#title: "Homework 3 lab 3"

#author: "KIM FAI CHAN"

#date: "9/28/2021"

setwd("~/Desktop/jeff's documents/Master of Economics in CCNY/ECONOMETRICS ECO B2000/homework 3 lab 3")

load("acs2017\_ny\_data.RData")

#For the k-nn classification, it's decided by personal Income Total, Housing Cost, the population was limited in NYC with the age between 21 and 65.

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"))

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)

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)

#after summary then below info pops up:

Bronx Manhattan Staten Island Brooklyn Queens

4820 5164 1892 12395 10955

prop.table(summary(cl\_data))

> prop.table(summary(cl\_data))

Bronx Manhattan Staten Island Brooklyn Queens

0.13683075 0.14659626 0.05371033 0.35187078 0.31099188

summary(train\_data)

norm\_inc\_tot norm\_housing\_cost

Min. :0.00000 Min. :0.00000

1st Qu.:0.01184 1st Qu.:0.02478

Median :0.02693 Median :0.96898

Mean :0.04249 Mean :0.58675

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

Max. :1.00000 Max. :1.00000

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 0.3457318

[1] 3.0000000 0.3482731

[1] 5.0000000 0.3595934

[1] 7.0000000 0.3684879

[1] 9.0000000 0.3726464

#The results are: The algorithm could be more accurate when k = 1. And higher the values of k, the lower the accurate.

#It is clear that as k increases after 1, the quality of algorithm to correctly classify the observations decreases. knowing more about variables allows us to refine the algorithm and the accuracy could become higher on classifying the test data. However, it seems more variables, the algorithm become more confused on toward difference direction.

cl\_data\_n <- as.numeric(cl\_data)

model\_ols1 <- lm(cl\_data\_n ~ train\_data$norm\_inc\_tot + train\_data$norm\_housing\_cost)

y\_hat <- fitted.values(model\_ols1)

mean(y\_hat[cl\_data\_n == 1])

[1] 3.485385

mean(y\_hat[cl\_data\_n == 2])

[1] 3.380373

mean(y\_hat[cl\_data\_n == 3])

[1] 3.762961

mean(y\_hat[cl\_data\_n == 4])

[1] 3.548343

mean(y\_hat[cl\_data\_n == 5])

[1] 3.635049

cl\_data\_n1 <- as.numeric(cl\_data\_n == 1)

model\_ols\_v1 <- lm(cl\_data\_n1 ~ train\_data$norm\_inc\_tot + train\_data$norm\_housing\_cost)

y\_hat\_v1 <- fitted.values(model\_ols\_v1)

mean(y\_hat\_v1[cl\_data\_n1 == 1])

[1] 0.1575926

mean(y\_hat\_v1[cl\_data\_n1 == 0])

[1] 0.1335395

#The tradeoff between classifying the training data and the test data comes from:

# what point do I think a tradeoff between better classifying the training data and doing worse at classifying the test data?

# a. how deep we could understand on the data;

# b. do I have the knowledge to identify the key useful variables in minimum level?

# c. do I have the ability to identify which are un-useful observations?

# Finally, use adequate useful classified observations to train the algorithm and keep a large enough test set to test the algorithm.

# However, in my understanding this would much more depends on the bias of observer,

# and this is decided by the knowledge of the observer to the data set.