Homework#3 Final

Thakur Prasad Ghimire

9/30/2021

1. Study group members name:

Thakur Ghimire

Nicholas Alonso

Bolina Oxilus

Diep luu

Joel rosas

setwd("C:/HW#3/Homework#3 Final")  
load("C:/HW#3/acs2017\_ny\_data.RData")

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) )  
}

norm\_varb <- function(X\_in) {  
 (max(X\_in, na.rm = TRUE) - X\_in)/( 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)  
data\_use\_prelim <- data.frame(norm\_inc\_tot,norm\_housing\_cost)  
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   
## 5110 5500 1971 12807 11376

prop.table(summary(cl\_data))

## Bronx Manhattan Staten Island Brooklyn Queens   
## 0.13899467 0.14960287 0.05361223 0.34835709 0.30943314

summary(train\_data)

## norm\_inc\_tot norm\_housing\_cost  
## Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.9500 1st Qu.:0.02216   
## Median :0.9744 Median :0.03131   
## Mean :0.9591 Mean :0.40903   
## 3rd Qu.:0.9890 3rd Qu.:0.97495   
## Max. :1.0000 Max. :1.00000

suppressMessages(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))\*100  
print(c(indx,correct\_rate))  
print(summary(pred\_borough))  
}

## [1] 1.00000 35.63447  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 991 1258 426 3467 2897   
## [1] 3.00000 36.22082  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 854 1176 325 3616 3068   
## [1] 5.00000 37.21651  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 768 1115 228 3708 3220   
## [1] 7.00000 38.02412  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 727 1035 168 3844 3265   
## [1] 9.00000 37.80285  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 654 1003 151 3901 3330

#Here, we will use the categorical variables poverty, family size, and household income for the second run to find the knn classification.

fam\_pov <- POVERTY + FAMSIZE  
norm\_houseinc <- norm\_varb(HHINCOME)  
norm\_fam\_pov <- norm\_varb(fam\_pov)  
data\_use\_prelim <- data.frame(norm\_houseinc,norm\_fam\_pov)  
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   
## 4790 4964 1919 12433 11119

prop.table(summary(cl\_data))

## Bronx Manhattan Staten Island Brooklyn Queens   
## 0.13598297 0.14092264 0.05447835 0.35295955 0.31565649

summary(train\_data)

## norm\_houseinc norm\_fam\_pov   
## Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.9234 1st Qu.:0.03089   
## Median :0.9544 Median :0.28378   
## Mean :0.9382 Mean :0.35720   
## 3rd Qu.:0.9759 3rd Qu.:0.65058   
## Max. :1.0000 Max. :0.99807

summary(test\_data)

## norm\_houseinc norm\_fam\_pov   
## Min. :0.2368 Min. :0.00000   
## 1st Qu.:0.9224 1st Qu.:0.03089   
## Median :0.9543 Median :0.28764   
## Mean :0.9375 Mean :0.36184   
## 3rd Qu.:0.9761 3rd Qu.:0.66216   
## Max. :1.0000 Max. :0.99807

suppressMessages(require(class))  
for (indx in seq(1, 9, by= 2)) {  
 pred\_borough1 <- knn(train\_data, test\_data, cl\_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)  
num\_correct\_labels <- sum(pred\_borough1 == true\_data)  
correct\_rate <- (num\_correct\_labels/length(true\_data))\*100  
print(c(indx,correct\_rate))  
print(summary(pred\_borough1))  
}

## [1] 1.00000 61.84591  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 1032 1150 361 3310 2804   
## [1] 3.0000 44.6575  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 955 1123 344 3388 2847   
## [1] 5.00000 43.17893  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 874 1055 301 3547 2880   
## [1] 7.00000 42.30103  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 842 972 252 3644 2947   
## [1] 9.0000 41.6426  
## Bronx Manhattan Staten Island Brooklyn Queens   
##

789 955 177 3717 3019

# For third run of knn we use the categorical variables of utility cost including gas,electricity,water,fulel and household income.

utility\_cost <- COSTELEC + COSTFUEL + COSTGAS + COSTWATR  
norm\_houseinc2 <- norm\_varb(HHINCOME)  
norm\_utility\_cost <- norm\_varb(utility\_cost)  
data\_use\_prelim <- data.frame(norm\_houseinc2,norm\_utility\_cost)  
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   
## 4790 4964 1919 12433 11119

prop.table(summary(cl\_data))

## Bronx Manhattan Staten Island Brooklyn Queens   
## 0.13598297 0.14092264 0.05447835 0.35295955 0.31565649

summary(train\_data)

## norm\_houseinc2 norm\_utility\_cost  
## Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.9234 1st Qu.:0.2231   
## Median :0.9544 Median :0.4460   
## Mean :0.9382 Mean :0.3937   
## 3rd Qu.:0.9759 3rd Qu.:0.5700   
## Max. :1.0000 Max. :0.9780

summary(test\_data)

## norm\_houseinc2 norm\_utility\_cost  
## Min. :0.2368 Min. :0.0000   
## 1st Qu.:0.9224 1st Qu.:0.2203   
## Median :0.9543 Median :0.4460   
## Mean :0.9375 Mean :0.3920   
## 3rd Qu.:0.9761 3rd Qu.:0.5685   
## Max. :1.0000 Max. :0.9645

suppressMessages(require(class))  
for (indx in seq(1, 9, by= 2)) {  
 pred\_borough2 <- knn(train\_data, test\_data, cl\_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)  
num\_correct\_labels <- sum(pred\_borough2 == true\_data)  
correct\_rate <- (num\_correct\_labels/length(true\_data))\*100  
print(c(indx,correct\_rate))  
print(summary(pred\_borough2))  
}

## [1] 1.000 79.866  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 1127 1160 473 3164 2733   
## [1] 3.00000 53.77151  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 1093 1120 418 3261 2765   
## [1] 5.00000 52.16588  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 1055 1045 324 3364 2869   
## [1] 7.00000 50.12129  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 1049 1050 268 3417 2873   
## [1] 9.0000 49.3589  
## Bronx Manhattan Staten Island Brooklyn Queens   
## 999 1067 238 3477 2876

Compared to the second and first run sequences, the output accuracy of the k-nn algorithm was considerably more significant in the third run.

Furthermore, the distribution of persons had a comparable count per person.

Finally, the k-nearest neighbor method is used to estimate a random individual's living location in our dataset. We begin by limiting our dataset to persons who live in New York (because pums data covers the whole state of New York), which allows us to identify 80% of our data as residents of one of the five boroughs of New York City. From the above probability summary table, 35% 0f people are living in Brooklyn, 31% in Queens,14% in Manhattan,13% in the Bronx and 5.4% people are in Istaten Islad .

We attempt to estimate the living location of the remaining 20% of our data using that probability distribution and an additional two or more factors of

our choosing.

The closeness of a random non-classified datapoint to other classified data points is used to make this prediction. For example, if the bulk of the residences in Manhattan have a very high utility cost, we may assume that a random datapoint with a high utility is in Manhattan. The accuracy of the second run, which included factors such as poverty, family size, and household income, was 61.85% for the first sequence. While the accuracy obtained from the third run, including variables such as the cost of various utilities and household income, yielded a score of 79.87 percent for the first sequence. As a result, the k-nn algorithm may make use of additional data or variables.