Assignment 4

Tanu Sreedharan TS3175

2/12/2020

# Part 1: Implementing a Simple Prediction Pipeline

Fit and evaluate two prediction models using linear regression. The aim of the models are to predict the number of days in a month an individual reported having good physical health (feature name: healthydays).

* Performing basic data cleaning.

physact <- read.csv("Data/class4\_p1.csv") %>%   
 drop\_na() %>%   
 janitor::clean\_names()  
str(physact)

## 'data.frame': 2195 obs. of 17 variables:  
## $ x : int 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 ...  
## $ chronic1 : int 2 2 1 2 1 2 2 2 2 2 ...  
## $ chronic3 : int 2 2 1 2 1 2 2 2 2 2 ...  
## $ chronic4 : int 2 2 2 2 2 2 2 2 2 1 ...  
## $ bmi : num 19.9 33.7 24.2 22.7 25.7 ...  
## $ tobacco1 : int 3 3 3 2 3 3 3 3 3 3 ...  
## $ alcohol1 : int 3 3 3 2 3 2 3 3 2 3 ...  
## $ gpaq8totmin : int 0 0 30 0 0 30 0 0 0 0 ...  
## $ gpaq11days : int 7 7 4 6 4 7 5 0 7 7 ...  
## $ habits5 : int 2 2 2 2 2 2 1 2 2 3 ...  
## $ habits7 : int 2 5 3 4 4 5 2 2 3 4 ...  
## $ agegroup : int 2 2 2 2 3 1 2 2 2 3 ...  
## $ dem3 : int 1 2 2 1 2 1 2 2 1 2 ...  
## $ dem4 : int 2 1 2 2 2 2 2 2 1 1 ...  
## $ dem8 : int 2 2 2 1 2 2 1 1 1 2 ...  
## $ povertygroup: int 1 1 1 5 2 6 4 3 2 2 ...  
## $ healthydays : int 30 27 30 30 23 30 30 30 30 0 ...

* Variable formats are all numeric
* Partitioning data into training and testing (use a 70/30 split)

set.seed(200)  
train.indices<-createDataPartition(y=physact$healthydays,p=0.7,list=FALSE)  
  
training<-physact[train.indices,]  
testing<-physact[-train.indices,]

## 1. Fit two prediction linear regression models using different subsets of the features in the training data

* Model 1: Include habit how active participant is, BMI, and whether or not the participant is born outside of the united states
* Model 2: Only include habit how active participant is

model.1 <- lm(healthydays ~ habits5 + povertygroup + agegroup, data=training)  
summary(model.1)

##   
## Call:  
## lm(formula = healthydays ~ habits5 + povertygroup + agegroup,   
## data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.4248 -0.3222 2.0746 3.9624 10.5316   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.0087 0.8449 37.885 < 2e-16 \*\*\*  
## habits5 -1.8878 0.2319 -8.140 8.12e-16 \*\*\*  
## povertygroup 0.5981 0.1161 5.151 2.92e-07 \*\*\*  
## agegroup -1.3968 0.2092 -6.675 3.44e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.369 on 1533 degrees of freedom  
## Multiple R-squared: 0.09388, Adjusted R-squared: 0.09211   
## F-statistic: 52.94 on 3 and 1533 DF, p-value: < 2.2e-16

model.2<-lm(healthydays ~ agegroup + povertygroup, data=training)  
summary(model.2)

##   
## Call:  
## lm(formula = healthydays ~ agegroup + povertygroup, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.0182 -0.0658 2.6350 4.0367 7.1381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28.3639 0.7316 38.772 < 2e-16 \*\*\*  
## agegroup -1.5507 0.2128 -7.288 5.02e-13 \*\*\*  
## povertygroup 0.7008 0.1178 5.947 3.38e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.524 on 1534 degrees of freedom  
## Multiple R-squared: 0.05472, Adjusted R-squared: 0.05349   
## F-statistic: 44.4 on 2 and 1534 DF, p-value: < 2.2e-16

## 2. Apply models to test data

* Model 1: Include how active participant is, BMI, and whether or not the participant is born outside of the united states
* Model 2: Only include how active participant is

model.1 <- lm(healthydays ~ habits5 + povertygroup + agegroup, data=testing)  
summary(model.1)

##   
## Call:  
## lm(formula = healthydays ~ habits5 + povertygroup + agegroup,   
## data = testing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.2289 0.2863 2.3296 3.8298 8.4468   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.0349 1.3623 22.781 < 2e-16 \*\*\*  
## habits5 -1.2962 0.3678 -3.525 0.000453 \*\*\*  
## povertygroup 0.3641 0.1820 2.001 0.045780 \*   
## agegroup -1.1652 0.3443 -3.384 0.000756 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.521 on 654 degrees of freedom  
## Multiple R-squared: 0.04785, Adjusted R-squared: 0.04349   
## F-statistic: 10.96 on 3 and 654 DF, p-value: 5e-07

model.2<-lm(healthydays ~ agegroup + povertygroup, data=testing)  
summary(model.2)

##   
## Call:  
## lm(formula = healthydays ~ agegroup + povertygroup, data = testing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.5255 0.2196 2.6401 3.9448 6.1103   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28.6782 1.1973 23.953 < 2e-16 \*\*\*  
## agegroup -1.3047 0.3450 -3.782 0.00017 \*\*\*  
## povertygroup 0.4304 0.1826 2.358 0.01868 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.586 on 655 degrees of freedom  
## Multiple R-squared: 0.02977, Adjusted R-squared: 0.0268   
## F-statistic: 10.05 on 2 and 655 DF, p-value: 5.036e-05

## 3. Describe one setting (in 1-2 sentences) where the implementation of your final model would be useful

The final model (model 1) would ideal be predicting the number of headays in a month a person feels healthy based on their physical activity behaviors, age group, and income. The implementation of the results could be informing programs that work with young low-income people to implement a normal physical activity routine to ensure that people feel healthier.

# Part II: Conducting an Unsupervised Analysis

Using the dataset from the Group assignment Part IIb (USArrests), identify clusters using hierarchical analysis.

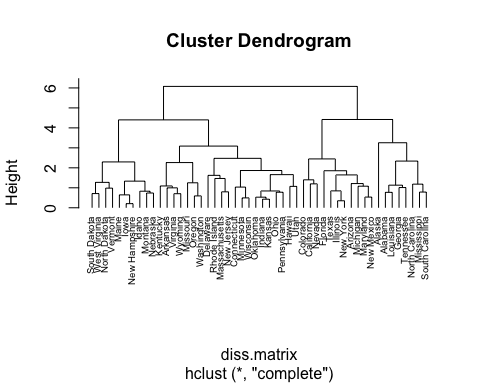
colMeans(USArrests, na.rm=TRUE)

## Murder Assault UrbanPop Rape   
## 7.788 170.760 65.540 21.232

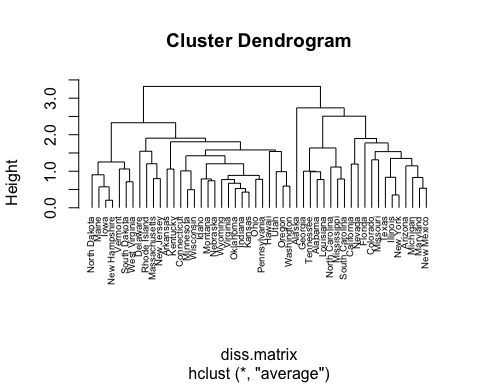
usa.clean<- na.omit(USArrests) %>%   
 scale()

* 1. Vary the choice for agglomeration method.
* For agglomeration method, use complete, average and single.
* For distance, use Euclidian as our features are continuous and numerical.

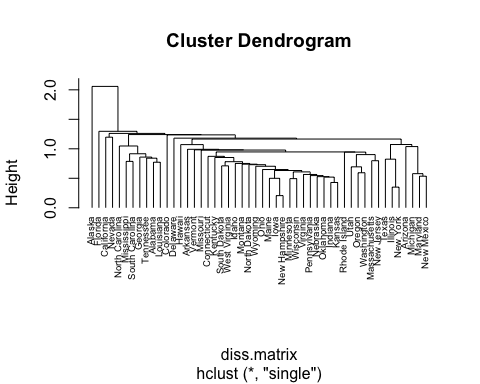
# Create Dissimilarity matrix  
diss.matrix <- dist(usa.clean, method = "euclidean")  
  
# Hierarchical clustering using Complete Linkage  
hc1 <- hclust(diss.matrix, method = "complete" )  
# Plot the obtained dendrograms  
plot(hc1, cex = 0.6, hang = -1)



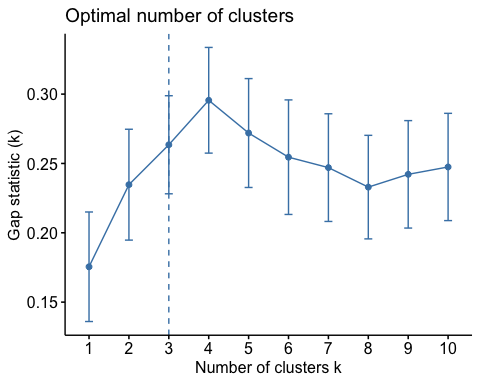
# Hierarchical clustering using Average Linkage  
hc2 <- hclust(diss.matrix, method = "average" )  
# Plot the obtained dendrograms  
plot(hc2, cex = 0.6, hang = -1)



# Hierarchical clustering using Single Linkage  
hc3 <- hclust(diss.matrix, method = "single" )  
# Plot the obtained dendrograms  
plot(hc3, cex = 0.6, hang = -1)



gap\_stat <- clusGap(usa.clean, FUN = hcut, nstart = 25, K.max =10, B = 50)  
fviz\_gap\_stat(gap\_stat)



* Determine the optimal number of clusters using a clear, data-driven strategy.
* Describe the composition of each cluster in terms of the original input features
* Do you see differences in results when you change the type of agglomeration method used?

For the complete linkage clustering, has 3 clear clusters. Two of them are about the same size but one is relatively large and it includes New York and New Jersey.

For the average linkage clustering, the clusters are a little similar to the compkete linkage. There is just more overlapping of states.

For the single linkage clustering, there was a lot more overlapping and unclear distinction between the states besides Alaska, Florida, Califronia, and Nevada.

All three methods produced different clusters. Each one I did changed, starting with the largest clusters to smallest the clusters.

Using the data-driven strategy of computing a gap statistic, the optimal number of clusters is 3.

# 5. Pretend that the data are from 2020 and not 1973. Describe one research question that can be addressed using the newly identified clusters. HINT: The clusters can be used as an exposure, an outcome or a covariate.

Using the data above, we can assess high medium and low arrest clusters and investigate the reasons behind those arrests. It could lead to more informed policing tactics and more awarness of social determinants of health within the world of law and policy.