.7777778 0.6666667 0.6666667 0.5555556 0.5555556
 [8] 0.5555556 0.5555556 0.5555556
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
 [1,]    3    1    7    8    9   14    6   10   12
 [2,]    5    4    2    6   11    7    3   14    1
 [3,]    4    6    2   11    5    7    3   14    1
 [4,]    6   11    4    2    7    3   14    1    9
 [5,]    7    6   11    3    4    2   14    1    9
 [6,]    7   14    9    3    1    8   12   11    6
 [7,]   12    9   10    8   14   13    1    7    3
 [8,]   14    9   12    8    7   10    3    1   11
 [9,]   14    7    9    3    8   11   12    1    6
[10,]   12    9   14   10    8   13    7    1    3
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
          [,8]     [,9]
 [1,] 21.91004 22.54263
 [2,] 48.19585 48.38016
 [3,] 33.66007 34.28936
 [4,] 23.76047 28.29982
 [5,] 19.27278 22.25489
 [6,] 19.30803 20.02598
 [7,] 21.90434 22.03089
 [8,] 18.19670 24.73297
 [9,] 19.94994 21.05327
[10,] 24.19421 25.47940
Levels: NonOwner Owner
Error Rate is  0.3

Performing k-NN with k=10
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner NonOwner
 [9] NonOwner NonOwner
```

```
attr(,"prob")
 [1] 0.6 0.7 0.7 0.7 0.7 0.5 0.6 0.5 0.5 0.6
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    3    1    7    8    9   14    6   10   12     2
 [2,]    5    4    2    6   11    7    3   14    1     9
 [3,]    4    6    2   11    5    7    3   14    1     9
 [4,]    6   11    4    2    7    3   14    1    9     5
 [5,]    7    6   11    3    4    2   14    1    9     8
 [6,]    7   14    9    3    1    8   12   11    6    10
 [7,]   12    9   10    8   14   13    1    7    3    11
 [8,]   14    9   12    8    7   10    3    1   11     6
 [9,]   14    7    9    3    8   11   12    1    6    10
[10,]   12    9   14   10    8   13    7    1    3    11
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
          [,8]     [,9]    [,10]
 [1,] 21.91004 22.54263 24.41311
 [2,] 48.19585 48.38016 55.74298
 [3,] 33.66007 34.28936 40.41089
 [4,] 23.76047 28.29982 29.72289
 [5,] 19.27278 22.25489 24.22313
 [6,] 19.30803 20.02598 21.85864
 [7,] 21.90434 22.03089 34.81436
 [8,] 18.19670 24.73297 26.65558
 [9,] 19.94994 21.05327 23.23446
[10,] 24.19421 25.47940 33.73366
Levels: NonOwner Owner
Error Rate is  0.1

Performing k-NN with k=11
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] Owner    NonOwner
attr(,"prob")
 [1] 0.6363636 0.7272727 0.7272727 0.7272727 0.6363636 0.5454545 0.5454545
 [8] 0.5454545 0.5454545 0.5454545
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
 [1,]    3    1    7    8    9   14    6   10   12     2     4
 [2,]    5    4    2    6   11    7    3   14    1     9     8
 [3,]    4    6    2   11    5    7    3   14    1     9     8
 [4,]    6   11    4    2    7    3   14    1    9     5     8
 [5,]    7    6   11    3    4    2   14    1    9     8    12
 [6,]    7   14    9    3    1    8   12   11    6    10     2
 [7,]   12    9   10    8   14   13    1    7    3    11     6
 [8,]   14    9   12    8    7   10    3    1   11     6    13
 [9,]   14    7    9    3    8   11   12    1    6    10     4
[10,]   12    9   14   10    8   13    7    1    3    11     6
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
          [,8]     [,9]    [,10]    [,11]
 [1,] 21.91004 22.54263 24.41311 25.67275
 [2,] 48.19585 48.38016 55.74298 57.30934
 [3,] 33.66007 34.28936 40.41089 42.02904
 [4,] 23.76047 28.29982 29.72289 30.08322
 [5,] 19.27278 22.25489 24.22313 28.91366
 [6,] 19.30803 20.02598 21.85864 24.46733
 [7,] 21.90434 22.03089 34.81436 35.39774
 [8,] 18.19670 24.73297 26.65558 26.71704
 [9,] 19.94994 21.05327 23.23446 26.34464
[10,] 24.19421 25.47940 33.73366 36.16352
Levels: NonOwner Owner
Error Rate is  0.3

Performing k-NN with k=12
 ------------------------------
 [1] Owner    Owner    Owner    Owner    Owner    Owner    NonOwner NonOwner
 [9] Owner    NonOwner
attr(,"prob")
 [1] 0.5833333 0.6666667 0.6666667 0.6666667 0.5833333 0.5833333 0.5000000
 [8] 0.5000000 0.5833333 0.5000000
```

```
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
 [1,]    3    1    7    8    9   14    6   10   12     2     4    11
 [2,]    5    4    2    6   11    7    3   14    1     9     8    12
 [3,]    4    6    2   11    5    7    3   14    1     9     8    12
 [4,]    6   11    4    2    7    3   14    1    9     5     8    12
 [5,]    7    6   11    3    4    2   14    1    9     8    12    10
 [6,]    7   14    9    3    1    8   12   11    6    10     2     4
 [7,]   12    9   10    8   14   13    1    7    3    11     6     2
 [8,]   14    9   12    8    7   10    3    1   11     6    13     2
 [9,]   14    7    9    3    8   11   12    1    6    10     4     2
[10,]   12    9   14   10    8   13    7    1    3    11     6     2
attr(,"nn.dist")
          [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
          [,8]     [,9]    [,10]    [,11]    [,12]
 [1,] 21.91004 22.54263 24.41311 25.67275 26.20095
 [2,] 48.19585 48.38016 55.74298 57.30934 62.21575
 [3,] 33.66007 34.28936 40.41089 42.02904 47.12452
 [4,] 23.76047 28.29982 29.72289 30.08322 34.98457
 [5,] 19.27278 22.25489 24.22313 28.91366 31.95121
 [6,] 19.30803 20.02598 21.85864 24.46733 25.19524
 [7,] 21.90434 22.03089 34.81436 35.39774 39.69043
 [8,] 18.19670 24.73297 26.65558 26.71704 31.17451
 [9,] 19.94994 21.05327 23.23446 26.34464 26.60470
[10,] 24.19421 25.47940 33.73366 36.16352 40.49802
Levels: NonOwner Owner
Error Rate is  0.3

Performing k-NN with k=13
 ------------------------------
 [1] Owner Owner Owner Owner Owner Owner Owner Owner Owner Owner
attr(,"prob")
 [1] 0.5384615 0.6153846 0.6153846 0.6153846 0.6153846 0.5384615 0.5384615
 [8] 0.5384615 0.5384615 0.5384615
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,]    3    1    7    8    9   14    6   10   12     2     4    11    13
 [2,]    5    4    2    6   11    7    3   14    1     9     8    12    10
 [3,]    4    6    2   11    5    7    3   14    1     9     8    12    10
 [4,]    6   11    4    2    7    3   14    1    9     5     8    12    10
 [5,]    7    6   11    3    4    2   14    1    9     8    12    10     5
 [6,]    7   14    9    3    1    8   12   11    6    10     2     4    13
 [7,]   12    9   10    8   14   13    1    7    3    11     6     2     4
 [8,]   14    9   12    8    7   10    3    1   11     6    13     2     4
 [9,]   14    7    9    3    8   11   12    1    6    10     4     2    13
[10,]   12    9   14   10    8   13    7    1    3    11     6     2     4
attr(,"nn.dist")
         [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
         [,8]     [,9]    [,10]    [,11]    [,12]    [,13]
 [1,] 21.91004 22.54263 24.41311 25.67275 26.20095 33.78239
 [2,] 48.19585 48.38016 55.74298 57.30934 62.21575 65.48954
 [3,] 33.66007 34.28936 40.41089 42.02904 47.12452 50.17171
 [4,] 23.76047 28.29982 29.72289 30.08322 34.98457 38.02631
 [5,] 19.27278 22.25489 24.22313 28.91366 31.95121 35.80740
 [6,] 19.30803 20.02598 21.85864 24.46733 25.19524 32.60061
 [7,] 21.90434 22.03089 34.81436 35.39774 39.69043 40.43315
 [8,] 18.19670 24.73297 26.65558 26.71704 31.17451 31.88291
 [9,] 19.94994 21.05327 23.23446 26.34464 26.60470 33.06297
[10,] 24.19421 25.47940 33.73366 36.16352 40.49802 41.39034
Levels: Owner
Error Rate is  0.6

Performing k-NN with k=14
 ------------------------------
 [1] Owner Owner Owner Owner Owner Owner Owner Owner Owner Owner
attr(,"prob")
 [1] 0.5714286 0.5714286 0.5714286 0.5714286 0.5714286 0.5714286 0.5714286
 [8] 0.5714286 0.5714286 0.5714286
attr(,"nn.index")
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,]    3    1    7    8    9   14    6   10   12     2     4    11    13
 [2,]    5    4    2    6   11    7    3   14    1     9     8    12    10
 [3,]    4    6    2   11    5    7    3   14    1     9     8    12    10
 [4,]    6   11    4    2    7    3   14    1    9     5     8    12    10
 [5,]    7    6   11    3    4    2   14    1    9     8    12    10     5
 [6,]    7   14    9    3    1    8   12   11    6    10     2     4    13
 [7,]   12    9   10    8   14   13    1    7    3    11     6     2     4
 [8,]   14    9   12    8    7   10    3    1   11     6    13     2     4
 [9,]   14    7    9    3    8   11   12    1    6    10     4     2    13
[10,]   12    9   14   10    8   13    7    1    3    11     6     2     4
      [,14]
 [1,]     5
 [2,]    13
 [3,]    13
 [4,]    13
 [5,]    13
 [6,]     5
 [7,]     5
 [8,]     5
 [9,]     5
[10,]     5
attr(,"nn.dist")
         [,1]      [,2]      [,3]      [,4]     [,5]     [,6]     [,7]
 [1,] 3.539774  4.124318  9.049309 12.676356 13.29248 20.95829 21.73131
 [2,] 2.823119 21.931712 23.062741 25.672553 26.01922 39.12493 43.56880
 [3,] 8.296987 10.516653 11.672618 11.884444 17.63434 24.03414 29.06682
 [4,] 5.000000  6.384356  9.897474 11.423222 12.36932 18.60108 22.27196
 [5,] 7.291090  9.730365 10.049876 14.458216 15.00000 15.49484 16.97174
 [6,] 7.904429  9.486833 12.568214 12.820296 14.88220 15.18157 18.42281
 [7,] 3.698648  6.260990  6.916647  8.520563 15.30621 16.72961 20.18118
 [8,] 6.276942  9.570789 12.172099 13.098091 14.46236 17.05286 17.81909
 [9,] 6.161169 11.556816 14.696938 17.368938 17.88743 18.26582 18.73393
[10,] 4.766550 11.422784 12.068140 12.280065 14.45683 18.65583 23.60085
         [,8]     [,9]    [,10]    [,11]    [,12]    [,13]    [,14]
 [1,] 21.91004 22.54263 24.41311 25.67275 26.20095 33.78239 48.66744
 [2,] 48.19585 48.38016 55.74298 57.30934 62.21575 65.48954 76.50124
 [3,] 33.66007 34.28936 40.41089 42.02904 47.12452 50.17171 61.22908
 [4,] 23.76047 28.29982 29.72289 30.08322 34.98457 38.02631 49.05548
 [5,] 19.27278 22.25489 24.22313 28.91366 31.95121 35.80740 42.96091
 [6,] 19.30803 20.02598 21.85864 24.46733 25.19524 32.60061 46.23949
 [7,] 21.90434 22.03089 34.81436 35.39774 39.69043 40.43315 62.09968
 [8,] 18.19670 24.73297 26.65558 26.71704 31.17451 31.88291 52.44740
 [9,] 19.94994 21.05327 23.23446 26.34464 26.60470 33.06297 46.28661
[10,] 24.19421 25.47940 33.73366 36.16352 40.49802 41.39034 61.72398
Levels: Owner
Error Rate is  0.6
```

Lesson 1 Question and Answer

1

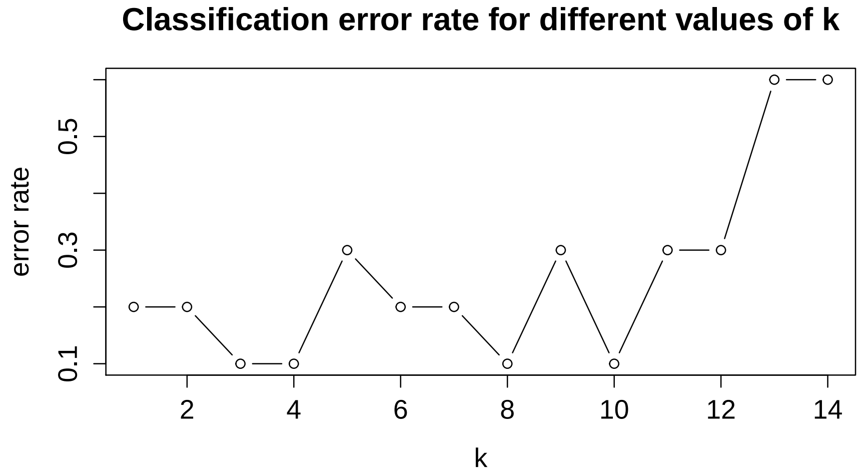
Try several different values of k, and report the classification error rate for each below.

```
> cbind("k"=1:n, "knn classification error"=knn.err, deparse.level=2)
```

```
       k knn classification error
 [1,]  1                      0.2
 [2,]  2                      0.2
 [3,]  3                      0.1
 [4,]  4                      0.1
 [5,]  5                      0.3
 [6,]  6                      0.2
 [7,]  7                      0.2
 [8,]  8                      0.1
 [9,]  9                      0.3
[10,] 10                      0.1
[11,] 11                      0.3
[12,] 12                      0.3
[13,] 13                      0.6
[14,] 14                      0.6
```

Here is a plot of the same data

```
> plot(knn.err, type="b",
+        main="Classification error rate for different values of k",
+        xlab="k", ylab="error rate")
```



2

What problems occur if you choose too small a value for k? Too large?

With a value for k too small we will classify in a way that is very sensitive to the local characteristcs of the training data.

With a value of k too large we essentially overfit, ignoring the information contained in the predictor variables. In the extreme with k equal the number of observations in the train data all test data is assigned to the most frequent class in the train data, Owner in the present case.