---

title: "Homework#3"

author: "Neshma Simon"

date: "10/6/2020"

output: html\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

## Study Group: Fareha, Stan, and Hertz

```

k-nn classification: First we write down a code in order to do k-nn classification. This code will help us predict the boroughs from the data, instead of the neighborhoods.

dat\_NYC <- subset(acs2017\_ny, (acs2017\_ny$in\_NYC == 1)&(acs2017\_ny$AGE > 20) & (acs2017\_ny$AGE < 66))

attach(dat\_NYC)

borough\_f <- factor((in\_Bronx + 2\*in\_Manhattan + 3\*in\_StatenI + 4\*in\_Brooklyn + 5\*in\_Queens), levels=c(1,2,3,4,5),labels = c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))

```

```

(0,1) interval: In this data we picked the household income as a variable to classify by the boroughs. Household income plays a significant role in the area, neighborhood people reside in. This code arranges the data to be in the (0,1) interval.

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) )

}

```

```

data fix:

is.na(OWNCOST) <- which(OWNCOST == 9999999)

housing\_cost <- OWNCOST + RENT

norm\_inc\_tot <- norm\_varb(INCTOT)

norm\_housing\_cost <- norm\_varb(housing\_cost)

```

```

dataframe created:This the data frame we will use to associate the hosuehold incomes with the different boroughs. It will direct the data to use household income as a variable to associate.

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)

```

```

80/20 split: one part to train the algo, then the other part to test how well it works for new data:

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]

```

```

k-nn algo run and compare against the simple means:

summary(cl\_data)

prop.table(summary(cl\_data))

summary(train\_data)

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))

}

```

## As we attempted to make it our own, we tried to modify the codes to see if there would be a signifcant change or shift.

```

Just as what was done previously, we tried to go a different route and see if this can work as a 60/40 split.

set.seed(12345)

NN\_obs <- sum(good\_obs\_data\_use ==1)

select1 <-runif(NN\_obs)<0.6

train\_data <- subset(dat\_use,select1)

test\_data <-subset(dat\_use,!select1)

cl\_data <-y\_use[select1] ##matrix of classes of the K

true\_data <-y\_use[!select1]

Then, we can do a k-nn algo and compare against the simple means as we have done before.

summary(cl\_data)

Bronx Manhattan Staten Island Brooklyn Queens

3641 3943 1433 9344 8250

prop.table(summary(cl\_data))

Bronx Manhattan Staten Island Brooklyn Queens

0.13682312 0.14817181 0.05384991 0.35113299 0.31002217

summary(train\_data)

norm\_inc\_tot norm\_housing\_cost

Min. :0.00000 Min. :0.00000

1st Qu.:0.01184 1st Qu.:0.02476

Median :0.02693 Median :0.96917

Mean :0.04268 Mean :0.58874

3rd Qu.:0.05219 3rd Qu.:0.97784

Max. :1.00000 Max. :1.00000

library(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 0.3480951

[1] 3.0000000 0.3527356

[1] 5.0000000 0.3670581

[1] 7.0000000 0.3744486

[1] 9.0000000 0.3797193

```

```

summary(cl\_data)

prop.table(summary(cl\_data))

summary(train\_data)

require(class)

for (indx in seq(1, 19, 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 0.3195239

[1] 3.0000000 0.3219162

[1] 5.0000000 0.3342864

[1] 7.0000000 0.3411717

[1] 9.0000000 0.3481736

[1] 11.0000000 0.3499825

[1] 13.0000000 0.3488738

[1] 15.0000000 0.3495157

[1] 17.0000000 0.3506827

[1] 19.0000000 0.3512078

summary(correct\_rate)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.3512 0.3512 0.3512 0.3512 0.3512 0.3512

mean(correct\_rate)

[1] 0.3512078

We increased our K value to increase to 19 from 15 to see the correlateion. After viewing the results, we were able to depict the direct relationship between the increase of K and the housing cost/total income of a specific borough.

We choose to look at the household incomes in the boroughs to see whether the code is a good determinant of how the household incomes of people correlates with the boroughs they reside in.

```