NRSG 741 Homework 8

Tommy Flynn

4/11/2018

The GitHub Repository can be found here https://github.com/tommyflynn/N741_Homework/tree/master/Flynn_HW_08

K-nearest neighbor Let's try a variation on the NHANES data set again.

```
# Create the NHANES dataset again
people <- NHANES %>% dplyr::select(Age, Gender, Diabetes, BMI, HHIncome, PhysActive, SleepTrouble)%>%
  mutate(Gender = as.numeric(Gender), Diabetes = as.numeric(Diabetes), BMI,HHIncome = as.numeric(HHIncor
  filter(!is.na(Age), !is.na(Gender), !is.na(Diabetes), !is.na(BMI), !is.na(HHIncome), !is.na(PhysActive
lut <- c("Yes" = "1", "No" = "0")
people$SleepTrouble <- as.numeric(lut[people$SleepTrouble])

#check the subset
glimpse(people)
```

```
## Observations: 7,037
## Variables: 7
## $ Age          <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 3...
## $ Gender       <dbl> 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 1, 2...
## $ Diabetes     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ BMI          <dbl> 32.22, 32.22, 32.22, 30.57, 27.24, 27.24, 27.24, ...
## $ HHIncome     <dbl> 6, 6, 6, 7, 11, 11, 11, 6, 12, 10, 11, 4, 6, 4, 1...
## $ PhysActive   <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 2...
## $ SleepTrouble <dbl> 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0...
```

Create the NHANES dataset again, just like we did in class, only using sleep trouble (variable name = SleepTrouble) as the dependent variable, instead of SleepTrouble. (I'm assuming you meann instead of Diabetes?)

Problem 1

What is the marginal distribution of sleep trouble?

```
# What is the marginal distribution of sleep trouble?
knitr::kable(tally(~ SleepTrouble, data = people, format = "percent"))
```

SleepTrouble	Freq
0	25.55066
1	74.44934

Problem 2

Apply the k-nearest neighbor procedure to predict SleepTrouble from the other covariates, as we did for SleepTrouble. Use k = 1, 3, 5, and 20.

```r
# Apply knn procedure to predict SleepTrouble

# Let's try different values of k to see how that affects performance
knn.1 <- knn(train = people, test = people, cl = people$SleepTrouble, k = 1)
knn.3 <- knn(train = people, test = people, cl = people$SleepTrouble, k = 3)
knn.5 <- knn(train = people, test = people, cl = people$SleepTrouble, k = 5)
knn.20 <- knn(train = people, test = people, cl = people$SleepTrouble, k = 20)
```

Now let's see how well these classifiers work overall

Problem 3

```r
# Calculate the percent predicted correctly
100*sum(people$SleepTrouble == knn.1)/length(knn.1)
```

```
## [1] 100
```

```r
100*sum(people$SleepTrouble == knn.3)/length(knn.3)
```

```
## [1] 91.99943
```

```r
100*sum(people$SleepTrouble == knn.5)/length(knn.5)
```

```
## [1] 88.4752
```

```r
100*sum(people$SleepTrouble == knn.20)/length(knn.20)
```

```
## [1] 78.57041
```

Problem 4

What about success overall?

```r
# Another way to look at success rate against increasing k
table(knn.1, people$SleepTrouble)
```

```
##
## knn.1    0    1
##     0 1798    0
##     1    0 5239
```

```r
table(knn.3, people$SleepTrouble)
```

```
##
## knn.3    0    1
##     0 1409  174
##     1  389 5065
```

```r
table(knn.5, people$SleepTrouble)
```

```
##
## knn.5    0    1
##     0 1209  222
##     1  589 5017
```

```r
table(knn.20, people$SleepTrouble)
```

```
##
## knn.20    0    1
##      0  434  144
##      1 1364 5095
```