# Homework_3

*Marjan Rezvani*

*10/5/2020*

## R Markdown

```
load('/Users/marjanrezvani/Documents/Fall2020/eco_stat/data/acs2017_ny/acs2017_ny_data.RData')
dat_NYC <- subset(acs2017_ny, (acs2017_ny$in_NYC == 1)&(acs2017_ny$AGE > 20) & (acs2017_ny$AGE < 66))
attach(dat_NYC)
#View(head(dat_NYC))

borough_f <- factor((in_Bronx + 2*in_Manhattan + 3*in_StatenI + 4*in_Brooklyn + 5*in_Queens), levels=c(

norm_varb <- function(X_in) {
  (X_in - min(X_in, na.rm = TRUE))/( max(X_in, na.rm = TRUE) - min(X_in, na.rm = TRUE) )
}
```

```
is.na(OWNCOST) <- which(OWNCOST == 9999999)
housing_cost <- OWNCOST + RENT
norm_inc_tot <- norm_varb(INCTOT)
norm_housing_cost <- norm_varb(housing_cost)
norm_poverty <- norm_varb(POVERTY)

data_use_prelim <- data.frame(norm_inc_tot,
                              norm_housing_cost,
                              norm_poverty)
good_obs_data_use <- complete.cases(data_use_prelim,borough_f)
dat_use <- subset(data_use_prelim,good_obs_data_use)
y_use <- subset(borough_f,good_obs_data_use)

set.seed(12345)
NN_obs <- sum(good_obs_data_use == 1)
select1 <- (runif(NN_obs) < 0.8)
train_data <- subset(dat_use,select1)
test_data <- subset(dat_use,(!select1))
cl_data <- y_use[select1]
true_data <- y_use[!select1]


summary(cl_data)
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens
##          4880          5250          1891         12416         10923
```

```
prop.table(summary(cl_data))
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens
##    0.13800905    0.14847285    0.05347851    0.35113122    0.30890837
```

```
summary(train_data)
```

```
##   norm_inc_tot     norm_housing_cost  norm_poverty
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.01191   1st Qu.:0.02493   1st Qu.:0.3234  
##  Median :0.02693   Median :0.96917   Median :0.7166  
##  Mean   :0.04265   Mean   :0.58972   Mean   :0.6450  
##  3rd Qu.:0.05219   3rd Qu.:0.97784   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
```

```
require(class)
```

```
## Loading required package: class
```

```
for (indx in seq(1, 9, by= 2)) {
  pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
  num_correct_labels <- sum(pred_borough == true_data)
  correct_rate <- num_correct_labels/length(true_data)
  print(c(indx,correct_rate))
}
```

```
## [1] 1.0000000 0.3876637
## [1] 3.0000000 0.3651505
## [1] 5.0000000 0.3737652
## [1] 7.0000000 0.3885826
## [1] 9.000000 0.387434
```

## adding educ_college

**but as you'll see the result, it doesn't help us to classify the boroughs better**

```
norm_poverty <- norm_varb(POVERTY)
norm_educ_college <- norm_varb(educ_college)
```

```
data_use_prelim <- data.frame(norm_inc_tot,
                              norm_housing_cost,
                              norm_poverty,
                              norm_educ_college)
good_obs_data_use <- complete.cases(data_use_prelim,borough_f)
dat_use <- subset(data_use_prelim,good_obs_data_use)
y_use <- subset(borough_f,good_obs_data_use)

set.seed(12345)
NN_obs <- sum(good_obs_data_use == 1)
select1 <- (runif(NN_obs) < 0.8)
train_data <- subset(dat_use,select1)
test_data <- subset(dat_use,(!select1))
cl_data <- y_use[select1]
true_data <- y_use[!select1]


summary(cl_data)
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens 
##          4880          5250          1891         12416         10923
```

```
prop.table(summary(cl_data))
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens
##    0.13800905    0.14847285    0.05347851    0.35113122    0.30890837
```

```
summary(train_data)
```

```
##   norm_inc_tot      norm_housing_cost  norm_poverty     norm_educ_college
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.01191   1st Qu.:0.02493   1st Qu.:0.3234   1st Qu.:0.0000
##  Median :0.02693   Median :0.96917   Median :0.7166   Median :0.0000
##  Mean   :0.04265   Mean   :0.58972   Mean   :0.6450   Mean   :0.2527
##  3rd Qu.:0.05219   3rd Qu.:0.97784   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

```
require(class)
for (indx in seq(1, 9, by= 2)) {
  pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
  num_correct_labels <- sum(pred_borough == true_data)
  correct_rate <- num_correct_labels/length(true_data)
  print(c(indx,correct_rate))
}
```

```
## [1] 1.0000000 0.3759476
## [1] 3.0000000 0.3591776
## [1] 5.0000000 0.3708936
## [1] 7.00000 0.37813
## [1] 9.0000000 0.3812313
```

**Now I going to try Cost_total, COSTFUEL combined with COSTWATER, COSTGAS and COSTELEC.**

## firstly fix up the data:

```
cost_total <- COSTELEC + COSTFUEL + COSTGAS + COSTWATR

norm_cost_total <- norm_varb(cost_total)


data_use_prelim <- data.frame(norm_inc_tot,
                              norm_housing_cost,
                              norm_poverty,
                              norm_educ_college,
                              norm_cost_total)
good_obs_data_use <- complete.cases(data_use_prelim,borough_f)
dat_use <- subset(data_use_prelim,good_obs_data_use)
y_use <- subset(borough_f,good_obs_data_use)

set.seed(12345)
NN_obs <- sum(good_obs_data_use == 1)
select1 <- (runif(NN_obs) < 0.8)
train_data <- subset(dat_use,select1)
test_data <- subset(dat_use,(!select1))
```

```
cl_data <- y_use[select1]
true_data <- y_use[!select1]


summary(cl_data)
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens 
##          4880          5250          1891         12416         10923
```

```
prop.table(summary(cl_data))
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens 
##    0.13800905    0.14847285    0.05347851    0.35113122    0.30890837
```

```
summary(train_data)
```

```
##   norm_inc_tot     norm_housing_cost  norm_poverty    norm_educ_college
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   
##  1st Qu.:0.01191   1st Qu.:0.02493   1st Qu.:0.3234   1st Qu.:0.0000   
##  Median :0.02693   Median :0.96917   Median :0.7166   Median :0.0000   
##  Mean   :0.04265   Mean   :0.58972   Mean   :0.6450   Mean   :0.2527   
##  3rd Qu.:0.05219   3rd Qu.:0.97784   3rd Qu.:1.0000   3rd Qu.:1.0000   
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   
##  norm_cost_total 
##  Min.   :0.0000  
##  1st Qu.:0.4079  
##  Median :0.5508  
##  Mean   :0.5860  
##  3rd Qu.:0.7768  
##  Max.   :1.0000
```

```
require(class)
for (indx in seq(1, 9, by= 2)) {
  pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
  num_correct_labels <- sum(pred_borough == true_data)
  correct_rate <- num_correct_labels/length(true_data)
  print(c(indx,correct_rate))
}
```

```
## [1] 1.0000000 0.4638181
## [1] 3.0000000 0.4362509
## [1] 5.0000000 0.4403859
## [1] 7.0000000 0.4430278
## [1] 9.000000 0.448771
```

**as you can see, there is more accuracy with using cost total which is combined of costs for water, fuel, electricity, and gas.**

```
norm_poverty <- norm_varb(POVERTY)
norm_educ_college <- norm_varb(educ_college)
#norm_advdeg <- norm_varb(educ_advdeg)

cost_total <- COSTELEC + COSTFUEL + COSTGAS + COSTWATR
norm_cost_total <- norm_varb(cost_total)
norm_FOODSTMP <- norm_varb(FOODSTMP)
```

```r
data_use_prelim <- data.frame(norm_inc_tot,
                              norm_housing_cost,
                              norm_poverty,
                              norm_educ_college,
                              norm_cost_total,
                              norm_FOODSTMP)
good_obs_data_use <- complete.cases(data_use_prelim,borough_f)
dat_use <- subset(data_use_prelim,good_obs_data_use)
y_use <- subset(borough_f,good_obs_data_use)

set.seed(12345)
NN_obs <- sum(good_obs_data_use == 1)
select1 <- (runif(NN_obs) < 0.8)
train_data <- subset(dat_use,select1)
test_data <- subset(dat_use,(!select1))
cl_data <- y_use[select1]
true_data <- y_use[!select1]


summary(cl_data)
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens
##          4880          5250          1891         12416         10923
```

```r
prop.table(summary(cl_data))
```

```
##         Bronx     Manhattan Staten Island      Brooklyn        Queens
##    0.13800905    0.14847285    0.05347851    0.35113122    0.30890837
```

```r
summary(train_data)
```

```
##   norm_inc_tot     norm_housing_cost  norm_poverty    norm_educ_college
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.01191   1st Qu.:0.02493   1st Qu.:0.3234   1st Qu.:0.0000
##  Median :0.02693   Median :0.96917   Median :0.7166   Median :0.0000
##  Mean   :0.04265   Mean   :0.58972   Mean   :0.6450   Mean   :0.2527
##  3rd Qu.:0.05219   3rd Qu.:0.97784   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##  norm_cost_total  norm_FOODSTMP
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.4079   1st Qu.:0.0000
##  Median :0.5508   Median :0.0000
##  Mean   :0.5860   Mean   :0.1757
##  3rd Qu.:0.7768   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

```r
require(class)
for (indx in seq(1, 9, by= 2)) {
  pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
  num_correct_labels <- sum(pred_borough == true_data)
  correct_rate <- num_correct_labels/length(true_data)
  print(c(indx,correct_rate))
}
```

```
## [1] 1.0000000 0.4758787
## [1] 3.0000000 0.4431427
## [1] 5.0000000 0.4461291
## [1] 7.0000000 0.4473926
## [1] 9.0000000 0.4507236
```

**I tried different variables to figure out which one would help to get more precise results to have a higher accuracy.**

**some of them like educ_college which tell us about having college degree or not, would not increase the accuracy. or another attribute like TRANWORK does not help us to get a better prediction either.**

**but as it obvious in the result, cost_total which includes cost of gas, electricity, water and fuel, would be considered as an effective variable to predict category of data.**

**in addition, interesting thing is that prediction using variable Foodstmp, the borough one which is Bronx, is more precise.**